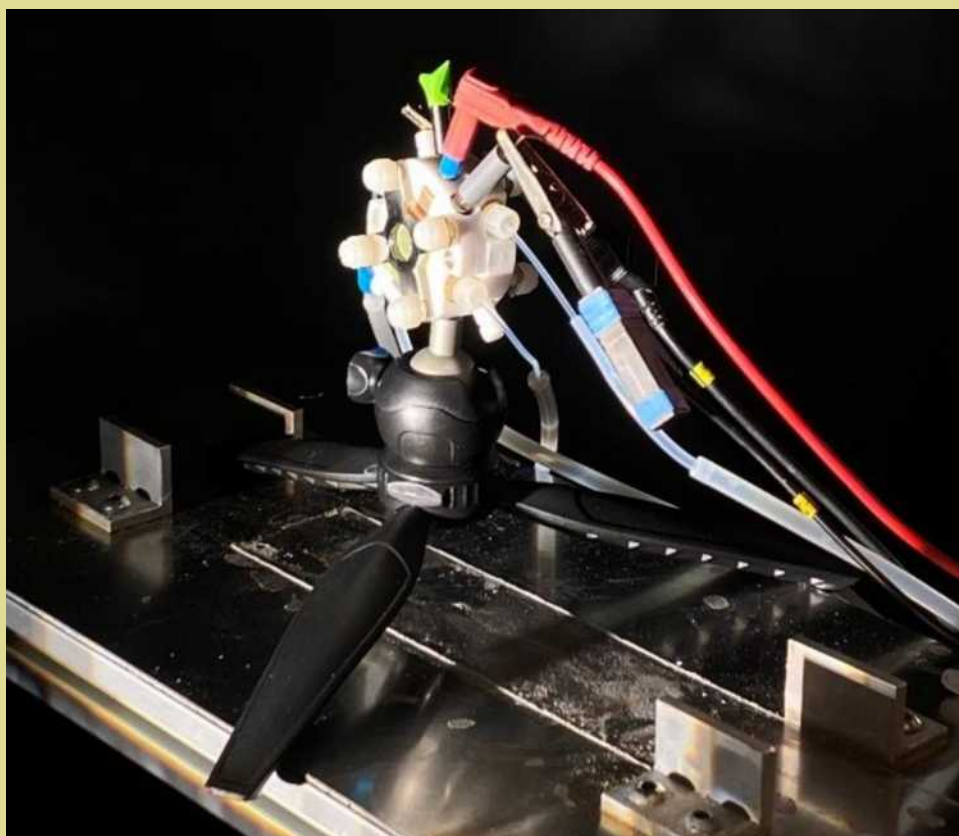


CHEMICAL ENGINEERING RESEARCH

*Reports of the 4th year research projects
in the Department of Chemical
Engineering at Imperial College London*

VOLUME 6



Edited by Erich A. Müller

February, 2024

CHEMICAL ENGINEERING RESEARCH

Reports of the 4th year research projects
in the Department of Chemical Engineering
at Imperial College London

Edited by Erich A. Müller

Volume 6

2024



© The Author(s) 2024.

Published by Imperial College London, London SW7 2AZ, UK

Contact details – e.muller@imperial.ac.uk

This book is a compilation of manuscripts created as part of a teaching assignment on the 4th year's Chemical Engineering CENG70001 course (Advanced Chemical Engineering Practice Research Project). Copyright in each paper rests with its authors. The author of each paper appearing in this book is solely responsible for the content thereof; the inclusion of a paper in the book shall not constitute or be deemed to constitute any representation by Imperial College London that the data presented therein are correct or sufficient to support the conclusions reached or that the experiment design or methodology is adequate.



The book is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0). Under this licence, you may copy and redistribute the material in any medium or format on the condition that: you credit the author, do not use it for commercial purposes and do not distribute modified versions of the work. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law. <https://creativecommons.org/licenses/by-nc-nd/3.0/>.

The information contained in this publication is being distributed without warranty of any kind, either expressed or implied. The responsibility for the interpretation and use of the material lies with the reader. In no event shall Imperial College London be liable for damages arising from its use.

E-ISBN 978-1-9160050-5-1

First published in 2024

Preface

This volume of *Chemical Engineering Research* collects the unedited research project reports written by 4th year undergraduates (Class of 2024) of the M.Eng. course on Chemical Engineering in the Department of Chemical Engineering at Imperial College London. The research project spans one term (Autumn) during the last year of the career. It emphasises independence, the ability to plan and pursue original project work for an extended period, produce a high-quality report, and present the work to an audience using appropriate visual aids. Students are also expected to produce a literature survey and to place their work in the context of prior art. The papers presented showcase the diversity and depth of some of the research streams in the department but only touch on a small number of research groups and interests. For a complete description of the research at the department, the reader is referred to the departmental website¹.

The papers presented are in no particular order, and a manuscript number identifies them. Some papers refer to appendixes and/or supplementary information which are too lengthy to include. These files are available directly from the supervisors (see supervisor index at the end of the book). Some reports are missing and being embargoed, as they contain confidential information. A few reports correspond to industrial internships, called LINK projects, in collaboration with Shell.

The cover figure corresponds to a photograph of a flow reactor for water splitting (taken from the work of Konrad Reents and Alexander Kovacs, manuscript 62).

London, February 2024

¹ <https://www.imperial.ac.uk/chemical-engineering>

Title Index

paper	Title	page
1	PrCa(5%)FeO ₃ Photocathodes Optimised Through Hole Transport Layers and Pt Catalyst	1
2	Viscoelasticity and Extensional Rheology of Concentrated Wormlike Micelles Solution	14
3	Analysis of Carbon Capture Readiness for Small-Scale Refuse Derived Fuel-to-Energy Power Plants based in the UK	22
4	Exploring Multi-Fidelity Bayesian Optimization and TuRBO-1 for Enhanced Engineering Solutions	32
5	Stability Study of Dual Drug Delivery Systems under Osmotic Stress	42
6	Enviro-Economic Analysis of Refrigeration Cycle Integration into Ground-Source Heat Pump-Supported Space Heating Systems	52
7	A Techno-Economic Analysis and Systematic Review of Blue and Green Hydrogen Production Technologies	62
8	Comparative Thermodynamic Efficiency Analysis of Acetylene-Ethane Separation Using Distillation and Absorption Process	(*)
9	Effects of Salts on Occurrence Domains of Triglycine Anhydrate and Dihydrate	72
10	A Techno-Economic Analysis of a Novel Process to Treat Pot Ale into Hexanoic Acid	80
11	Enviro-Economic Assessment of a Scaled-Up Hydrogenolysis Process for the Treatment of Polypropylene Waste	90
12	In Situ DRIFTS Investigation of CO ₂ Adsorption & Desorption on Carbon Nitride Based Materials	(*)
13	Prediction of Thermodynamic Properties and Phase Behaviour of CANDU Nuclear Reactor Fluid Coolant using the SAFT-VR Mie Equation of State	100

paper	Title	page
14	Electrochemical Reduction of CO ₂ : Insights into Cobalt Single-Atom Catalysts via a Decoupled Two-Step Synthesis	110
15	Data-driven Modelling and Prediction of Complex Systems Using Neural ODEs	120
16	Tabular and Deep Q-Learning for Optimal Control of a Commercial HVAC System	130
17	A Flexible Calcium Ion Holographic Sensor for Wound Monitoring via Smartphone Readout	140
18	Effect of 2D thickness on the performance of 2D/3D organic-inorganic metal halide perovskite solar cells	150
19	CO ₂ Capture Using Adsorption: an Outreach Project	160
20	Neural networks to simulate and optimise a Pressure-Vacuum Swing Adsorption process	170
21	Cold Chain Integration of Liquefied Natural Gas Supply Chains	180
22	A Machine Learning Platform for the Optimisation and Innovation of Ionizable lipids for Efficient RNA Delivery	190
23	Turning a new page on PAGE: Investigating the effect of oligonucleotide structure on gel mobility	199
24	Modelling the Solubility of Cholesterol in Primary Alcohols using the SAFT- γ Mie Equation of State	209
25	Optimisation of Ionic Liquids in a Closed-Loop Dye Recycling Process	219
26	Engineering Magnetically Steerable Biohybrid Cells	229
27	Data driven modelling using time series recurrent neural networks (RNN) for glycosylation prediction in mAbs	239
28	Optimizing Crude Distillation Units: An Exploration of Neural Network Surrogates and Evolutionary Algorithms	(*)

paper	Title	page
29	The Role of Ground Source Heat Pumps in UK Domestic Heat Decarbonisation	248
30	Experimental Design to Investigate Aqueous Amino-Acid Solvents for CO ₂ Capture	258
31	Recovery of Materials (Li, Mn, Ni, Co) from Lithium-Ion Battery Cathode	268
32	Derivative-free Optimisation of Neural Networks in Reinforcement Learning for Process Control	278
33	Characterisation of a Complex Mixture of Tri-, Di- and Monoacylglycerols from Ethanolysis of Sunflower Oil by NMR Spectroscopic Techniques	288
34	Prediction of Aqueous Solubility of Polycyclic Aromatic Hydrocarbons and their Derivatives by ML-QSPR Modelling	298
35	The Development of Criteria for Predicting Breakthrough and Optimising Adsorbent Use in Gas Phase Carbon Dioxide Adsorption	(*)
36	Development of a Unified Kinetic Model for the Hydrothermal Carbonisation (HTC) of Microalgal Biomass	308
37	Feasibility Analysis of Metal-Organic Frameworks (MOFs) and Zeolites for Direct Air Capture (DAC)	(*)
38	Machine Learned Equation of State for the 2D Lennard-Jones Fluid	318
39	Optimising Detergent Formulation: A Bi-Objective Computer-Aided Molecular Design Approach	328
40	Recovery of 5-(hydroxymethyl)furfural (HMF) from Effluents	338
41	Numerical Modelling and Performance Assessment of Spectral Beam Splitting Based Concentrated Photovoltaic-Thermal Collector	348
42	The Effect of Hydrodynamics on the Transesterification of Sunflower Oil to Produce FAME	358
43	Screening the chemical space: An ML Approach to Predicting the Prices of Chemical Compounds	368

paper	Title	page
44	Utilizing the SAFT- γ Mie GC Equation of State to Assess the pH-Dependent Solubility of Active Pharmaceutical Ingredients	377
45	Techno-economic Analysis for the ZESTY Process in Sweden, UK and USA	(*)
46	On the Solubility and Recovery of Gypsum with Ionic Liquids	387
47	Lead-free halide Ternary Perovskite Composites for CO ₂ Photocatalytic Reduction to Solar Fuels	397
48	Environmental Impact Assessment of Sustainable Aviation Fuels Against Planetary Boundaries	408
49	Superhydrophobic Cotton: Fabrication and Application in Oil/Water Separation	416
50	Investigating the Electrochemical Behaviour of Graphitic Carbon Cathodes in Aluminium Dual-ion Batteries	426
51	CFD Modelling of Air Entrainment Mechanisms in a Plunging Jet	436
52	Filling in the Cracks: An Investigation into Surface Patterning via Wrinkling	446
53	Exploring the Influences of Impurities and Silica Nano-templates on Diglycine Crystallisation: A Comprehensive Study for Innovative Crystal Engineering	456
54	Catalytic Performances of UiO-66(Hf) under Various Synthesis Conditions for the Methylation of DHA	466
55	Interpretable Supply Chain Optimisation for Inventory Management Problems with Genetic Decision Trees	476
56	In-Situ FTIR Spectroscopy of Interactions Between High-Pressure CO ₂ and Porous Liquids	486
57	High-Flux Ethanol-Water Separation via Mildly Reduced Graphene Oxide Membranes	496
58	Assessing Pyrolytic Carbon Derived from Methane Pyrolysis as an Anode Material	506

paper	Title	page
59	Using Agent-Based Modelling to Investigate Effects of the Socioeconomic Climate on the UK Power Sector	516
60	Direct Visualisation of Surfactant Flooding in Micromodels	526
61	Ultrathin Graphene Oxide Based Membranes with Tailored Graphitic Domain for Organic Solvent Nanofiltration	536
62	Operation and Modelling of a New Reactor for Solar Water Splitting	546
63	Scaling up <i>in vitro</i> glycosylation reactions in cell-free systems using filtration: a preliminary assessment	556
64	Machine Learning Approaches for Periodic Separation Systems Modelling	(*)
65	A Comparative Study of the Pure Gas Permeability of PIM-1 Membranes and Other Polymers for CO ₂ Separation	565
66	Model-based Design Space and Flexibility Analysis for Carbon Capture Adsorbent Screening	575
67	Support Vector Machines Practice on Design Space Identification	585
68	Impact of Predicted Data on ML-QSPR Predictions of Lower Flammability Limits for Pure Compounds	595
69	Techno-economic Assessment of a Novel Hybrid PV-T and Heat Pump System for Household Heating	605
70	Supply Chain Optimisation for Plasmid DNA	615
71	Analysis of Morphology and Microstructure of the Lignin-derived Mesoporous Anode for Sodium-ion Batteries and Sodium Storage Mechanism	625
72	Utilising Graph Neural Networks in the Glycomic Analysis of N-Glycan Biomarkers for the Diagnosis of Colorectal Cancer	634
73	Differentiable Equations of State for Machine Learning Thermodynamic-Property Prediction	644

paper

Title

page

(*) These papers have been removed by request of the authors and/or supervisors

Author and Supervisor index at the end of the book.

PrCa(5%)FeO₃ Photocathodes Optimised Through Hole Transport Layers and Pt Catalyst

Di Wen, Zhenran Zhang

Department of Chemical Engineering, Imperial College London, U.K.

Abstract

There is continuously increasing demand for green, safe, and efficient energy across the world since many countries and multinational companies have committed themselves to a net zero pathway. Solar energy is an optimal substitute for the conventional fossil fuel where Solar-to-hydrogen (STH) conversion offers a reliable storage method. PEC water splitting, a highly efficient and cost-effective method of generating hydrogen, is investigated in this article. A state-of-the-art perovskite oxide photocathode based on Ca-doped PrFeO₃ is developed and optimized using a number of hole transport layers (HTLs) and Pt catalyst. The FTO / cNiO_x (annealed at 400 °C) / PrCa(5%)FeO₃ / 3-layer Pt (deposited at 400 °C) gives the best photocurrent of ~21 $\mu\text{A cm}^{-2}$ when testing PEC performance under the N₂ environment. Data of multiple characterization methods including SEM, UV-Vis, and XRD managed to justify the experimental results. Overall, the cNiO_x layer is an effective HTL when PEC measurements are carried out in atmospheric condition whereas it fails to reduce charge recombination in the hydrogen evolution reaction (HER). When the cNiO_x is annealed at 600 °C, it loses its function as an HTL. When Pt is deposited onto PFO as a catalyst at 100 °C, it can effectively improve the selectivity of HER.

Keywords: Solar-to-hydrogen, PEC water splitting, perovskite oxide, Ca-doped PrFeO₃

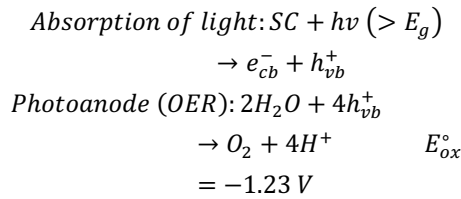
Introduction

Ever since the Paris Agreement, over 140 countries have set up their net zero targets.¹ Various laws and policies have been laid out. The demand for decarbonisation has prompted huge interest in the development of low carbon technologies and renewable substitutes of fossil fuels. Solar energy is a potential solution to the continuously increasing energy needs. However, the utilisation of solar radiation is severely hindered by its intermittent nature. An efficient and safe storage method is required to retain the excess solar energy generated in the daytime.² Solar-to-hydrogen (STH) energy conversion has been regarded as a promising method to store solar energy through water splitting reaction.^{3,4} A high purity of H₂ can be obtained since H₂ and O₂ are readily separated through water decomposition. The green hydrogen H₂ itself is also an efficient and renewable fuel which generates zero

carbon emissions. It demonstrates a superior gravimetric energy of 120 MJ/kg comparing to that of gasoline (44 MJ/kg).⁵ (Hydrogen has a higher gravimetric heating value (141.9 MJ kg⁻¹) than most of the conventional fossil fuels (methane 55.5 MJ kg⁻¹, gasoline 47.5 MJ kg⁻¹, diesel 44.8 MJ kg⁻¹, and methanol 20.0 MJ kg⁻¹) There are three main approaches to STH energy conversion via water splitting, which are the photovoltaic-electrolysis (PV-EC), photocatalytic (PC), and photoelectrochemical (PEC) ways As a highly developed technology, PV-EC system has already been partially commercialised. Among the three technologies, PEC cell is ranked in the middle in terms of overall efficiency, complexity, and choice of material.⁶ PEC water splitting thus is not only highly efficient but also relatively simple and cost-effective. This article gives more insights into the STH energy conversion using PEC water splitting.

Background

According to the Nernst Equation, water is converted into oxygen and hydrogen (i.e. Gibbs energy = 237.2 kJ/mol) when a minimum energy of 1.23 eV is applied under atmospheric temperature and pressure (i.e. 298 K and 1 bar). The solar irradiance with a wavelength of around 1000 nm provides the same amount of energy.^{2,6,7} Nevertheless, an energy greater than the theoretical minimum is required to drive the reaction in practice as a result of the energy loss in PEC water splitting.² The energy losses accounts for about 0.8 eV, including the potential loss because of the electrode and contact resistances as well as the electron-hole recombination. Thus, in practice, an energy of ~2.0 eV is needed to initiate PEC water splitting.⁸ The reactions for water splitting are shown below.



The STH conversion efficiency (Z_{STH}) is normally employed to quantify the PEC performance of solar cells.⁷⁻¹⁴ By definition, the STH efficiency is the amount of chemical (H_2) energy generated per unit

$$\begin{aligned}
 \eta_{STH}(\%) &= \left[\frac{\text{Chemical energy produced}}{\text{Solar energy input}} \right] \\
 &= \left[\frac{\text{Rate of } H_2 \text{ Production} \times \Delta G_{H_2O \rightarrow H_2 + \frac{1}{2}O_2}}{\text{Total incident solar power} \times \text{Electrode Area}} \right] \\
 &= \left[\frac{(\text{mmol } H_2 \text{ per s}) \times (237\,000 \text{ J mol}^{-1})}{P_{Total}(\text{mW cm}^{-2}) \times \text{Area}(\text{cm}^2)} \right]_{AM\,1.5G} \\
 \eta_{STH}(\%) &= \left[\frac{J_{sc}(\text{mA cm}^{-2}) \times (1.23 \text{ V}) \times \eta_F}{P_{Total}(\text{mW cm}^2)} \right]_{AM\,1.5G}
 \end{aligned}$$

Where J_{sc} is the generated photocurrent density, η_F is the Faradaic efficiency of O_2 or H_2 production (i.e. the efficiency of holes and electrons contributing to OER or HER)

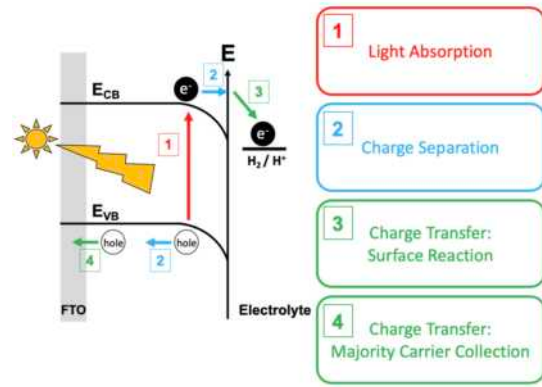
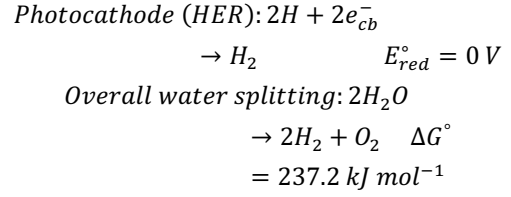


Figure 1 :PEC water splitting mechanism

A photocatalyst must have a band gap energy of ~2.0 eV to decompose water due to the significant overpotentials of the two half equations. electrons (e_{cb}) and holes (h_{vb}^+) are generated by solar irradiation and drive the overall reaction.⁶

incident solar energy. It is obtained under normalised solar irradiance with a value of one sun (100 mW/cm²).^{8,14} The Air Mass 1.5 global (AM 1.5 G) filter is normally used.

N-type metal oxide and p-type non-oxide photocathodes have been intensively visited in the field of PEC water splitting. For example, the studies on using materials including Si, GaP, and InGaP as photocathodes are well-established.¹⁵ However, the research on the novel p-type metal oxide semiconductor photocathodes is relatively limited, most of which focus on the Cu-based photocathodes. The p-type Cu-based metal oxides are regarded as reliable photocathodes due to their wide bandgaps as well as favourable band edges correlated to the water splitting redox couples. Nevertheless, the potential of binary and ternary copper-based oxides is limited by the chemical instability against reduction and non-ideal

optoelectronic properties. With buried p-n junctions, protective layers, and nanostructures, Cuprous oxide Cu_2O possesses promising PEC performance.¹⁶ It has a band gap of 2.0 eV and gives a theoretical Z_{STH} of 18%.⁶

PrFeO_3 is a state-of-the-art metal oxide photocathode. Perovskite oxide has favourable band gaps for solar illumination absorption and stability for aqueous applications.¹⁷ However, its PEC performance is still hindered by the hole-electron recombination and high overpotential.

Aim

The main objective of this project is to optimize the calcium doped (5%) praseodymium orthoferrite PrFeO_3 (PFO). It is made up of two aspects, in which the first one is to reduce the recombination of charge carriers (i.e. holes and electrons) using a range of hole transport layers (HTLs). The other aspect is to boost the selectivity and hence the PEC performance using platinum as a photocatalyst. This project aims to give a potential option for an efficient, robust and environmentally friendly PFO-based photocathode for PEC water splitting.

Methodology

Fluorine-doped tin oxide (FTO) substrate preparation

FTO glass substrates were cut to 2.7 cm x 1.5 cm and placed in a staining jar. The substrates were cleaned by the solution of Hellmanex detergent in deionised (DI) water, under ultrasonic water bath for 10 minutes. To avoid the contamination by detergent, the substrates were then rinsed by DI water for 10 times to ensure that there is no new bubble formed. The substrates were subsequently treated with acetone and isopropanol for 10 minutes of ultrasonification with each solution. After drying carefully with the radiation of hot plate (80°C), a further 20 minutes of UV-Ozone treatment was carried out to increase the wettability of the FTO surface for the immediately following spin coating

steps. An ohmmeter was used to check the side of FTO with non-zero resistance, and the FTO side was put upwards.

Compact NiO_x layer (cNiO_x)

0.01 g of Solaronix Ni-Nanoxide slurry (nickel oxide nanoparticle paste) was dissolved in 1g (equivalent to 1260 μL) ethanol and vigorously stirred for 20 mins. Spin coating was carried out on the FTO side at 2000 rpm with 2000 rpm/s acceleration for 30 seconds with 0.05ml solution. 3.5 bar of N_2 was used for vacuum for the spin coater. Annealing temperatures of 400°C and 600°C were both investigated. The maximum temperature of hot plate was 500°C. Therefore, the films were heated on the hot plate of 400°C for 30 minutes when the annealing temperature condition was set to 400°C. For the temperature condition of 600°C, these films were calcinated in the tube furnace for 30 minutes.

Mesoporous NiO_x layer (mpNiO_x)

0.1 g of Solaronix Ni-Nanoxide slurry was dissolved in 0.5 g ethanol and mixed with the solution prepared by 0.5 g ethanol and a varied mass of Triton X-100 (TX100), to achieve different overall mass ratio of 1:1, 1:2, 1:5 and 1:10 for TX100: ethanol. The mixture was stirred for 20 minutes. This was followed by spin coating of the resulting mixture on the annealed cNiO_x layer at 2000 rpm with 2000 rpm/s acceleration for 30 seconds with 0.05ml solution. The annealing temperature and time were 400°C and 30 minutes.

Compact MoO_x layer (cMoO_x)

10 nm of MoO_x layer was deposited by thermal evaporation on the cleaned FTO substrate.

PrCa(5%) FeO_3 layer

4 ml Tetrahydrofuran (THF) was extracted by syringe with needle in nitrogen environment, which was added to 2ml TX100 and stirred for 1 hour. Simultaneously, 0.9 g citric acid powder, 0.5 g $\text{Fe}(\text{NO}_3)_3 \cdot 9\text{H}_2\text{O}$ and 0.0146 g $\text{Ca}(\text{NO}_3)_2 \cdot 4\text{H}_2\text{O}$ solids were mixed with 0.51 g $\text{Pr}(\text{NO}_3)_3 \cdot 6\text{H}_2\text{O}$ solid.

2ml DI water was added to the solid mixture immediately to prevent the change in the composition of the hydrates and the solution was stirred for 1 hour. 4ml of the polymer solution were then added to the inorganic solution and stirred for 3 hours. 0.05ml solution was used per each sample for spin coating at 2000rpm with 2000 rpm/s acceleration for 30 seconds. The coated films were calcinated in three different temperatures (600°C, 700°C and 800°C) in the tube furnace for 2 hours. This method was adapted from Freeman et al.¹⁷

Platinum nanoparticles

8.3mg K₂PtCl₄ and 223.5mg trisodium citrate were dissolved in 20 ml H₂O. 0.6 ml of 10mM NaBH₄ solution was then added to the solution as a reducing agent and form Pt nanoparticles.¹⁸ A 200nm filter was used on the syringe to filtrate the large Pt particles or agglomerates. Spin coating was performed subsequently using 0.02ml at different spin speed (2000rpm, 3000rpm and 6000rpm) and evaporating the solvent at two different temperatures (400°C and 100°C) for 30 minutes.¹⁸ The coating procedure was repeated for adding more Pt layers.

Characterisation methods

Photoelectrochemical measurements (PEC)

A PEC cell was set up with a three-electrode configuration consisted of an Ag/AgCl reference electrode, a Pt counter electrode and a working electrode. The tips of the reference electrode and counter electrode were immersed in 0.1M Na₂SO₄ aqueous electrolyte with pH 12, after adding NaOH solution of pH 14 to tune the pH value. The glass side of the sample (without coated layers) was faced to the light source while the other side was in contact with the electrolyte. A 0.28 cm² mask was used to control the illumination area and simulated sunlight was introduced by the LOT Quantum Design lamp with the filter to control the light intensity to 100 mWcm⁻² (AM 1.5).

To be able to compare the PEC performances in different cell conditions, Nernst Equation at room temperature and pressure was applied to convert the potentials to reversible hydrogen electrode (RHE):

$$E_{RHE} = E_{Ag/AgCl} + 0.059 * pH + 0.197$$

The IVIUM potentiostat was connected to the IVIUM software and the light was chopped at a rate of 2 second. The applied external potential was swept from +1.4V to -0.6V V_{RHE} linearly at a scan rate of 10mVs⁻¹.

Ultraviolet-Visible Spectroscopy (UV-Vis)

The UV-Vis spectroscopy was investigated through a Shimadzu 2500I spectrophotometer. The Kubella-Munk function was evaluated for wavelengths from 300 to 800 nm.

X-Ray Diffraction (XRD)

X-ray diffraction (XRD) was carried out to analyse the crystalline structure of perovskite layers. A PANalytical X'Pert Diffractometer (Cu Ka, $\lambda = 1.54$ Å) was employed at 40 kV and 40 mA. Diffraction patterns were taken for 2θ values from 20 – 80° in a slope of 0.0170° and then processed by Highscore software.

Scanning Electron Microscopy (SEM)

Field Emission - Scanning Electron Microscopy (FE-SEM) was investigated using ZEISS LEO 1525 with an accelerating voltage of 3KeV.

Results and discussion

PEC measurements

Effect of cNiO_x and mpNiO_x HTLs

In order to reduce the electron-hole recombination and to increase the photocurrent density, the performance of HTLs were investigated by PEC measurements with the photoactive layer, 5% calcium doped PFO.

The mass ratio of the surfactants to ethanol was varied in the mpNiO_x layer with the structure of FTO/cNiO_x/mpNiO_x/PFO and the FTO/PFO configuration without the hole transport layers was used as a reference. Figure 2(a) and 2(b) showed that mpNiO_x made with a 1:5 or 1:10 mass ratio between polymer TX100 and EtOH demonstrated the best PEC performance in air while there was no photocurrent can be observed when the ratio reached 1:2 and 1:1. Figure 2(b) illustrated the photocurrent density when applied a flat baseline to Figure 2(a) (i.e. the total current density subtracting the dark current density at 0.43V V_{RHE})

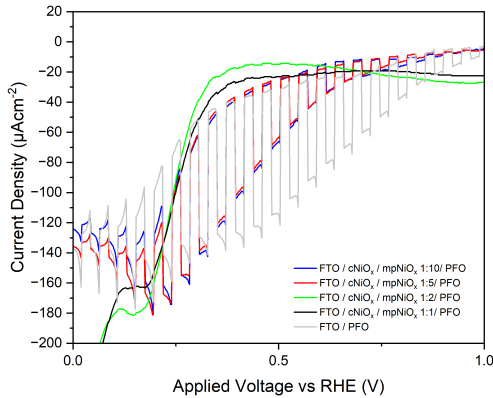


Figure 2(a): mpNiO_x made with different mass ratio between polymer TX100 and EtOH

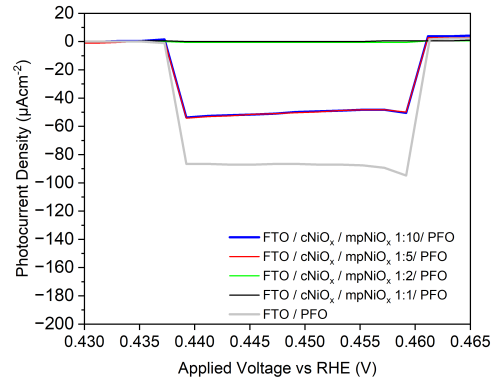


Figure 2(b): mpNiO_x made with different mass ratio between polymer TX100 and EtOH with flat baseline

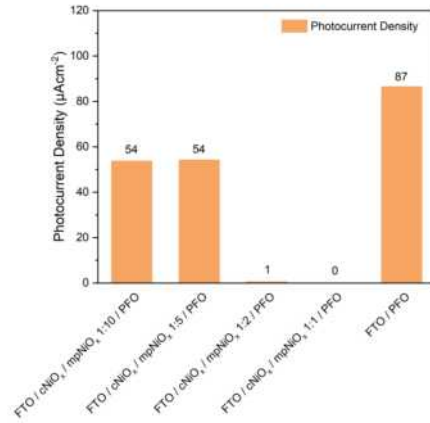


Figure 2(c): Bar chart of mpNiO_x made with different mass ratio between polymer TX100 and EtOH

Figure 2(c) demonstrated the best photocurrent with mpNiO_x (54 μAcm⁻²) is still much lower compared to the sample without the hole transport layer (87 μAcm⁻²). Nevertheless, the effect of cNiO_x as HTL alone was required to be investigated to conclude the effects of mpNiO_x on the performance. Figure 3(a) and 3(b) illustrated the mpNiO_x layer reduced the electron-hole recombination as well as the photocurrent density. The trend was clearer from figure 3(c), adding a cNiO_x layer to the structure improved the photocurrent density in air, from 87 μAcm⁻² to 106 μAcm⁻², while further coating a 1:10 mpNiO_x layer on cNiO_x layer reduced it to 54 μAcm⁻².

². Therefore, mpNiO_x was not a desired HTL in that configuration and cNiO_x was promising.

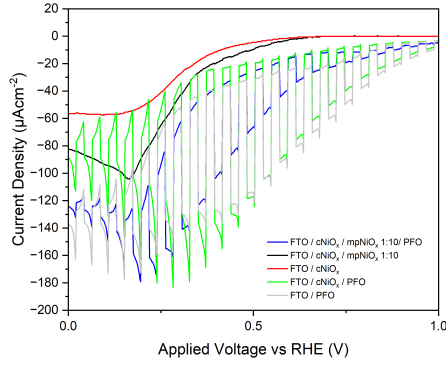


Figure 3(a): mpNiO_x and cNiO_x comparisons

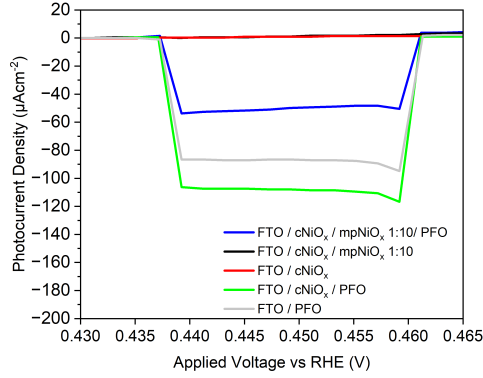


Figure 3(b): mpNiO_x and cNiO_x comparisons with flat baseline

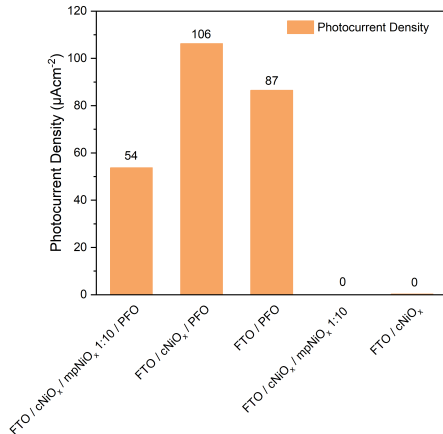


Figure 3(c): Bar chart of mpNiO_x and cNiO_x comparisons

Effect of cMoO_x HTL

The structure with a thin layer of 10nm cMoO_x as HTL was investigated to compare its PEC measurements with the cNiO_x layer. The results in figure 4 indicated that the poor performance of the cMoO_x layer as the photocurrent density was limited within 2 μAcm⁻², which was significantly lower than the samples coated with cNiO_x and might be due to incompatible structure with PFO.

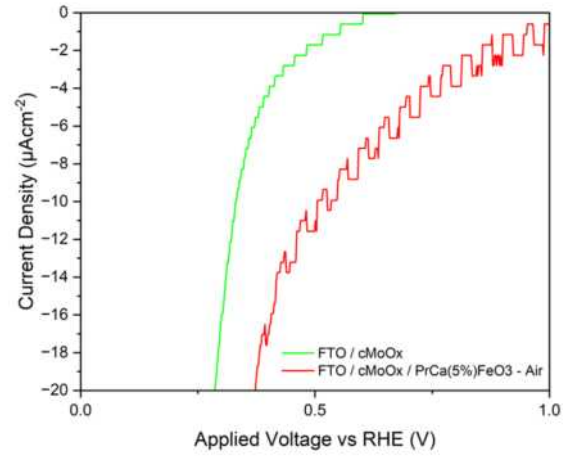


Figure 4: cMoO_x PEC performance

Effect of calcination temperature of the photoactive layer

Temperature of calcination of the photoactive layer, 5% calcium doped PFO, was studied to optimise the configuration and to improve photocurrent density. From figure 5(a) and 5(b), with interested voltage 0.43V vs. RHE, 600°C was the optimal temperature since it demonstrated the highest photocurrent density with 106 μAcm⁻² and it was the lowest temperature which saved the energy.¹⁷ Apart from the optimal temperature, the photocurrent density with calcination temperature of 800 °C was almost zero. A possible reason for the PEC performance was the cNiO_x cannot withstand the temperature of 800 °C and form cracks, which required further justification from SEM results.

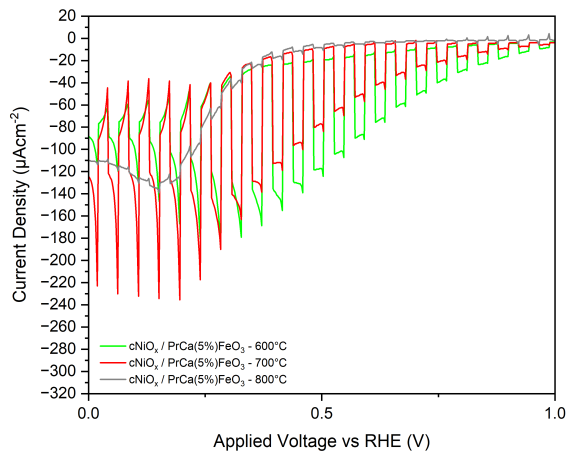


Figure 5(a): Change temperature for calcination

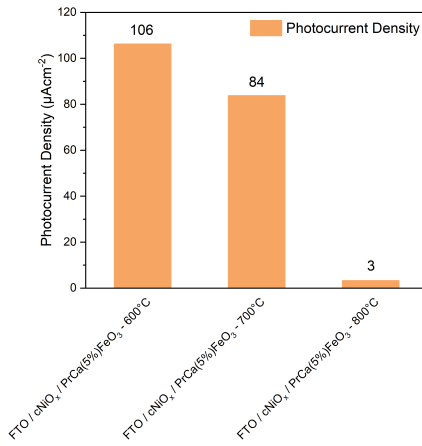


Figure 5(b): Bar chart for change temperature for calcination

Effect of spin coating speed for Pt

The spin coating speed would influence the thickness of the coating layers and thus affected the distribution pattern of Pt nanoparticles. From Figure 6(a) and 6(b), The spin coating speed of 3000 rpm showed a 25% improvement to the photocurrent density compared to 2000 rpm and slightly reduced the electron-hole recombination. The photocurrent density by using 6000 rpm was almost identical to 3000 rpm and therefore 3000 rpm was chosen as it was more energy efficient.

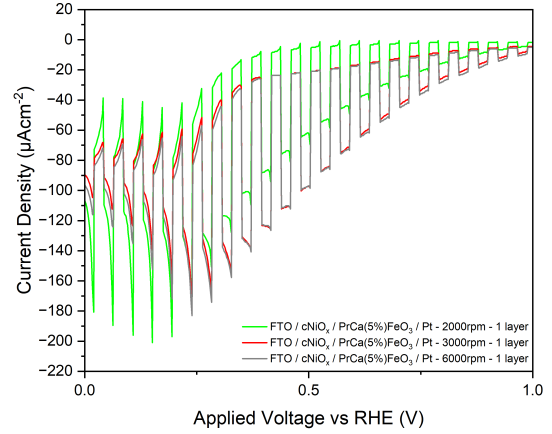


Figure 6: Different spin speed

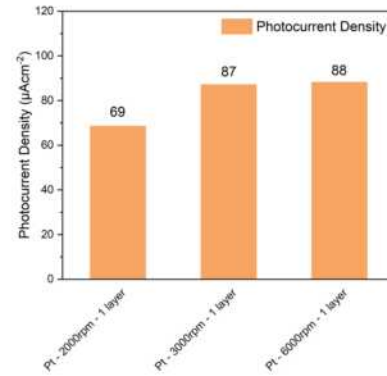


Figure 6: Bar chart for different spin speed

Effect of number of coating layers of Pt

To coat the Pt nanoparticles more evenly with more nanoparticles on the PFO surface, the number of coating layers of Pt was varied and studied. Figure 7 illustrated the photocurrent density in air increased as the number of coating layers increased, with 27% improvement from 1 layer to 2 layers of Pt and a

further 10% improvement from 2 layer to 3 layers.

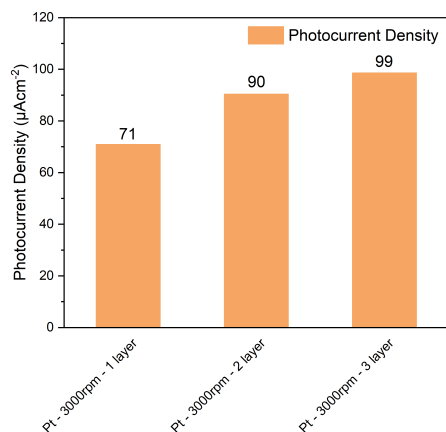


Figure 7: Bar chart for different coating layers

PEC measurements in nitrogen environment were performed to exclude the photocurrent density of all the other reactions including oxygen reduction reaction (ORR) and demonstrated the selectivity to hydrogen evolution. The trend in nitrogen environment was still hold and can be concluded from Figure 8. However, there was no clear improvement after adding the Pt layers since difference between the photocurrent with Pt layers and without Pt layer was within the range of experimental error.

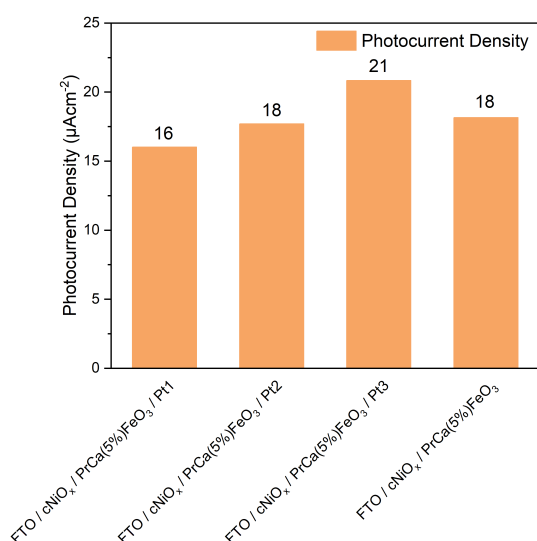


Figure 8: Bar chart for different coating layers in N₂

To improve the performance of Pt layers, the temperature to evaporate the water after spin coating was lower to 100°C to reduce the effect of Ostwald ripening and the agglomerations. The temperature of cNiO_x annealing was increased to 600°C to be consistent with the calcination temperature of 5% Ca doped PFO layer simultaneously. The PEC results under nitrogen environment were shown on Figure 9. Compared the photocurrent density of sample 1,2,3 with sample 5 in Figure 9, it illustrated Pt did improve the photocurrent since there was an over 50% improvement on the photocurrent density from 6 to 9 μAcm⁻² at least. In addition, after changing to the new temperatures, the 2 layers of coating of Pt demonstrated the best performance. The cNiO_x might not be effective since there is no obvious difference between the results of sample 4 and sample 6.

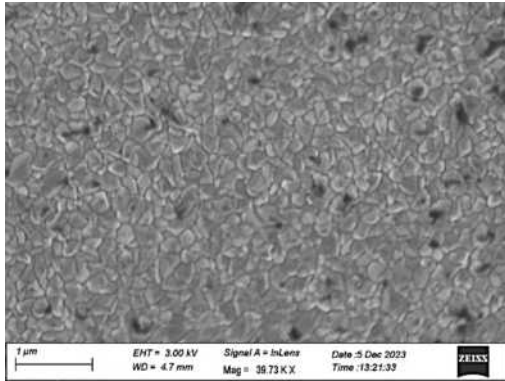
Optimal photocathodes

A collection of the optimal photocathodes under N₂ condition with FTO/PFO and FTO/cNiO_x/PFO as reference was shown in Figure 10(a), 10(b) and 10(c). The best photocurrent density was achieved by the sample with lower temperature (100°C) to evaporate the solvent and higher temperature (600°C) to anneal the cNiO_x. It could be explained by even though in the lower temperature, the performance of Pt was improved and resulted higher selectivity, the cNiO_x might lost its function as HTL to reduce the electron-hole recombination.

Characterisation Results

SEM Results

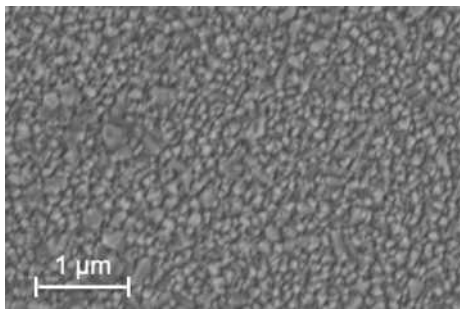
FTO



The image demonstrates the scanning electron microscope (SEM) result for FTO-coated glass. The marble-like pattern of FTO as well as the grain boundaries are shown. FTO has an average crystallite size of $\sim 0.2 \mu\text{m}$.

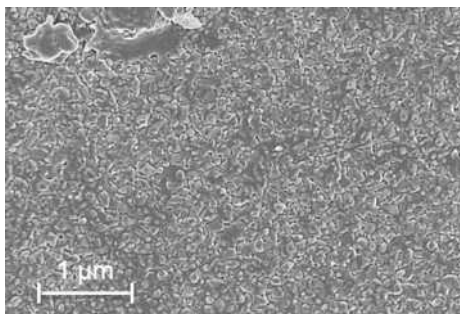
FTO/cNiO_x

cNiO_x annealed at 400 °C



The SEM image of FTO/cNiO_x composite has a similar pattern as the bare FTO one. However, it has a smaller crystallite size than that of FTO.

cNiO_x annealed at 600 °C

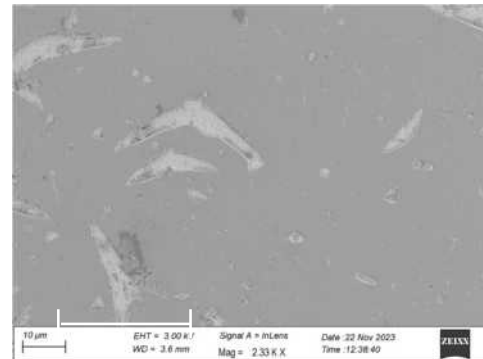


The SEM image of FTO/cNiO_x composite demonstrates a similar morphology as the former one. The only difference in the preparation of sample is the annealing temperature of cNiO_x is increased from 400 to 600 °C. cNiO_x particles start

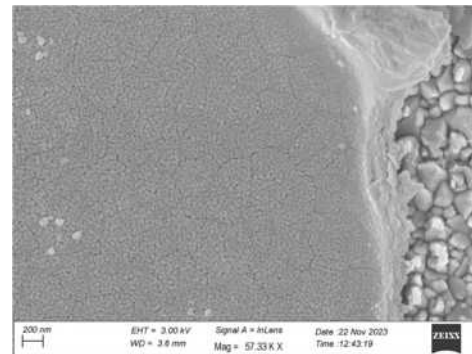
to agglomerate at a higher temperature due to the Ostwald ripening. As a result, cNiO_x loses its preferred function as a HTL and fails to reduce charge carrier recombination. Moreover, cNiO_x might diffuse into the PFO layer and thus deactivate PFO. Therefore, the PEC performance for samples with cNiO_x (600 °C) is non-ideal.

FTO/cNiO_x/PFO/2-layer Pt

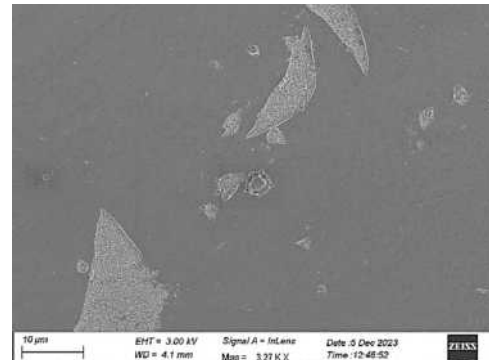
Pt deposited at 400 °C



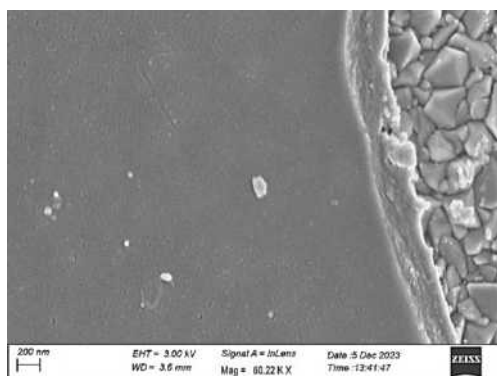
Pt deposited at 400 °C



Pt deposited at 100 °C



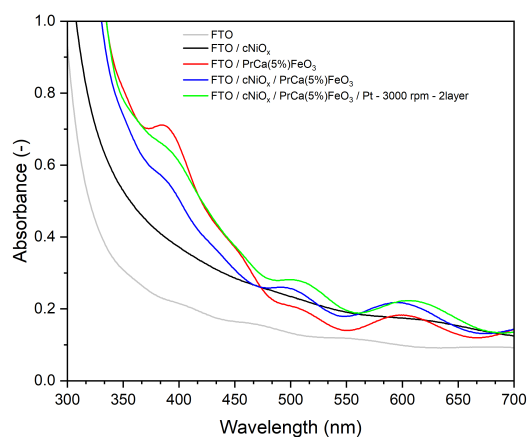
Pt deposited at 100 °C



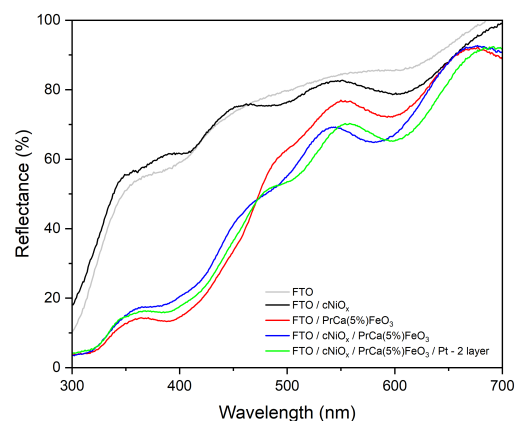
The four images from **figure X to X** refer to the FTO/cNiO_x/PFO/2-layer Pt photocathode which gives ideal PEC performance. As these images shown, there are more cracks on the surface with Pt (400 °C). Also, more Pt agglomerations can be observed with a higher deposition temperature for Pt. The lower layers fail to mechanically support the PFO and Pt nanoparticles are spread less evenly in **figure X and X**. Therefore, the samples with Pt (100 °C) demonstrate better PEC performance.

UV-Vis Results

Absorbance

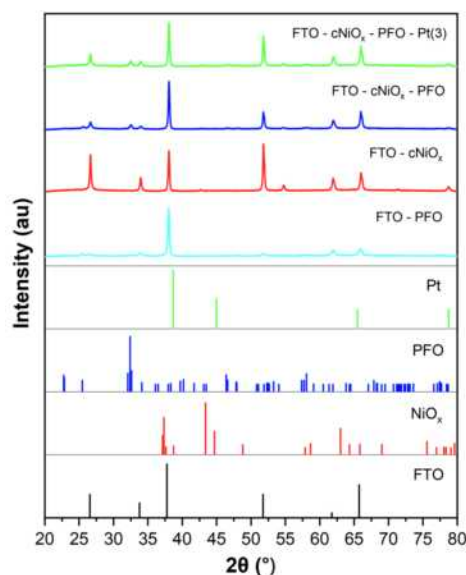


Reflectance



The absorbance and reflectance characteristics were investigated using Ultraviolet-visible (UV-Vis) spectroscopy. Absorbance increases as the wavelength of incident light decreases from 700 to 300 nm. Absorbance is inversely proportional to the reflectance which is desired. The absorption of light rapidly increase as the curve enters the ultraviolet region (wavelength between 100 and 400 nm). The fluctuation of curves possibly indicates the iridescence which is the phenomenon gradually changing colour due to the change of angle of illumination. It is also partially because of the intrinsic wave pattern of the incident light.

XRD Results



X-ray diffraction was investigated for samples with different layers. The XRD data and reference data for each material were analysed and plotted using OriginLab 2022. Then all the plots are combined and shown in the XRD graph. There is not much difference between the plots generated by four different samples. The six reference peaks of FTO are clearly demonstrated on all four plots. The reference peaks for NiO_x and PFO are partially shown (i.e. NiO_x: 2θ = ~43.0°, ~78.0°; PFO: 2θ = ~25.5°, ~32.5°, ~57.5°) while the Pt peaks cannot be easily identified. The sharp reflections of FTO peaks dominate which indicates the high crystallinity and possible desired orientation effects. The sharp peaks of FTO indicate a low Full Width at Half Maximum (FWHM) value. Since the crystallite size of a material is inversely proportional to its FWHM value, it can be concluded that FTO has a relatively large crystallite size which is consistent with our SEM result.¹⁹ On the contrary, tiny Pt nanoparticles with diameters less than 50 nm result in a large FWHM value and hence negligible peaks. The XRD patterns of FTO – PFO, FTO – cNiO_x – PFO, and FTO – cNiO_x – PFO – Pt(3) composites all partially match with the diffraction pattern of PFO with JCPDS no. 00-047-0065. Similarly, the peaks of all three samples with cNiO_x layer partially match with the known cubic phase of NiO (ICSD: 024014) and rhombohedral phase of NiO₂ (ICSD: 078698).^{20, 21} It demonstrates that PFO, NiO, and NiO₂ might not have the crystallinity as high as that of FTO.

Conclusion

This study has investigated the effects of the cNiO_x, mpNiO_x and cMoO_x hole transport layers, the calcination temperature of calcium doped PFO, spin coating speed, number of coating layers of platinum with different temperatures to evaporate the solvent of Pt nanoparticle solution and presented a collection of the optimal photocathodes.

In conclusion, the cNiO_x layer is an effective HTL when PEC measurements are performed in atmospheric condition, while it fails to reduce charge recombination in the HER. mpNiO_x and

cMoO_x layers are incompatible with PFO at 600°C and gives poor PEC performance. However, when the cNiO_x is annealed at 600 °C, it loses its function as an HTL as well. For tuning the temperature of evaporating the solvent in Pt coated film at 100 °C, it can effectively improve the selectivity of HER. The structure of FTO / cNiO_x (400 °C) / PrCa(5%)FeO₃ / 3-layer Pt (400 °C) gives the best PEC performance under the N₂ environment among all the tested configurations with a photocurrent of 21 μA cm⁻².

Outlook

The preparation method for cNiO_x / PFO based photocathode can be further optimized. For example, the cNiO_x layer might possess a higher activity to reduce electron-hole recombination when the annealing temperature is lower than 400 °C. Furthermore, it would be beneficial if more advanced characterization method such as transmission electron microscopy (TEM) can be used. Finally, more HTLs and catalysts can be examined for the PFO photocathode.

References

- [1] United Nations. *For a livable climate: Net-zero commitments must be backed by credible action*. United Nations. <https://www.un.org/en/climatechange/net-zero-coalition>.
- [2] Wang, G.; Ling, Y.; Wang, H.; Xihong, L.; Li, Y. Chemically Modified Nanostructures for Photoelectrochemical Water Splitting. *Journal of Photochemistry and Photobiology C-photochemistry Reviews* **2014**, *19*, 35–51. <https://doi.org/10.1016/j.jphotochemrev.2013.10.006>.
- [3] Fan, R.; Mi, Z.; Shen, M. Silicon Based Photoelectrodes for Photoelectrochemical Water Splitting. *Optics Express* **2019**, *27* (4), A51. <https://doi.org/10.1364/oe.27.000a51>.

- [4] Luo, Z.; Wang, T.; Gong, J. Single-Crystal Silicon-Based Electrodes for Unbiased Solar Water Splitting: Current Status and Prospects. *Chemical Society Reviews* **2019**, *48* (7), 2158–2181. <https://doi.org/10.1039/c8cs00638e>.
- [5] Zhu, Y.; Zhou, W.; Zhong, Y.; Zhong, Q.; Chen, X.; Liu, M.; Zhou, W. A Perovskite Nanorod as Bifunctional Electrocatalyst for Overall Water Splitting. *Advanced Energy Materials* **2017**, *7* (8), 1602122–1602122. <https://doi.org/10.1002/aenm.201602122>.
- [6] Kim, J. H.; Hansora, D.; Sharma, P.; Jang, J.-W.; Lee, J. S. Toward Practical Solar Hydrogen Production – an Artificial Photosynthetic Leaf-To-Farm Challenge. *Chemical Society Reviews* **2019**, *48* (7), 1908–1971. <https://doi.org/10.1039/c8cs00699g>.
- [7] Kalanur, S. S.; Duy, L. T.; Seo, H. Recent Progress in Photoelectrochemical Water Splitting Activity of WO₃ Photoanodes. *Topics in Catalysis* **2018**, *61* (9-11), 1043–1076. <https://doi.org/10.1007/s11244-018-0950-1>.
- [8] Bak, T.; Nowotny, J.; Rekas, M.; Sorrell, C. C. Photo-Electrochemical Hydrogen Generation from Water Using Solar Energy. Materials-Related Aspects. *International Journal of Hydrogen Energy* **2002**, *27* (10), 991–1022. [https://doi.org/10.1016/s0360-3199\(02\)00022-8](https://doi.org/10.1016/s0360-3199(02)00022-8).
- [9] Walter, M. G.; Warren, E. L.; McKone, J. R.; Boettcher, S. W.; Mi, Q.; Santori, E. A.; Lewis, N. S. Solar Water Splitting Cells. *Chemical Reviews* **2010**, *110* (11), 6446–6473. <https://doi.org/10.1021/cr1002326>.
- [10] Döschner, H.; Geisz, J. F.; Deutsch, T. G.; Turner, J. A. Sunlight Absorption in Water – Efficiency and Design Implications for Photoelectrochemical Devices. *Energy Environ. Sci.* **2014**, *7* (9), 2951–2956. <https://doi.org/10.1039/c4ee01753f>.
- [11] Henning Döschner; Young, J. B.; Geisz, J. F.; Turner, J. A.; Deutsch, T. G. Solar-To-Hydrogen Efficiency: Shining Light on Photoelectrochemical Device Performance. *Energy and Environmental Science* **2016**, *9* (1), 74–80. <https://doi.org/10.1039/c5ee03206g>.
- [12] *Solar Energy Materials and Solar Cells | Scholars Portal Journals*. journals.scholarsportal.info. <https://journals.scholarsportal.info/browse/09270248/v92i0004> (accessed 2023-12-14).
- [13] Grimes, C. A.; Varghese, O. K.; Ranjan, S. *Light, Water, Hydrogen*; 2008. <https://doi.org/10.1007/978-0-387-68238-9>.
- [14] Dotan, H.; Mathews, N.; Hisatomi, T.; Grätzel, M.; Rothschild, A. On the Solar to Hydrogen Conversion Efficiency of Photoelectrodes for Water Splitting. *The Journal of Physical Chemistry Letters* **2014**, *5* (19), 3330–3334. <https://doi.org/10.1021/jz501716g>.
- [15] María Isabel Díez- García; Verónica Celorrio; Calvillo, L.; Tiwari, D.; Gómez, R.; Fermín D. J. YFeO₃ Photocathodes for Hydrogen Evolution. *Electrochimica Acta* **2017**, *246*, 365–371. <https://doi.org/10.1016/j.electacta.2017.06.025>.
- [16] Li, C.; He, J.; Xiao, Y.; Li, Y.; Delaunay, J. Earth-Abundant Cu-Based Metal Oxide Photocathodes for Photoelectrochemical Water Splitting. *Energy and Environmental Science* **2020**, *13* (10), 3269–3306. <https://doi.org/10.1039/d0ee02397c>.
- [17] Freeman, E.; Kumar, S.; Thomas, S.; Pickering, H.; Fermín D. J.; Eslava, S. PrFeO₃ Photocathodes Prepared through Spray Pyrolysis.

ChemElectroChem **2020**, *7* (6), 1365–1372.
<https://doi.org/10.1002/celec.201902005>.

[18] Freeman, E.; Kumar, S.; Celorrio, V.; Park, M. S.; Kim, J. H.; Fermin, D. J.; Eslava, S. Strategies for the Deposition of LaFeO₃ Photocathodes: Improving the Photocurrent with a Polymer Template. *Sustainable Energy & Fuels* **2020**, *4* (2), 884–894. <https://doi.org/10.1039/c9se01103j>.

[19] Tiwari, S.; Kumar, S.; Ganguli, A. K. Role of MoS₂/RGO Co-Catalyst to Enhance the Activity and Stability of Cu₂O as Photocatalyst towards Photoelectrochemical Water Splitting. *Journal of Photochemistry and Photobiology A: Chemistry* **2022**, *424*, 113622.
<https://doi.org/10.1016/j.jphotochem.2021.113622>.

[20] Cairns, R.W., Ott, E., *J. Am. Chem. Soc.*, **55**, 527, (1933)

[21] Hirano, A., Kanno, R., Kawamoto, Y., Takeda, Y., Yamaura, K., Takano, M., Ohyama, K., Ohashi, M., Yamaguchi, Y., *Solid State Ionics*, **78**, 123, (1995)

Viscoelasticity and Extensional Rheology of Concentrated Wormlike Micelles Solution

Farhad Lukman, Ruishan Han

Department of Chemical Engineering, Imperial College London, U.K.

Abstract

The use of wormlike micelles (WLMs) solutions has seen a growth in recent years due to the remarkable rheological and viscoelastic properties of these fluids, comparable to the more expensive ultra-high molecular weight polymer solutions. To better understand these complex fluids, numerous research had been done to study the structural and rheological properties of WLMs solutions. Though, one main gap remained in these studies: the extensional rheology— a major flow character that will allow a better characterisation of these fluids for real life use. In this paper, we studied cetyltrimethylammonium bromide/sodium salicylate (CTAB/NaSal) solution, by fixing CTAB concentration to 0.1M and using different NaSal concentration ratio, R of 1, 2, 3, 4, and 5. Formation of WLMs in solutions was confirmed by employing previous molecular structure and shear rheology, alongside our own shear rheology and viscoelasticity study before proceeding with extensional rheology study. Viscoelasticity studies showed that storage and loss modulus intersection moved towards higher shear rate until $R=3$, before moving towards lower shear rate. Thus, indicated that viscosity of the solution increased until a maximum at $R=3$ and decreased at higher R values. Extensional rheometer-on-a-chip was used to measure extensional viscosity with various extensional rates at constant temperature. It was found with $R < 5$, CTAB/NaSal solution exhibited a general tension thinning curve similar to its shear viscosity whereas $R = 5$ exhibited a tension thickening pattern before following the general tension thinning behaviour as found in shear rheology studies. In addition, our research indicated that extensional viscosity was significantly higher than shear viscosity in all R , which was consistent with previous rheological studies.

Keywords: Wormlike micelles, extensional rheology, CTAB/NaSal solutions, viscoelasticity

1.Introduction and Background

Amphiphilic molecules garnered attention by researchers due to the highly complex nature of the molecules when suspended in water. It had been known that these molecules were able to self-organise into many different aggregates, with various kinds of geometry, which affected the rheology of the fluid significantly. Wormlike micelles (WLMs) stood out amongst others due to their vast applications including fracturing technology in oil industry, template synthesis of different nanoobjects, micellar copolymerization of hydrophilic and hydrophobic monomers^[1].

WLMs were formed by the self-organisation of amphiphiles into an elongated and extremely flexible aggregates, capable of forming a network of transient and highly entangled chains. Hence, the term 'living polymers' was often be associated with WLMs solutions since they mimic the behaviour of water-soluble polymers but apart from the transient nature of the WLMs in solution. This was due to the hydrophobic interactions holding WLMs together are extremely weak relatively to the covalent bonds that bond polymer molecules together, causing the micelles to break and form

constantly. Thus, WLMs solutions are extremely susceptible to change in temperatures and concentrations due to the specific conditions required to form these micelles. Hence, the rheological properties of WLMs micelles were more complex and less predictable than any polymer solutions.

The formation of WLMs is relatively simple, requiring only three major components: water, a cationic surfactant, and an ionic salt. This, combined with WLMs astounding viscoelastic behaviour comparable to high molecular weight polymers, has attracted numerous research in understanding WLMs.

Cetyltrimethylammonium bromide (CTAB) is one of the surfactants that has widely been known to form WLMs at a critical micelle concentration of 1.0M. The addition of ionic salt such as sodium salicylate (NaSal) allows the formation of WLMs at a lower concentration of CTAB. The presence of high electronegative phenyl group in salicylate ion reduces the repulsions between polar head of CTAB, allowing CTAB molecules to pack closer to each other forming thus forming wormlike micelles at lower concentration.

In recent years, researchers have mostly focused their interest in studying the molecular structures of various WLMs solutions, and the macroscopic behaviour or the rheology of these solutions specifically when imposed by a shear stress. These studies, especially the shear rheology, are important to better characterise these living polymers and allow a greater understanding on how they behave in real-life situations. Though shear rheology of CTAB/NaSal has been extensively studied, these studies have failed to address an important component to fully characterise the rheology of these fluids: their extensional rheology.

With the recent rise of interest in WLMs solutions, we aim to bridge the gap in rheological studies by employing CTAB/NaSal solution to understand their extensional flow behaviours, which to the best of our knowledge has not been investigated previously. This research primary goal is to study the extensional rheology of CTAB/NaSal solutions and using established studies to compare the behaviour of these complex fluids.

2. Methodology

2.1 Production of CTAB/NaSal Solution

CTAB, manufactured by ThermoSCIENTIFIC™ and NaSal, salts, manufactured by Merck KGaA™ were with purity of 99.5% respectively, without the need for further refinement or processing. 5 CTAB solutions with concentration 0.1M were made by measuring a weight of 0.3665g to produce a 0.1L solution each. Before addition of deionised water in each beaker, 0.1601g, 0.3202g, 0.4803g, 0.6404g, and 0.8005g of NaSal salts were measured and added to beakers containing CTAB to produce concentration ratio, R of 1, 2, 3, 4, and 5.

Deionised water was added to the 100ml line, and solution were then mixed at least overnight under a constant stirring speed of 400 rpm and temperature of 35 °C before any rheological studies to allow solution to equilibrate.

2.2 Confirmation of WLM Structure Formation

Physical inspection of solution produced, was used as the primary method of confirming the formation of WLMs by comparing the physical attributes of the solution with multiple available literatures.

This was done to qualitatively confirm the structures formation before proceeding with shear viscosity studies that would allow us to quantitatively confirm the formation of said structures.

2.3 Rotational Shear Rheometry

Shear rheology studies were mainly done on Anton Parr MCR 302 and ThermoSCIENTIFIC™ MARS 60 shear rheometer with a coaxial cylinder as the measuring geometry, where samples were loaded in the gap between cylinders.

2.4 Shear Viscosity and Viscoelastic Behaviour

To study the behaviour of CTAB/NaSal solution, different shear rates were imposed ranging from 0.01 to 100 s⁻¹ at a fixed temperature of 25 °C. Graph of the viscosity obtained were then plotted on a logarithmic axes of shear viscosity (Pa s) against shear rate (s⁻¹).

In addition, small amplitude oscillatory test was done on the fluid within the linear viscoelastic region where samples were sheared in an oscillatory manner about the equilibrium position at a fixed amplitude. The experiments were carried out by manipulating the temperature from 15 to 30°C and varying frequency of oscillations and measuring the maximum shear stress obtained. The measurement obtained was used by the software to calculate the storage (G') and loss (G'') moduli by using the equations:

$$G = \frac{\sigma_{max}}{\epsilon_{max}} \quad (1)$$

Where G is the modulus of spring, σ_{max} is the max strain and ϵ_{max} is the maximum measure stress.

Then, the moduli can then be calculated using:

$$G' = \frac{G(\omega\tau)^2}{1 + (\omega\tau)^2} \quad (2)$$

$$G'' = \frac{\eta\omega}{1 + (\omega\tau)^2} \quad (3)$$

Where $\omega\tau$ is the product of angular frequency and the relaxation time of the solution after the imposed stress.

The analysis of the viscoelastic behaviour was done by plotting the G' and G'' against frequency of oscillation on the same graph. This in turn gave us an insight on how fluid behaviour changes with different shear rates.

2.5 Extensional Rheometry

All extensional rheology studies were done on RheoSense e-VROC™ microfluidics chip. The channel of the chip was engineered with hyperbolic

divots that caused fluid to expand and contract thereby producing a constant extensional flow around the region. The presence of this hyperbolic gap caused a significant pressure drop of the fluid, which was a characteristic of extensional fluid flow, instead of a constant pressure drop when shear is the only significant attribute.

Micro Electronic Mechanical Systems (MEMS) pressure sensors were equipped in throughout the channel especially upstream and downstream of the gap, which allows the measurement of pressure drop, subsequently extensional viscosity to be done.

Calculation of apparent extensional viscosity was done fully by the proprietary software made by RheoSense™, which made use the following modified viscosity equation:

$$\eta_e = \frac{\Delta P}{\epsilon_H \dot{\epsilon}} \quad (4)$$

Where ΔP is the pressure drop caused by the extensional flow through the hyperbolic divots, ϵ_H is the Hencky strain attributed by the fluid contraction/expansion through the divots and $\dot{\epsilon}$ is the apparent extensional rate of fluid passing through the channel.

The e-VROC chip used in all extensional studies had the same value of Hencky strain of 2.0130, which was calculated by the equation:

$$\epsilon_H = \ln \frac{w_c}{w_t} \quad (5)$$

Where w_c is the width of the main flow channel of 2.994 mm, and w_t is the smallest width of the main contraction/expansion zone with a value of 0.400 mm.

Fluid was loaded into the chip by use of 1ml syringe attached to the inlet of the chip, and covered in a temperature controlled thermal jacket, connected to Thermocube bath. The syringe plunger was attached to a pusher block which acted like a pump, delivering the required flow rates as specified by the user using the included software.

To ensure measurement of extensional viscosity to be as accurate as possible, the first extensional rate was run for an extended period, roughly 1500 seconds, compared to the 50 to 500 seconds for subsequent extensional rates. This was done to 'prime' the sensor, thereby removing any fluids and/or air bubbles present in the flow channel, disrupting the pressure drop measurements.

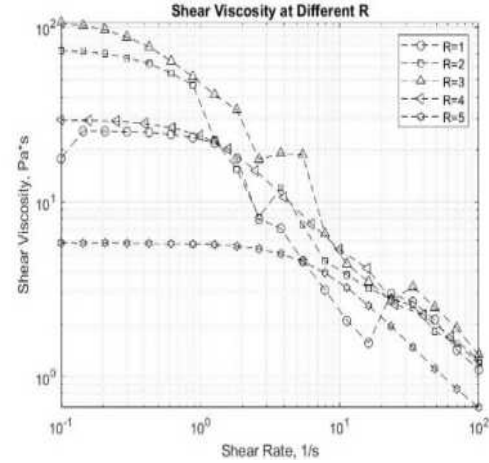


Figure 1: Shear Viscosity vs. shear Rate for Different R

3. Results and Discussion

3.1 Shear Viscosity

From Figure 1, for R=1-5, shear-thinning behaviour was shown for all ratio, shear viscosity generally decreases with shear rates. However, several waves showed up at shear rates of 16.2/s for R=1 and 2.64/s for R=2 and 3 respectively. Additionally, an extra fluctuation took place at shear rates of 25.1/s for R=3, while there was no fluctuation was detected by the rotational rheometer at R=5, meaning the worm-like micelle structure was not affected by rotational movement. This phenomenon indicated that there would be shear-induced structure, which CTAB molecules aggregated under rotational movement. Furthermore, a maximum shear viscosity was reached at R=3, 106600 mPa*s particularly. Once R became larger than 3, generally the shear viscosity started to decrease, possibly due to the excess of NaSal dissolved in CTAB aqueous solution causing extra repulsion between the salt ions and CTAB molecules, making the worm-like micelle structure to destruct. In addition, from visually observation, high NaSal solution might even break caused the CTAB molecule itself to be decomposed forming bromine in the solution, as the solutions shown in figure 1, were appeared to be brownish colour for R=4 and R=5.

3.2 Viscoelastic Behaviour

The exchange of viscous and elastic behaviour of a fluid was determined by measurement of storage modulus and loss modulus. At 25°C, the results

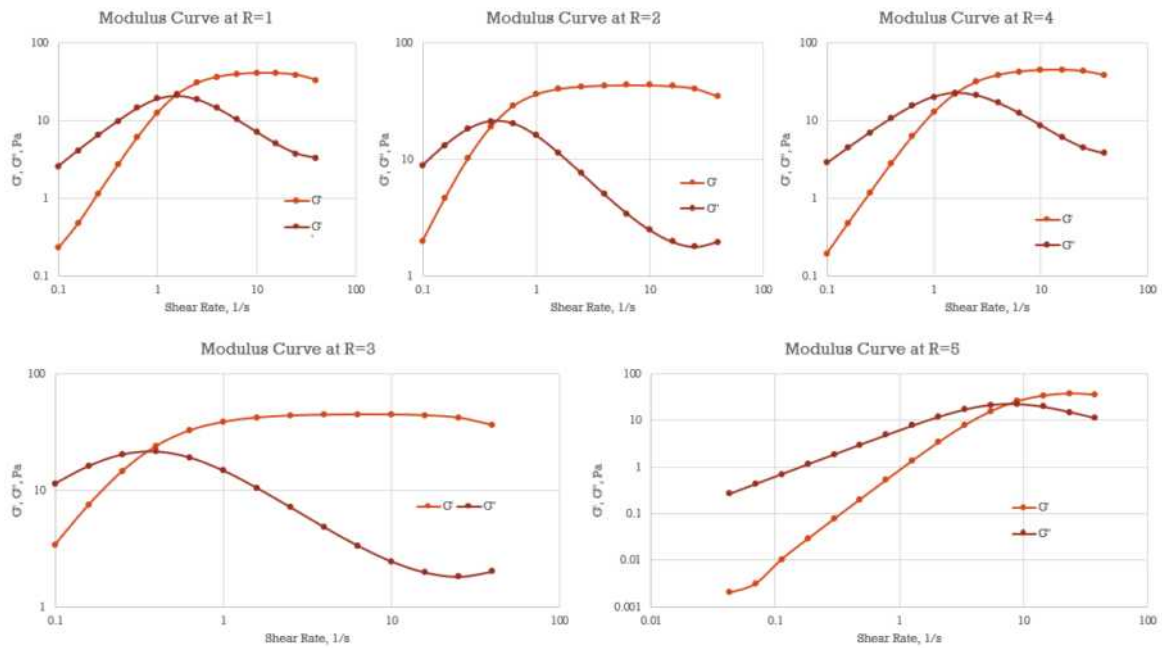


Figure 2-6: Storage & Loss Modulus vs. Shear Rate at Different R for Temperature=25°C

illustrated in figure 2-6 had shown that the left most intersection point of G' and G'' curve occurred at $R=3$, indicating that the solution with $R=3$ would be the most elastic solution. Apart from $R=3$, the intersection point shifted to left and right with increasement in shear rates, for $R<3$ and $R>3$ respectively. Meanwhile, the largest plateau zone of G' , which kept stationary after shear rate of 1/s, was also found on the curve of $R=3$, meaning that the fluid with $R=3$ gave the most elastic behaviour among all 5 testing samples, the viscoelastic

behaviour started to stabilise at low shear rates. By combining this observation with shear viscosity, the limitation of shear rheology and viscoelastic behaviour tended to show up at $R=3$, in another word, this ratio for high concentration CTAB solution might be an indicator for significant changing in properties.

Besides, temperature also played an important role in viscoelastic properties. For each R, the rightward shifting of intersection point of the two moduli

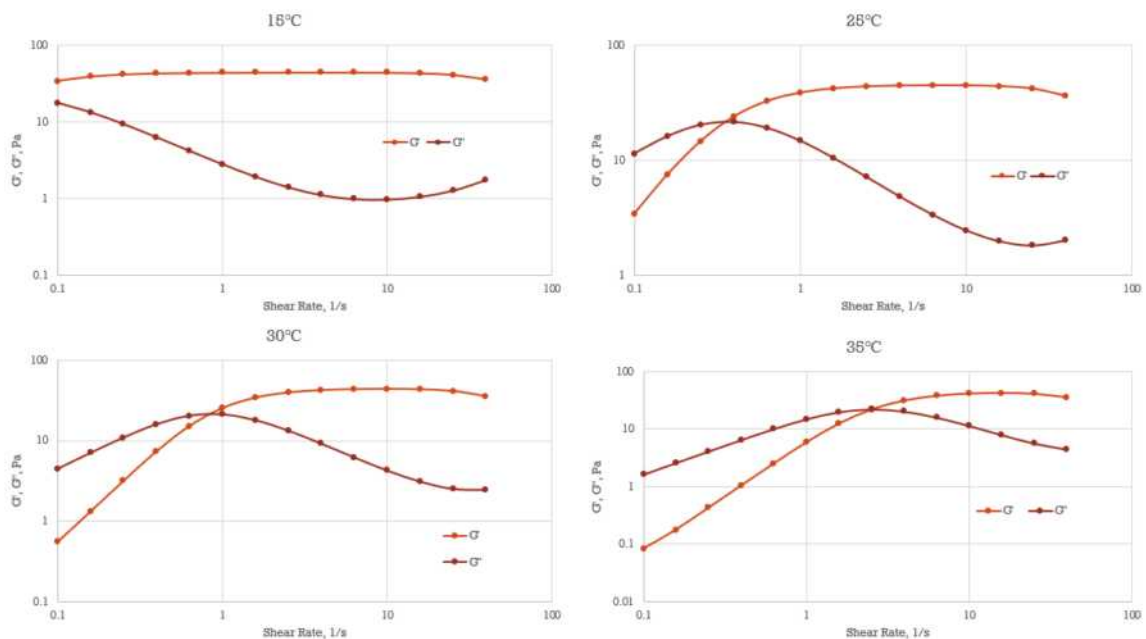


Figure 7-10: Storage & Loss Modulus vs. Shear Rate at Different Temperature for $R=3$.

curve indicated that higher temperature would encourage more viscous behaviour, as more energy got absorbed by the micelle molecule, causing the crosslinking and intermolecular interaction between molecules to be destructed. For some extreme cases, such as for R=3 at 15°C, the G' and G'' curve did not intersect, meaning that there would be negligible viscoelastic behaviour and no fluid change from viscous to elastic regime.

3.3 Extensional Viscosity

Ranging from R=1 to R=5, a general negative relationship between extensional viscosity and extensional rate was shown in figure 11, however, except for R=5, a relatively small amount of increase occurred at low extensional rates. Particularly, the viscosity started to decrease at 2.5, 2.5, 1.4 and 4.6/s corresponding to R=1, 2, 3 and 4 respectively. The solution with R=5 generated a curve with a relatively unique trend, having a second maximum viscosity point following the increasing after the decreasing at low extensional rates where the experiments started. Though the rising in viscosity might be considered as due to possible tension-induced structure, similar to shear-induced structure, the uncertainty at low shear rates caused by the sensor chip used in the experiments should be also taken into account. Nevertheless, the second convex shape curve shown in the curve representing R=5 might be trustworthily illustrating the nature of the fluid. Particularly due to the high shear and extensional viscosity of solution with R=2 and 3, the test at high extensional rates could not be carried out, since the fluids would stick inside the sensor chip not only damaging the sensor but breaking the gastight syringe, alternative methods measuring extensional viscosity might need to be discovered.

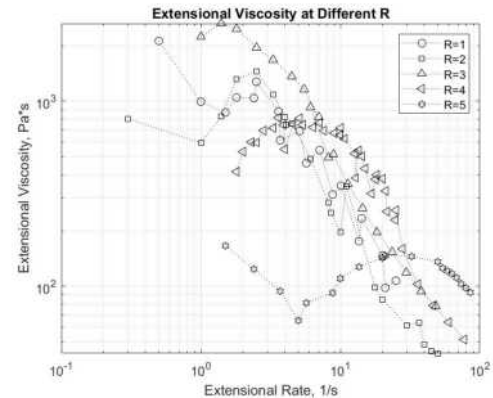


Figure 11: Extensional Viscosity vs. Extensional Rate at Different R

Same as shear viscosity, the maximum extensional viscosity was also found at R=3 while the lowest was at R=5. Although in shear viscosity analysis, the curve of R=2 gave the second highest shear viscosity, the extensional viscosity curve shown in figure 11 illustrating that the curve for R=1 and 2 had similar results at extensional rates greater than 1.4/s. At high extensional rates, the gap between extensional viscosity for different R tended to become narrower, apart from R=5, where a relatively more significantly higher extensional viscosity was shown in figure 11. This phenomenon indicates that extensional viscosity might not be necessarily strongly correlated with shear viscosity.

Temperature effects on extensional viscosity were also studied for solutions with R=1 and 4.

Temperature would not have significant effects on the fluid viscosity at extensional rates higher than 27.2 and 9/s corresponding to R=1 and R=4 respectively. Below the threshold rates, at R=1, the extensional viscosity showed an increasing trend with temperature, while the opposite was observed for R=4.

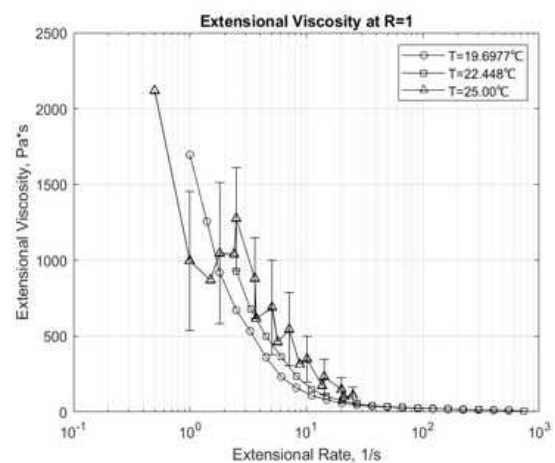
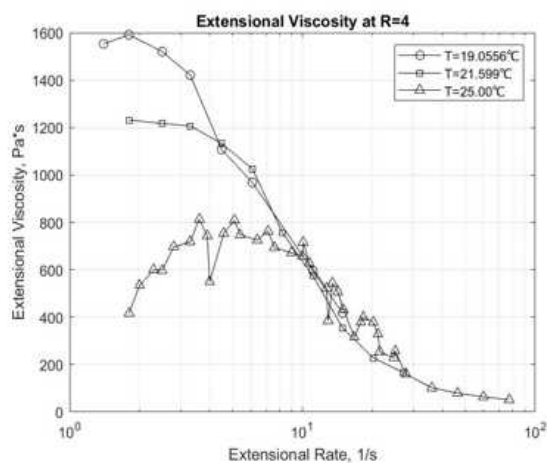


Figure 12-13: Extensional Viscosity vs. Extensional Rate at Different Temperature for R=1 and 4.

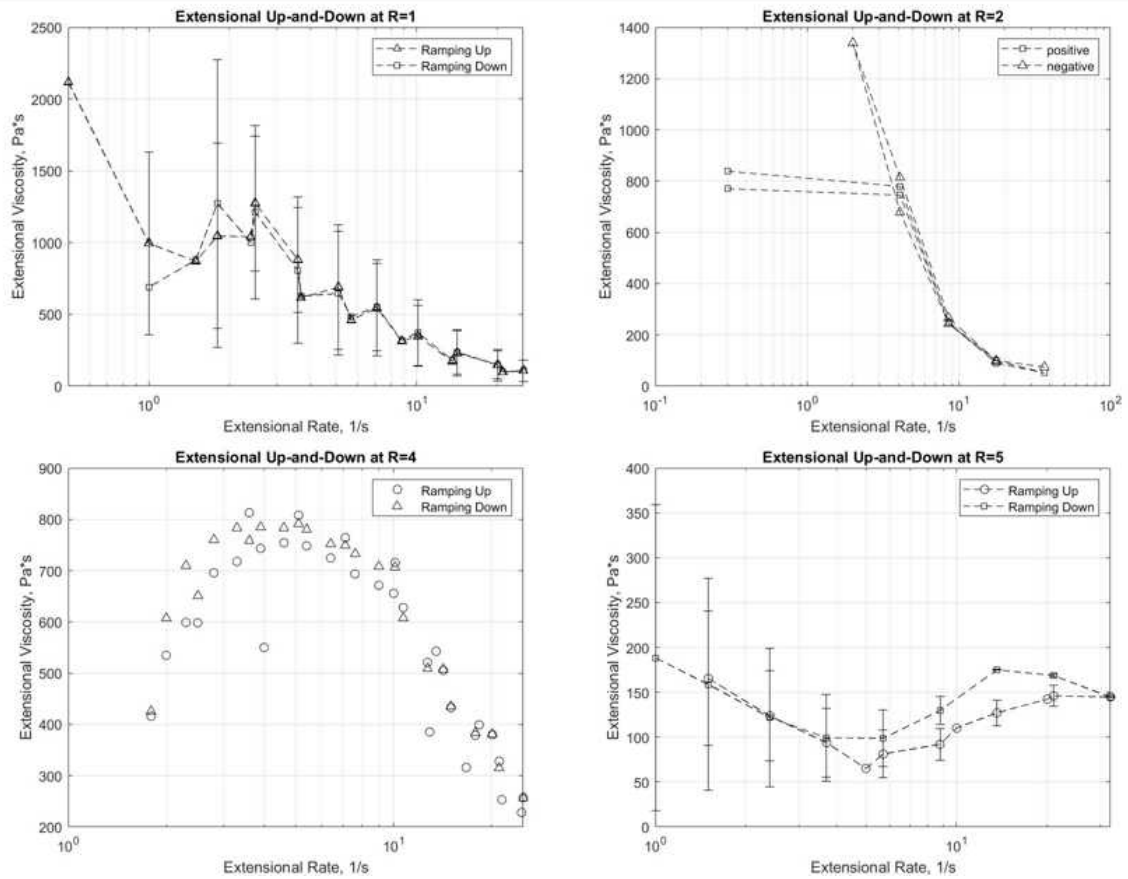


Figure 14-17: Extensional Up-and-Down Curve at Different R.

3.4 Extensional Up-and-Down

Compared to shear stress thixotropic behaviour, the hysteresis effects in extensional viscosity would

also be worthy to research. Instead of naming 'extensional thixotropic' experiments, extensional up-and-down was defined to the test to study the hysteresis behaviour. For $R=1, 2$ and 4 , there would

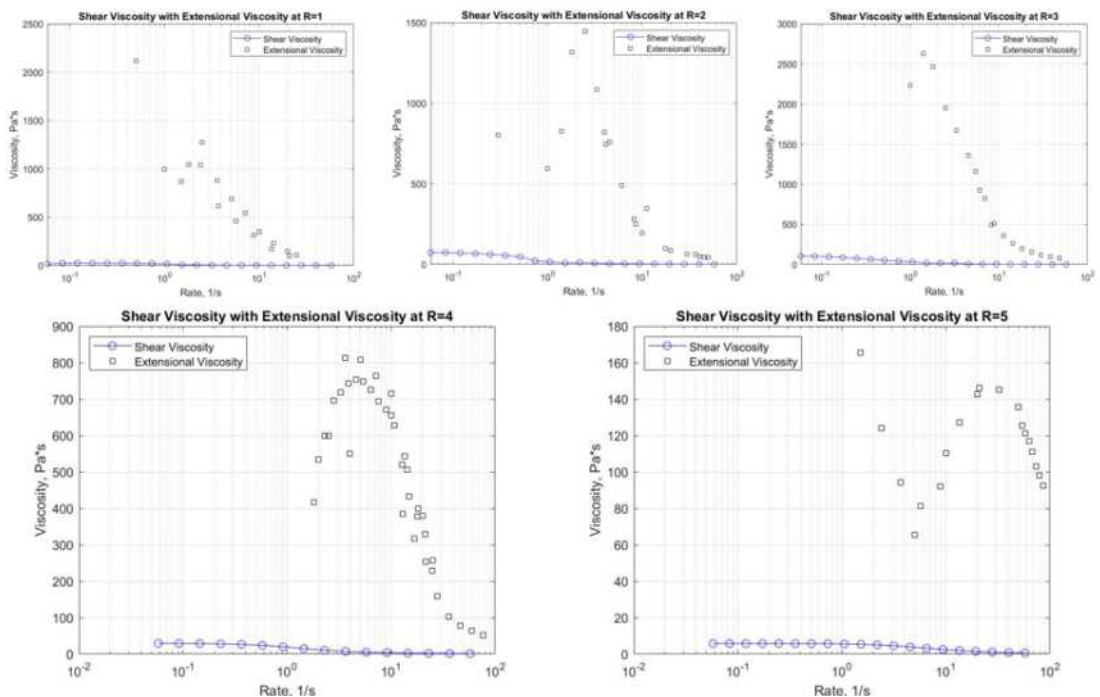


Figure 18-22: Shear and Extensional Viscosity vs. Rate at Different R.

not be apparent differences between the up and down curves shown in figure 14-16, at extensional rates above 1.5, 4.1 and 10.7 respectively. Meanwhile, the results at $R=5$ indicated that the micelle structure might get reformed and even elongated during ramping down, corresponding to the deviation between extensional rates of 3.7 to 32.4 shown in figure 17. In particular, significant difference between positive and negative experiments showed in figure 15. Nevertheless, for other ratio, the difference between up and down curves at low rates could possibly indicating there would be hysteresis effects, while the uncertainty of sensor chip may not be neglected.

3.5 Compare Shear and Extensional Viscosity

Finally, shear and extensional viscosity were compared based on the Trouton ratio equation, where the shear rates were divided by $\sqrt{3}$ to match extensional rates. As shown in figure 19, 20 and 21, at high rates for $R=2,3$ and 4, two curves are getting closer, indicating extensional viscosity approaching the shear viscosity, indicating the effect of viscoelastic behaviour might become negligible.

4. Conclusion

In conclusion, all solutions are tension-thinning at low extensional rates, apart from $R>4$, where might be tension-thickening. Also, temperature does not have significant effects at high extensional rates, however, it might affect the extensional viscosity at low rates. Solutions with $R>4$ would give different results in up-and-down experiments. Both shear viscosity and extensional viscosity would have maximum at $R=3$ for CTAB concentration=0.1M. Besides the main conclusions for extensional rheology, viscoelastic behaviour is not negligible. At temperature between 20-35, intersection points showed up, indicating there was exchange of viscous and elastic behaviour.

5. Acknowledgment

We would like to appreciate Prof. Paul Luckham for guiding us through the entire project, and Shawn J. H. Lew, Guangyao Lyu and Patricia Carry for characterisation and giving instructions on instrument using.

6. References

1. Barnes, H., J.F. Hutton, & K. Walters. (1989). *An Introduction to Rheology*. Elsevier.
2. Berret, J.-F. (2004). *Rheology of Wormlike Micelles: Equilibrium Properties and Shear Banding Transition*. Retrieved December 1, 2023, from <https://arxiv.org/pdf/cond-mat/0406681.pdf>
3. Cates, M. E., & Candau, S. J. (1990). Statics and Dynamics of Worm-like Surfactant Micelles. *Journal of Physics: Condensed Matter*, 2, 6869.
4. Das, N. C., Cao, H., Kaiser, H., Warren, G. T., Gladden, J. R., & Sokol, P. E. (2012). Shape and Size of Highly Concentrated Micelles in CTAB/NaSal Solution by Small Angle Neutron Scattering (SANS). *Langmuir*(28), 11962-11968.
5. Irfan, M., Usman, M., Mansha, A., Rasool, N., Ibrahim, M., Rana, U. A., & Siddiq, M. Z.-U.-H. (2014). *TheScientificWorldJournal. Thermodynamic and spectroscopic investigation of interactions between reactive red 223 and reactive orange 122 anionic dyes and cetyltrimethyl ammonium bromide (CTAB) cationic surfactant in aqueous solution.*, 2014(540975), 8.
6. Kim, W.-J., & Yang, S.-M. (2000). Effects of Sodium Salicylate on the Microstructure of an Aqueous Micellar Solution and Its Rheological Responses. *Journal of Colloid and Interface Science*(232), 225-234.
7. Malvern Instruments Limited. (2016). *A Basic Introduction to Rheology*. Worcestershire, United Kingdom.
8. Nodoushan, E. J., Yi, T., Lee, Y. J., & Kim, N. (2019). Wormlike Micellar Solutions, Beyond the Chemical Enhanced Oil Recovery Restrictions. *Fluids*, 4(3), 173.
9. Oda, R., Narayanan, J., Hassan, P. A., Manohar, C., Salkar, R. A., Kern, F., & Candau, S. J. (1998). Effect of Lipophilicity of the Counterion on the Viscoelasticity of Micellar Solutions of Cationic Surfactants. *Langmuir*, 14, 4364-4372.
10. Raghavan, S. R., & Feng, Y. (2017). Wormlike Micelles: Solutions, Gels, or Both? *Soft Matter Series No. 6*, pp. 9-26.
11. RheoSense. (n.d.). *E-VROCTM VISCOMETER — TECHNICAL NOTE*. Retrieved December 1, 2023, from

<https://www.rheosense.com/extensional-viscosity-application-note>

12. Shibaev, A. V., S., O. A., K., K. E., I., K. A., M., A. T., Novikov, V. V., & Philippova, O. E. (2022). Universal Character of Breaking of Wormlike Surfactant Micelles by Additives of Different Hydrophobicity. *Nanomaterials*, 12(24), 4445.

Analysis of Carbon Capture Readiness for Small-Scale Refuse Derived Fuel-to-Energy Power Plants based in the UK

Hannah Bai and Shiwaan Dharma Sena

Department of Chemical Engineering, Imperial College London, U.K.

Abstract

With a growing global interest for sustainable energy coupled with the challenge of underutilised waste, there is a rising recognition to the significance of generating low-carbon energy via waste-to-energy power plants. This paper investigates the feasibility of retrofitting a post-combustion carbon capture plant using monoethanolamine (MEA) technology to a small-scale UK waste-to-energy power plant utilising refuse derived fuel. The overall power generation and carbon capture system is simulated on Aspen Plus V11, modelled with a capture efficiency of 95%. Both the energy requirements and the economic potential were explored as part of the feasibility study. At the given capture efficiency, when retrofitting a post combustion carbon capture (PCC) plant, an energy penalty of 53.3% is imposed, leading to a net energy output of 5.95MW_e and a decreased plant efficiency to 31%. A comparison between refuse derived fuel (RDF) power plants and conventional fuel plants such as coal and natural gas combined cycle (NGCC) power plants revealed that the former has a significantly larger energy penalty owing to the high energy consumption of the reboiler in conjunction with the low thermal efficiency of the RDF fuel. The economic potential is estimated to be -£2.44million for the integrated carbon-capture plant at this energy capacity. The unfavourable economic prospects coupled with the substantial energy penalties pose challenges to the feasibility of this simulated waste-to-energy plant.

1. Introduction

In the face of escalating concerns about climate change and the urgent need to mitigate greenhouse gas emissions, carbon capture has emerged as a pivotal solution for a more sustainable future. As nations worldwide aim to transition towards sustainable energy solutions, the exploration of carbon capture technologies becomes a focal point. Concurrently, with the recent developments at COP28, where the phaseout of fossil fuels faced reconsideration, the increasing reliance on technologies like carbon capture to mitigate emissions becomes evident. As we strive towards achieving the ambitious net-zero target by 2050, it becomes imperative to recognise that emitting zero CO₂ is only one aspect of the problem. Equally vital is the removal of existing CO₂ from the atmosphere, marking it a crucial component in the pursuit of a sustainable future.

Moreover, rapid urbanisation and population growth have contributed to a surge in waste generation, placing immense strain on the existing waste management infrastructure. Landfills are now facing their limitations, with finite capacity and environmental consequences. The current trajectory indicates that London's landfill capacity is anticipated to reach its threshold by 2026 (London Assembly, n.d.), necessitating a re-evaluation of waste management strategies. As the demand for efficient and sustainable waste disposal solutions intensifies, the UK finds itself at a critical juncture to explore alternatives that not only address the immediate challenge of waste disposal but also align with broader environmental goals.

Among the escalating waste management challenges, the utilisation of RDF emerges as a noteworthy alternative, gaining prominence in the UK's pursuit of sustainable waste disposal practices. RDF is derived from the processing of municipal solid waste (MSW), transforming non-recyclable materials into a

valuable energy resource. This approach not only diverts waste from landfills but also harnesses its energy potential, contributing to the reduction of reliance on traditional fossil fuels.

Although there are references to RDF incineration plants for power generation in literature, there are limited resources exploring the integrations of an RDF-to-energy plant with a carbon capture and storage system. Successfully bringing together an RDF-to-energy facility with integrated carbon capture technology, not only promotes the utilisation of RDF but also enables the production of low carbon energy. An investigation of the economic viability and energy penalty will offer valuable insights into the feasibility and readiness of carbon capture for RDF-to-energy power plants, with an aim to advance industrial capabilities.

2. Background

2.1. RDF

MSW represents the diverse range of discarded materials from households and institutions, comprising of everyday items like packaging, food scraps, appliances, and more. The variability in MSW composition is intricately linked to diverse socioeconomic factors. For instance, affluent areas often exhibit lower food waste while lower-income regions may have a higher organic content (Chavando et al., 2022). The variation extends beyond local environments, resonating on a global scale. This highlights the need for a nuanced examination of MSW compositions originating from diverse regions. Given that RDF is derived from MSW, its composition is dictated by the components of the waste. Table 1 presents a summary of proximate analysis, ultimate analysis, and lower heating values (LHV) of RDF from various global locations.

Table 1: RDF Compositions Across the World

RDF	Location	Proximate analysis (wt%)				Ultimate analysis (wt%)					LHV (MJ/kg)
		Moisture Content	Ash Content	Volatile Matter	Fixed Carbon	C	H	N	O	S	
1	UK	5.9	12.9	70.0	11.2	58.7	8.4	1.0	16.0	0.4	24.9
2	Pakistan	8.8	8.3	78.3	9.5	54.2	4.7	0.8	30.4	0.0	22.1
3	EU	1.6	9.1	80.7	8.7	48.4	6.9	0.4	35.0	0.3	22.2
4	Kazakhstan	1.5	8.2	86.7	3.6	58.3	9.9	0.6	22.8	0.2	23.4
5	Spain	8.5	26.0	70.4	3.6	46.8	5.4	1.1	20.4	0.3	11.4

1 - (Materazzi et al., 2015), 2 - (Mehdi et al., 2020), 3 - (Alfè et al., 2022), 4 - (Botakoz Suleimenova et al., 2022), 5 - (Garcia et al., 2021)

The production of RDF from MSW, requires a drying process to reduce its moisture content. Simultaneously, the waste undergoes fragmentation, and any inert materials are extracted to enhance its calorific value (Jannelli and Minutillo, 2007). Following processing, the resultant RDF serves as a fuel source for energy production through combustion. This energy can be harnessed in a Combined Heat and Power (CHP) plant, where steam turbines, coupled with a generator, convert it into electrical energy, thermal energy is released via the cooling water (Environ Consultants Ltd., n.d.).

2.2. Post Combustion Carbon Capture

Amidst the continuously growing demand for power generation in the UK, reducing the carbon-footprint of industries dependent on fossil fuels is crucial for a cleaner and more carbon-friendly future. Carbon capture and storage not only facilitates the continued usage of existing infrastructure but also represents one step closer towards a net-zero nation. Among the three primary methods of carbon capture – oxy-fuel combustion, pre-combustion carbon capture, and post-combustion carbon capture (PCC) – this paper solely focuses on PCC due to its promising industrial advancements. PCC can easily be retrofitted and implemented to an existing chemical plant without much disturbance to its current infrastructure, making this the most economically favourable method. In this process, fossil fuels are conventionally combusted for energy generation, while carbon dioxide in the effluent gas stream is captured before being discharged to the atmosphere.

2.3. Solvent Selection

Monoethanolamide (MEA), an amine-based solvent commonly utilised for post combustion carbon dioxide absorption due to its high reactivity and low cost (Li et al., 2016), is selected as the aqueous solvent for this study.

In addition, MEA stands out as one of the most commercially ready technologies for minimising the amount of CO₂ released to the atmosphere (Jung et al., 2013). A significant amount of energy is required for MEA solvent regeneration when retrofitting a carbon capture and storage (CCS) system, in turn imposing a heavy energy penalty with a notable decrease in plant efficiency (Luis, 2016). As a result, extensive research is being conducted within this field in attempt to

minimise the energy-intensive CCS process and enhance its thermal efficiency.

In line with ongoing CCS investigations, this study focuses on the economic feasibility of retrofitting and optimising a MEA-CO₂ PCC plant on a waste-to-energy (WtE) facility.

3. Methodology

3.1. Power and Heat Generation Plant

The power and heat generation plant consists of two main sections, RDF incineration and then the power generation section.

The feed of RDF was selected based of the 25 MSW incinerators with energy recovery in the UK (Nixon et al., 2013). Due to the relatively small scale of the plant the capacity of the second smallest plant was taken which is 30 ktpa. However, due to the plant operating for 8000 hrs/yr the capacity was scaled up in accordance to 37 ktpa. As a result, a feed of 4600 kg/hr was chosen.

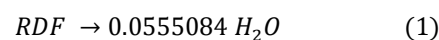
3.1.1. Characteristics of RDF

In the context of modelling RDF as a non-conventional fuel in ASPEN, it is essential to input appropriate values for proximate analysis, ultimate analysis, and LHV. RDF 1 from Table 1 as the chosen input, selected for its relevance due to the location (UK) and its relatively high LHV. The emphasis on LHV is particularly crucial, considering that it is positively correlated with the amount of energy it releases.

3.1.2. Drying

The RDF incineration modelled in ASPEN Plus shown in Figure 1 is similar to that shown in the ASPEN Plus user guide (Aspentech, 2013) however, the inputs were designed specifically for the RDF plant.

The first section of the incineration involves pre-treating the RDF to reduce its moisture content and therefore increase its heating value. To model this we assume the reaction for coal drying applies to RDF as well, where 1 mole of RDF produces 1g of water.



The RDF then enters a flash column where RDF is separated from the other components present.

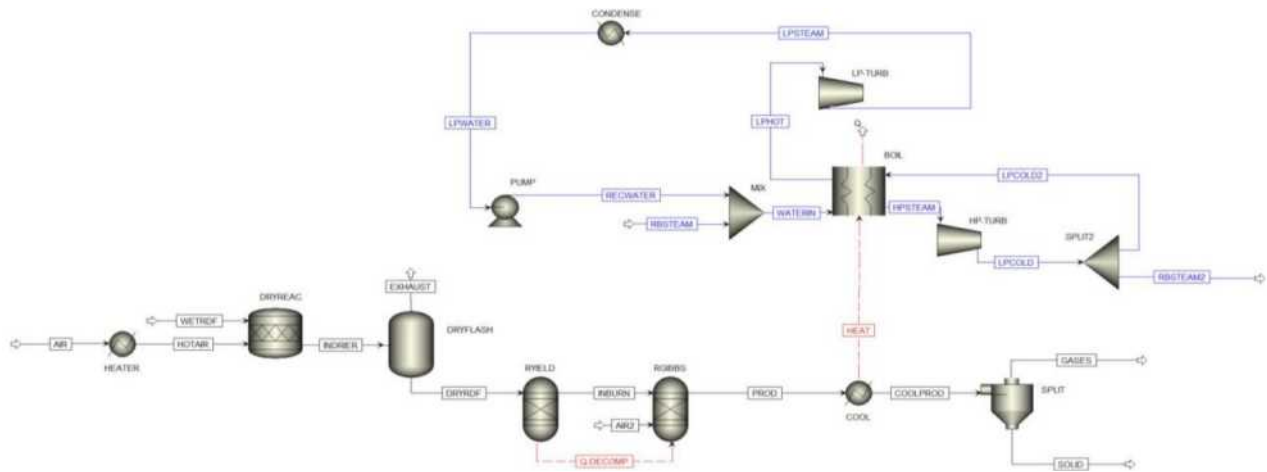


Figure 1: Aspen Plus Flowsheet for RDF-to-energy plant

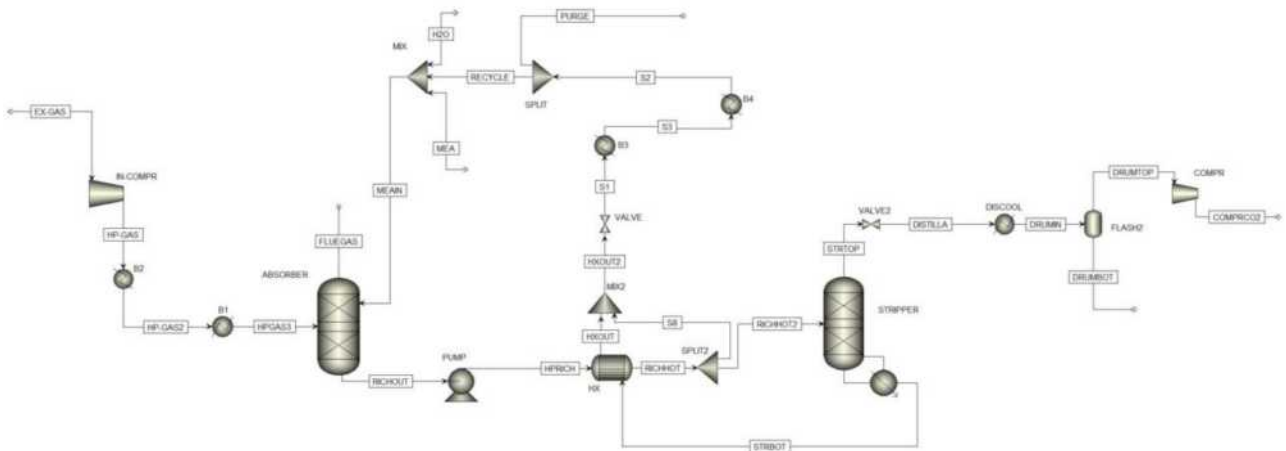


Figure 2: Aspen Plus Flowsheet of PCC plant

3.1.3. Combustion

Given the non-conventional nature of RDF, a two-reactor system in ASPEN PLUS is necessary for its incineration. Illustrated in the figure is the sequential operation of a RYIELD reactor followed by a RGIBBS reactor. The RYIELD reactor initiates the decomposition of RDF into its elemental components, aligning with the fuel's ultimate analysis. The heat generated during the decomposition is directed to the RGIBBS reactor, denoted by heat stream 'Q-DECOMP', supplying the required energy for combustion. The decomposed RDF then enters the RGIBBS reactor, where combustion occurs. The RGIBBS block operates by considering all possible products and establishes chemical equilibrium by minimising Gibbs free energy, removing the need for specifying the reaction stoichiometry.

A sensitivity analysis was conducted to find a suitable trade-off between the carbon dioxide, oxygen and carbon monoxide compositions whilst varying the air feed stream. This is to ensure that complete combustion takes place whilst ensuring that the mass flowrate of carbon dioxide is not too high. Subsequently, an air flowrate of 50000 kg/hr was chosen.

The product stream leaving from the RGIBBS reactor is at a temperature of 1515°C. This stream is then cooled down to 600°C whilst subsequently providing a heat duty of 18.07MW to the boiler in the power generation cycle.

3.1.4. Exhaust Gas-Solid Separation

Following the reactor, the stream is used to heat up the boiler in the energy generation section whilst subsequently getting cooled itself. This cooled stream then enters a splitter where the solid ash gets removed of the bottom and the exhaust gases come off the top. The composition of the exhaust gas stream is outlined in table below. This stream is the input stream for the PCC process.

Table 2: Outlet RDF Plant Gas Composition

Gas	Mass Fraction
N ₂	0.713
CO ₂	0.176
H ₂ O	0.063
O ₂	0.045
SO _x , NO _x , CO, H ₂	0.003

3.1.5. Power Generation

A Rankine steam cycle has been employed for this plant to generate power. Two turbines are used for this process, a high-pressure (HP) turbine and a low pressure (LP) turbine and these both operate with isentropic efficiencies of 90% (Fröhling, Unger and Dong, n.d.). The boiler is modelled as a heat exchanger which is powered by the heat energy released from the hot combustion gases.

Firstly, the water stream that enters the boiler is heated to 600°C and a pressure of 100 bar and is then passed through the HP turbine where it is discharged at 10 bar and 286°C. This creates an electrical output of 5.01MW. Following from this the stream is split where 17500 kg/hr is diverted to power the reboiler in the stripper whilst the remainder is sent back to the boiler which then heats it up to 600°C. Then it is passed through the LP turbine from which it gets discharged at 0.1bar and 88°C. This releases a further electrical output of 2.89MW, meaning that the total electrical output provided by the plant is 7.90MW.

The low-pressure steam undergoes complete condensation, reaching a temperature of 45°C. Following this, the condensed steam is pumped back to the boiler at a pressure of 10 bar, along with the recycled steam.

There is a recycle loop of steam used in the power generation which is depicted by the two streams “RBSTEAM” and “RBSTEAM2” as seen in the flowsheet in Figure 1. This steam leaves the system at 286°C and enters back the system at 179°C, in the meantime it provides the duty that is required for the reboiler in the stripper in the PCC process.

3.2. Post Combustion Carbon Capture Plant

The Aspen Plus simulation for PCC with a capture efficiency of 95% can be observed on Figure 2. An overview of the carbon capture process are as follows:

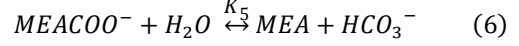
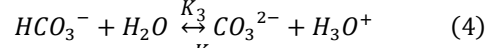
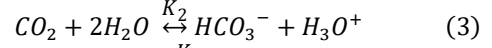
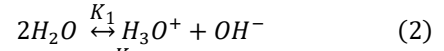
- Following the incineration plant, exhaust gas is compressed and cooled to 40°C before entering the bottom of the absorber column.
- MEA entering from the top of the absorber absorbs CO₂ in the exhaust gas stream, with remaining flue gases rising and exiting from the top of the column.
- The MEA rich CO₂ leaves the bottom of the absorber column and is then pumped through a heat exchanger before entering the stripper column.
- Inside the stripper, CO₂ is removed from MEA. Lean MEA is then recycled and fed back to the absorber whilst CO₂ exits the top of the stripper to be further processed to allow for transportation and storage.

3.2.1. Reaction Mechanism

Vapor-liquid equilibrium and mass transfer from one phase to the other is crucial for Aspen Plus operation simulations.

The vapour phase is described via the Soave-Redlich-Kwong equations of state whilst the formation of ionic species in the liquid phase causes the system to be highly non-ideal. Consequently, this requires the liquid phase to be modelled via the activity coefficient Electrolyte-NRTL model. This model uses the local electronegativity and strong like-ion repulsion assumption (Moioli et al., 2012).

The governing equilibrium reactions for the MEA-CO₂-H₂O electrolyte system are highlighted in equations below, with all species in an aqueous solution (Soltani et al., 2017).



Reaction 2 represents the ionisation of water, and the formation and dissociation of bicarbonate are identified in Reactions 3 and 4 respectively. Lastly, Reactions 5 and 6 reflects the reaction of molecular MEA with CO₂ in aqueous solution.

In addition to these five equations, the following kinetically controlled reactions govern the reaction mechanism as shown in Equations 7 and 8.



3.2.2. Rate Based Model

A rate-based approach, aligned with Errico et al (2016) was employed to rigorously model the various mass, heat, and energy transfer phenomena intrinsic to the carbon capture process.

3.2.3. Cooler

The exhaust gases are cooled from 162°C to 40°C via a two-stage cooling process prior to entering the absorber. This is achieved by firstly contacting the exhaust gas stream with the air stream utilised for RDF drying in the incineration plant, where heat is rejected to warm the air stream. The exhaust gas stream is then cooled to 40°C by passing through a heat exchanger.

3.2.4. Absorber

Table 3 below details the absorber design specifications modelled in Aspen Plus. The optimum column dimensions were obtained using sensitivity analysis, avoiding the possibility of hydraulic infeasibility if too small of an absorber diameter was selected. The internal structured packing of choice was Sulzer MELLAPAK™ 250Y due to its high specific surface area, allowing for an increased absorbance per unit area of packing material (Notz et al., 2012). The recycled lean MEA stream entering the top of the absorber has a lean loading of 19% and a MEA concentration of 30wt%, which are in accordance with values stated in literature (Alie et al., 2005). Optimal MEA concentration and lean loading values were enforced using design specifications within Aspen Plus where MEA and water compositions in the recycle stream were specified. By optimising these values, the total regeneration energy of the overall process is minimised, with the stripper reboiler consuming the most energy.

Table 3: Optimised Absorber Specifications

Design Variable	Specification
Pressure (bar)	0.75 (isobaric)
Packing Type	MELLAPAK™ 250Y
Height (m)	12
Diameter (m)	2.5
Number of Stages	20

3.2.5. Rich/Lean Heat Exchanger

The cold CO₂-rich solvent stream leaving the absorber is heated to 94°C by passing through a shell and tube heat exchanger, where the stream is contacted with the hot recycled lean solvent stream leaving the stripper reboiler.

3.2.6. Stripper

Table 4 details the stripper design specifications modelled in Aspen Plus. Like the absorber, Sulzer MELLAPAK™ 250Y was selected as the internal packing to facilitate adequate CO₂ desorption.

Table 4: Optimised Stripper Specifications

Design Variable	Specification
Pressure (bar)	2.01 (isobaric)
Packing Type	MELLAPAK™ 250Y
Height (m)	4.5
Diameter (m)	2.317
Number of Stages	25
Boil Up Ratio	0.131

The stripper reboiler was designed as a kettle-type reboiler which operates at 123°C. Both sensitivity analysis and Aspen Plus aided optimisation functions were run to minimise reboiler energy consumption. As a result, table 4 details the most economically favourable design specifications. At the given lean loading values and MEA concentration as specified in Section 3.2, the minimum energy consumption of the reboiler was determined to be 3.98GJton⁻¹(CO₂), in accordance with literature values outlined to be 4.00GJton⁻¹(CO₂) (Soltani et al., 2017).

The reboiler duty provides the heat of CO₂ desorption, the vaporisation of water as well as the

sensible heat (Lin & Rochelle, 2014). The reboiler heating requirement is supplied by bleeding steam between the high-pressure and low-pressure turbines in the energy generation plant, where total condensation is assumed with no further sub-cooling. As a result, a ratio of $1.94 \frac{kg(steam)}{kg(CO_2 \text{ captured})}$ was obtained for the given capture efficiency of 95%, aligning with values reported by Idem, Gelowitz and Tontiwachwuthikul of $1.9 - 2.5 \frac{kg(steam)}{kg(CO_2 \text{ captured})}$ (Idem et al., 2009).

3.2.7. CO₂ Compression

For carbon dioxide to be transported and further stored, CO₂ must be compressed to a pressure above its critical pressure of 73.8 bar. In this study, a discharge pressure of 110bar was chosen in accordance with research conducted by Goto, Yogo and Higashii (2013).

3.3. Heat Integration

To improve the economic potential of the process, heat integration was carried out in order to minimise the total amount of heat duty required for the heating and cooling of streams, by transferring heat between streams. Hot and cold streams are to be identified and in the case of this plant, 1 cold stream and 3 hot streams were identified as shown in the heat exchanger network, Figure 3.

An analysis under the first law of thermodynamics was carried out to determine the minimum duty requirement after assuming full heat integration. It is found that $Q_{min} = -8.64MW$, indicating that heat needs to be removed from the process. Subsequently, a further analysis under the second law of thermodynamics was conducted which states that heat can only flow from hot streams to cold streams and not vice versa. A minimum allowable temperature difference of 10K is applied to ensure that there is enough driving force for effective heat transfer. When conducting a Grand Composite Curve, no pinch point was present meaning that heat can be transferred across all the streams involved.

Upon completion, a total of 0.68MW is integrated for a final cooling duty of 8.64MW which results in a 7.3% improvement over the non-integrated system.

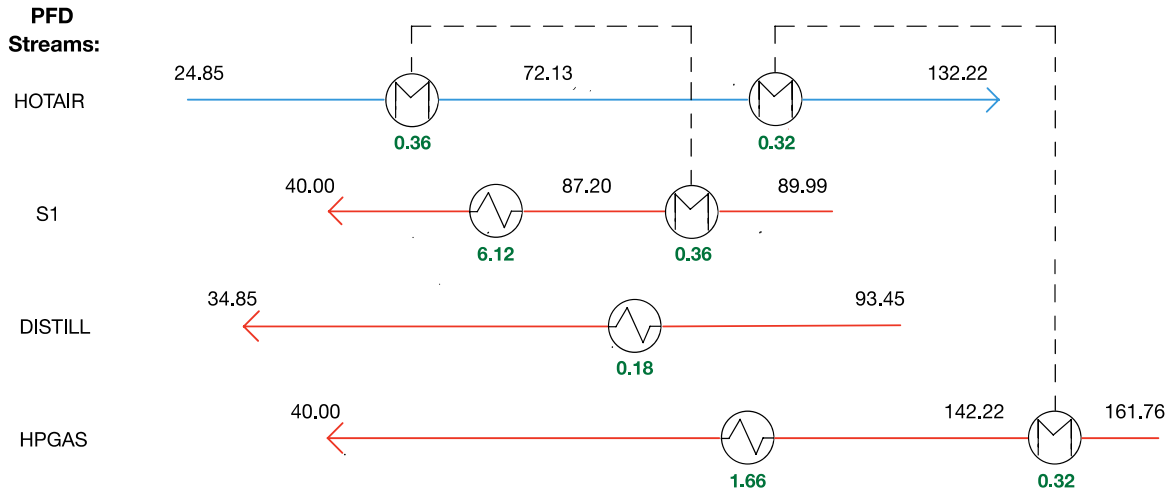


Figure 3: Heat Exchanger Network

4. Economic Study

To explore the economic feasibility of retrofitting a post combustion carbon capture plant on the RDF incineration system, both the plant's capital costs, and utility costs were explored.

4.1. Equipment Costing

Manual equipment costs were interpolated with process design correlations (Douglas, 1988), it should be noted that the Chemical Engineering Plant Cost Index (CEPCI) was employed for updated costings in reflection to the year of this study ($CEPCI_{1968} = 113.7$, $CEPCI_{2023} = 800.8$) (University of Manchester, 2023). Process units were costed for the UK with an exchange rate of $\$1.00 = \pounds0.79$ for 2023.

4.2. Sample Costings – Compressor & Stripper

The installed cost of the 110bar gaseous compressor was calculated using Guthrie correlations in Equation 9.

$$\text{Installed Cost, \$} = 517.5 \times bhp^{0.82} (2.11 + F_c) \quad (9)$$

Where Bhp is the brake horsepower, measured at 2415 on Aspen Plus V11 and F_c is the correlation factor.

The equipment cost of the stripper was broken into its constituent components, the column and the reboiler.

$$\text{Installed Cost, \$} = 101.9D^{0.82}H^{0.802}(2.18 + F_c) \quad (10)$$

$$\text{Installed Cost}_{reb}, \$ = 101.3A^{0.65}(2.29 + F_c) \quad (11)$$

$$\text{where } A = \frac{Q}{U\Delta T_{LM}} \quad (12)$$

Equation 10 represents the installed cost for the column, D is given as the column diameter in feet, H is the column height in feet and F_c is the correlation factor dependent on the column pressure and shell material.

Equation 11 specifies the installed cost of the stripper reboiler, where A is the heat transfer area of the reboiler in ft^2 , $U\Delta T_{LM}$ was approximated as 11,250 $\text{BTU/hr}\cdot\text{ft}^2$ from Douglas. F_c is given as 1.35 for a kettle type reboiler, the reboiler heat duty Q was extracted from Aspen Plus as 33,473,700 BTU/hr .

CEPCI index from the 1968 and 2023 were then used to update costings in reflection to the year of this study, giving a final installed compressor cost of approximately \$6,840,000 and stripper cost of \$775,000 prior to GBP conversion.

4.2.1. PCC Plant

Table 5 details the installed costs of all major process units within the PCC plant.

Table 5: Installed Cost for PCC Plant

Process Unit	Installed Cost [£]
Absorber	418,969
Stripper	617,118
Compressors	6,493,183
Heat Exchangers	180,637
Coolers	383,496
Pressure Vessels	58,016

The total installed cost for retrofitting the simulated PCC was £8,183,000. The most notable contribution attributing to the systems compressors, accumulating to over 80% of the total cost as shown on Figure 4. The compressor operating at 110bar, amassed to an installed cost of £5,450,000 alone. It was impossible to minimise the cost of the compressors due to the pressure requirements within the pipeline system and final CO_2 gaseous outlet. Despite running computer-aided optimisation and sensitivity analysis to achieve the most economically favourable column specifications, both the absorber and stripper columns combined still constitute to over 12% of the overall installed cost.

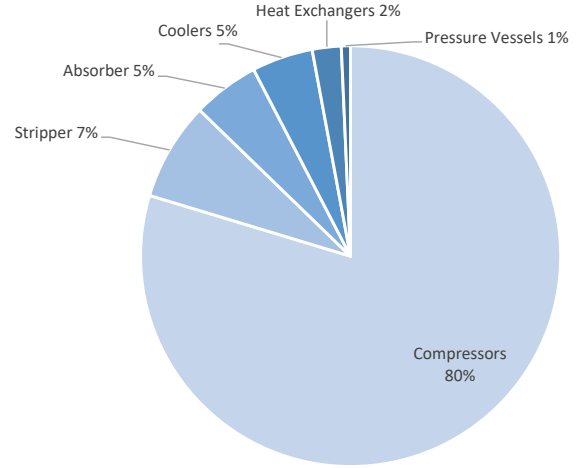


Figure 4: Installed Cost of PCC Plant

4.2.2. WtE Plant

The installed cost for the WtE plant was evaluated similarly to the PCC plant, Table 6 displays the equipment breakdown for all major process units, again the installed cost of mixers and splitters were deemed negligible. The installed cost for the deployment of a WtE plant is calculated to be £8,950,000.

Table 6: Installed Cost for RDF Plant

Process Unit	Installed Cost [£]
Fuel Incinerator	1,594,867
Furnace	3,156,500
Pressure Vessels	331,178
Turbines	3,302,118
Coolers	424,815
Cyclone Separator	114,055
Pumps	22,848

4.3. Annualising Installed Costs

Installed costs are considered as capital expenditures in this study. The following formula was utilised, where d and n denote discount rate and expected life of plant in years respectively:

$$\text{Annualised CAPEX} = \text{CAPEX} \times \frac{d}{1 - (1 + d)^{-n}} \quad (13)$$

The annualisation of capital expenditures was executed with a conservative strategy, employing a 12% discount rate. This decision was prompted by the challenging market conditions prevailing in the UK, characterised by notably high interest rates of 5.25%. The expected life of the plant was considered to be 20 years. After applying the formula, annual capital expenditures of £1.09 million and £0.99 million are calculated for the PCC and RDF plants, respectively. This results in a total capital expenditure of £2.08 million.

4.4. Operational Expenditures

Utilities are essential to provide the required energy to operate various units and processes within the plant. For the PCC plant, there are three utilities required: cooling water, electricity and steam.

Cooling water is used for the 3 coolers within the process which amounts to a consumption of 406 tonnes/hr at £0.85 per tonne (Driver et al., 2022) resulting in a cost of £2.76 million per year.

The electricity required for the plant is mainly required for the compressors with minor amounts required by the pumps and this will be provided by the RDF to energy plant. This electricity, 1.95MW, will be negated from the revenue generated by the plant which amounts to £1.88 million per year at 12p per kWh (Scottish Power, 2022).

Steam is required for the reboiler, however, this is covered by the steam that is bled from the energy plant. An additional operational expense to take into account involves the RDF feed, totalling 4.6 tonnes/hr. The RDF incurs a cost of £90 per tonne, resulting in an annual feed expense of £3.31 million.

4.5. Revenue Generation

The sole revenue source is the electricity generated in the power generation section of the RDF to energy plant. While the total power output from this section is 7.90MW, the PCC plant necessitates 1.95MW, leaving a revenue-generating capacity of 5.95MW. This produces a total annual energy output of 47.6GWh, consequently, the total revenue generated amounts to £5.71 million.

This low revenue generation may be attributed to RDF's inefficiency as a fuel coupled with the electricity requirement of the PCC plant.

4.6. Economic Potential

The combined annual cost sums to be £8.15million which exceeds the revenue generated, leaving the economic potential to be -£2.44million.

5. Energy Penalty

As a result of integrating CCS on a WtE plant, the energy consumption of the plant increases drastically contradictory to the main purpose of a power plant by imposing an energy penalty on the overall system. One way of calculating the impact of CCS is by introducing the energy consumption of capture per MWh of electrical energy, this is calculated with parameters detailed in Section 3.2.7 using equation 14 below (Soltani et al., 2017).

$$E \left(\frac{GJ}{MWh} \right) = A \left(\frac{GJ}{t_{CO_2}} \right) \times \frac{B \left(\frac{t_{CO_2}}{MWh} \right)}{\eta} \times \frac{C}{100} \quad (14)$$

E is the regeneration energy consumption for capture per MWh of electricity produced, A is the energy consumption of the reboiler, B is the amount of CO₂ generated per MWh of thermal energy produced, η is the power plant efficiency and C is the capture rate of CO₂.

Table 7: Energy Penalty per MWh Comparison

Fuel Type	GJ/ t _{CO2}	GJ/ MWh
RDF	3.98	4.52
CCGT	3.98	1.50
Coal (bituminous) - fired	3.90	2.33

Table 7 demonstrates the continuation of Soltani, Fennell and Dowell's work on the regeneration energy consumption between combined cycle gas turbine (CCGT) and coal-fired (bituminous) power plants with 600MW_{th} capacity at a carbon capture rate of 90%.

By comparing the simulated RDF plant with Soltani, Fennell and Dowell's study, integrating PCC on an RDF plant incurs the either the same or a greater energy penalty per tonne of CO₂ captured when compared to CCGT (3.98 GJ/t_{CO2}) and coal-fired power plants (3.90 GJ/t_{CO2}). However, RDF power plants suffer a much greater energy penalty with regards to electrical energy produced, with the regeneration energy consumption for capture per MWh of electricity produced being more than double in comparison to CCGT power plants (1.50 GJ/MWh). This is explained due to the poor efficiency of the stimulated RDF power plant (31%), which is significantly lower than the efficiency of the CCGT and coal-fired power plant which is stated to be 48% and 60% respectfully.

In addition, energy penalty can be measured by determining the difference in overall power output of the plant before and after retrofitting PCC. For the simulated 95% capture efficiency, an energy penalty of 53.3% was imposed. As demonstrated in Table 8, the energy penalty for the WtE plant is more than two-fold in comparison to other fuel-powered plants observed in literature with a given capture efficiency of 90%.

Table 8: Energy Penalty % Comparison

Fuel Type	Energy Penalty (%)
RDF	53.3
NGCC	21
Coal (bituminous) - fired	29

6. Discussion

In this study, the primary objectives of assessing carbon capture readiness involve an initial analysis of both technical and economic feasibility. As depicted on Figure 3, retrofitting a PCC plant continues to enable over 12.8MW of power generation. Nevertheless, it is important to highlight that even with the optimisation of the reboiler duty to 3.98GJton⁻¹(CO₂), aligned with literature values, the overall system still imposes a substantial energy penalty of 53.3%. A noticeable reduction in the power plant's electricity output is observed before and after retrofitting PCC.

Additionally, a significant steam flowrate is bled between the series of turbines, as illustrated in Figure 5, leading to a decrease in the overall efficiency of the power plant. This phenomenon could be explained due to the substantial energy regeneration demand of the selected MEA solvent. Nevertheless, when compared to traditional power plants such as coal and NGCC, RDF-to-energy plants incur a notably higher energy penalty, which could be attributed to the lower efficiency of the RDF fuel. The lower efficiency of the fuel results in a higher amount of CO₂ released for a given MWh of electrical energy generated. Consequently, the energy required for CO₂ capture is inherently higher.

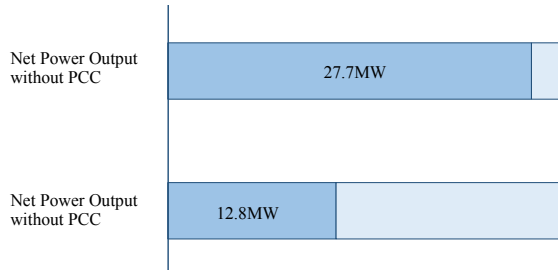


Figure 5: Power Output with and without PCC

It is crucial to acknowledge that while these comparisons are adjusted for their respective CO₂ and electrical outputs, there exists empirical distinctions in the design parameters among the three different types of power plants. Firstly, both CCGT and coal-fired power plants boast a significantly higher capacity of 600MW compared to the simulated RDF plant. The overall efficiency of the RDF plant is likely to improve with a proportional increase in capacity. Additionally, if the capture efficiency of other fuel-powered plants were enhanced to 95%, the energy consumption would inherently rise. Despite these variations, the conclusions drawn still provide valuable insights for ongoing and future research.

The combined economic potential of the RDF-to-energy plant with the PCC retrofit is -£2.44 million, indicating the economic infeasibility of such a plant. Even assuming the existence of an original RDF plant, the retrofitted PCC plant remains economically unviable, with an economic potential of -£1.45 million. To enhance economic feasibility, two potential exploration avenues are identified: increasing revenue generation and reducing costs.

In this study, only revenue generated from electricity production was considered. However, there was also thermal energy that was ejected in the cooling water which could be used for the heating of additional buildings which could generate additional revenue (Environ Consultants Ltd., n.d.).

When considering costs, it can be noticed that the expense of the RDF feed is notably high. Exploring a potential collaboration with a waste-producing company could present a mutually beneficial opportunity. By collecting waste from this company, both parties stand to gain. The waste-producing company can realise savings on disposal costs and landfill tax, while the RDF to energy plant would see a substantial reduction in its operational expenditures.

7. Conclusion

In this investigation, a 12.8MW waste-to-energy plant, incorporating a retrofitted MEA-based post-combustion carbon capture plant, was modelled using Aspen Plus V11. The simulation targeted a small-scale UK plant processing 4600kg/hr of RDF, aligning with literature key operating parameters. The introduction of PCC resulted in an overall increase in the plant's energy consumption, imposing an energy penalty of 53.3% on the system. The carbon capture efficiency was set at 95%, with a specified MEA concentration and lean loading values of 30wt% and 19% respectively. The overall plant efficiency was determined to be 31%, and the optimal reboiler duty was established to be 3.98GJton⁻¹(CO₂), supplied by a flow of high temperature steam bled between the HP and LP turbines at a ratio of $1.94 \frac{kg(steam)}{kg(CO_2 \text{ captured})}$. The findings revealed that despite comprehensive optimisation on the overall system, the implementation of PCC to a small-scale RDF plant had a substantial impact on overall plant performance. Whilst 5.95MW_e of electrical energy was generated, the inherently low efficiency of RDF fuel led to a significantly higher energy penalty per MWh of electrical energy compared to other conventional fuels. Although there was a positive energy generation, the overall cost of the system surpassed the revenue generated from electrical energy. Consequently, the study demonstrated a negative economic potential of -£2.44million, highlighting severe challenges in terms of economic feasibility.

References

- Alfè, M., Gargiulo, V., Porto, M., Migliaccio, R., Le Pera, A., Sellaro, M., Pellegrino, C., Abe, A.A., Urciuolo, M., Caputo, P., Calandra, P., Loise, V., Rossi, C.O. and Ruoppolo, G. (2022). Pyrolysis and Gasification of a Real Refuse-Derived Fuel (RDF): The Potential Use of the Products under a Circular Economy Vision. *Molecules*, [online] 27(23), p.8114. doi:<https://doi.org/10.3390/molecules27238114>.
- Alie, C. *et al.* (2005) 'Simulation of CO₂ capture using MEA scrubbing: A flowsheet decomposition method', *Energy Conversion and Management*, 46(3), pp. 475–487. doi:10.1016/j.enconman.2004.03.003.
- Aspentech, *Getting Started Modeling Processes with Solids*, A. Technology, Editor. 2013, Aspen Technology.
- Botakoz Suleimenova, Berik Aimbetov, Daulet Zhakupov, Shah, D. and Yerbol Sarbassov (2022). Co-Firing of Refuse-Derived Fuel with Ekibastuz Coal in a Bubbling Fluidized Bed Reactor: Analysis of Emissions and Ash Characteristics. *Energies*, 15(16), pp.5785–5785. doi:<https://doi.org/10.3390/en15165785>.
- Chavando, J.A.M., Silva, V.B., Tarelho, L.A.C., Cardoso, J.S. and Eusébio, D. (2022). Snapshot review of refuse-derived fuels. *Utilities Policy*, 74, p.101316. doi:<https://doi.org/10.1016/j.jup.2021.101316>.
- Douglas, J.M. (1988) *Conceptual design of Chemical Processes*. New York, New York: McGraw-Hill Book Co.
- Driver, J.G., Hills, T., Hodgson, P., Sceats, M. and Fennell, P.S. (2022). Simulation of direct separation technology for carbon capture and storage in the cement industry. *Chemical Engineering Journal*, [online] 449, p.137721. doi:<https://doi.org/10.1016/j.cej.2022.137721>.
- Environ Consultants Ltd., (n.d.). *UK Refuse Derived Fuel (RDF) Energy from Waste - Environ Consultants LTD*. [online] Available at: <https://environltd.co.uk/uk-refuse-derived-fuel-energy-from-waste/> [Accessed 27 Nov. 2023].
- Errico, M. *et al.* (2016) 'Model calibration for the carbon dioxide-amine absorption system', *Applied Energy*, 183, pp. 958–968. doi:10.1016/j.apenergy.2016.09.036.
- Fröhling, W., Unger, H.-M. and Dong, Y. (n.d.). *THERMODYNAMIC ASSESSMENT OF PLANT EFFICIENCIES FOR HTR POWER CONVERSION SYSTEMS*. [online] Available at: https://inis.iaea.org/collection/NCLCollectionStore/_Public/33/033/33033053.pdf
- García, R., González-Vázquez, M.P., Rubiera, F., Pevida, C. and Gil, M.V. (2021). Co-pelletization of pine sawdust and refused derived fuel (RDF) to high-quality waste-derived pellets. *Journal of Cleaner Production*, 328, p.129635. doi:<https://doi.org/10.1016/j.jclepro.2021.129635>.
- Get paid for generating Green Electricity* (N/A) ScottishPower. Available at: <https://www.scottishpower.co.uk/smart-export-guarantee> (Accessed: 14 December 2023).
- Goto, K., Yogo, K. and Higashii, T. (2013) 'A review of efficiency penalty in a coal-fired power plant with post-combustion CO₂ Capture', *Applied Energy*, 111, pp. 710–720. doi:10.1016/j.apenergy.2013.05.020.
- Idem, R., Gelowitz, D. and Tontiwachwuthikul, P. (2009) 'Evaluation of the performance of various amine based solvents in an optimized multipurpose technology development pilot plant', *Energy Procedia*, 1(1), pp. 1543–1548. doi:10.1016/j.egypro.2009.01.202.
- Jannelli, E. and Minutillo, M. (2007) 'Simulation of the flue gas cleaning system of an RDF incineration power plant', *Waste Management*, 27(5). doi: 10.1016/j.wasman.2006.03.017.
- Jung, J. *et al.* (2013) 'Advanced CO₂ capture process using mea scrubbing: Configuration of a split flow and phase separation heat exchanger', *Energy Procedia*, 37, pp. 1778–1784. doi:10.1016/j.egypro.2013.06.054.
- Li, K. *et al.* (2016) 'Systematic study of aqueous monoethanolamine (mea)-based CO₂ Capture Process: Techno-economic assessment of the MEA process and its improvements', *Applied Energy*, 165, pp. 648–659. doi:10.1016/j.apenergy.2015.12.109.
- Lin, Y.-J. and Rochelle, G.T. (2014) 'Optimization of advanced flash stripper for CO₂ Capture using piperazine', *Energy Procedia*, 63, pp. 1504–1513. doi:10.1016/j.egypro.2014.11.160.
- Luis, P. (2016) 'Use of monoethanolamine (MEA) for CO₂ capture in a global scenario: Consequences and alternatives', *Desalination*, 380, pp. 93–99. doi:10.1016/j.desal.2015.08.004.
- Materazzi, M., Lettieri, P., Mazzei, L., Taylor, R. and Chapman, C. (2015). Fate and behavior of inorganic constituents of RDF in a two stage fluid bed-plasma gasification plant. *Fuel*, 150, pp.473–485. doi:<https://doi.org/10.1016/j.fuel.2015.02.059>.
- Miqdad Mehdi et al 2020 IOP Conf. Ser.: Earth Environ. Sci. 565 012096

Moioli, S., Pellegrini, L.A. and Gamba, S. (2012) 'Simulation of CO₂ capture by mea scrubbing with a rate-based model', *Procedia Engineering*, 42, pp. 1651–1661. doi:10.1016/j.proeng.2012.07.558.

Nixon, J.D., Wright, D.G., Dey, P.K., Ghosh, S.K. and Davies, P.A. (2013). A comparative assessment of waste incinerators in the UK. *Waste Management*, [online] 33(11), pp.2234–2244. doi:https://doi.org/10.1016/j.wasman.2013.08.001.

Notz, R., Mangalapally, H.P. and Hasse, H. (2012) 'Post combustion CO₂ capture by reactive absorption: Pilot plant description and results of systematic studies with mea', *International Journal of Greenhouse Gas Control*, 6, pp. 84–112. doi:10.1016/j.ijggc.2011.11.004.

Soltani, S.M., Fennell, P.S. and Mac Dowell, N. (2017) 'A parametric study of CO₂ capture from gas-fired power plants using monoethanolamine (MEA)', *International Journal of Greenhouse Gas Control*, 63, pp. 321–328. doi:10.1016/j.ijggc.2017.06.001.

University of Manchester. (2023) *Chemical Engineering Plant Cost Index, Fluid mechanics*. Available at: https://personalpages.manchester.ac.uk/staff/tom.rodgers/Interactive_graphs/CEPCI.html?reactors%2FCEPCI%2Findex.html (Accessed: 08 December 2023).
www.london.gov.uk. (n.d.). *Trash or treasure? Time to rethink our waste habits | London City Hall*. [online] Available at: <https://www.london.gov.uk/press-releases/assembly/trash-or-treasure#:~:text=Local%20authorities%20in%20London%20collected>.

Exploring Multi-Fidelity Bayesian Optimization and TuRBO-1 for Enhanced Engineering Solutions

Jeremy Kohn and Vincent Verkammen

Department of Chemical Engineering, Imperial College London, U.K.

Abstract

Bayesian Optimization (BO) has emerged as a pivotal tool for effectively navigating the optimization of intricate, high-dimensional functions, especially in cases where derivative information is unavailable. This study presents a deep dive into the intricacies of BO, starting with a foundational algorithm, and progressing towards integrating advanced methods such as multi-fidelity and TuRBO-1 to tackle the key challenges of high computational cost and the dimensionality curse, often encountered in machine learning and engineering applications. In this context, the study aims to deliver a framework for two robust, scalable, and efficient BO methods to offer a comprehensive toolset for real-world applications, while also providing insights into the inner dynamics and behaviour of the algorithms. Specifically, a novel acquisition function is developed featuring a γ hyperparameter to provide a more flexible and nuanced trade-off between the cost and covariance at different fidelities. The study demonstrates the effectiveness of the proposed multi fidelity framework in achieving lower computational cost, and highlights the practicality of tuning γ to adapt to any problem. Accurate optima of various test functions in 2, 5, and 10 dimensions were achieved, consistently beating the cost output of the old methodology. Moreover, the TuRBO-1 method achieved a complementary solution when dealing with higher dimensions by prioritizing effective local optimisation within dynamically controlled trust regions, which allows it to adeptly handle the intricacies of complex, high-dimensional spaces. The proposed research therefore highlights the practicality of two adaptable strategies designed to address a wide range of challenges, which lays solid foundation for future research and application of these frameworks.

1. Introduction

Bayesian Optimization (BO), an active learning framework, has emerged as a critical tool in complex systems design, where direct evaluations of complex, often non-linear systems are prohibitively expensive (Shahriari *et al.*, 2016). This is particularly true when dealing with black-box functions, whose underlying relationships are unknown or too intricate to be explicitly defined. Black-box objective functions are commonly found in fields such as robotics, automated machine learning (AutoML), engineering, and especially in the rapidly evolving domain of the chemical sciences (Terayama *et al.*, 2021). Indeed, processes in this field can be characterized by high-dimensional spaces with many governing variables, such as reaction conditions and material properties. The traditional approach of grid search or random sampling is often computationally prohibitive due to the high cost of evaluations, either in terms of experimental resources or computational time.

Unlike traditional methods, Bayesian Optimization utilizes a probabilistic model, typically Gaussian Processes, to create a surrogate model of the objective function, which quantifies uncertainty in areas where the function is sampled less. This surrogate model, unlike the underlying objective function, is computationally efficient to evaluate, which is the key to BO's practicality and effectiveness, in particular with regards to expensive and complex functions. The surrogate model's predictive capabilities enable navigation of the search space by estimating the outcomes of various configurations, thereby focusing the exploration on areas with the highest potential for improvement. This allows for a more targeted exploration,

significantly reducing the number of evaluations needed to reach optimal or near-optimal solutions. As such, BO incorporates a crucial balance between exploration of new, potentially promising areas, and exploitation of known high-performing regions. This balance is key in environments where each evaluation is costly, ensuring that resources are utilized in the most effective manner.

The practical implications of Bayesian Optimization in chemical engineering and other sciences are far-reaching. In an industry where precision, safety, and efficiency are paramount, BO has the potential to revolutionize the way chemical processes are optimized.

One notable application is the early-stage process development of pharmaceutical compounds. In a recent study, Braconi *et. Al* (Braconi and Edouard Godineau, 2023) optimised sustainable reaction conditions for C-N coupling using copper catalysts and non-hazardous solvents through Bayesian Optimization. BO was able to efficiently explore a vast reaction space of over 138,000 possible experiments, using only 80 simulations. This represents an exploration of less than 0.05% of the total space, effectively highlighting its efficiency in identifying optimal conditions in complex chemical processes by leveraging its probabilistic model to iteratively refine and direct the search towards the most promising areas. A similar study was able to navigate the vast space of potential drug-like molecules to enable the discovery of antimalarial compounds, and molecules with targeted activity against pulmonary fibrosis, where it outperformed traditional greedy search methods (E. O. Pyzer-Knapp, 2018).

A different practical application is the optimization of wind farm layouts to maximize sustainable power output (Bempedelis and Magri, 2023). The authors highlight how BO, and the use of Gaussian processes enabled the capture and exploitation of complex flow dynamics, which are usually overlooked in simpler wake models. Bayesian Optimization is also used to optimize machine learning systems and server performance, as demonstrated in real-world applications at Facebook where BO was utilized for the optimization of a ranking system and server compiler flags (Letham et al., 2019). Examples within literature underscore the versatility of BO in addressing pressing challenges in chemical engineering and beyond, making it an indispensable tool in the advancement of the field. In this context, the main objective of this project is to achieve a working, scalable, and efficient BO algorithm capable of optimizing a wide range of multimodal and multidimensional test functions. Moreover, the aim is to expand the BO capabilities by incorporating advanced techniques such as Multi-Fidelity and TuRBO-1 for enhanced performance and reduced computational cost. As such, the objective is to achieve robust solutions which can be applied in the field, and contribute towards tackling the key challenge of computationally expensive problems. Finally, this paper aims to provide sensitivity-analysis on key parameters to gain valuable insights into the algorithm’s behaviour, and lay the foundations for its use in real-world scenarios.

2. Background

BO is fundamentally about devising a surrogate model to navigate an expensive black-box function that is potentially non-differentiable. Mathematically, the objective function to be minimised is

$$\min_{x \in X} f(x) \quad (1)$$

where f represents the unknown, expensive-to-evaluate function. Despite this, the function f can be probed by costly evaluations at various points within its domain X , with the intent to minimize f utilizing the least number of evaluations.

Viewed as a sequence of decisions, BO requires selecting the next point- or batch of points- in the domain for evaluation in each iteration, guided by prior observation. To manage this effectively, a representation of the uncertainty about f is updated progressively with new data. Gaussian Processes (GP) are ideally suited for this role.

Surrogate models lie at the heart of BO and are used to model the black-box function. GPs are a natural choice for this model, as they estimate the function’s value across its domain and, importantly, provide a predictive posterior distribution that reflects the potential range of function values. They are defined by a mean function $\mu(x)$ and a covariance function $k(x, x')$:

$$f(x) \sim GP(\mu(x), k(x, x')) \quad (2)$$

The mean function $\mu(x)$ represents the average predicted output of the model. The covariance function $k(x, x')$, also known as the kernel, is parametrised by a variance (σ^2) and lengthscale (θ). This function, which can be selected from common types such as the Radial Basis Function (RBF) and Matérn kernels, defined as

$$k_{\text{Matern}}(d) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \cdot \left(\sqrt{2\nu} \frac{d}{\theta} \right)^\nu \cdot K_\nu \left(\sqrt{2\nu} \frac{d}{\theta} \right) \quad (3)$$

$$k_{\text{RBF}}(d) = \exp \left(\frac{-d^2}{2\theta^2} \right) \quad (4)$$

where Γ is the gamma function, d is the Euclidian distance between two points in the input space, K_ν is the modified Bessel function of the second kind (Weisstein, E. W., 2023), and ν is a parameter from the covariance that controls the smoothness of the function. The Matérn kernel becomes equivalent to the RBF kernel as ν approaches infinity. The kernel dictates how function values correlate across the input space, and it encapsulates the assumptions regarding the function’s variability and smoothness characteristics. To ensure the model aligns effectively with the observed data, the tuning of hyperparameters is essential. This tuning process aims to maximise the log-likelihood, which is expressed as

$$\log[p(Y|X)] = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log(\det[K(\theta)]) - \frac{1}{2} Y^T [K(\theta)]^{-1} Y \quad (5)$$

where Y represents the vector of observed target data, X is the matrix of input data where each row is an input vector, n is the number of observations, $K(\theta)$ is the covariance matrix derived from the covariance function $k(x, x')$ with θ representing the hyperparameters (variance σ^2 and lengthscale θ), and the term $Y^T [K(\theta)]^{-1} Y$ represents the ‘goodness of fit’. Maximizing this log-likelihood function is crucial for refining the model, providing the best statistical explanation for the observed data under the GP model assumptions.

This optimisation process is fundamental in fine-tuning the surrogate model. Such refinement enhances the model’s predictive accuracy and strengthens the reliability of uncertainty quantification. The improved model becomes instrumental in the subsequent stage of selecting query points. This crucial step employs acquisition functions, a methodological approach designed to direct the choice of subsequent points to be sampled from the objective function.

Figure 1 illustrates the Gaussian Process post hyperparameter tuning. The figure shows the mean function $\mu(x)$ as a solid blue line, with the shaded region indicating the confidence interval. This interval, based on the GP's predictive variance σ^2 quantifies uncertainty. It facilitates a strategic balance in exploring and exploiting the search space for the function's minima.

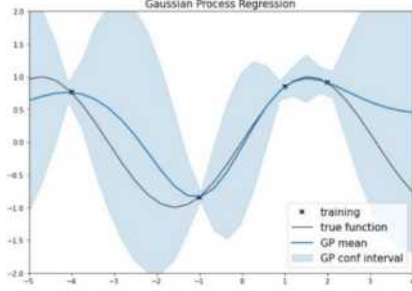


Figure 1 – Gaussian Process Regression (D.R. Chanona, 2023)

The acquisition function integrates both the mean ($\mu(x)$) and the uncertainty (σ^2) projections derived from the Gaussian process. This integration is key to striking a balance between exploration-investigating new, potentially promising areas of the function's domain, and exploitation- focusing on regions known to yield high values. Such a precisely calibrated approach significantly boosts the efficiency and efficacy of the exploration process. Consequently, this leads to a quicker convergence towards the optimal point of the function, thereby enhancing the overall effectiveness of the model.

The landscape of acquisition functions is diverse, each with unique benefits and limitations, as detailed by (Shahriari et al., 2015).

Among these functions, the Expected Improvement (EI), Probability of Improvement (PI), and Upper Confidence Bound (UCB) are defined as follows

$$EI(x) = E[\max(f(x) - y_{best}, 0)] \quad (6)$$

$$PI(x) = P(f(x) > y_{best}) \quad (7)$$

$$UCB(x) = \mu(x) + \beta\sigma(x) \quad (8)$$

where y_{best} represents the best observed value, β is the bonus for exploration and σ is the standard deviation. EI quantifies the anticipated improvement over the current best observation, incorporating both mean and variance from the Gaussian Process. This method balances exploration and exploitation. PI, on the other hand, assesses the likelihood of surpassing the current best observation, focusing more on exploitation. UCB merges the predicted mean and variance, adjusting exploration and exploitation through the parameter β , with higher values favouring exploration. The selection of these acquisition functions is contingent on the optimization problem's specific needs, particularly in balancing exploration of new search areas against exploiting known optimal regions.

Optimizing the acquisition function is a critical component of BO. This process often employs gradient ascent or evolutionary strategies to navigate complex, high-dimensional spaces.

Subsequently, optimisation algorithms like ADAM (Brownlee, 2021) are applied to maximize the acquisition function, aiding in the selection of the next point evaluation.

A pivotal extension of the BO framework lies in the concept of multi-fidelity optimization. It is an approach that introduces an additional layer of efficiency by leveraging a range of data sources of varying accuracy and cost – the so-called fidelities. It combines high-fidelity models, which are accurate but expensive to evaluate, and low-fidelity models, which are less accurate but cheaper and faster to compute. By combining insights from these different levels of fidelity, BO can make more informed decisions about where to allocate resources for evaluation. In practice, this is achieved by extending the GP model to incorporate fidelity as an additional input dimension. This extension affects both the kernel and mean function, allowing them to interpret and integrate data across various fidelity levels effectively. The acquisition function is adapted to evaluate not just the predictive performance at each point, but also to consider the varying computational costs associated with different fidelity levels.

Bayesian Optimization relies heavily on the ability to construct a global model that is accurate enough to eventually uncover a global optimizer. This task presents significant challenges due to the curse of dimensionality and the heterogeneous nature of the function. In contrast to the approach presented by multi-fidelity, another innovative strategy in reducing computational costs and achieving sample-efficient optimization is the application of trust regions (TR). Trust Region Bayesian Optimization (TuRBO), unlike traditional BO which operates across the entire search space, conducts BO within multiple local trust regions. This novel use of local BO, combined with dynamically adjusting trust regions, effectively addresses some of the key challenges that have hindered the success of conventional global optimization methods, providing a more efficient and focused approach to optimization. The key TR parameters that should be fine-tuned to the bounds and dimensionality of the problem, are the success (τ_{succ}) and failure (τ_{fail}) threshold, the minimum (L_{min}) and maximum (L_{max}) diameter of the ellipsoidal TR, the TR centre ($centre_{TR}$) and TR radius (r_{TR}).

3. Methodology

Initial BO

The initial phase of this research involved the development of a Bayesian Optimization algorithm from scratch. This foundational step was essential to gain a thorough understanding of the underlying mechanics of BO and to set the stage for more advanced implementations. Starting from basic principles, the objective was to gradually incorporate more sophisticated techniques to achieve a robust and well-rounded algorithm.

Three n-dimensional test functions were selected for the optimization trials due to their varying levels of complexity and landscapes, offering a comprehensive evaluation platform for the BO algorithm. Each represents a unique optimization challenge, from simple convex shapes to more complex, multimodal landscapes. In order of increasing complexity:

- Sphere function, smooth and convex with a single global minimum:

$$f(x) = \sum_{i=1}^n x_i^2 \quad (9)$$

- Rosenbrock function, known for its long, narrow, parabolic shaped flat valley containing the global minimum:

$$f(x) = \sum_{i=1}^{n-1} [100(x_{i+1} - x_i^2)^2 + (1 - x_i)^2] \quad (10)$$

- Styblinski-Tang Function, defined by a complex landscape with multiple local minima:

$$f(x) = \frac{1}{2} \sum_{i=1}^n x_i^4 - 16x_i^2 + 5x_i \quad (11)$$

Gaussian Process Regression was utilized as the surrogate model due to its flexibility and proficiency in estimating uncertainty in predictions, and multiple kernels were explored, including the Radial Basis Function, Matérn kernel, and a combined kernel consisting of RBF, white kernel (introducing a noise term), and constant kernel (which scales the overall variance), each imparting distinct characteristics to the surrogate model.

Three acquisition functions were implemented and analysed: Expected Improvement, Probability of Improvement, and Upper Confidence Bound.

Additionally, two distinct search methods were used: whole space sampling, where the algorithm explored the entire domain of the objective function, and focused sampling, which concentrated the search around regions that were already identified as promising.

The BO process started by generating a randomly distributed set of 5 / 10 / 20 (for 2D / 5D / 10D) initial samples (in accordance with the chosen search strategy) to initialize the gaussian process by computing the objective at each point. It then enters its iterative phase, where 15 / 40 / 80 iterations are performed utilizing the acquisition function's

balance between exploration and exploitation. At each evaluation, the surrogate is updated, and the GP refines its understanding of the objective function's behaviour, thus converging towards the optimum.

The outcomes such as the found optimum, the number of iterations, and the overall runtime were then analysed for different combinations of acquisition function, kernel, and search strategy, effectively conveying their strengths and weaknesses with regards to different test functions.

It is worth noting that an early stopping mechanism was introduced to enhance computational efficiency, particularly in higher dimensions where the time started increasing rapidly. This mechanism halts the process when the improvement of objective value falls below a specified threshold over a set number of consecutive iterations (defined as patience), effectively avoiding unnecessary time loss when the optimum has already been found, or when the algorithm gets stuck in a local optimum.

Multi-fidelity

Building on the foundational knowledge gained from the initial algorithm, the research progressed to a more advanced stage with the implementation of multi-fidelity techniques. This approach significantly enhances the BO framework by incorporating evaluations at various levels of fidelity z , striking a balance between accuracy and computational cost. In this research, the possible fidelity levels are defined as $1 \leq z \leq 10$.

Similar to the initial BO implementation, the multi-fidelity version was tested using the Sphere, Rosenbrock, and Styblinski-Tang functions. A key aspect however, is the modelling of these objective functions to account for varying fidelities. The objective is augmented to incorporate fidelity levels, creating a composite function that simulates the objective across different levels of precision. This is achieved by scaling the input parameters of the objective function based on the fidelity level. The higher the fidelity, the more the objective function reflects the true, high-resolution behaviour of the system being modelled. Conversely, at lower fidelities, the function provides a coarser approximation. The scale used in this study is

$$f(x, z) = g(x \times \text{scale}(z)) \quad (12)$$

where f is the fidelity augmented objective function, and g is the original. The scale is defined as

$$\text{scale}(z) = \frac{z}{z_*} \quad (13)$$

where z_* represents the highest fidelity level (10). This approach allows for a simulated fidelity-adjusted objective, where evaluations at the highest fidelity result in the true objective optimum.

Similar to the simple BO algorithm, a GP model is used as surrogate. In this context, a covariance function that takes fidelity levels into account has been structured as a combination of a spatial kernel

k_s (Matérn) and a fidelity kernel k_f . This allows the model to consider both the distance between points in the input space and the difference in their fidelity levels. The kernel reflects that the similarity between points in the spatial domain is modulated by their corresponding fidelity levels. The adopted approach was to use a product of these kernels:

$$k((x, z), (x', z')) = k_s(x, x') \times k_f(z, z') \quad (14)$$

Assuming z_* is the highest fidelity, the hypothesis is that the correlation induced by the fidelity kernel is maximal when $z = z_*$ and decreases as z deviates from z_* . To model this, a decreasing function of the absolute difference between z and z_* is used as an exponential decay

$$k_f(z, z_*) = \exp(-\lambda|z - z_*|) \quad (15)$$

where λ is a non-negative parameter that controls the rate of decay—larger values of λ mean that fidelity levels significantly different from z_* will have a smaller influence on the kernel computation.

Perhaps the most important aspect of the multi-fidelity algorithm lies in the modification of the acquisition function. As a starting point, an existing acquisition function from literature was used (Savage et al., 2023)

$$x_{t+1}, z_{t+1} = \operatorname{argmax}_{(x,z) \in X \times Z} \frac{\mu_{f_t}(x, z_*) \beta^{\frac{1}{2}} \sigma_{f_t}(x, z_*)}{\mu_{\lambda_t}(x, z) \sqrt{1 - k((x, z), (x, z_*))^2}} \quad (16)$$

where $\mu_{f_t}(x, z_*)$ is the predictive mean and $\sigma_{f_t}(x, z_*)$ is the predictive standard deviation of the objective function at the highest fidelity z_* . They respectively represent the best estimate of the function's output, and the uncertainty in the model's predictions given the highest available fidelity simulations. $\mu_{\lambda_t}(x, z)$ is the predictive mean of the cost associated with a simulation at fidelity z . In this study, the cost has been modelled as the square of fidelity level, to simulate cost-aggressive applications and penalize high fidelity, as real-world problems often employ multi-fidelity when functions are very expensive and resources are limited. $\sqrt{1 - k((x, z), (x, z_*))^2}$ quantifies the loss of information when choosing a lower fidelity level compared to the highest one since the covariance measures the similarity between current and high-fidelity evaluations. A high covariance suggests that information gained at a lower fidelity is highly relevant to higher fidelities, thus indicating less information loss. The parameter β acts as the exploration bonus as it is multiplied by the predictive standard deviation, scaling the influence of uncertainty in the acquisition function, thus governing the trade-off between exploration and exploitation.

The variables in the denominator are instrumental in fidelity selection, as they effectively

determine cost and loss/gain in accuracy of choosing one fidelity over the other, whereas the other terms are only functions of x since z_* is fixed. As such, it was theorized that a better and more flexible trade-off management was needed for these terms, achieved through weighing them with hyperparameter γ :

$$x_{t+1}, z_{t+1} = \operatorname{argmax}_{(x,z) \in X \times Z} \frac{\mu_{f_t}(x, z_*) \beta^{\frac{1}{2}} \sigma_{f_t}(x, z_*)}{\gamma \mu_{\lambda_t}(x, z) + (1 - \gamma) \sqrt{1 - k((x, z), (x, z_*))^2}} \quad (17)$$

This novel approach allows for more precise control on the algorithm's fidelity selection strategy through an improved and more flexible trade-off, which is key when tackling different problems with different goals, especially given specific budget constraints. It dynamically adjusts the weight given to the computational cost and the benefit of exploring at different fidelity levels. When γ approaches 1, the acquisition function gives more weight to minimizing the computational cost associated with evaluating the objective function at a particular fidelity level, which means the algorithm will prefer points that are cheaper to evaluate. Inversely, when γ approaches 0, the algorithm emphasizes the loss of information due to choosing a lower fidelity level. It will therefore prioritize points that are expected to provide more accurate information, even if they are more expensive to evaluate, resulting in higher fidelity selection. As such, γ provides an additional layer of flexibility. Note that the cost and covariance values have been normalized to allow for reasonable sensitivity towards γ , since these values differ in magnitude.

A final important component in the context of the multi-fidelity algorithm is the fidelity selection function, which operates as an extension of the acquisition function. Its task is to iterate over possible fidelity levels for each candidate sample point to maximize the acquisition value, given the current budget and the remaining number of iterations.

TuRBO-1

The project further addressed a significant challenge: achieving global convergence in high-dimensional problems using the TuRBO-1 algorithm, conducted alongside the development of the multi-fidelity approach. Utilizing GPJax, a Gaussian Process library which has gained a lot of appreciation in recent years. GPJax was chosen for its advanced features, including GPU acceleration and Just-In-Time (JIT) compilation, which substantially enhance computational efficiency. Additionally, its mathematical coding closely aligns with the mathematical expressions in textbooks, particularly regarding log likelihood calculations.

The methodology involves initializing the TuRBO-1 algorithm with a quasi-random Halton sequence, ensuring improved coverage of the solution space. The number of initial samples is

adjusted based on the problem's dimensionality: 5 samples for 2D, 10 for 5D, and 20 for 10D problems. This variation accounts for the exponential increase in the search space as dimensions grow. The primary focus of the function evaluations lies in the Bayesian Optimization (BO) iterations, particularly critical in higher dimensions. Once the initial function evaluations are completed, the centre of the Trust Region (TR) of ellipsoidal shape, is strategically positioned around the most promising initial sample. The TR radius is meticulously fine-tuned according to the problem's dimensionality and bounds of the problem.

The dataset \mathcal{D}_i is then updated with these initial samples, followed by the generation of an optimised posterior utilising the Matern5/2 kernel. After experiment with various options in the initial algorithm, the Matern5/2 kernel, and Expected Improvement (EI) as the acquisition function were selected. These were found to effectively balance exploration and exploitation, and provided the best fit for the smoothness of the Gaussian Process.

The Bayesian optimisation loop proceeds as follows

1. Sample $\{100d\}$ samples within the Trust Region using the GPJax PRNGKey(42) which is a pseudo random sampling, where d is the number of dimensions.
2. Evaluate the EI from all the samples on the posterior
3. Extract point with maximum EI on posterior function
4. Choose the next query point x_i , by maximizing the acquisition function α , using the surrogate model \mathcal{M}_i conditioned on the dataset \mathcal{D}_i for a number of samples:

$$x_i = \underset{x}{\operatorname{argmax}} \alpha(x; \mathcal{D}_i, \mathcal{M}_i)$$

5. Obtain new observations by evaluating the objective function at x_i , yielding $y_i = f(x_i)$.
6. Expand the dataset with the new observation: $\mathcal{D}_{i+1} = \mathcal{D}_i \cup \{(x_i, y_i)\}$.
7. Update TR centre and Radius depending on whether there was an improvement in the objective function value. (TR Dynamics discussed in depth in next section)
8. Update the surrogate model with the new dataset to produce \mathcal{M}_{i+1} .

The iterative process continues until reaching a predefined stopping criterion, such as a specific number of function evaluations.

This was repeated for a total of 3 experiments, intended to replicate the effect of TuRBO, which runs multiple TRs on the search space. The only main difference to note between TuRBO-1 and TuRBO is that in TuRBO-1 the Trust Regions are not generated simultaneously, each with their own posterior, but rather one after the other. This inevitably leads to longer computational efforts.

After establishing a better understanding of TuRBO-1, the next step is to cover the dynamics and hyperparameters behind the TR that guide the TR. (Park, J. 2020)

TR Dynamics and hyperparameters

In TR optimisation algorithms, the dynamics of the TR play a pivotal role in guiding the search process through complex solution landscapes efficiently. This section discusses the adaptive nature of the TR dynamics, emphasizing its key mechanisms and underlying logic, which was inspired by (Eriksson, 2020).

In the experiments conducted, the TuRBO-1 algorithm was configured with specific hyperparameters as outlined in the theoretical background: $\tau_{succ} = 2$, $\tau_{fail} = \frac{d}{q}$, $L_{min} = 2^{-4}$, $L_{max} = 10$, and $r_{TR} = 2.0$.

Here, d represents the number of dimensions, q denotes the batch size, L_{min} and L_{max} denotes the minimum and maximum length of the TR, and τ indicates the thresholds for success and failure. The algorithm's adaptability is crucial, relying on its recent performance. An 'improvement'—a superior objective function value at iteration $n+1$ —leads to an increment in the success counter and an update of the Trust Region (TR) centre to this new optimal value. (Chen, 2016)

Expansion of half the search area is considered when the current region shows promise, potentially yielding superior outcomes. This expansion is activated when the success counter equals or exceeds the τ_{succ} threshold. In contrast, if the objective function value is unchanged or lower, the algorithm's failure counter increases by 1. The Trust Region contracts by half whenever τ_{fail} reaches its threshold. This iterative process continues until the prescribed number of Bayesian Optimization (BO) iterations is completed, or when the Trust Region length reaches its minimum or maximum limits, as extreme sizes render the sampling ineffective.

It should be highlighted that these TR hyperparameters were meticulously adjusted for different bounds and dimensions due to the initial challenges in scaling the domain effectively. (Diouane, 2022)

4. Results and Discussion

Initial BO

The initial algorithm was used to study the impact of different acquisition functions, kernels and search

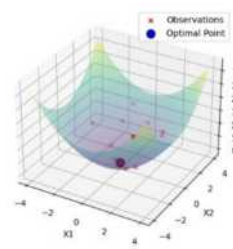


Figure 2 – Sphere function

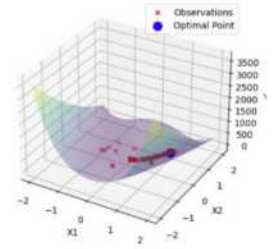


Figure 3 – Rosenbrock function

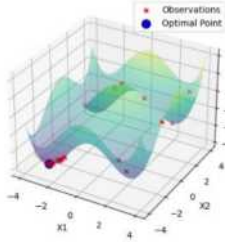


Figure 4 – Styblinski Tang function BO

illustrate an example of the results where the combined kernel, expected improvement, and whole space sampling strategy were used.

However, in the case of higher dimensions, where the volume of the search space increases exponentially and data becomes sparse, the Matérn kernel proved to be more advantageous compared to the RBF kernel. Its inherent roughness and ability to capture less smooth variations in the function landscape provided a more robust model of the objective function, leading to better exploration in vast and complex search spaces. Particularly with a ν parameter of 2.5, it provides a compelling choice due to its capacity to model functions that are sufficiently smooth but have areas of abrupt change, making it suitable for the selected test functions.

As for acquisition functions, a balance between exploration and exploitation becomes crucial; hence, EI was slightly preferred due to its ability to flexibly balance this tradeoff. The Upper Confidence Bound was also a solid option due to its tunable parameter, which can be adjusted to emphasize exploration in early iterations to avoid local optima, gradually shifting towards exploitation as the algorithm converges. This was particularly practical for the Styblinski-Tang function

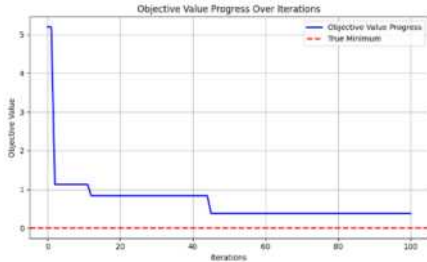


Figure 5 - 10D Sphere function BO

Overall, the algorithm had difficulties finding the right optimum for the more complex 10D functions, which motivated the use of more sophisticated approaches.

Multi-fidelity

The application of the multi-fidelity BO algorithm demonstrated marked variances in optimal behaviour across the selected test functions. Indeed, each test function required a distinct tuning of hyperparameters β and γ , which suggests that BO's performance is not only a function of its internal mechanics but is also deeply contingent on the nature of the optimization landscape it navigates.

It is important to note that all the presented results have been obtained in 10-dimensional functions, as lower dimensions yielded less informative results as the optimum was very often reached at low fidelities due to the simplicity.

For the sphere function, a simple and convex landscape, the algorithm exhibited rapid and accurate convergence towards the global minimum. Due to the objective's less complex nature, it did not necessitate aggressive exploration, resulting in a relatively low β value of 1. Additionally, it was found that a higher γ value of 0.75 was preferred, resulting in overall lower fidelities, despite a general trend of increasing fidelity with iteration number, as can be seen from the high-fidelity evaluations concentrated around the optimum. The found optimum was 0.59, which is very close to the true optimum of 0. For reference, the achieved simulated cost was 1,793, which is lower than the old approach (2,301) for a similar level of accuracy.

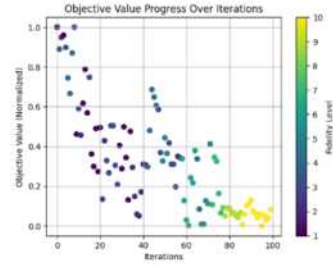


Figure 6 – Objective value over iterations for the 10D sphere function with $\beta=1$, $\gamma=0.7$, optimum=0.59, simulated cost=1,793

The Rosenbrock function, with its narrow, curved valley, presented a greater challenge. The multi-fidelity approach proved advantageous, as initial lower-fidelity evaluations provided valuable insights into the broader landscape, guiding the search towards the valley where it started choosing higher fidelities, and reached an optimum of 2.36, again very close to the true optimum. The γ term of 0.53 was a conservative approach, as it did not greatly emphasize or penalize either term. Interesting enough, it was found that even by modifying this parameter to emphasize higher fidelities, the general behaviour would not deviate much, as the acquisition function still intelligently navigates all the other parameters to achieve an optimal tradeoff – in this case much higher fidelities where unjustifiable as the improvement would always be suboptimal. Moreover, the beta parameter was higher (1.5), as exploration was encouraged to navigate the valley. Overall, the optimization path demonstrated a more gradual convergence, often requiring exploration at various fidelities before zoning in on the valley. The algorithm's adaptability to the challenging topology of the Rosenbrock function was evident in its dynamic fidelity selection and thorough exploration, which was better than the old approach which achieved this through higher cost.

The most complex of the test functions, Styblinski-Tang, with its numerous local minima, put the algorithm's exploration capabilities to the test. The results indicate that the algorithm successfully avoided premature convergence to local minima by leveraging the explorative aspect of the acquisition function. This was done through a higher beta parameter (1.8), and a lower gamma parameter (0.32). Overall, the fidelity levels were higher, which is to be expected for a more complex function, and the optimum found was -384.03, compared to the real optimum of -390. This was achieved with a cost of 7,593, significantly improving the old acquisition function's implied cost of 9,021.

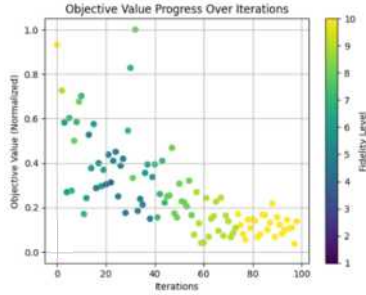


Figure 7 – Objective value over iterations for the 10D Styblinski-Tang function with $\beta=1$, $\gamma=0.6$, optimum=-379.93, simulated cost=7,593

Overall, the sensitivity analysis highlights that there is no one-size-fits-all parameter setting; instead, parameters must be adapted to the characteristics of the objective function being optimized. This is a particularly interesting feature, especially in the context of the developed acquisition function which allows for a more flexible cost trade-off management through γ . It is hypothesized that “simpler” functions benefit from a lower β and higher γ , effectively encouraging exploitation at low fidelity, whereas more “complex” functions benefit from a higher β and lower γ , which has inverse implications. The novel acquisition function has proven to be more effective than the old one, provided that the hyperparameters are well-tuned. The findings emphasize the importance of an adaptive and context-sensitive approach, as flexibility is an invaluable tool in real-world scenarios, where each problem comes with a different set of constraints and goals. It was found that the new method achieved a better trade-off between cost and accuracy, consistently yielding optimal results with less simulated cost.

Finally, these results underline the need for further research into automated or semi-automated methods for hyperparameter tuning in multi-fidelity BO. Developing strategies to predict optimal parameter values based on preliminary assessments of the objective function's characteristics could significantly enhance the efficiency and applicability of BO in diverse fields.

TuRBO-1

In this section, the TuRBO-1 algorithm is evaluated across various test functions and dimensions. Its performance is compared with the initial BO algorithm, a Random Search, and the sophisticated multi fidelity approach.

Starting with the simplest unimodal sphere function (Equation 9), characterized by a singular global minimum at the origin. This function presented intriguing results that deviated from the initial expectations. While the TuRBO-1 algorithm was anticipated to excel in unimodal functions due to the simplicity of the function, the dynamic TR updating mechanism and propensity to converge towards the centre, the outcomes were somewhat unexpected. The algorithm displayed a marked tendency for exploration over exploitation, particularly in high-conditioning scenarios of unimodal functions. This was evident in the behaviour of the TuRBO-1, which prioritized discovering new promising regions rather than quickly converging to the global optimum. This is especially true when performing global optimisation in the 5D and 10D case.

A critical illustration of this phenomenon is provided in Figure 8. The figure captures a scenario where all three TRs initially positioned outside the global optimum gradually converge towards the centre over several iterations. Notably, the red crosses in the figure highlight the algorithm's exploratory nature, underscoring that the high conditioning of the function does not significantly influence its sampling decisions.

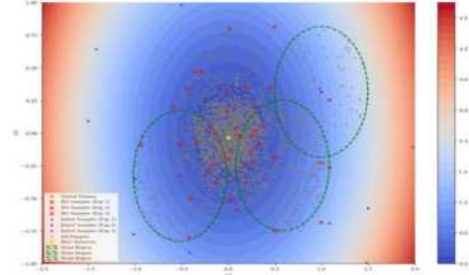


Figure 8 – Contour plot of the 2D Sphere function

In all three experiments, the initial Halton Sequence is represented by dark-coloured samples, with the best initial point used as the centre of the TR, indicated by the dashed green circle. Grey samples within the TR are evaluated for Expected Improvement (EI), with the highest EI point utilized as a function evaluation, marked by a red cross. The global minimum is also highlighted yellow.

Moreover, Figure 9 presents a comparative analysis of the logarithmic regret between the TuRBO-1 algorithm and the Random Search (RS) strategy. A key observation is the dashed red line at the fifth evaluation, marking the transition point where initial sampling evaluations cease, and BO iterations begin. This demarcation provides valuable insights into the adaptive response of the TuRBO-1 algorithm as it shifts from initial exploration to more focused optimization efforts.

Nevertheless, the TuRBO-1 algorithm outperforms the RS and initial BO algorithm.

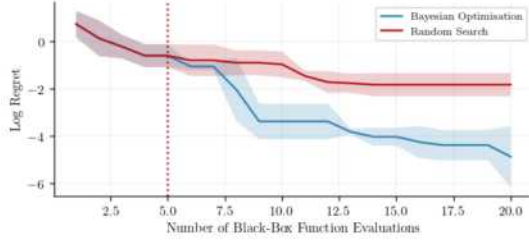


Figure 9 – Logarithmic regret plot of the RS and TuRBO-1

A summary table of all the algorithms, and runtimes for the 5D and 10D case are shown in *Table 1*.

Table 1 – 5D and 10DSphere function results for different methods.

5D	BO From Scratch	Random Search	TuRBO-1	Multi Fidelity
Optimum found	0.056	0.09	0.036	0.047
Runtime	<1m	<1m	5m	2m
10D	BO From Scratch	Random Search	TuRBO-1	Multi Fidelity
Optimum found	0.275	0.773	0.07	0.69
Runtime	2m	<1m	11m	3m

In the 5D and 10D sphere function case, a near global optimum was achieved with TuRBO-1 at both occasions, although runtime increased linearly with dimensionality. This is primarily due to TuRBO-1's single TR BO framework.

In comparative analyses, both the multi-fidelity approach and TuRBO-1 exhibited enhanced performance. However, the high conditioning nature of the sphere function resulted in no definitive superiority of TuRBO-1. The algorithm's inherent exploratory behaviour did not significantly contribute to outperforming other methods in this context, underscoring the impact of the function's characteristics on optimization efficiency.

The research further delved into the realm of multimodal functions by testing the TuRBO-1 algorithm on the Styblinski-Tang function (Equation 10). This function, with its inherent nonlinearity and nonconvexity, is particularly relevant to the field of chemical engineering. It possesses a global minimum at the point $(-2.904, -2.904)$, with the objective function value at this minimum being -39.166 multiplied by the dimensionality (d), translating to approximately -78 , -195 and -390 in the 2D, 5D and 10D respectively. in Figure 10, illustrating the 2D case, shows that the TuRBO-1 algorithm achieved global convergence within a limited number of iterations, effectively navigating through local optima to find the global minimum with a minimal change in the objective function value. This outcome exemplifies the algorithm's proficiency in handling complex multimodal landscapes, highlighting its potential

applicability in challenging real-world scenarios where optimal solutions are sought despite multiple local optima.

In the analysis of the 5D and 10D Styblinski-Tang function, the TuRBO-1 algorithm consistently achieved near-global optimum results. The

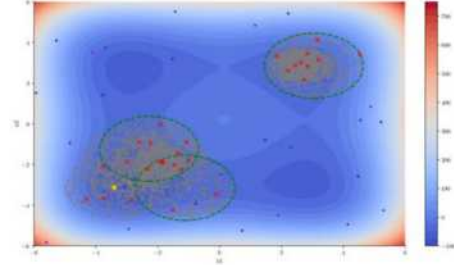


Figure 10 – Contour plot of the 2D Styblinski-Tang function

function's landscape, characterized by numerous local optima, favours an exploratory approach, which significantly contributed to TuRBO-1's superior performance over other models. This advantageous outcome is comprehensively documented in *Table 2* presents the results for the 5D and 10D scenarios, highlighting the relatively high running times of the TuRBO-1 model, attributable to its single TR design. These running times could be optimized within the TuRBO-1 framework by fine-tuning the minimum (L_{min}) and maximum (L_{max}) TR limits. This adjustment would allow the TR to terminate more efficiently when the input space becomes less effective, followed by the generation of a new TR around the next most promising initial sample.

Table 2 – 5D and 10D Styblinski-Tang function results for different methods.

5D	BO From Scratch	Random Search	TuRBO-1	Multi Fidelity
Optimum found	-178	-160	-190	-188
Runtime	<1m	<1m	7m	3m
10D	BO From Scratch	Random Search	TuRBO-1	Multi Fidelity
Optimum found	-230	-215	-390	-380
Runtime	1m	<1m	14m	6m

Conclusions and Outlook

This study has successfully developed two distinct yet complementary Bayesian Optimization (BO) algorithms: the multi-fidelity approach and the TuRBO-1 method. These algorithms represent significant improvements in the field of optimization, especially in the context of high-dimensional and computationally intensive problems.

The multi-fidelity framework has been an effective tool for optimizing the range of tested functions. By integrating an innovative γ hyperparameter in the acquisition function, the

computational cost and accuracy are balanced much more accurately, and in a way that allows for tuning the algorithm to the problem at hand. The strength of it lies in the fact that it provides a much more flexible approach in tackling functions with different levels of complexity and behaviour, especially in 10-dimensional spaces. As Bayesian Optimization continues to evolve, the multi-fidelity approach presents exciting avenues for future research. One key area lies in the dynamic tuning of hyperparameters. Current methodologies typically involve manual or static hyperparameter settings, which, although effective, may not always optimally exploit the potential of multi-fidelity models. Future research could focus on developing algorithms that dynamically adjust hyperparameters like β and γ in response to real-time feedback during the optimization process (Fucci *et al.*, 2022). Moreover, another promising direction is exploring different methodologies in the acquisition function, which is an ongoing process in literature (Takeno *et al.*, 2023) (Song, Chen and Yue, 2019).

The TuRBO-1 algorithm excels in high-dimensional optimization, adeptly navigating complex landscapes and evading early local optima, as demonstrated in the Styblinski-Tang function. Its dynamic trust region (TR) mechanism surpasses traditional Bayesian Optimization methods in balancing exploration and exploitation, marking it as an effective tool for complex optimization tasks. However, its single TR approach limits computational efficiency. Introducing simultaneous local TRs, each with a distinct posterior, could enhance its application in high-dimensional spaces, maximizing TuRBOs potential for real-world problems.

References

- Fucci, D., Romano, S., Baldassarre, M., Caivano, D., Scanniello, G., Thuran, B. and Juristo, N. (2022). A Longitudinal Cohort Study on the Retainment of Test-Driven Development. doi:<https://doi.org/10.1145/nnnnnnn.nnnnnnn>.
- Takeno, S., Fukuoka, H., Tsukada, Y., Koyama, T., Shiga, M., Takeuchi, I. and Karasuyama, M. (2023). Multi-fidelity Bayesian Optimization with Max-value Entropy Search and its parallelization. [online] Available at: <https://arxiv.org/pdf/1901.08275.pdf> [Accessed 14 Dec. 2023].
- Song, J., Chen, Y. and Yue, Y. (n.d.). A General Framework for Multi-fidelity Bayesian Optimization with Gaussian Processes. [online] Available at: <https://proceedings.mlr.press/v89/song19b/song19b.pdf> [Accessed 14 Dec. 2023].
- Shahriari, B., Swersky, K., Wang, Z., Adams, R.P. and de Freitas, N. (2016). Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proceedings of the IEEE*, 104(1), pp.148–175. doi:<https://doi.org/10.1109/jproc.2015.2494218>.
- Terayama, K., Sumita, M., Tamura, R. and Tsuda, K. (2021). Black-Box Optimization for Automated Discovery. *Accounts of Chemical Research*, 54(6), pp.1334–1346. doi:<https://doi.org/10.1021/acs.accounts.0c00713>.
- Braconi, E. and Edouard Godineau (2023). Bayesian Optimization as a Sustainable Strategy for Early-Stage Process Development? A Case Study of Cu-Catalyzed C–N Coupling of Sterically Hindered Pyrazines. *ACS Sustainable Chemistry & Engineering*, 11(28), pp.10545–10554. doi:<https://doi.org/10.1021/acssuschemeng.3c02455>.
- Savage, T., Basha, N., McDonough, J., Matar, O., Antonio, E. and Chanona, D. (2023). Machine Learning-Assisted Discovery of Novel Reactor Designs. [online] Available at: <https://arxiv.org/pdf/2308.08841.pdf>.
- Bempedelis, N. and Magri, L. (2023). Bayesian optimization of the layout of wind farms with a high-fidelity surrogate model. [online] Available at: <https://arxiv.org/pdf/2302.01071.pdf>.
- Pyzer-Knapp, E.O. (2018). Bayesian optimization for accelerated drug discovery. 62(6), pp.2:1–2:7. doi:<https://doi.org/10.1147/jrd.2018.2881731>.
- Letham, B., Karrer, B., Ottoni, G. and Bakshy, E. (2019). Bayesian Analysis (0000) 00, Number 0. [online] Available at: <https://arxiv.org/pdf/1706.07094.pdf> [Accessed 13 Dec. 2023].
- Jason Brownlee (2017). Gentle Introduction to the Adam Optimization Algorithm for Deep Learning. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/>.
- Weisstein, E.W. (n.d.). *Modified Bessel Function of the Second Kind*. [online] mathworld.wolfram.com. Available at: <https://mathworld.wolfram.com/ModifiedBesselFunctionoftheSecondKind.html>.
- Diouane, Y., Picheny, V., Riche, R.L. and Perrotolo, A.S.D. (2022). TREGO: a trust-region framework for efficient global optimization. *Journal of Global Optimization*, [online] 86(1), pp.1–23. doi:<https://doi.org/10.1007/s10898-022-01245-w>. J. Snoek, H. Larochelle, and R. P. Adams. Practical Bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, pages 2951–2959, 2012.
- Shahriari, B., Swersky, K., Wang, Z., Adams, R.P. and de Freitas, N. (2016). Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proceedings of the IEEE*, 104(1), pp.148–175. doi:<https://doi.org/10.1109/jproc.2015.2494218>.
- Eriksson, D., Pearce, M., Gardner, J.R., Turner, R. and Poloczek, M. (2020). *Scalable Global Optimization via Local Bayesian Optimization*. arXiv.org. doi:<https://doi.org/10.48550/arXiv.1910.01739>.
- Park, J. (2020). Contextual Bayesian optimization with trust region (CBOTR) and its application to cooperative wind farm control in region 2. *Sustainable Energy Technologies and Assessments*, 38, p.100679. doi:<https://doi.org/10.1016/j.seta.2020.100679>.
- Chen, R., Menickelly, M. and Scheinberg, K. (2016). *Stochastic Optimization Using a Trust-Region Method and Random Models*. [online] arXiv.org. doi:<https://doi.org/10.48550/arXiv.1504.04231>.

Stability Study of Dual Drug Delivery Systems under Osmotic Stress

Inny Yeung and Barbara Simonetou

Department of Chemical Engineering, Imperial College London, U.K.

Abstract It was required to investigate the stability of vesicles under osmotic stress to be employed as dual-drug delivery systems. The effects of lipid architecture and encapsulated cargo were studied using DPPC and DOPC, with two fluorescent drug mimics - calcein and methylene blue. The stability was analysed through release assays of the fluorescent dyes and deviations in size for vesicles in sucrose and KCl buffers of various concentrations. DPPC, which is a saturated lipid, was determined to be more optimal for application in a drug delivery system, as it had lower degree of passive leakage, despite larger changes in size to counter changes in osmotic pressure. Buffer type and concentration were found to have minimal effect on the release efficiencies for both DPPC and DOPC vesicles. However, significant differences in the release efficiencies of both encapsulated cargos were noted, with calcein having a lower passive leakage in all conditions.

Keywords: Vesicles, stability, dual-drug delivery system, osmotic pressure, calcein, methylene blue

Introduction

Widely employed in various applications, liposomes, or vesicles¹, exhibit outstanding properties, particularly in drug delivery systems, where they first gained traction in the 1970s². They are recognised as promising and adaptable drug carriers due to their biocompatibility, their capability to encapsulate both hydrophilic and hydrophobic therapeutic agents, and their ability to safeguard encapsulated substances from physiological degradation². With the ability to selectively transport payloads to specific sites using passive or active targeting, liposomes mitigate systemic side effects, boost the maximum-tolerated dosage, and amplify therapeutic benefits³. Unstable vesicles, however, pose a significant risk as the potential premature release of drugs heightens toxicity risks to healthy tissues and may limit the treatment's effectiveness. Hence, maintaining vesicle stability is crucial for ensuring the efficacy, safety, and targeted delivery of drugs.

In this investigation, vesicle stability was characterised by changes in size and passive leakage of encapsulated cargo through the membrane. While several factors influence vesicle stability, variations in osmotic pressure within the body are particularly relevant in drug delivery systems, and therefore a crucial consideration. Consequently, this study investigated the stability of vesicles with two encapsulated drug mimics (calcein and methylene blue) under osmotic stress for two lipids, DPPC and DOPC, with varying properties.

The properties and release mechanisms of these two encapsulated molecules were characterised, and the effect of lipid architecture, buffer type and concentration on vesicle stability was explored. The study aimed to determine the limit of drug concentration that can be encapsulated such that vesicles have minimal leakage after being introduced to the bloodstream, while also distributing the required dose at the target site. These results were deemed significant in the construction of a drug delivery system that can trigger the release of temperature, pH or light responsive cargos, whilst also minimising the side effects. This would require the introduction of the fewest number of vesicles possible, as vesicles can clog blood vessels or get attacked by the immune system, whilst also delivering the necessary drug dosage.

Background

Liposomes are nano-sized to micro-sized vesicles composed of a phospholipid bilayer, that structurally adopt a spherical or multi-layered spherical shape⁴. Phospholipids, typically composed of a glycerol backbone, two hydrophobic fatty acid tails, and a hydrophilic phosphate group are amphipathic⁵. In an aqueous environment, driven by the hydrophobic effect, phospholipids will spontaneously arrange themselves into a double-layered structure, known as the phospholipid bilayer, with the hydrophilic heads on the outside and hydrophobic tails pointing towards the inside⁶. This structural similarity, which mimics the structure of natural cell membranes, aids in their integration with biological systems and helps reduce the likelihood of immune responses or toxicity when used in drug delivery systems. For this reason, phospholipids are extensively used in liposomes.

Additionally, lipid bilayers may exhibit different phase behaviours depending on lipid tail interactions within the bilayer structure⁷. There are two primary phases: a solid (gel) phase, and a liquid phase, which are characterised by lipid saturation. Saturated phospholipids, such as DPPC, result in straight, unknicked tails, that can be packed closer together in a crystalline-like matrix, thereby maximising the intermolecular interactions between tails and decreasing bilayer fluidity. Unsaturated phospholipids, such as DOPC, which has two carbon-carbon double bonds, present with crooked, knicked tails, resulting in fewer intermolecular interactions and increased bilayer fluidity. Hence, it is expected that vesicles made with DOPC will have a higher degree of membrane fluidity, whilst DPPC membranes will be more viscous.

The lipid saturation will also influence the vesicle's ability to withstand variations in osmotic pressure⁸, which is defined as the amount of force applied to a solution, preventing the movement of solvent across a semipermeable membrane⁹. DOPC vesicles, owing to their fluid nature, can deform more easily to counteract changes in osmotic stress on the membrane.

Water is transported across a membrane by osmosis, which describes the spontaneous net movement of water molecules across a semi-permeable membrane from an area of high-water potential to low water potential until a state of equilibrium is reached⁹. In isosmotic media,

equilibrium is achieved and there is no net movement of water in or out of the vesicles, and they are stable in size. When vesicles are placed in hyperosmotic media, where the concentration of solute is higher on the outside, there will be a net movement of water out of the vesicle, which results in the vesicles shrinking. Conversely, when they are placed in hypoosmotic media, where the concentration of solute is higher on the inside, there will be a net movement of water into the vesicles, causing them to swell. Equilibrium is reached when the force of water on the hyperosmotic side of the membrane is equal to the force of diffusion on the hypoosmotic side of the membrane.

Osmolarity is a property of a solution that considers the number of particles formed when a substance is dissolved in water. Whilst molarity is a measure of the concentration of a solution in terms of moles per volume, osmolarity is a measure of the number of particles per volume⁷. This means that while 2 solutions could have the same molarity, they could have different osmolarities.

When ionic compounds are dissolved in water, they dissociate to form cations and anions, which contribute towards the osmolarity. Hence, it would be expected that the osmolarity of KCl buffers to be higher than for sucrose buffers of the same molarity. In theory, it was also be expected that if a 5mM sucrose solution corresponds to 5 mOsm/L, a 5mM KCl solution would be 10 mOsm/L as KCl dissociates into K^+ and Cl^- ions⁷.

Furthermore, the rate at which drugs diffuse from the vesicle's membranous lipid bilayer, is another critical aspect related to the drug delivery systems under investigation¹⁰. Understanding the release, or leakage mechanism, of the encapsulated cargo is especially important, as ideally no passive cargo release, or leakage, to the surroundings should occur upon transport of the vesicle from the site of administration to the target. Cell membranes act as biological barriers, selectively restricting the passage of certain molecules based on their permeability¹¹. The cell membrane is semi-permeable, with only small uncharged molecules able to diffuse freely through the phospholipid bilayers, in a process known as simple diffusion⁶. Other mechanisms for transport across the cell membrane exist, and may be active or passive, depending on their energy consumption¹¹. However, most pertinent to this study is simple diffusion as a transport mechanism. In this mode of transport, nonpolar molecules freely diffuse across the lipid bilayer in a process driven by a difference in concentration⁷.

The rate of diffusion across a cell membrane directly relates to this concentration gradient, but is also influenced by other factors, such as the molecule's solubility and acidity, as represented by its logP and pKa respectively¹¹. Hence, to explore the passive leakage mechanism across the membrane, specific to this investigation, it is necessary to consider the logP and pKa properties of the drug mimic cargos - calcein and methylene blue. Calcein and methylene blue are both self-quenching dyes. In other words, they are nonfluorescent at high concentrations, and fluorescent at low concentrations, as demonstrated in figure 1. This property enables them to be useful indicators for vesical leakage, as related to vesicle stability. In addition, these 2 molecules were chosen for their drug-like structures, namely, the abundance of aromatic rings and polar groups. Furthermore, their different excitation and emission wavelengths are crucial to enable them to be used in a study for dual drug delivery.

The partition coefficient, logP, is a measure of the hydrophilicity (or hydrophobicity) of a molecule. It is measured as the ratio of concentrations of a compound that has dissolved into an organic solvent phase and into an aqueous water phase. Most commonly, the organic partitioning solvent used is octan-1-ol. A negative logP value indicates that the compound has a higher affinity for the aqueous phase, meaning it is more hydrophilic. Conversely, a positive logP value is representative of a higher concentration of the substance in the lipid phase, meaning the compound is more hydrophobic. A lower logP is indicative of higher membrane permeability, and it is easier for the molecule to diffuse through the lipid bilayer. logP values for methylene blue has been reported ranging from -1.1 to -0.62, whilst calcein has a logP of 1.56¹². It would be therefore expected that methylene blue leaks out or diffuses through the lipid bilayer more readily than calcein. The pKa value represents the acidity of a molecule. It is the negative log of the equilibrium constant for dissociation in acid-base reactions, which is the concentration of the conjugate base, multiplied by the concentration of hydrogen ions, divided by the concentration of the acid at equilibrium. A lower pKa value indicates a stronger acid. Calcein has a pKa of 2.1, whilst this value is 3.14 for methylene blue¹². Since both cargos were maintained at the physiological pH of 7.4, which is higher than their pKa values, the cargos should not be in their protonated/acidic forms.

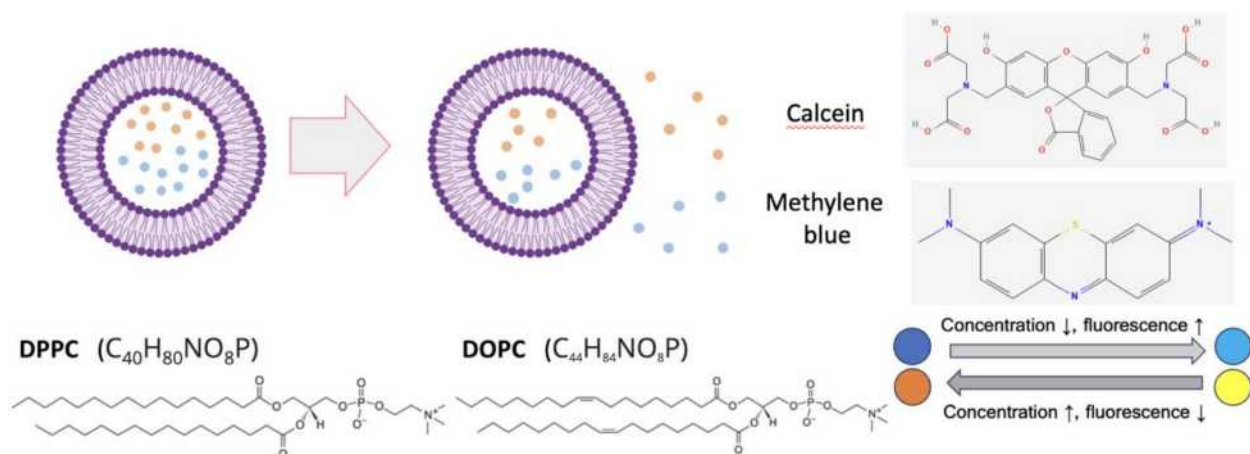


Figure 1. Schematic showing leakage from vesicles containing 2 drug like dyes (calcein and methylene blue) and their structures. Vesicles are made from DPPC or DOPC and diagram showing self-quenching properties of both dyes: As concentration decreases, the fluorescence increases.

Methods

Since the properties of methylene blue are not well documented, it was necessary to characterise its fluorescence and absorbance properties. 200 μ L of 30 concentrations between 0mM and 250mM methylene blue was placed in a fluorimeter well plate and a clear absorbance well plate. The absorbance spectra and fluorescence spectra were then measured to determine the concentration range of methylene blue that would be ideal for the investigation.

Lipid films were prepared by drying 200 μ L of the 25 mg/mL DPPC or DOPC in chloroform stock under nitrogen flow on glass to evaporate the chloroform. The films were then dessicated overnight in a vacuum and rehydrated with 500 μ L of either 500mM calcein (20mM HEPES, pH 7.4) or 60mM methylene blue (20mM HEPES, pH 7.4). The vials were vortexed to ensure they were well mixed, before being heated above the transition temperatures of the lipids, at 70 $^{\circ}$ C. To improve the encapsulation efficiency of the dyes, four freeze thaw cycles were undertaken, where the vials were cooled in liquid nitrogen until frozen, then left in the hot plate for at least five minutes. The lipid dispersions were then extruded through a 100nm polycarbonate membrane 21 times at 70 $^{\circ}$ C to create small unilamellar vesicles. Finally, size exclusion chromatography (SEC) was employed to filter the sample and remove any unencapsulated dyes. This involved adding the samples dropwise to a 5mL SEC column made from 0.4g of Sephadex G50, hydrated with 12mL of column buffer, which was 500mM sucrose, 20mM HEPES for encapsulated calcein and 60mM sucrose, 20mM HEPES for encapsulated methylene blue. Four fractions, each 300 μ L, were then combined and pipetted up and down to ensure mixing before being mixed with 12 different buffers of varying sucrose and KCL concentrations for measurements in the fluorimeter and dynamic light scatterer (DLS).

Release assays with the 12 buffers were run overnight using a Cary Eclipse fluorimeter. A well plate with 96 wells was utilised, with 20 μ L of sample mixed with 180 μ L of buffer in each well. 3 trials were conducted for the sample in all buffers, and 2 trials of only buffer were tested as a control. Lastly, a gas

permeable sealing membrane was stuck onto the well plate to minimise evaporation. Calcein fluorescence was measured at excitation wavelength (λ_{ex}) of 495nm and an emission wavelength (λ_{em}) of 515nm, whilst methylene blue fluorescence was recorded at λ_{ex} = 668nm and λ_{em} = 688nm.

The maximum release was then obtained by lysing the vesicles with 2 μ L of the detergent Triton X-100, and measuring the total fluorescence under the same voltage conditions. The release efficiency was then calculated using equation (1), where F_s is the fluorescence at a specific time, F_0 is the initial fluorescence of vesicles in isosmotic buffer, and F_t is the maximum fluorescence after adding Triton⁷.

$$\text{Release efficiency [\%]} = \frac{F_s - F_0}{F_t - F_0} \times 100\% \quad (1)$$

The stability of vesicles was also quantified by their size deviation from reference values using a Malvern Zetasizer DLS. 20 μ L of vesicles were mixed with 980 μ L of each buffer in a polystyrene cuvette and covered with parafilm overnight. Reference values were taken to be the size of vesicles in isosmotic buffer before lysis, which refers to 350mM sucrose, 20mM HEPES buffer for calcein, and 60mM sucrose, 20mM HEPES for methylene blue. A fluorescent filter was used as both calcein and methylene blue molecules are fluorescent dyes at low concentrations. In addition, the 13% sucrose dispersant setting was used for the sucrose buffers, whilst a water dispersant was used for the 0mM sucrose and KCl buffers.

A freezing point depression osmometer was used to test the osmolarity of various buffers used and to identify the isosmotic reference point for both encapsulated cargos. 25 μ L of each buffer was pipetted into an Eppendorf tube and attached to the measuring head, before being lowered into the cooling aperture, which started the supercooling process. The sample was supercooled to a predetermined temperature below the expected sample freezing point, which is -6.2 $^{\circ}$ C for the Loser Type 7M osmometer. A cooled pre-wetted needle with ice crystals was then automatically inserted into the sample to initiate freezing. The heat of fusion from the crystallisation process increased the sample temperature until a plateau point was achieved, where the liquid solid equilibrium was maintained¹³. This plateau was taken as the true freezing point of the sample. A

microprocessor then calculated the osmolarity by comparing the freezing point measured with the freezing point of distilled water and 2 other standard solutions.

Results and Discussion

Absorbance

It was decided to use methylene blue at a 60mM concentration with 20mM HEPES to rehydrate the lipid films based on supplementary figures 1 and 2. Since the dye is diluted at least 30 times when encapsulated (1:3 dilution in the column and 1:10 in the fluorimeter) and further diluted by the movement of water molecules when mixed with buffer, it was necessary to use a concentration of methylene blue that is non fluorescent even after being diluted 30 times, so that further dilutions result in fluorescence. It was seen that methylene blue is quenched at concentrations above 35mM, so this was the minimum stock concentration that was required. Based on the fluorescence and absorbance data of various concentrations of methylene blue, 60mM methylene blue was used, which corresponded to an intermediate level of fluorescence.

The absorbance spectrum was also used to confirm the excitation and emission wavelengths of methylene blue, which were 668nm and 688nm, respectively.

Osmolarity

Figure 2 shows all the osmolarity measurements recorded, for buffer concentrations ranging between 0-250mM for methylene blue and 0-1000mM for calcein.

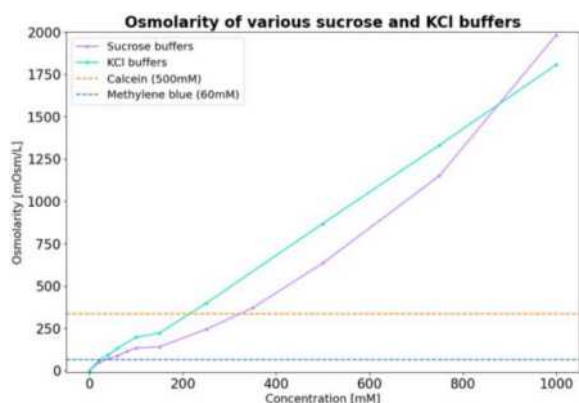


Figure 2. Osmolarity of sucrose and KCl buffers used with both calcein and methylene blue. KCl has higher osmolarity for all concentrations except 1000mM. Osmolarity of 500mM calcein, 20mM HEPES was found to be 337 mOsm/L, and closest to 350mM sucrose solution. Osmolarity of 60mM methylene blue, 20mM HEPES was measured as 66 mOsm/L, and isosmotic reference point was therefore taken to be 60mM sucrose solution.

The osmolarity of KCl buffers was higher than those of sucrose buffers at almost all concentrations, except at 1000mM, where sucrose was found to have a higher osmolarity. This is likely due to measurements having been taken using a freezing point depression osmometer, which compares the freezing temperature of the solution to reference standards to produce an osmolarity reading.

When solutes are dissolved, the freezing point of the resulting solution will be lower than that of the solvent on its own due to changes in chemical potential of the solvent. The degree to which the freezing point is

depressed is directly correlated to the molarity of a solution through a cryoscopic constant, which is concentration dependent¹⁴. Since sugars are well known to have cryoprotectant properties, they can prevent ice formation on biological tissues^{15, 16}. As such, with a higher concentration of sucrose, a lower temperature is required to freeze the sample, and therefore, a larger temperature difference between the freezing point and the reference value is recorded, producing a larger osmolarity reading. Repeat measurements should be taken to confirm the trends observed and alternative methods of measuring osmolarity should also be explored to negate the cryoprotectant effects of sucrose.

Further investigation is also required into factors that affect osmolarity, as the conversion between osmolarity and molarity is not linear as expected. The ratio of osmolarities between the investigated sucrose and KCl buffers of the same molarity is smaller at low concentrations, and appears to deviate to a higher degree with increasing concentration.

As shown in figure 2, 500mM calcein, 20mM HEPES solution was found to have an osmolarity of 337 mOsm/L, which was most closely balanced with the 350mM sucrose buffer, which had an osmolarity of 371 mOsm/L. This was therefore taken to be the isosmotic reference point for vesicles with encapsulated calcein. Similarly, 60mM sucrose solution was considered to be isosmotic for vesicles with 60mM methylene blue, 20mM HEPES solution, which had an osmolarity of 66 mOsm/L.

Release assays

The release profiles of both lipids with calcein and methylene blue in different buffers were measured in the fluorimeter and analysed to determine trends in vesicle stability in terms of passive leakage. In drug delivery contexts, it is crucial to minimise the passive leakage over time.

Vesicle leakage was found to be independent of time for both lipids with encapsulated calcein in sucrose and KCl buffers of all concentrations, as seen in figure 3 and supplementary figure 3. This suggested that the release occurred instantaneously after the vesicles were mixed with the buffers and before the well plate was placed in the fluorimeter.

For DPPC, the maximum release was observed for vesicles in 0mM solution at 10.8%. The minimum release in sucrose buffers was 1.2% for 1000mM solution compared to 1.7% for KCl buffers, which occurred in the 500mM solution. In hypoosmotic media (below 350mM sucrose), the vesicles in sucrose buffers of all concentrations consistently have a slightly higher release than those in KCl buffers. The opposite effect is observed for vesicles in hyperosmotic media, where those in KCl buffers are then observed to have higher releases.

In DOPC, the releases for sucrose buffers were higher than those in KCl buffers of the same molarity at all concentrations, though the difference was smaller at both extremes. A larger range of releases was observed, with the minimum being 0.4% for 1000mM sucrose, and

20.8% for 0mM sucrose. DOPC was found to be more sensitive to changes in buffer concentration due to the increased membrane fluidity arising from its unsaturated lipid tails, therefore, making DOPC more likely to have membrane pores through which leakages can occur.

With both lipids, the maximum release occurred for vesicles in 0mM buffer, which is expected as water is hypoosmotic relative to the encapsulated calcein. The net movement of water into the vesicles causes them to

swell in an attempt to equilibrate the difference in osmotic pressure, which results in a slight increase in the fluorescence as the calcein is diluted. However, as the vesicle continues to swell to a point where the membrane is no longer able to withstand the osmotic stress, the distance between adjacent lipid molecules will increase, inadvertently forming pores in the membrane, through which the calcein is able to leak out.

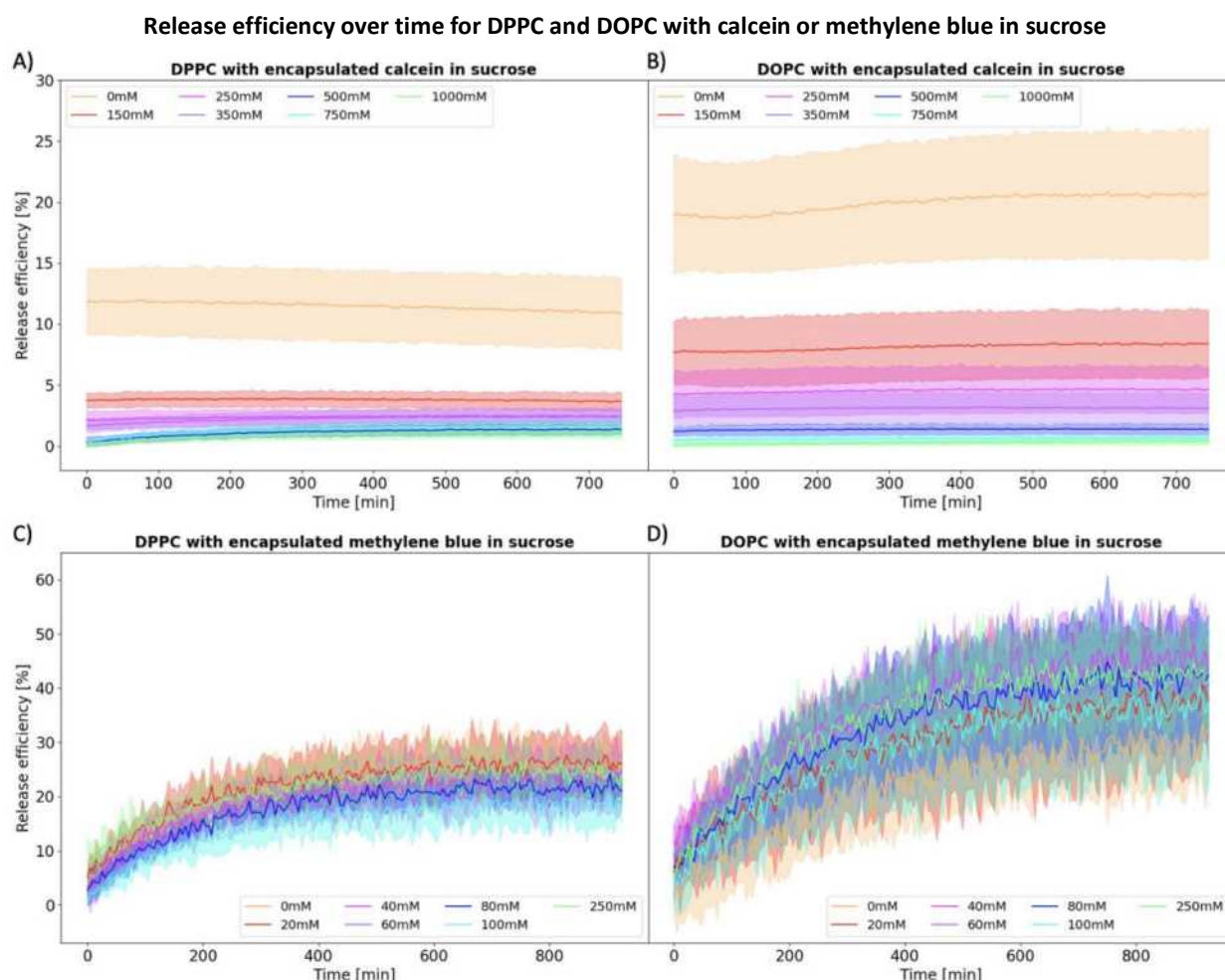


Figure 3. Release efficiency over time for both DPPC and DOPC with encapsulated calcein and methylene blue in sucrose buffers of varying concentrations. (A) Time-dependent release profile for DPPC vesicles with 500mM calcein encapsulated in sucrose. Maximum release was achieved for 0mM (orange) at 10.8% and was steady over time. (B) Time-dependent release profile DOPC vesicles with 500mM calcein encapsulated in sucrose. Maximum release is 20.8% for 0mM (orange). Larger range and spread of release efficiencies observed for DOPC due to increased membrane fluidity. (C) Time-dependent release profile for DPPC vesicles with 60mM methylene blue encapsulated in sucrose. Release increases non-linearly over time and plateaus for all concentrations to an average of 23.1%. (D) Time-dependent release profile for DOPC vesicles with 60mM methylene blue encapsulated in sucrose. Release increases non-linearly over time and reaches a plateau later in comparison to DPPC. There is a wider range of releases observed, ranging between 32.6% and 43.2% on average.

Furthermore, in both cases, comparing the final release efficiencies, shown in figure 4, for both lipids in various concentrations of sucrose and KCl buffers, it was noted that the presence of ions in the buffer had minimal effect on the vesicle leakage. The final release had an inverse effect with sucrose concentration, but no clear trend was observed for KCl.

The release profile of methylene blue was found to be extremely different from that of calcein for both DPPC and DOPC in all buffers.

Comparing the release profiles of calcein and methylene blue using figure 3, sucrose concentration did not have much effect on the DPPC vesicle leakage as the same trend was observed for all concentrations and they all resulted in a final release between 19% and 26%. In contrast, for DOPC, the final release ranged between 34% and 46%, which corresponded to the 0mM sucrose and 250mM sucrose buffers. However, there was also no clear trend between the sucrose concentration and final release observed in DOPC vesicles. Furthermore, there was a wider range of releases observed for DOPC

than DPPC, which was likely the result of the combined effect of increased membrane fluidity and permeability of methylene blue.

The release profiles of both lipids in KCl was consequently explored to investigate whether the passive leakage of methylene blue could be reduced in an ionic buffer. The results can be seen in supplementary figure 3. For DPPC, the release ranged between 20% and 36%, for 500mM KCl and 1000mM KCl respectively. For DOPC, the final release had a smaller range between 37% and 39%, but this was not reflective of the maximum release that was recorded before the final

time. More repeats of this condition are needed as large errors were noted in the releases. The release profile for both lipids were not improved using KCl and there is a very similar trend for both lipids encapsulating methylene blue in KCl and sucrose. Larger differences in the final releases were noted across the buffer concentrations for KCl than sucrose, but more investigation is needed to determine if there is a trend in the normalised release with varying KCl concentration. The results suggest that KCl does not counteract osmotic pressure as well as sucrose.

Final release efficiency against concentration for DPPC and DOPC with calcein and methylene blue in all buffers

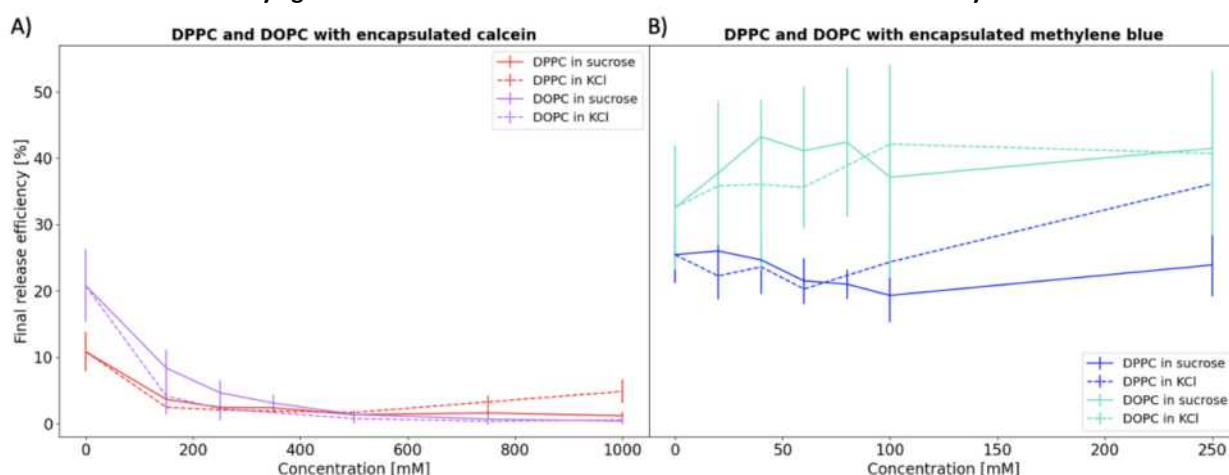


Figure 4. Release efficiency against concentration for 500mM calcein and 60mM methylene blue in DPPC and DOPC vesicles. Both lipids have higher release efficiencies with encapsulated methylene blue due to its lower logP value. The difference between DPPC and DOPC is larger with methylene blue than with calcein. (A) Final release efficiency against concentration for DPPC and DOPC with encapsulated 500mM calcein in sucrose and KCl buffers of varying concentrations. At hypoosmotic conditions, DOPC has more leakage than DPPC. Both lipids in sucrose buffers have higher releases in hypoosmotic conditions. In hyperosmotic conditions, DPPC vesicles in KCl have the highest release and DPPC has a higher release efficiency than DOPC in hyperosmotic conditions. (B) Final release efficiency against concentration for both DPPC and DOPC with encapsulated 60mM methylene blue in sucrose and KCl buffers of varying concentrations. DOPC consistently has a higher release at all concentrations than DPPC due to increased membrane fluidity. At hypoosmotic conditions, both lipids in sucrose buffers have higher release efficiencies than those in KCl buffers of the same concentration. In hyperosmotic conditions, a higher release efficiency is observed for both lipids in KCl than in sucrose.

From Figure 4, DPPC (blue), in hypoosmotic conditions (below 60mM for methylene blue), vesicles in sucrose buffers were found to have higher releases than those in KCl at the same molarity. However, in hyperosmotic conditions, vesicles in sucrose buffers were found to have smaller releases than those in KCl. In contrast, for DOPC, the trend is unclear as the release efficiencies between sucrose and KCl fluctuate. Under 90mM and above 240mM, DOPC vesicles in sucrose have elevated releases when compared with those in KCl, but the observation is reversed between these 2 points.

The final release efficiencies for methylene blue were higher than those for calcein at all concentrations. This was explained by the difference in logP of methylene blue and calcein. As methylene blue has a lower logP value, it is able to diffuse through transient pores in the membrane more readily, thus becoming diluted in the external environment and fluorescing.

There was a smaller effect of sucrose concentration on the methylene blue release than with calcein, as the range of release efficiencies at different concentrations is smaller. No significant trends in concentration and final release efficiency were recognised. The type of

buffer had a more significant effect in methylene blue encapsulated vesicles than with calcein vesicles.

Size

The dynamic light scatterer (DLS) was used to measure vesicle stability in terms of vesicle size deviations between a measured sample and its corresponding reference vesicle size.

The average size of vesicles in the liquid (DOPC) or gel (DPPC) phase, and in either type of buffer (KCl or sucrose) concentration, was compared to that of vesicles of the same phase, and in the same buffer (KCl or sucrose), but under isosmotic buffer concentration conditions. In the case of calcein encapsulated vesicles in sucrose, for instance, the osmometer data presented earlier suggests that the osmolarity of the calcein cargo is most closely balanced with the osmolarity of the 350mM sucrose buffer. The reference vesicle size for calcein encapsulated DOPC (or DPPC) vesicles in sucrose, therefore, corresponds to that of the calcein encapsulated DOPC (or DPPC) vesicles in the 350mM sucrose buffer. For the case of calcein cargo in KCl, the osmolarity of the cargo is most closely balanced with the osmolarity of the 250mM KCl buffer. Similarly, for the

methylene blue cargo the isosmotic sucrose and KCl buffer concentrations, correspond to 60mM and 40mM, respectively. Supplementary Table 1 summarises these results.

Further, these reference concentrations are important to determine the concentration range for hypo and

hyperosmotic media, as established within our context. More specifically, buffer concentrations below the isosmotic concentration and above, refer to hypoosmotic and hyperosmotic environments, respectively.

Effect of buffer concentration on average size for vesicles with encapsulated calcein and methylene blue

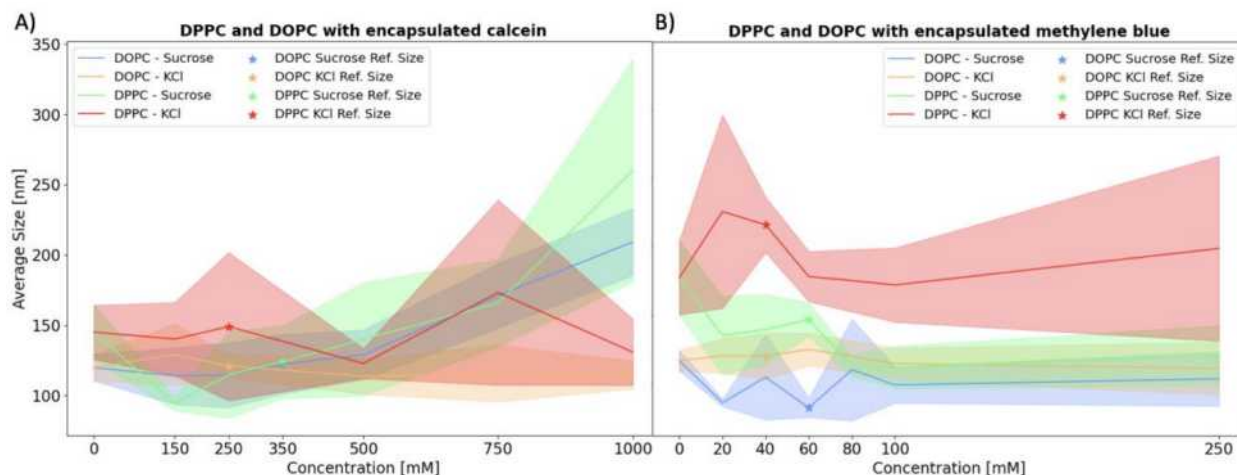


Figure 5. (A) and (B) respectively refer to calcein and methylene blue encapsulated vesicles, respectively. The figures summarise the average vesicle size [nm] for each lipid phase (fluid - DOPC or gel - DPPC), in either type of buffer (KCl or sucrose). The reference average vesicle size for each lipid phase and type of buffer is indicated using a star symbol. This reference helps to distinguish the concentration range for hypo and hyperosmotic environments as related to the sucrose and KCl buffers. (A) suggests that DOPC in hypoosmotic media is more stable in sucrose, whereas in hyperosmotic media it is more stable in KCl. DPPC on the other hand, seems to be more stable in KCl in both hypo and hyperosmotic media. Overall, DOPC vesicles seem to be more stable than DPPC, and DPPC vesicles on average tend to be larger in size than DOPC vesicles. In the case of methylene blue cargo, (B) suggests that DOPC vesicle stability is unaffected with regards to buffer type and osmotic media. Since, no clear trend for whether DOPC in sucrose or DOPC in KCl, is more stable in hypo or hyperosmotic environments. The same may be observed for DPPC. Overall, DOPC vesicles seem to be more stable than DPPC vesicles in the ionic (KCl) buffer and once again, DPPC vesicles seem to be larger than DOPC vesicles.

Figure 5A and 5B refer to calcein and methylene blue cargos, respectively. Both figures summarise the average vesicle size [nm] for each lipid phase (fluid - DOPC or gel - DPPC), in either type of buffer (KCl or sucrose). The reference average vesicle size for each lipid phase and type of buffer is indicated using a star symbol, which further helps to distinguish the concentration range for hypo and hyperosmotic environments as related to the sucrose and KCl buffers. Using Figure 5A to compare the change in vesicle size in the sucrose (blue) buffers alone, vesicles appear to be more stable under hypoosmotic media, as the (vertical) difference in size between vesicles with respect to the reference size is smaller than it is for vesicles in hyperosmotic media. This may be explained by the fact that in hypoosmotic conditions there's a net flow of water into the vesicles, to equalize the osmotic pressure difference between external and interior environments. This phenomenon causes the vesicles to swell but also potentially reduces the stress applied on the lipid bilayer, thereby enhancing stability. However, this is not always observed. In KCl (orange), DOPC vesicles seem to be more stable under hyperosmotic media. Supplementary Figure 4 compares vesicle sizes of DOPC in sucrose and KCl buffers, for the calcein cargo, and shows these trends clearly. Comparing DPPC in sucrose (green) and KCl (red), in both hypo and hyperosmotic media, the vesicles seem to be more stable

in KCl. This result is also illustrated in Supplementary Figure 5. Furthermore, a comparison of DOPC (blue) and DPPC (green) in sucrose suggests that vesicles in the fluid phase (DOPC) are more stable in size than in the gel phase (DPPC) in both hypo and hyperosmotic media. This may be expected as lipids in the fluid phase are mobile compared to the gel phase and may therefore more easily adjust to accommodate changes in the environment. In the gel phase, lipids are more rigidly packed, so it's possible that alterations in buffer conditions can have a more pronounced effect on their size. Supplementary Figure 6 summarises this finding. Finally, as shown in Supplementary Figure 7, the result that DOPC (orange) is more stable than DPPC (red) seems to be true for the KCl buffer as well. In the case of the methylene blue cargo, similar comparisons can be made using Figure 5B. A comparison of DOPC vesicles in sucrose (blue) or KCl (orange) does not seem to yield a clear trend regarding size variations in hypo or hyperosmotic media. Therefore, unlike in the case of the calcein cargo, the media does not seem to affect vesicle size in a particular way. This can be closely observed in Supplementary Figure 7, and may be explained by considering that methylene blue leaks out more than calcein, as suggested by the release assay studies. Hence, in the case of methylene blue, the osmotic pressure difference that drives the change in size decreases over time, so the net water movement is not

enough to make significant changes in the size of vesicles.

Likewise, consideration of DPPC in sucrose (green) and DPPC in KCl (red) also does not seem to suggest a clear trend for changes in vesicle size in hypo or hyperosmotic environments. This comparison is also shown in Supplementary Figure 8. Furthermore, like in the case of the calcein cargo, a comparison of DOPC (blue) and DPPC (green) in sucrose suggests that vesicles in the fluid phase (DOPC) are more stable in size than in the gel phase (DPPC) in both hypo and hyperosmotic media. Supplementary Figure 9 illustrates this. Similarly, as shown in Supplementary Figure 10, the result that DOPC (orange) is more stable than DPPC (red) seems to be true for the KCl buffer as well.

The PDI, or Polydispersity Index, is an important parameter to be considered in drug delivery applications, as it provides an indication of the quality of the average particle size measurement with respect to the size distribution¹⁷. A low PDI indicates a more uniform size distribution, which is desirable for ensuring reliable behaviour and performance of the delivery system¹⁷. Generally, for the purposes of drug delivery applications, a PDI value below 0.4 is an acceptable measure¹⁸. Correlograms for their respective PDI measurements followed the expected trend, indicating mono sized dispersion. An example of such a correlogram can be found in Supplementary Figure 12.

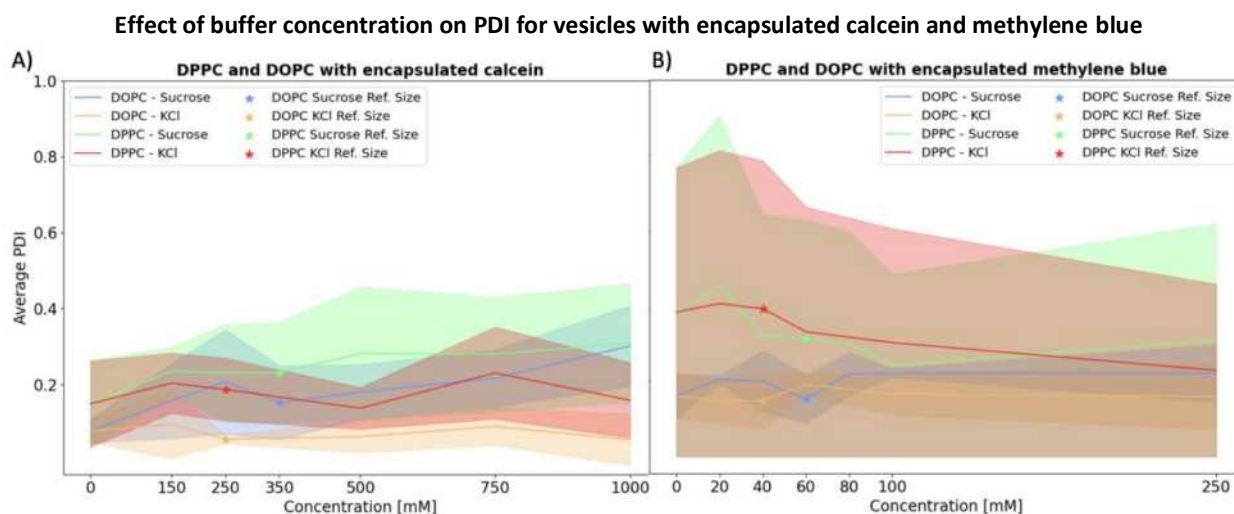


Figure 6. (A) and (B) refer to calcein and methylene blue encapsulated vesicles, respectively. The plots summarise the average PDI values for each lipid phase (fluid - DOPC or gel - DPPC), in either type of buffer (KCl or sucrose). The reference average PDI for each lipid phase and type of buffer is indicated using a star symbol. This reference helps to distinguish the concentration range for hypo and hyperosmotic environments as related to the sucrose and KCl buffers. (A) suggests that overall PDI values for DOPC are smaller in KCl than they are in sucrose, regardless of hypo or hyperosmotic environment. The same trend is observed in the case of PDI values for DPPC. Overall, it may be observed that while the PDI values for the liquid phase (DOPC) change more than they do for the gel phase (DPPC) in both hypo and hyperosmotic media, the liquid phase line is always below the gel phase line, suggesting lower PDI values. This is the case in both KCl and sucrose buffers. (B) suggests that overall PDI values for DOPC are smaller in KCl than they are in sucrose. In the case of DPPC, the opposite is observed, most PDI values seem to be higher for KCl than sucrose. Comparing DOPC and DPPC in sucrose, and in KCl, there does not seem to be a clear trend with respect to how PDI changes in the hypo and hyperosmotic media. Overall, however it is observed that PDI values for the liquid phase are smaller than the gel phase.

Key findings from Figure 6A, which relate to calcein encapsulated vesicles, suggest that overall, the PDI values for DOPC are smaller in KCl (orange) than they are in sucrose (blue), regardless of hypo or hyperosmotic environment. A similar trend is observed in the case of PDI values for DPPC. Furthermore, it may be observed that while the PDI values for the liquid phase (DOPC) fluctuate more than they do in the gel phase (DPPC), in both hypo and hyperosmotic media, PDI values for the liquid phase (blue and orange) are always smaller than in the gel phase. Overall, average PDI values for both lipids and in both buffers seem to be below or around the 0.4 reference PDI value mentioned previously.

Figure 6B, as related to methylene blue encapsulated vesicles, suggest that overall PDI values for DOPC are smaller in KCl than they are in sucrose. However, the opposite is observed for DPPC. Comparing DOPC and DPPC in sucrose, and in KCl, there does not seem to be a clear trend with respect to how PDI changes with

media. Once again, however it is observed that PDI values for the liquid phase are smaller than the gel phase. This can be expected as DOPC's fluid nature enables it to withstand changes in osmotic pressure better, and therefore experience a smaller degree of deformation, which is reflected in a lower PD value.

It is important to note that even though the average PDI measurements for both lipids in all buffers did not significantly exceed the 0.4 reference value, it is advisable to repeat measurements due to large differences in results recorded.

Conclusions

Vesicle stability was analysed by considering the extent of passive leakage and the deviations in size. The release profiles showed that vesicles with methylene blue leak over time, which was not seen with encapsulated calcein. This suggests that DPPC and DOPC vesicles with encapsulated calcein in both sucrose and KCl buffers are more stable than those with methylene blue,

as the final release is lower, and the release is stable (unchanged) over time. Vesicles had lower leakages in sucrose than KCl for methylene blue. However, despite this observation, no clear trend was established between the methylene blue release and concentration for both sucrose and KCl with both lipids.

Buffer concentration was found to have a higher effect in DOPC with calcein cargos, but no concentration dependence was observed for methylene blue with both lipids.

It was also concluded that DPPC is better than DOPC for methylene blue encapsulation as less passive leakage is observed. For both calcein and methylene blue, DPPC, which has a higher degree of phospholipid saturation, and less membrane fluidity, would be the more suitable choice for drug delivery purposes.

Key findings from the DLS explained the effect of lipid architecture and buffer on changes in vesicle size. For DOPC with encapsulated calcein, vesicles in hypoosmotic sucrose media and hyperosmotic KCl media were more stable. This distinction of buffer and environment was not clear for DOPC with methylene blue. Since size seemed to change in a similar manner in both types of media for both buffers.

For DPPC vesicles with calcein, they were more stable in KCl regardless of osmolarity, whereas, for methylene blue, there was no notable trend regarding stability in different buffers and environments. In calcein, it seemed as though DOPC vesicles were always more stable than DPPC (since smaller deviations in vesicle size were observed). In methylene blue, DOPC vesicles were only more stable in KCl.

However, despite DPPC vesicles having larger deviations between the reference size and final size with both encapsulated cargos in all buffers, this was the consequence of the membrane expanding to accommodate changes in osmotic stress.

For drug delivery applications, factors such as drug retention time, maximum release at target site, vesicle circulation time, likelihood to aggregate and macrophage sensitivity are all relevant in selecting the appropriate delivery system. Compromises are therefore required to maximise the benefits of the selected system.

Future experiments would focus on encapsulating both calcein and methylene blue in the same system to investigate vesicle stability under different types of buffer and concentrations. It is also imperative to explore their release mechanisms to see how the passive leakage changes when multiple drugs are combined, as dual drug delivery systems often involve the simultaneous or sequential release of two different drugs. It is also worth exploring the incorporation of PEG polymers and to make mixtures of lipids in various phases to increase vesicle stability. Other lipid architectures can also be explored, such as other saturated lipids with different carbon chain lengths should be investigated, to see if carbon chain length is another factor that influences vesicle stability.

The next step would then be to apply this research to co-delivery systems of different drugs or treatments and compare the performance against existing dual drug delivery systems.

Acknowledgements

The authors thank Dr. Ignacio Gispert for the immense help he provided during the project and are grateful to members of the Membrane Biophysics Group for their support.

References

- ¹ Nsairat, H., Khater, D., Sayed, U., Odeh, F., Al Bawab, A. and Alshaer, W. (2022). Liposomes: structure, composition, types, and clinical applications. *Heliyon*, [online] 8(5). doi:<https://doi.org/10.1016/j.heliyon.2022.e09394>.
- ² Shade, C.W. (2016). Liposomes as Advanced Delivery Systems for Nutraceuticals. *Integrative medicine (Encinitas, Calif.)*, [online] 15(1), pp.33–36. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4818067/>.
- ³ Liu, P., Chen, G. and Zhang, J. (2022). A Review of Liposomes as a Drug Delivery System: Current Status of Approved Products, Regulatory Environments, and Future Perspectives. *Molecules*, [online] 27(4). doi:<https://doi.org/10.3390/molecules27041372>.
- ⁴ Rai, M., Ingle, A.P., Bansod, S. and Kon, K. (2015). Chapter 9 - Tackling the Problem of Tuberculosis by Nanotechnology: Disease Diagnosis and Drug Delivery. In: M. Rai and K. Kon, eds., *Nanotechnology in Diagnosis, Treatment, and Prophylaxis of Infectious Diseases*. [online] Boston: Academic Press, pp.133–149. Available at: <https://www.sciencedirect.com/science/article/abs/pii/S9780128013175000098> [Accessed 14 Dec. 2023].
- ⁵ Ahmed, S., Ahmed, O. and Shah, P. (2018). *Biochemistry, Lipids*. [online] National Library of Medicine. Available at: <https://www.ncbi.nlm.nih.gov/books/NBK525952/> [Accessed 14 Dec. 2023].
- ⁶ ScienceDirect. (n.d.). *Hydrophobic Effect*. [online] Available at: <https://www.sciencedirect.com/topics/pharmacology-toxicology-and-pharmaceutical-science/hydrophobic-effect#:~:text=The%20hydrophobic%20effect%20discussed%20earlier> [Accessed 14 Dec. 2023].
- ⁷ Gispert, I., Hindley, J.W., Pilkington, C.P., Shree, H., Barter L.M.C., Ces, O. and Elani, Y. (2022). Stimuli-responsive vesicles as distributed artificial organelles for bacterial activation. *Proceedings of the National Academy of Sciences of the United States of America*, 119(42). doi:<https://doi.org/10.1073/pnas.2206563119>.
- ⁸ Sparr, E. and Wennerström, H. (2001). Responding phospholipid membranes--interplay between hydration and permeability. *Biophysical Journal*, [online] 81(2), pp.1014–1028. doi:[https://doi.org/10.1016/S0006-3495\(01\)75759-1](https://doi.org/10.1016/S0006-3495(01)75759-1).
- ⁹ Foflonker, F. (2023). *Osmotic pressure*. [online] Encyclopedia Britannica. Available at:

<https://www.britannica.com/science/osmotic-pressure> [Accessed 14 Dec. 2023].

assessments, *Drug Delivery*, 28:1, 973-984, DOI: [10.1080/10717544.2021.1927245](https://doi.org/10.1080/10717544.2021.1927245).

¹⁰ Nakhaei, P., Margiana, R., Bokov, D.O., Abdelbasset, W.K., Jadidi Kouhbanani, M.A., Varma, R.S., Marofi, F., Jarahian, M. and Beheshtkhoo, N. (2021). Liposomes: Structure, Biomedical Applications, and Stability Parameters With Emphasis on Cholesterol. *Frontiers in Bioengineering and Biotechnology*, [online] 9, p.705886. doi:<https://doi.org/10.3389/fbioe.2021.705886>.

¹¹ Le, J. (2020). *Drug Absorption*. [online] MSD Manual Professional Edition. Available at: <https://www.msdmanuals.com/professional/clinical-pharmacology/pharmacokinetics/drug-absorption> [Accessed 14 Dec. 2023].

¹² go.drugbank.com. (n.d.). *Methylene blue trihydrate* | *DrugBank Online*. [online] Available at: <https://go.drugbank.com/salts/DBSALT001852> [Accessed 14 Dec. 2023].

¹³ Advanced Instruments. (n.d.). *Freezing Point Depression Theory*. [online] Available at: <https://www.aicompanies.com/education-training/knowledge-center/freezing-point-depression-theory/> [Accessed 15 Oct. 2023].

¹⁴ Peverati, R. (2022). *14.2: Colligative Properties*. [online] Chemistry LibreTexts. Available at: [https://chem.libretexts.org/Bookshelves/Physical_and_Theoretical_Chemistry_Textbook_Maps/The_Live_Textbook_of_Physical_Chemistry_\(Peverati\)/14%3A_Properties_of_Solutions/14.02%3A_Colligative_Properties](https://chem.libretexts.org/Bookshelves/Physical_and_Theoretical_Chemistry_Textbook_Maps/The_Live_Textbook_of_Physical_Chemistry_(Peverati)/14%3A_Properties_of_Solutions/14.02%3A_Colligative_Properties).

¹⁵ Chen, S., Ren, J. and Chen, R. (2019). Cryopreservation and Desiccation Preservation of Cells. *Comprehensive Biotechnology*, 5(Third Edition), pp.157–166. doi:<https://doi.org/10.1016/b978-0-444-64046-8.00451-1>.

¹⁶ ScienceDirect. (n.d.). *Cryoprotectant*. [online] Available at: <https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/cryoprotectant#:~:text=Nonpermeable%20cryoprotectants%2C%20such%20as%20sugars> [Accessed 14 Dec. 2023].

¹⁷ Danaei, M.; Dehghankhold, M.; Ataei, S.; Hasanzadeh Davarani, F.; Javanmard, R.; Dokhani, A.; Khorasani, S.; Mozafari, (2018). M.R. Impact of Particle Size and Polydispersity Index on the Clinical Applications of Lipidic Nanocarrier Systems. *Pharmaceutics* 10, 57. <https://doi.org/10.3390/pharmaceutics10020057>

¹⁸ Iqra Rahat, Syed Sarim Imam, Md. Rizwanullah, Sultan Alshehri, Mohammad Asif, Chandra Kala & Mohamad Taleuzzaman (2021) Thymoquinone-entrapped chitosan-modified nanoparticles: formulation optimization to preclinical bioavailability

Enviro-Economic Analysis of Refrigeration Cycle Integration into Ground-Source Heat Pump-Supported Space Heating Systems

Elinor Lewis, Thao Vy Nguyen-Phuong

Department of Chemical Engineering, Imperial College London, U.K.

Abstract

This paper examines the integration of waste heat from the refrigeration cycle of a Sainsbury's store into the building heating, ventilation, and air conditioning (HVAC) system to investigate the potential of electricity consumption and carbon footprint reductions. Case studies were proposed with different configurational integration concepts: (0) no integration of the waste heat; (1) indirect integration into the ground; (2A) direct integration via a heat exchanger into the primary HVAC loop after the ground-source heat pump (GSHP); (2B) direct integration via a heat exchanger into the primary HVAC loop before the GSHP. The results then were compared to existing Sainsbury's store performances based on available historical data.

All cases considered reduced both the cost required to provide space heating, and the carbon dioxide emissions produced, compared to the base case. The most beneficial case was the indirect integration, case 1, whereby the refrigeration waste heat is directed into the ground near the supermarket, then extracted by the GSHP with an increased theoretical coefficient of performance (COP). This case however, due to the requirement of a GSHP, will be difficult to retrofit to existing stores. The primary integration cases 2A and 2B, do not suffer this drawback, and can readily be implemented into existing stores, including those operating with a gas boiler, providing a reduction in cost and emissions of the space heating systems. Our analysis also emphasises the value of government incentives to make renewable energy solutions and waste heat integration economically competitive to traditional technologies.

Keywords: Waste Heat Integration, Refrigeration, Ground-source Heat Pump, Supermarket

1. Introduction and Background

Supermarkets in the United Kingdom consume 3% of the country's total electricity, corresponding to 1% of the greenhouse gas (GHG) emissions. Refrigeration systems within these supermarkets use between 30-60% of the electricity, thus meaning supermarket refrigeration systems use approximately 1% of total UK electricity (Tassou, et al., 2011). As the concerns of rising global temperatures and the connection to the burning of fossil fuels continue to rise, focus has been put onto reducing GHG emissions, hence the UK has a target to reduce emissions to 77% of 1990 levels by 2035 (Sunak MP, 2023). As supermarkets, and notably their refrigeration systems, contribute significantly to these GHG emissions, reducing this energy consumption particularly, is an important area of focus to reach this target. One method to aid this is to integrate the refrigeration system into space heating, to utilise waste heat energy of condensing the refrigerant. This is particularly significant as the heating, ventilation, and air conditioning (HVAC) systems can use up to 35% of supermarkets electricity demand (Tassou & Ge, 2008). A study done in 2011 (Ge & Tassou, 2011) suggests that heat recovery from CO₂ refrigeration systems can satisfy up to 40% of the space heating demand within supermarkets, thus reducing the overall energy requirements. Sainsbury's have gone on to surpass this estimate. Sainsbury's Olney store fulfils all the space heating demand using an integrated refrigeration heating and cooling system (Sainsbury's, n.d.). However, this is a small store, typically associated with a higher proportion of refrigeration to space heating demand, meaning this integration is much easier to achieve. This does however prove the value and validity of integrating refrigeration and space heating systems.

An additional way for supermarkets to reduce their greenhouse gas emissions from the HVAC systems is to implement a ground source heat pump (GSHP) to provide a portion of the space heating demand. The temperature of the ground remains steady all year round, between 7°C and 12°C in Britain (NERC, 2011). In wintertime when the surface temperature is below this, the GSHP can provide heating to the supermarket by pumping a working fluid into boreholes within the ground, which then extracts thermal energy. This is then utilised by the supermarket for space heating. Conversely, in summertime the GSHP can be used to meet the cooling demand as the working fluid deposits thermal energy into the ground (Monschauer, et al., 2022), this is also essential to avoid excessive fluctuation of the ground temperature, and to ensure it remains within the expected temperature range (Dalpane, et al., 2016). This summertime mode of operation has the added benefit that thermal storage of the excess heat can be extracted by the GSHP in winter improving the coefficient of performance (COP) (Maidment, 2013). The COP is a metric of performance where the rate of thermal heat delivered by the system (\dot{Q}_{th}) is compared to the electrical power input (kW) of the heat pump, as shown in equation 1. Typical values of the COP are between 3 and 5 (Maidment, 2013).

$$COP = \frac{\text{Rate of Heat Delivered (kW}_{th})}{\text{Power Input (kW)}} \quad (1)$$

Ground-source heat pumps have been implemented into Sainsbury's stores to aid net zero targets (Silverman, 2020), more specifically the one on Kings Lynn Hardwick Road. This store is located near the Norfolk coast, opened in late 2012 and has a sales area of 72,196 m². Instead of a gas boiler, as would be a conventional practise, the space heating is delivered by a GSHP.

Reaching net zero is not achieved with a singular solution. Many different approaches must be used in tandem, investigating the coupling of proven technologies, a GSHP with a refrigeration waste heat integration system is beneficial. The expectation is that the waste heat integration lowers electricity consumption of the GSHP, thus lowers GHG emissions to a greater extent than each solution separately.

The analysis of this paper was conducted using energy consumption data collected from 1st January to 6th December 2022 within Sainsbury's stores. This study analyses the performance of GSHP and refrigeration integration into the HVAC system in different configurations. The cost and carbon dioxide emissions of all cases is compared to each other, as well as to a typical store obtaining space heating from a gas boiler.

2. Benchmarking Analysis

To gain an insight into the space heating demand of Sainsbury's supermarkets, a comparison of the Kings Lynn store, GSHP-supported, was made to other stores operating with a conventional gas boiler, before considering refrigeration integration. The stores of comparison were Hayes, Lincoln, Wandsworth, and Washington. All have a comparable sales area but opened earlier than Kings Lynn, details are found in table 1.

Table 1: Overview of Sainsbury's stores analysed.

Store	Type	Opened	Sales area (m ²)	Location
Kings Lynn	GSHP	2012	72,196	Midlands
Hayes	GAS	1993	76,129	London
Lincoln		1991	75,678	Midlands
Washington		1977	69,963	North East
Wandsworth		1987	73,369	London

To directly compare the energy consumption used by the different stores, the daily space heating energy consumption in 2022 (E_{daily}) was normalised by the store area (A_{store}), and heating degree day (HDD). The HDD calculated with equation 2, considers how much colder the external temperature is compared to the desired internal temperature of the store which is the baseline temperature of 15°C (Met Office, 2023). The final equation to determine the normalised daily space heating energy demand (E_{norm}) is given by equation 3.

$$HDD = T_{\text{baseline}} - T_{\text{daily ave}} \quad (2)$$

$$E_{\text{norm}} = \frac{E_{\text{daily}}}{A_{\text{store}} \cdot HDD} \quad (3)$$

In figures 1-5 all stores exhibit a positive correlation between E_{norm} and HDD. This is as expected, as it is typical to invest more in space heating when it is colder externally. Another point of significance would be that for all HDD values E_{norm} is much less for Kings Lynn than for the other stores. This is due to the higher efficiency associated with GSHPs compared to gas boilers (Calvillo, et al., 2023) meaning much less energy is required to achieve the same amount of heating. Washington appears to be anomalous to the other gas stores, requiring much more energy for heating for the same HDD, this could be due to its comparative age as

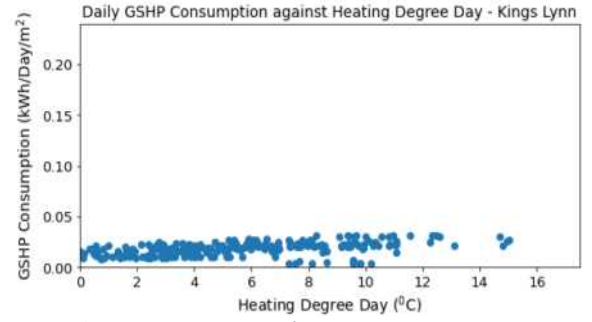


Figure 1: E_{norm} against HDD for Kings Lynn

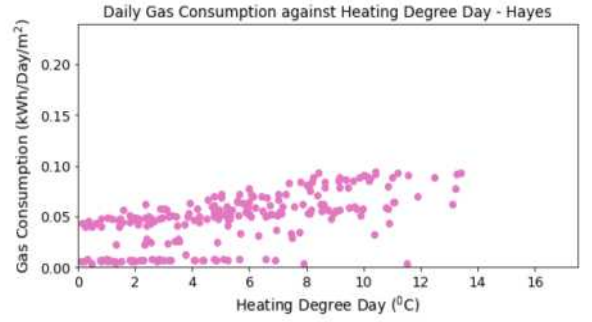


Figure 2: E_{norm} against HDD for Hayes

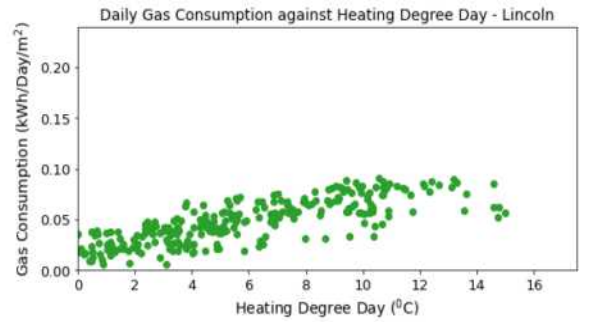


Figure 3: E_{norm} against HDD for Lincoln

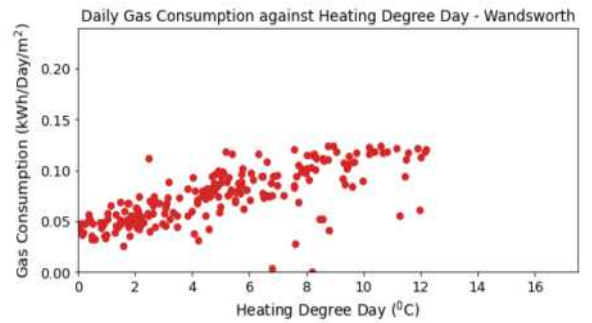


Figure 4: E_{norm} against HDD for Wandsworth

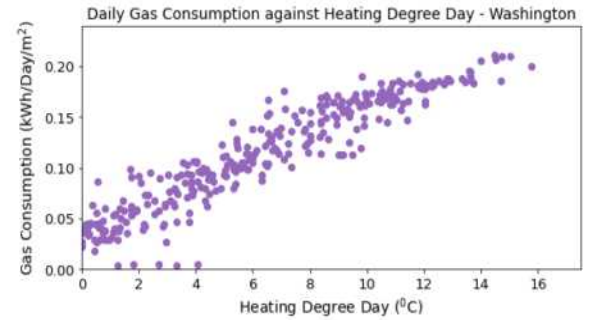


Figure 5: E_{norm} against HDD for Washington

older buildings are associated with lower energy efficiencies (ONS, 2022).

2.1 Carbon Dioxide Emission Comparison of GSHP and Gas Stores

To provide further insight into the different modes of operation, the annual normalised carbon dioxide emissions ($\text{kgCO}_2 \text{ m}^{-2} \text{ HDD}^{-1} \text{ year}^{-1}$) were compared. The total normalised energy consumption for space heating in 2022 was calculated, then multiplied by an emission factor, EF, of $0.19338 \text{ kgCO}_2 \text{ kWh}^{-1}$ for GSHP and $0.2 \text{ kgCO}_2 \text{ kWh}^{-1}$ for gas (GOV UK, 2022) to determine the kg of CO_2 released by the energy consumption, shown by equation 4.

$$\text{Annual CO}_2 \text{ emission} = EF * \sum E_{\text{norm}} \quad (4)$$

All gas stores produced much more CO_2 than Kings Lynn on a normalised basis. The lowest store, Lincoln produced double the CO_2 emissions of Kings Lynn, whereas Washington produced over four times as much CO_2 as Kings Lynn, shown in figure 6. The lower emissions of Kings Lynn are because of the lower energy consumption of the GSHP due to the higher efficiency discussed previously, and the influence of the slightly lower emission factor of electricity.

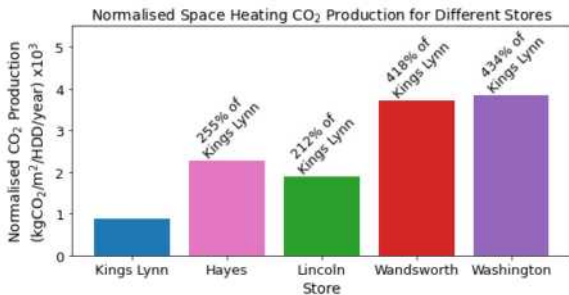


Figure 6: Comparison of the normalised CO_2 production from space heating in different stores

2.2 Economic Comparison of GSHP and Gas Stores

Analogously, an economic comparison was conducted by multiplying the annual normalised energy consumption ($\text{kWh m}^{-2} \text{ HDD}^{-1} \text{ year}^{-1}$) by the unit price of each energy type, as shown in equation 5. Electricity was taken at a representative price of $\text{£}0.30 \text{ kWh}^{-1}$ and gas at $\text{£}0.08 \text{ kWh}^{-1}$.

$$\text{Annual Cost} = \text{unit price} * \sum E_{\text{norm}} \quad (5)$$

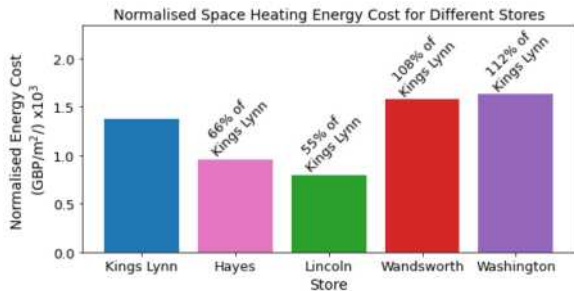


Figure 7: Comparison of the normalised cost production from space heating in different stores

The cost of heating the gas stores was comparable or more economical than the GSHP store. Hayes and Lincoln were around half as costly, whilst Wandsworth

and Washington were around 10% more costly than Kings Lynn, shown in figure 7. This is because despite the higher efficiency, the electricity required to operate the GSHP is 3.75 times more expensive than gas. The two stores more costly to heat than Kings Lynn were the two oldest, Washington and Wandsworth. This is unsurprising as the lower energy efficiency of older buildings mean they require more energy to achieve the same degree of heating and hence are more expensive to heat.

Overall, regarding carbon dioxide emissions the GSHP store vastly outperforms the gas boiler stores, highlighting their relevance to meet net zero targets. Economically however, the cost of the GSHP store is less competitive.

3. Refrigeration Integration Cases

There are many ways in which the refrigeration system can be integrated into the HVAC system. Different cases, as well as the base case of no integration for comparison, are described. Several commonalities exist throughout all cases. The space heating duty required by the building is assumed constant at 500 kW as per the maximum load for the control system. Consequently, the building supply and return temperatures are set at 45°C and 30°C , respectively.

The control volume for the GSHP heat pump is the primary heat pump loop (orange) and the GSHP extraction loop (purple) on all the schematics below.

3.1 Base Case – No Integration of Refrigeration System

In the base case, the building HVAC system is supported by the ground-source heat pump, only. The base case provides the basis to further comparison, the configuration is shown in figure 8.

‘The use of heat pumps, namely Ground Source Heat Pumps (GSHPs), has increased significantly in recent decades worldwide due to their low carbon footprint and their ability to extract heat from the ground for building heating and cooling in different climatic typologies.’ (Xian Li, et al., 2023)

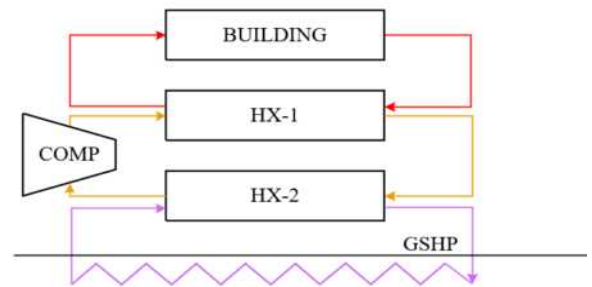


Figure 8: Base case schematics (red: HVAC loop, orange: primary heat pump loop, purple: GSHP heat pump loop)

3.2 Case 1 – Indirect Integration

In case 1, shown in figure 9, the residue condensation heat from the refrigeration cycle is supplied into the ground. Here, as the temperature difference is reduced between the inlet and outlet of the heat pump system, the COP is expected to be higher, and thus the compressor work is expected to be reduced due to the

improved efficiency of the system. This is the current configuration of the Sainsbury's store in Kings Lynn.

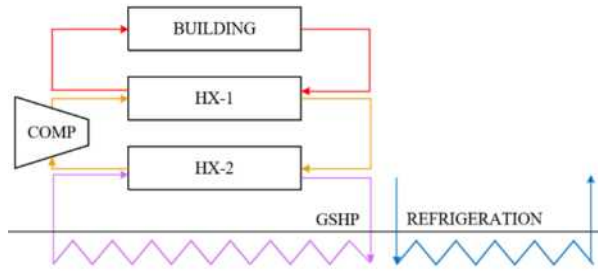


Figure 9: Case 1 - real case schematics (red: HVAC loop, orange: primary heat pump loop, purple: GSHP heat pump loop, blue: refrigeration loop)

3.3 Case 2A – Primary heat exchanger integration after GSHP

Case 2A, shown in figure 10, integrates the waste heat into the primary HVAC loop after the GSHP, to directly reduce the heating demand supplied by the GSHP.

The expected benefit of this configuration is the reduction in the workload of the heat pump compressor, with a small variation in the COP.

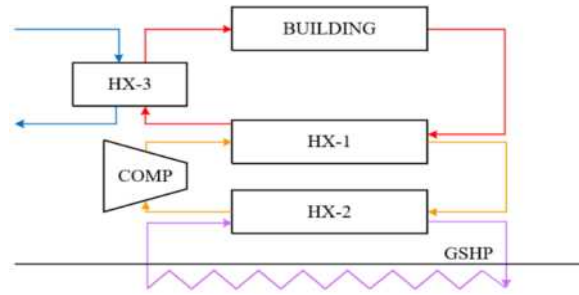


Figure 10: Case 2A schematics (red: HVAC loop, orange: primary heat pump loop, purple: GSHP heat pump loop, blue: refrigeration loop)

3.4 Case 2B – Primary integration before GSHP

Similarly, to case 2A, here the refrigeration cycle is integrated in the primary thermodynamic loop of the HVAC system. The difference is the order of reservoirs: refrigeration system before the GSHP, as shown in figure 11.

This configuration is expected to result in similar benefits as of case 2A.

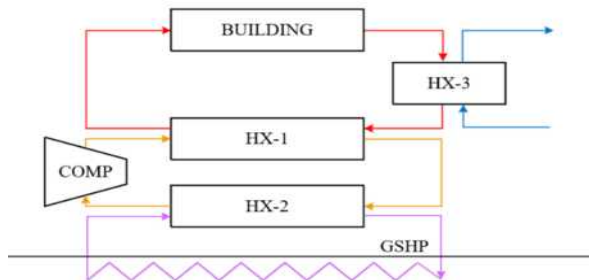


Figure 11: Case 2B schematics (red: HVAC loop, orange: primary heat pump loop, purple: GSHP heat pump loop, blue: refrigeration loop)

4. Integration Analysis Methods

4.1 Refrigeration Waste Heat

As measured telemetry data was not available for the refrigeration cycle a primary model of the system was developed in Aspen Plus, shown in figure 12. The refrigeration system was modelled as a subcritical R744 (CO₂) cycle using specified operating conditions in conjunction with thermophysical data from (NIST, 2021). Only the intermediate temperature refrigeration system was included. The Aspen method selected was REFPROP. This was developed by NIST to provide thermodynamic and transport properties of industrially important refrigerants including CO₂ (Aspen Plus, 2019). Several key assumptions were made, such as the isentropic efficiency of the compressor being 0.8 (Rasmussen & Kurz, 2009). The development of the Aspen model ultimately revealed a condenser duty of 280.00 kW, which is required to lower the inlet temperature of 92.23°C (stream 7-in) to the outlet temperature of 0.55°C (stream 8). The duty is the maximum amount of waste heat extractable for heating in the HVAC system, and the temperatures provide constraints to the integration cases. Streams 7-in and 7-out are the inlet and outlet to the heat exchanger that is integrated into the HVAC system. HVAC-IN and HVAC-OUT and the HVAC supply and return temperatures of 30°C and 45°C.

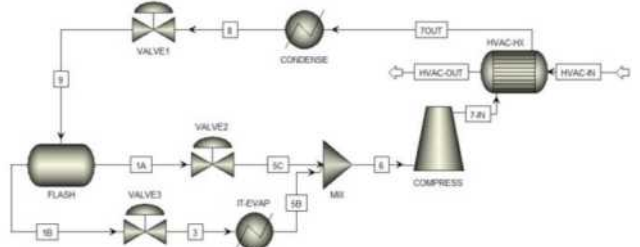


Figure 12: Aspen model created to determine refrigeration waste heat duty.

4.2 Compressor Workload and COP

Due to the lack of historical data for the GSHP, all calculations were based on a theoretical approach with the guidance of engineers from the Sainsbury's team.

4.2.1 Base case

As explained in section 3.1 the base case is the GSHP system without the integration of the refrigeration cycle. The theoretical COP for this case is found to be 7.52 using equation 6 according to Carnot's theorem (Sidebotham, 2022).

$$COP_{th} = \frac{T_{hot}}{T_{hot} - T_{cold}} \quad (6)$$

where T_{hot} is the outlet temperature of the heat pump compressor, and T_{cold} is the ground temperature from where the heat is extracted. T_{hot} is kept constant at 50°C due to a 5°C approach temperature difference with the return temperature to the building of 45°C. The ground temperature is assumed to be 7°C. With the heating duty assumed to be a constant 500 kW provided solely by the heat pump, $Q_{HVAC} = Q_{HP} = 500.00$. The work of the heat pump was calculated to be 66.53 kW by rearranging equation 7.

$$Q_{HP} = W * COP \quad (7)$$

4.2.3 Case 1

Case 1 is the current configuration of the Sainsbury's store in Kings Lynn as defined in section 3.2. The heat from the refrigeration cycle is exerted into the ground to raise its temperature from 7°C to an estimated 15°C.

The theoretical COP from equation 6 is calculated to be 9.23 by substituting in 50°C as the outlet temperature of the compressor and 15°C as the ground temperature.

From equation 7, the work of the heat pump is calculated to be 54.15 kW.

4.2.3 Case 2

In cases 2A and 2B, the refrigeration cycle is integrated into the primary thermodynamic cycle to extract heat for space heating in the HVAC system of the building.

The surplus heat of the refrigeration cycle provides a maximum of 280.00 kW (section 4.1), which can only be fully extracted by subcooling the working fluid, CO₂, to 0.55°C. Thus, exploitation of the latent heat of 197.22 kJ kg⁻¹ would be required at the dew point of 9.98°C at 45 bar (NIST, 2021), the operating pressure of the refrigeration cycle.

If the maximum available heat of 280 kW were to be able to be extracted, the energy required by the GSHP would be 220 kW. This would correspond to a heat pump work of 29.27 kW, with a COP_{th} of 7.52 unchanged from the base case. Unfortunately, this is unfeasible.

4.2.3.1 Case 2A

The primary integration of the refrigeration cycle after the heat pump from the GSHP is case 2A, as previously described in section 3.3. The COP for the heat pump changes, as the outlet of the compressor cannot reach 50°C like in previous cases.

As explained in the previous section, the configuration has a constraint due the temperatures of the HVAC system. This governs the amount of heat the refrigeration system can supply to the building. Due the second law of thermodynamics, no temperature crossover can occur to maintain a driving force and avoid a pinch point (Sidebotham, 2022). In the pinch point analysis, an approach temperature difference of 5°C was used. Via iterative calculations on energy balances the minimum outlet temperature, that satisfies temperature crossover constraints, was found to be 48.68°C. Therefore, using the fundamental heat transfer equation below (equation 8), the amount of waste heat integrated from the refrigeration cycle was found to be 44.04 kW.

$$Q = F * C_p * \Delta T \quad (8)$$

where Q stands for heat in kW, F is the flowrate of the working fluid in the refrigeration cycle modelled in Aspen, 3000 kg hr⁻¹, C_p is the average heat capacity of CO₂ over the range of 50 – 92°C at 45 bar, (NIST, 2021), and ΔT is the temperature difference of the inlet and outlet streams for the refrigeration cycle loop from the heat exchanger connecting to the HVAC system.

This vast difference from the ideal case is due to large disparity between sensible and latent heat, as the

temperature is restricted to remain above the pinch point. It is, therefore, not possible to exploit the latent heat at the specified operating conditions.

Thus, the heat pump needs to supply $Q_{HVAC} - Q_{frig} = 500.00 - 44.04 = 455.96 \text{ kW}$.

The theoretical COP is improved slightly to 7.72 as T_{hot} is lowered to 48.68°C and T_{cold} remains at 7°C. And thus, using equation 7, the work of the heat pump is 59.05 kW.

4.2.3.2 Case 2B

In case 2B, the supply temperature from the building of 30°C again, limits the outlet temperature of the CO₂ refrigeration loop, here the refrigeration heat exchanger outlet temperature is 35°C, which gives a maximum available heat to be 57.87 kW. The heat capacity here is estimated to be the average over the range of 35 – 92°C, explicitly, 1,21 kJ kg⁻¹ K⁻¹. Thus, the remaining heat that the GSHP and its heat pump needs to provide is $Q_{HVAC} - Q_{frig} = 500.00 - 57.87 = 442.13 \text{ kW}$. Since the COP_{th} is 7.52, the work of the heat pump is 58.83 kW.

5. Results and Discussion of Refrigeration Integration

5.1 Results of Compressor Work and COP

Overall, following the constraints for each case and the configurations, the base case requires a theoretical work of 66.6 kW, case 1 requires 54.2 kW, whilst with constraints, case 2A requires 60.9 kW and 2B 58.8 kW, as summarised in table 2. Case 1 leads the largest reduction in compressor work, 18.6% less than the base case, whilst case 2 configurations lead to a reduction of around 10%. Case 1 is also most efficient with a COP of 9.23.

Table 2: Summary of the theoretical heat pump work requirements for all cases

	Base	1	2A	2B
Q _{HVAC} (kW)	500	500	500	500
Q _{fridge} (kW)	-	-	42.7	57.9
Q _{GSHP} (kW)	500	500	457	442
COP _{GSHP}	7.52	9.23	7.52	7.52
W _{GSHP} (kW)	66.6	54.2	60.9	58.8
W _{total} Saving (%)	-	18.6	8.54	11.6

5.2 Discussion of Refrigeration Integration

5.2.1 Effect on Heat Pump Coefficient of Performance and Compressor Work

Shown in table 2, the integration of the refrigeration cycle causes a reduction in compressor work in *all* cases, compared to the base case.

In case 1, the reduction in compressor work is due to an increase in the compressor efficiency (COP) thanks to the more favourable operating conditions of the heat pump. By utilising the heat from the refrigeration cycle, the temperature of the ground is raised. Thus, the load of the heat pump is lowered by reducing the temperature difference between the reservoirs of the heat pump, and therefore easing the heat transfer from the cold to the hot reservoir.

For both scenarios of case 2, the coefficient of performance is unchanged or deviates only slightly from the base case. The reduction in compressor work is due to the integration of the refrigeration cycle in the primary loop. Thus, there is a reduction in the amount of heat required from the ground by the heat pump, consequently, the amount of work that the compressor must execute.

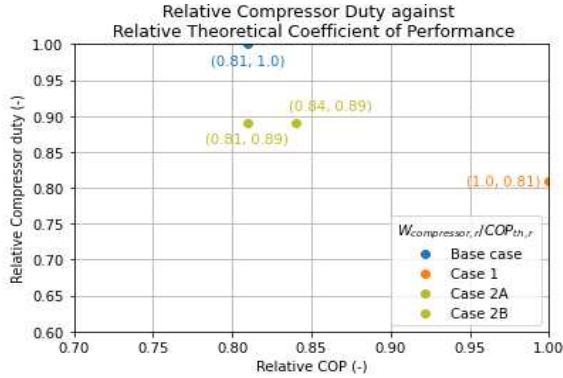


Figure 13: Relative theoretical coefficient of performance against relative compressor duty

Figure 13 shows all cases compared, relative to each other. As demonstrated, case 1 operates at the highest compressor efficiency, whilst all other cases operate at only around 81% of the efficiency of case 1. Compared to the base case, case 2B operates at the same efficiency, whilst case 2A introduces a 3% increase, with case 1 leading to a 23% increase in compressor efficiency.

Large-scale R717 heat pumps operate in a range of 0.50 – 13.00 MW (Aguilera, et al., 2022), whilst the heat pumps discussed in this study operate in the range of 54.1 – 66.6 kW. Yet, as the failure of the compressor is the costliest in the operation of a heat pump (Madani & Roccatello, 2014), it is sensible to mention that the upscaling of such cases for larger stores should be mindful of the compressor load. An increase of 23 % in compressor duty of the base compared to case 1 can reduce the lifetime of the compressor, and thus cause undesirable costs, whilst also causing degradations over the expected performance. From this perspective, case 1, the current structure of the Kings Lynn store, is the most desirable configuration.

5.2. Carbon Dioxide Emission Comparison

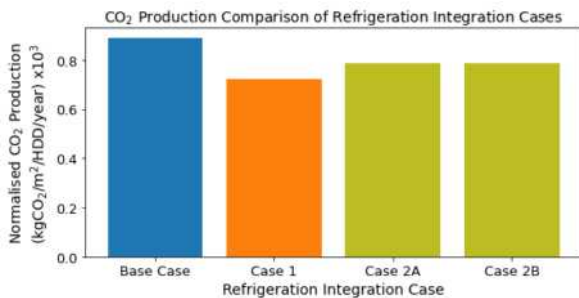


Figure 14: Effect of refrigeration integration on CO₂ production from the Kings Lynn space heating.

As expected, when refrigeration integration is implemented, the compressor work of the GSHP is reduced, and the normalised CO₂ production is reduced by the same amount, as shown in figure 14. The Kings

Lynn case, case 1, previously produced just 30% the carbon dioxide emissions compared to the average gas store on a normalised basis. The base case produces 37% of the emissions of the averaged gas stores, whilst cases 2A and 2B produce 33%. This demonstrates that

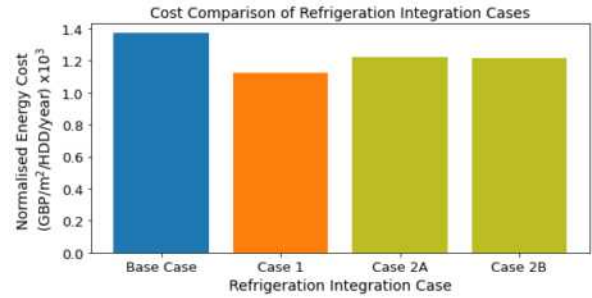


Figure 15: Effect of refrigeration integration cases on cost of Kings Lynn space heating

refrigeration integration is valuable to reducing the carbon dioxide emissions of a GSHP as all cases lowered the emissions compared to the base case.

5.3. Economic Comparison

Similarly, to the carbon dioxide emissions, as the compressor work is reduced with each case of refrigeration integration the normalised cost of space heating also decreases to the same extent compared to the base case, shown in figure 15.

The base case costs 36% more than an average gas store to provide space heating on a normalised basis. This decreases, yet remains more than the average gas stores by 11% for case 1, 21% for case 2A and 20% for case 2B. Despite the decrease in cost, all cases of the GSHP and refrigeration integration are still more costly to heat than the average gas store.

5.3.1 Renewable Heat Incentive (RHI)

The non-domestic renewable heat incentive (RHI) is a governmental initiative to motivate businesses, public sector, and non-profit organisations to reduce electricity consumption supplied from the grid. This supports the transformation of the UK towards net zero targets (UK Gov, n.d.).

The eligibility criteria for the RHI are listed in table 3, which are all met by the configurations discussed in this study.

Table 3: Summary of the eligibility criteria for the non-domestic RHI (as of 2023/24)

	Required	Acquired
Capacity	100 kWth <	yes
COP	2.90 <	yes
SPF	2.50 <	yes

The incentive can be claimed for only ‘naturally occurring energy’ and ‘must not be designed to provide cooling or to use heat which has been expelled from a building or from a process which generates heat’ (Ofgem, 2022). According to section 8.13 of the governmental specifications, if the amount drawn from the ground is measurable for simultaneous operations, the natural heat drawn from the ground is still eligible for incentive claim. Hence all configurations are eligible to a certain amount. As summarised in table 4, the amount of incentive for the cases varies, slightly. The highest

amount can be claimed via the base case and case 1, followed by case 2A, and lastly case 2B. Case 2A is a 3% increase compared to the smallest, case 2B, while the base and case 1 are an 13% increase.

Table 4: Summary of RHI for all cases

	Base	1	2A	2B
Q_{HP} (kW)	500	500	457	442
RHI (10^3£)	19.8	19.8	18.1	17.5

The GSHP in Kings Lynn became accredited when the store opened in 2012, as such all eligible heat output is at a tariff rate of $\text{£}0.0452 \text{ kWh}^{-1}$ (Ofgem, 2022). The eligible heat output was assumed to be the total heat output, Q_{HP} , of the GSHP as a best-case estimate. To determine the effective cost of running the GSHP with the RHI scheme, the annual savings achievable are determined by multiplying the eligible heat output by the tariff rate. This is subtracted from the annual electricity cost for the GSHP, then normalised as before.

As shown by figure 16 below, the RHI leads to a reduction in energy cost for all cases. As the incentive is dependent on GSHP heat output, the cases where this is larger, the base case and case 1, achieve greater savings.

One note of significance is that with the RHI schemes, all cases of integration become more cost effective or highly competitive to the average comparable gas boiler store. With the RHI scheme, the base case reduces to 19% more costly than the average gas store on a normalised basis, but case 1 now costs 6% less, case 2A 5% more, and case 2B 6% more.

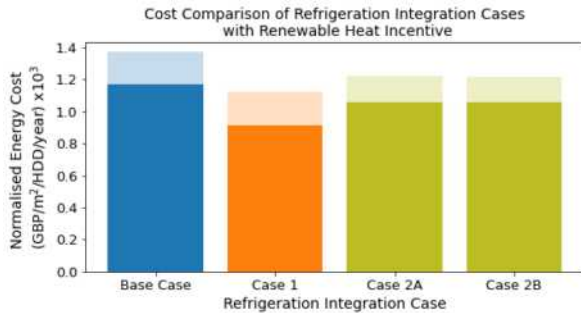


Figure 16: Effect of refrigeration integration cases on cost of running the Kings Lynn space heating with (bold) and without (shadow) the renewable heat incentive.

5.3.2 Carbon Emissions Tax

It is likely for the UK to introduce a scheme such as shadow carbon pricing, whereby CO_2 emissions are taxed (Ferrovial, n.d.), thus an analysis considering this was conducted. An expected price of $\text{£}100$ per tonne of CO_2 emitted was applied to the total annual costs. This was applied separately and in conjunction with the renewable heat incentive and normalised, as before.

All integration cases are cost competitive to the average gas store, only the base case remains more costly. Due to the greater carbon dioxide emissions, as discussed in section 2.1, gas stores would be impacted to a greater extent by a carbon tax. Table 5 summarises the heating costs of the taxed integrated cases with and without the RHI to the gas stores.

Table 5 also highlights the importance of government incentives to making renewable energy

solutions cost competitive to traditional technologies. Only with these incentives do the integration cases become cost competitive to the gas stores. This is noteworthy as it is unlikely that businesses will move towards renewable solutions to lower their GHG emissions unless it financially benefits them.

Table 5: Summary of the heating costs of the taxed integrated cases with and without the RHI to the gas stores

	Base	1	2A	2B
Cost vs. Average Gas Store (Carbon Tax)	+17%	-10%	-2%	-2%
Cost vs. Average Gas Store (Carbon Tax and RHI)	+3%	-23%	-14%	-14%

6. Conclusions

It was found that all feasible cases of refrigeration integration led to a reduction in the electricity demand required to provide space heating in a large UK Sainsbury's store. This reduced electricity demand leads to a decrease of the carbon dioxide emissions and cost of operation by the same percentage. The integration cases considered include: (1) indirect integration whereby heat is deposited into the ground to enhance compressor efficiency in the GSHP; (2A,B) direct integration of the refrigeration system into the building HVAC system via a heat exchanger to lower the heating demand of the GSHP. Case 1 was found to reduce the GSHP electricity consumption the most, by 19%.

Comparisons of the integration cases were made to the base case of a GSHP, and stores providing space heating with a typical gas boiler. It was shown that when including the savings achievable with the Renewable Heat Incentive scheme of the UK government, all GSHP refrigeration integration cases became cost competitive to a traditional gas store. This was not quite the case without this scheme applied, highlighting the importance of governmental schemes to motivate businesses to achieve net zero emission targets. The integration cases become even more economically competitive to gas stores, when considering the RHI scheme and a carbon tax, which the UK government will likely introduce in the near future. With both these government incentives applied, the best-case refrigeration integration, case 1, costs 23% less than the average gas store to provide the same amount of normalised heating.

According to existing literature, the reduction of electricity consumption is expected by the integration of waste heat. In Canada in a similar study to this (Reddick, et al., 2020), thermal demands of the building were linked to potential heat sources coupled with a pinch analysis. The conclusion of said paper was, that the combination of greywater and heat pumping reduces the electricity costs by 53%. Furthermore, additional solar thermal collectors can reduce the consumption up to 64%. In another study (Dhole & Linnhoff, 1993), the Total Site Heat Integration method was used with mathematical optimisation, and an extended pinch analysis. In the paper waste heat and renewable energies were integrated into an industrial site and were shown to reduce CO_2 emissions.

The study was conducted by guidance from the Sainsbury's engineering team on a theoretical basis due

to the lack of historical data for the refrigeration and GSHP systems. Therefore, the confidence of the findings is limited. Typical COP values range between 3 and 5 (Maidment, 2013), whilst the theoretical values in this paper range between 7 and 9, and hence the benefits drawn are likely overestimations. Yet, previous works and research support that integration of waste heat is beneficial for both the reduction of electricity consumption and the carbon emissions of the system, as found in this project. Further studying of the topic is highly suggested as explained in sections 7.2 and 7.3 below, both from a large and small-scale perspective. Pressing matters in the net zero transition of the UK demand an urgent action in the energy strategy.

6.1 Retrofitting Refrigeration Integration Cases

There is much difficulty associated with retrofitting ground source heat pumps due to the requirement of vast amounts of space or underground boreholes to extract the energy held within the ground (UK Gov, 2022). This presents a hinderance to case 1, as it cannot be easily implemented in existing stores without a GSHP. Furthermore, retrofitted GSHPs are associated with lower efficiencies (Breembroek, 2002). Hence the reductions in compressor work and improvement of COP will occur but less impactfully than in case 1.

The variations of case 2A and B however, could potentially be installed to existing gas stores or stores with other heating configurations much more easily with heat exchangers. This reduces the demand of the gas boiler, by supplying a portion of the space heating demand by primary refrigeration integration. This reduces the amount of gas required, and henceforth the associated carbon dioxide emissions and the cost.

The limitations of retrofitting suggest that case 2A should be implemented to existing stores in the short term to quickly reduce costs and emissions. However, in the long term and for all new buildings the configuration of case 1 should be installed or retrofitted.

7. Outlook

7.1 Analysis Limitations

Due to the brevity of the project and the assumptions made, the model built for the system has limitations. Firstly, the assumption of constant building heating demand can be improved by introducing diurnal and seasonal changes. Coupled with sensitivity, consumer and error analysis, an extended model can be built based on time-varying demand.

Additionally, time-varying pricing can be introduced to reflect industrial economy mechanisms. Capital costs associated with the integration of the cases should also be considered.

7.2 Alternatives and Whole-System Approach

Many studies have been conducted on the topic of thermal storage. In a study of (Ohannessian & Sawalha, 2014), similarly, cases of refrigeration were proposed and modelled. It is discussed that a GSHP operating as a brine thermal storage unit has a higher COP than a GSHP operating as a connecting heat pump. Another suggestion in the same paper is that a

supermarket system with heat recovery performs significantly more efficiently than the ones relying solely on the GSHP. In summary, supermarkets with GSHP can further reduce their energy consumption by changes in the system. Furthermore, with thermal storage alternatives, the dependency of the system on grid supply can reduce the need to stabilise with fossil fuels or nuclear energy. This strategy is called load shifting from the grid to the respective consuming units. Current thermal storage alternatives include hydrogen boilers, hybrid heat pumps with natural gas boilers (Hoseinpoori, et al., 2023), electrochemical batteries (Ghilardi, et al., 2023).

Furthermore, other alternatives can be further investigated such as thermal storage (Li, et al., 2021) and a potential hydrogen technology (Aunedi, et al., 2023). A creation of an energy mix was found (Hoseinpoori, et al., 2023) to be a great solution to increase independency from uncontrollable factors and to increase energy security for any system.

In a more detailed model, system flexibility can also be investigated, which is '*the ability of the system to reliably and cost-effectively manage the variability and uncertainty of demand and supply across all relevant timescales*' (IEA, 2018). Thus, from an energy security perspective by analysing system flexibility, the pressing demands of net zero targets can be aided to be met through a smoother transition period.

7.3 Store-Specific Recommendations

Above, the potential of further work on large-scale implementation scenarios were employed on a systematic aspect. In the case of small-scale opportunities, other configurations can be further explored. Different working fluid performances can also be further studied like in the work of (Radulovic, et al., 2023) on refrigeration cycle fluids, where similarly to this study, the COP and compression work is analysed.

As another example, further improvement can be made by modelling the ground temperature change caused by the heat extracted from the refrigeration cycle. Another potential is to implement control systems on built models to see if the physical implementation of said cases are feasible.

As discussed here, a healthy amount of potential lies in the study of heat integration. It is encouraged for this topic to be further examined, as from small-scale perspective localised costs can be reduced, and the energy security of a store can be established with less dependency on district energy supplies. On a large scale, by tools discussed above, a smoother systematic transition can be established to support the net zero targets by 2050 for the UK.

7.4 Proposed Third Case of Refrigeration Integration

As discussed in sections 4.2.3.1 and 4.2.3.2, the full duty available from the refrigeration system is not exploitable via primary integration with the HVAC system due to second law of thermodynamic constraints preventing temperature crossover. As such in cases 2A and 2B only a fraction of available refrigeration waste heat is integrated into the HVAC system. A proposed third integration case should be developed and

investigated as an additional mode of operation, to exploit much more of the available waste heat.

In case 3, the case 2 configuration is altered with a heat pump instead of heat exchanger, shown in figure 17. This configuration is expected to overcome temperature crossover via an additional working fluid, and thus increase the amount of heat extracted from the refrigeration cycle with a trade-off in the workload of the additional heat pump compressor. The drawback is the lack of efficiency during the summertime period. Furthermore, the work required for this heat pump operation is unknown and thus the overall benefit is also unknown.

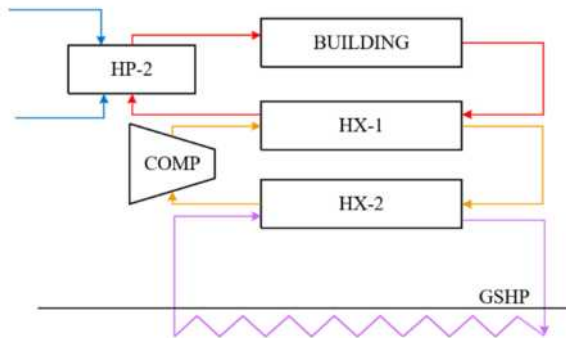


Figure 17: Proposed Third Case of Refrigeration Integration, (red: HVAC loop, orange: primary heat pump loop, purple: GSHP heat pump loop, blue: refrigeration loop).

A quantitative benefit of the configuration is that with a double heat pump system, the amount of heat extracted from the refrigeration cycle can be controlled based on demand or supply. By providing control over a range of independent heat sources, the dependency of the system on a particular heat source reduces. Therefore, the system becomes more independent from factors like weather conditions or electricity price variations. However, the savings achievable with the renewable heat incentive will be lowered, as any heat extracted with this second heat pump is not naturally occurring and is ineligible for the RHI.

7.5 Guidance for Sainsbury's

It is strongly recommended for Sainsbury's to collect detailed GSHP and refrigeration data from existing stores, to increase the confidence of further analysis.

As argued in section 2.2, older buildings have a lower energy efficiency. Not relying on national infrastructure plans and expecting the same efficiency for current stores, it is highly recommended to use the Total site Heat Integration method referenced above in section 6, to utilize all available heat sources and heat sinks.

It is clear that the GSHP-supported store in Kings Lynn performs significantly better compared to the averaged gas boiler stores with the available government schemes. Therefore, a GSHP with refrigeration integration can be an appropriate solution for the energy strategy of Sainsbury's stores. It was found by (Staffell, et al., 2012) that the capital costs for ground source heat pumps are estimated to be in the region of £2,500 – £5,000. Installation costs are estimated to be £500 – £800 per kW of operation. This means for a heat pump of 54.15

kW the overall costs would be assumed to be £48,320 for the worst-case scenario. Coupled with the Boiler Upgrade Scheme, another governmental incentive to support homes and non-domestic buildings (Ofgem, n.d.), a support of £7,500 can be claimed, reducing the costs to £41,320. This amount is comparable to the electricity consumption reduction for Kings Lynn compared to older buildings, as was discussed and shown in section 2.2 figure 7.

Acknowledgements

The authors would like to thank Max Bird for his continued support and guidance throughout the completion of this research. We also extend our thanks to Salvador Acha, Greig Horton and Paul Arrowsmith for technical insight.

References

- Aguilera, J. J. et al., 2022. A review of common faults in large-scale heat pumps. *Renewable and Sustainable Energy Reviews*, Volume 168, p. 112826.
- Aspen Plus, 2019. *Aspen Plus Help*, Bedford: AspenTech.
- Aunedi, M. et al., 2023. System-driven design and integration of low-carbon domestic heating technologies. *Renewable and Sustainable Energy*, 187(9), p. 113695.
- Breembroek, G., 2002. *RETROFITTING WITH HEAT PUMPS IN BUILDINGS*, s.l.: IEA Heat Pump Centre, Sittard (Netherlands).
- Calvillo, C. F. et al., 2023. Technology pathways, efficiency gains and price implications of decarbonising residential heat in the UK. *Energy Strategy Reviews*, Volume 48, p. 101113.
- Dalpane, P., Acha, S. & Nilay, S., 2016. Operational and Economic Analysis of GSHP Coupled with Refrigeration Systems in UK Supermarkets. *ASHRAE*.
- Dhole, V. R. & Linhoff, B., 1993. Total site targets for fuel, co-generation, emissions, and cooling. *Computers & Chemical Engineering*, 17(Supplement 1), pp. S101-S109.
- Ferrovial, n.d. *Shadow Carbon Pricing*. [Online] Available at: <https://www.ferrovial.com/en-gb/sustainability/environment/carbon-footprint/reducing-emissions/shadow-carbon/#:~:text=Shadow%20Carbon%20Pricing%20is%20a,Agreement%20to%20establish%20carbon%20prices> [Accessed November 2023].
- Ge, Y. T. & Tassou, S. A., 2011. Performance evaluation and optimal design of supermarket refrigeration systems with supermarket model "SuperSim". Part II: Model applications. *International Journal of Refrigeration*, 34(2), pp. 540-549.
- Ghilardi, A. et al., 2023. *Pumped Thermal Energy Storage for Multi-Energy Systems Optimization*. Västerås, Sweden, s.n.
- GOV UK, 2022. *Greenhouse gas reporting: conversion factors 2022*. [Online] Available at: <https://www.gov.uk/government/publications/greenhouse-gas-reporting-conversion-factors-2022> [Accessed November 2023].

- Hoseinpoori, P. et al., 2023. Comparing alternative pathways for the future role of the gas grid in low-carbon heating system. *Energy Strategy Reviews*, Volume 49, p. 101142.
- IEA, 2018. *Status of Power System Transformation 2018*, Paris: IEA.
- Li, Z. et al., 2021. Applications and technological challenges for heat recovery, storage and utilisation with latent thermal energy storage. *Applied Energy*, Volume 283, p. 116277.
- Madani, H. & Roccatello, E., 2014. A comprehensive study on the important faults in heat pump system during the warranty period. *International Journal of Refrigeration*, Volume 48, pp. 19-25.
- Maidment, G., 2013. *Ground source heat pumps*, London: The Chartered Institution of Building Services Engineers.
- Met Office, 2023. *Annual Heating Degree Days - Projections (12km)*. [Online] Available at: <https://climatedataportal.metoffice.gov.uk/datasets/TheMetOffice::annual-heating-degree-days-projections-12km/about> [Accessed November 2023].
- Monschauer, Y., Wetzels, D., Laura, C. & Birol, F., 2022. *The future of Heat Pumps*, s.l.: International Energy Agency.
- NERC, 2011. *British Geological Survey: Temperature and Thermal Properties (Basic)*, s.l.: National Environmental Research Council.
- NIST, 2021. *NIST Chemistry WebBook*. [Online] Available at: <https://webbook.nist.gov/chemistry/> [Accessed January 2023].
- Ofgem, 2022. *Non-Domestic Renewable Heat Incentive Guidance Volume 1: Eligibility and how to apply*, London: Ofgem.
- Ofgem, 2022. *Non-Domestic RHI tariff table 2022-23*. [Online] Available at: <https://www.ofgem.gov.uk/publications/non-domestic-rhi-tariff-table-2022-23> [Accessed November 2023].
- Ofgem, n.d. *Boiler Upgrade Scheme (BUS)*. [Online] Available at: <https://www.ofgem.gov.uk/environmental-and-social-schemes/boiler-upgrade-scheme-bus> [Accessed December 2023].
- Ohannessian, R. & Sawalha, S., 2014. *THERMAL ENERGY STORAGE POTENTIAL IN SUPERMARKETS*. London, 3rd IIR International Conference on Sustainability and the Cold Chain.
- ONS, 2022. Age of the property is the biggest single factor in energy efficiency of homes. *Office for National Statistics*, 6 January.
- Radulovic, J., Bull, J. & Buick, J. M., 2023. Chapter 18 Investigation of Working Fluid Performance in a Refrigeration Cycle. In: *Energy and Sustainable Futures: Proceedings of the 3rd ICESF, 2022*. s.l.:Springer Proceedings in Energy.
- Rasmussen, P. C. & Kurz, R., 2009. *Centrifugal Compressor Applications - Upstream and Midstream*. s.l., Proceedings of the Thirty-Eighth Turbomachinery Symposium.
- Reddick, C., Sorin, M., Bonhivers, J.-C. & Laperle, D., 2020. Waste heat and renewable energy integration in buildings. *Energy and Buildings*, Volume 211, p. 109803.
- Sainsbury's, n.d. *Reduce carbon emissions*. [Online] Available at: <https://www.about.sainsburys.co.uk/sustainability/better-for-the-planet/carbon> [Accessed November 2023].
- Sidebotham, G., 2022. *An Inductive Approach to Engineering Thermodynamics*. 1st ed. s.l.:Springer.
- Silverman, D., 2020. Imperial partnered with Sainsbury's on sustainability research. *Imperial College London News*, 12 October.
- Staffell, I., Brett, D., Brandon, N. & Hawkes, A., 2012. A review of domestic heat pumps. *Energy & Environmental Science*, p. 9297.
- Sunak MP, T. R. H. R., 2023. *PM recommit UK to Net Zero by 2050 and pledges a "fairer" path to achieving target to ease the financial burden on British families*, London: Prime Minister's Office, 10 Downing Street.
- Tassou, S. A., Ge, Y., Hadaway, A. & Marriott, D., 2011. Energy consumption and conservation in food retailing. *Applied Thermal Engineering*, 31(2-3), pp. 147-156.
- Tassou, S. & Ge, Y., 2008. Reduction of Refrigeration Energy Consumption and Environmental Impacts in Food Retailing. *Handbook of Water and Energy Management in Food Processing*.
- UK Gov, 2022. *Evidence update of low carbon heating and cooling in non-domestic buildings*, s.l.: s.n.
- UK Gov, n.d. *Non-domestic Renewable Heat Incentive (RHI)*. [Online] Available at: <https://www.gov.uk/non-domestic-renewable-heat-incentive> [Accessed December 2023].
- Xian Li, H. et al., 2023. *An integrated framework of ground source heat pump utilisation for high-performance buildings*, s.l.: s.n.

A Techno-Economic Analysis and Systematic Review of Blue and Green Hydrogen Production Technologies

Achour Tala and Fukushima Naoki

Department of Chemical Engineering, Imperial College London, U.K.

Abstract: To remain on course to meet the 1.5°C climate goal, industries will be required to substantially decrease their carbon emissions. Currently, only 0.7% of hydrogen is produced from low-carbon sources. To achieve these targets, technologies such as blue and green hydrogen must be a part of the decarbonisation strategies. A systematic review of the literature on these technologies has been carried and costing data have been collected for steam methane reforming with and without carbon capture, alkaline electrolyzers (AWE), proton exchange membrane electrolyzers (PEM), and solid oxide electrolyser cells (SOEC). These data are used in a model to harmonise the levelised costs of each technology, and project the costs of AWE and PEM electrolyzers in 2030. The harmonisation outcome highlights the importance of standardisation, and detailed and uniform cost reporting in literature. A sensitivity analysis on the blue hydrogen harmonisation identified the natural gas cost and carbon taxation to be the factors with the greatest overall impact on the LCOH with 21.24% and 7.82% deviation from the base LCOH. A Projection based on the learning rate approach indicate an LCOH reduction of 24% and 36% for AWE and PEM respectively, highlighting the potential of PEM to become an attractive investment option, although challenges associated material could limit cost reductions to \$4.45/kg H₂.

Introduction

In order to mitigate further climate change and remain on course to limit the increase in global temperature to 1.5°C, the Intergovernmental Panel on Climate Change (IPCC) stated that immediate, rapid, and deep reductions in global greenhouse gas (GHG) emissions are required in all sectors this decade, with global net zero CO₂ emissions reached in the early 2050s¹. Given that approximately 58% of net global GHG emissions in 2019 came from the energy sector and industry¹, it is imperative that society decarbonises these sectors by transitioning away from energy derived from fossil fuels.

Hydrogen, produced with minimal GHG emissions or from renewable energy sources, is one possible solution for making this transition, as it has multiple potential uses, especially in the energy, industrial and transportation sectors. In 2022, almost 95Mt of hydrogen was produced worldwide, however low emission production only accounted for 0.7% of the total, with the rest produced from natural gas without carbon capture, utilisation and storage, unabated coal, and naphtha reforming². Production via steam methane reforming typically emits around 9kg of CO₂ per kg of H₂ produced² and production via coal gasification emits around 20kg of CO₂ per kg of H₂ produced⁵⁰, therefore supply through low carbon emissions technology is required for major decarbonisation. Blue and green hydrogen have been widely investigated as potential solutions. Blue hydrogen consists of retrofitting grey hydrogen technologies, predominantly steam methane reforming (SMR), with carbon capture and storage (SMR+CCS). It is known as ‘low-carbon hydrogen’. Green hydrogen on the other hand, uses water electrolysis systems powered by renewable energy to generate an electrochemical reaction splitting water molecules into oxygen and hydrogen, and therefore producing zero GHG emissions. Typical technologies consist of Alkaline Water Electrolyzers (AWE), Proton Exchange Membranes (PEM), and Solid Oxide Electrolyser Cells (SOEC)³.

This study systematically reviews the existing academic literature on green and blue hydrogen production technologies to understand their present status and associated challenges and identify areas of development for large-scale deployment of low emissions hydrogen. It intends to provide useful insights that will direct future research towards addressing these limitations. The research places a particular focus on costing information to assess the competitiveness of low-carbon hydrogen production. Current literature reviews predominantly focus on technical aspects with little critical analysis of literature costs^{27, 44, 45, 101}. Consequently, an economic model aimed at reporting and harmonising the literature data is developed, and the projection of each technology's future costs using predicted input parameters is produced. For the first time in this subject area to the authors' knowledge, levelised costs of hydrogen production (LCOH) are harmonised to facilitate direct comparisons across different papers. This aims to address the oversight in current literature where the LCOHs are compared across papers without considering the impact of location-specific factors such as electricity and natural gas prices, capacity factor, and carbon tax, all of which have significant contributions to the LCOH.

Methods

i. Systematic Review

A systematic research strategy was created and implemented to identify the relevant cost-focused literature on hydrogen production. Search terms were produced and refined through multiple iterations, with each set targeting a specific aspect of the hydrogen industry. These topics include green hydrogen production, blue hydrogen production, hydrogen storage and transport, hydrogen applications in industry and hydrogen for power generation. The subsequent step involved screening the abstracts using a questionnaire and categorising them based on their relevance as follow:

- **High Relevance:** The abstract provided direct information about the technology and economics of (topic)
- **Medium Relevance:** The abstract provided indirect information about the technology and economics of (topic)
- **Low Relevance:** The abstract provides supporting material that might be utilised to contextualise (topic)

Papers categorised as highly relevant were reviewed in depth, and summary paragraphs about the technologies and costs were produced for each study using a newly developed and more focused questionnaire. Additionally, references within the papers and grey literature were consulted to validate assumptions and provide techno-economic background. The literature data for each technology type was then reported and analysed using the economic model below. The search terms and abstract screening questionnaire are provided in the supplementary information SI 1. This serves as a comprehensive resource detailing the methodology of the systematic review.

ii. Economic Analysis

The papers reviewed utilised the Levelized Cost of Hydrogen (LCOH) or Net Present Value (NPV) as economic metrics for hydrogen production. Nonetheless, the majority of the literature predominantly utilised the LCOH, with only a few articles using NPV. Accordingly, LCOH was selected as an economic metric to comprehensively capture the costs associated with each technology across their lifetime. This avoids uncertainties associated with hydrogen selling prices and potential co-produced product.

Cost Escalation

An escalation model based on the Chemical Engineering Plant Cost Index (CEPCI)⁴ was developed to compare historically calculated levelized costs. The LCOH values were converted into USD using the annualised mean exchange rates sourced from the International Monetary Fund⁵. This conversion was applied from the initial currency used for the year of costing presented or the year of publication of the study. The values are subsequently escalated from the base year to the comparison year, selected as 2022, using equation (1):

$$\text{Cost (2022)} = \text{Cost (base year)} \times \frac{\text{CEPCI(2022)}}{\text{CEPCI(base year)}} \quad (1)$$

Harmonisation

The harmonisation of LCOH values across different technologies aims to standardise the costs under uniform techno-economic conditions. In the analysis, intrinsic values, such as specific capital costs (CAPEX), operational costs (OPEX), and efficiency, remain identical to the values reported in their

respective papers, while extrinsic parameters are modified and made consistent across all papers for each technology and energy source. The harmonised parameters include the discount rate, cost of electricity or natural gas, and the operating hours.

Harmonisation: Blue Hydrogen

The method, initially presented by Hazrat et al.²⁷, has been adapted and applied to evaluate the cost of blue hydrogen production, as shown in equation (2). The LCOH is defined as the total lifetime cost normalised by the total hydrogen production²⁸. Expenditures encompass the total capital cost $CAPEX$ and the annual operating costs $OPEX_{annual}$, divided into variable and fixed costs.

$$LCOH = \frac{CAPEX \times CRF + OPEX_{annual}}{m_{H_2} \times 8760 \times CF} \quad (2)$$

Here, m_{H_2} represents the hourly hydrogen production rate and CF denotes the capacity factor, indicating the percentage of operating hours in a year.

The capital cost is annualized using the capital recovery factor (CRF), calculated by the plant lifetime n and the discount rate i ^{28, 29, 30}:

$$CRF = \frac{i(1+i)^n}{(1+i)^n - 1} \quad (3)$$

The fixed operating cost is determined by the annual operating and maintenance (O&M) costs, while the variable cost is defined as the electricity, natural gas costs and carbon tax costs.

$$OPEX_{annual} = OPEX_{fixed} + OPEX_{variable} \quad (4)$$

$$OPEX_{variable} = (\dot{E}_{CO_2} \times Tax_{CO_2} + \dot{E}_{el} \times C_{el} + \dot{E}_{NG} \times C_{NG}) \times 8760 \times CRF \quad (5)$$

\dot{E}_{CO_2} denotes the rate of carbon emissions in ton/hr, \dot{E}_{el} and \dot{E}_{NG} the rate of electricity and natural gas consumption kWh/hr and kg/hr, and Tax_{CO_2} , C_{el} , and C_{NG} represent their respective costs in \$/ton CO₂, \$/kWh, and \$/kg.

Harmonisation: Green Hydrogen

Similarly, the method presented by Scheepers et al.²⁸, has been modified to calculate the cost of green hydrogen production through water electrolysis. Unlike blue hydrogen, the costs utilised are power specific. The plant's power-specific capital cost includes the initial investment cost, $CAPEX_{inv}$, of the stack, and balance-of-plant (BOP) associated with its procurement and installation, as well as the total stack replacement costs, REPEX, over its lifetime.

$$CAPEX = CAPEX_{inv} + REPEX \quad (6)$$

The replacement cost is determined by the number of replacements required throughout the plant's lifetime, and the stack unit cost, $CAPEX_{stack}$, as defined by

Equation (7), with LT_{el} denoting the electrolyser lifetime in hours.

$$REPEX = CAPEX_{stack} \times \frac{n \times 8760 \times CF}{LT_{el}} \quad (7)$$

The operating cost $OPEX_{annual}$ is determined by the power-specific O&M costs, while the variable costs is restricted to electricity only. Notably, water costs are excluded from the LCOH for two reasons: several papers lack clear information about water costs, and literature indicates that water typically accounts for only 1-2% of the LCOH^{31, 32, 33} making its impact negligible on the final result. Combining all the previously defined parameters, the LCOH is ultimately calculated using Equation (8), incorporating the efficiency of the plant efficiency of the plant η in percentage, cost of electricity C_{el} in \$/kWh, and lower heating value of hydrogen LHV_{H_2} in kWh/kg.

$$LCOH = \frac{LHV_{H_2}}{\eta} \left(\frac{CAPEX \times CRF + OPEX_{annual}}{8760 \times CF} + C_{el} \right) \quad (8)$$

Cost Projection

The learning curve approach is employed to assess the projection of the electrolyser capital expenditure. This technique outlines the process of learning by doing, which indicates innovation by production that is sparked by the competition between companies⁷. The learning curve refers to the reduction in production cost as a result of accumulated knowledge. It quantifies the relationship between the CAPEX of a technology and the cumulative capacity as follow³⁴:

$$CAPEX(t_2) = CAPEX(t_1) \times \left(\frac{P(t_2)}{P(t_1)} \right)^{-b} \quad (9)$$

$CAPEX(t_1)$ and $CAPEX(t_2)$ represent the capital cost in year 1 and 2 respectively, $P(t_1)$ and $P(t_2)$ the cumulative capacity of the electrolyser in year 1 and 2 respectively, and b , the learning parameter.

The learning parameter can be derived from the learning rate, which denotes the proportion of cost reduction per electrolyser unit for every doubling of capacity.

$$LR = 1 - 2^{-b} \quad (10)$$

Using equation (9), the projected CAPEX values are assessed and integrated into the LCOH equations.

Results & Discussion

I. Demographics Of The Systematic Review

The papers were initially evaluated and categorised into the primary focus topics. In the systematic review, 407 papers were evaluated, with 103 classified as highly relevant. Among these, 38 papers focused on green hydrogen, 30 on blue hydrogen, while 35 papers covered the remaining topics. Due to time limitations, this study's scope was narrowed down to blue and green hydrogen exclusively. Out of these papers, 30 were not particular to a single technology, 20 of which originated from the green

hydrogen search terms. In this subset, 16 papers discussed AWE, 14 focused on PEM, 12 on SOEC, and 1 on Anion Exchange Membrane (AEM). The uneven focus observed can be attributed to the market deployment status of electrolyser technologies. AWE, being the most commercially established production method⁷, is more prevalent in the literature, while less commercialised technologies currently lack available data. For instance, AEM, still in its early development stages⁷, exhibits minimal coverage.

Conversely, 33 papers exclusively focused on a single technology: with 20 on SMR with or without CCS, 1 on AWE, 6 on PEM, 4 on SOEC, and 3 on other innovative developments. The abundance of papers on SMR and SMR+CCS can be attributed to the technological maturity of SMR and the potential of CCS to serve as a short-term solution on already existing plants. For green hydrogen, the shift in focus can be recognised by the technical maturity. Less mature technologies such as PEM and SOEC⁷, attract larger research attention investigating their techno-economic potential, thereby contributing to their advancement on the Technology Readiness Level (TRL) scale. In contrast, AWE which has reached its final stage of technological maturity, demands less R&D for improvement. Lastly, four papers discussed other topics such as cost projections. SI 5 compiles all 407 screened papers, along with short summaries of their contents and classification.

II. Technology Overview & Analysis

As each technology has its own characteristics affecting the LCOH, it is imperative to consider each in turn to evaluate their economic feasibility. In this section, only the main technologies are evaluated.

Blue Hydrogen

In 2021, approximately 62% of hydrogen was produced by SMR without the use of carbon capture, utilisation, and storage². If all the existing plants are retrofitted with carbon capture technologies, it could lead to the capture of 710-880 Mt per year of CO₂², therefore it is seen as an interim solution until the green production technologies develop in scale and efficiency. It also offers a sustainable prospect to fossil fuel producers such as Canada, Iran, Qatar, Norway, the Russian Federation, and the United States⁶.

In the SMR process, natural gas undergoes initial pre-treatment to remove any sulphur and chlorine to prevent any catalyst poisoning downstream. In a pre-reformer, any C₂+ hydrocarbons or olefins are converted into methane as well as CO₂, CO and H₂. This is fed with steam into a reformer to produce synthesis gas (syngas), a mixture of CO₂, CO, H₂ and residual CH₄, which is subsequently fed into a shift reactor. This converts the CO and H₂O into H₂ and CO₂, which is fed into the pressure swing adsorption (PSA) unit, recovering around 85-90% of H₂ at a

purity greater than 99.9%. Therefore, CO₂ can be captured at three possible locations in the process: 1) the reformer flue gas, 2) the shifted syngas, and 3) the PSA tail gas³⁵. For merchant plants, where hydrogen production is not integrated with the production of ammonia or methanol, capturing the CO₂ from the syngas stream can lead to an emissions reduction of up to 60%, at a cost of \$53 per tonne of CO₂ (/tCO₂). Emissions reductions can reach up to 90% if capture also occurs at the reformer flue gas, however this increases the capture cost to \$80/tCO₂⁶. The TRL of SMR+CCS is currently an 8³⁶.

Levelised costs of hydrogen production were generally in the range of \$1.5-\$5.0/kg H₂, with multiple papers having higher values, with the maximum cost found being \$8.88/kg H₂³⁷. However, these outliers are associated with small scale hydrogen production plants, as economies of scale play a large part in the LCOH³⁸. As with SMR plants with unabated emissions, the most significant factor in the LCOH is the natural gas price with Argyris et al.³⁹ reporting that fuel costs accounted for 50-60% of the LCOH. This agrees with grey literature, where the IEA states that 45-75% of the LCOH is due to natural gas costs⁶. Given the sensitivity of the natural gas price to geopolitical tensions, as observed after the beginning of the Russian conflict with Ukraine, future conflicts may also cause fluctuations in gas prices and therefore LCOH for SMR+CCS plants.

The other significant factor effecting the LCOH was found to be the cost of CCUS technologies. CO₂ capture and compression require significant amounts of thermal and electrical energy and thus auxiliary utility systems may need to be employed to meet the demand, incurring high costs³⁷. Large investment and research costs will also be required to construct pipelines and storage locations as large-scale infrastructure is required for transport and storage⁴⁰ and currently does not exist. Sub-surface rock formations are possible storage locations, however, it must be ensured that the stored CO₂ does not leak back into the atmosphere or oceans⁴¹. Although CCUS has been used in oil production for enhanced oil recovery, currently only 2 commercial scale hydrogen production plants operate with CCS due to its technological immaturity⁴². Currently, the most mature and commercially available capture technology is absorption, which produces a CO₂ stream with purities greater than 95%, however have the disadvantages of requiring large equipment, energy intensive absorbent regeneration, and corrosion of equipment if using amine-based solvents³⁷. Other capture technologies include membranes, cryogenic separation, and adsorption; however, they are not commercially available and require further development before they can compete with absorption³⁷.

Sorption-enhanced SMR is a novel process which has the advantages of a low reforming temperature, the lack of the need for multiple shift reactors and subsequent purification steps and producing high purity CO₂ streams which can be captured without the need for further processing such as absorption³⁶. However, to supply the high calcination heat to regenerate the CaO sorbent without emitting CO₂ emissions requires energy intensive processes such as oxy-fuel combustion or an indirectly heated calciner. A feasibility study conducted by Yan et al.³⁶ found LCOHs for the process to range from \$3.08-\$4.46/kg H₂, which is fairly competitive to the SMR+CCS process, however, given a low TRL of 4, no economic assessments have been carried out to investigate if it is viable at large scale.

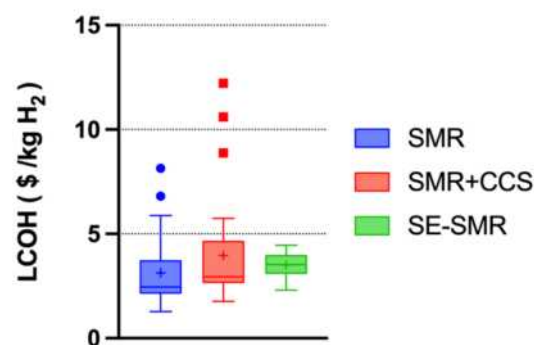


Figure 1. Cost ranges for grey and blue technologies

Costs found in grey literature generally agree with those from academic literature, however, tend to be on the lower end of the range at \$1.64-3.14/kg H₂⁷. The majority of the costs were in the range of \$3.00-5.00/kg H₂ and are displayed in Figure 1.

Green Hydrogen

Alkaline Electrolyser

Alkaline water electrolyzers (AWE) have been used since 1920 and therefore are the most mature technology for water-splitting, accounting for around 70% of the total market share⁴³. They typically operate at temperatures of 60-80°C and pressures of 5-30 bar³³. The advantages include high durability, large scale operation, and low cost due to its inexpensive materials, unlike the noble metals required in proton exchange membrane electrolyzers³³. However, they operate at low efficiencies (58-70% LHV)⁴⁴, current densities (0.2-0.6A/cm²)⁴⁵ and partial load range (40%-100%)⁴⁶. In addition, due to the low operating pressures, additional compression is required to increase the product hydrogen pressure to the required levels for current transportation, incurring high costs. One option for operating an AWE is to supply the electricity through the grid. However, depending on the proportion of fossil fuels used to generate the electricity, it can have higher emission rates than SMR or roughly equal rates to hydrogen production using coal at 24kg eCO₂/kg H₂². Therefore, for low emissions electrolytic production, electricity from green sources must be utilised. The LCOH for alkaline electrolyzers utilising grid electricity varied between

\$4.73-\$14.30/kg H₂. The large range can be attributed to the varying electricity costs in different regions. For instance, Khatiwada et al.⁴⁷ uses a levelised electricity cost of \$0.121/kWh whilst⁴⁸ utilised \$0.014/kWh, a difference of almost an order of magnitude. Given that the cost of electricity accounts for 47%-78% of the LCOH⁴⁴, a large disparity is expected.

The main challenge of alkaline electrolyzers is their low partial load range, meaning that they are unable to operate below a certain load factor. This is particularly important if the electricity is supplied from wind turbines or photovoltaic cells (PV), due to their intermittent nature meaning that the electrolyser will have a low-capacity factor, and therefore produce little hydrogen. For example, Pagani et al. reported a capacity factor of 42-43% for an offshore wind farm and only 13% for an onshore PV farm. To address this challenge, the system can either be connected to the grid to supply electricity when the renewable source is not generating power, or to a battery, which can store excess electricity generated by the turbines or PV cells. Superchi et al.⁴⁹ found that by adding a battery to the electrolyser system, the capacity factor could be increased by up to 10%. In the same paper, it was found that utilising multiple lower capacity modules in series, in this case four 1MW modules, led to a higher capacity factor than a system with one module of a 4MW capacity, at 64% and 62% respectively.

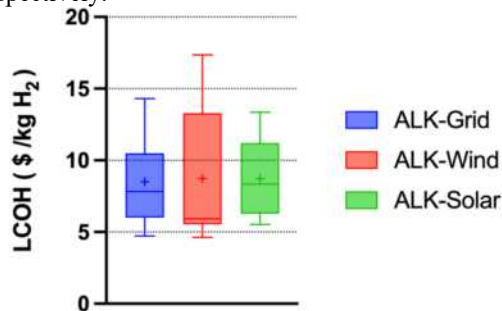


Figure 2. Cost ranges for AWE by energy source

The ranges of LCOH found in literature for onshore wind, offshore wind and PV powered AWE are \$4.63-\$7.25/kg H₂, \$11.85-\$17.36/kg H₂, and \$8.50-\$13.36/kg H₂ respectively. Although offshore wind farms have the highest capacity factor amongst the three sources, at around 45% compared to 27% and 13% for onshore wind and PV respectively³², the larger installation costs, leads to higher LCOHs. Moreover, solar-powered electrolyzers can incur large costs due to their low-capacity factor requiring a high number of cells to meet the energy demands. Conversely, if the electrolyser is connected to the grid, the PV+grid pairing results in a lower LCOH range of \$5.53-\$9.51/kg H₂ due to the increased hydrogen production offsetting the PV capital costs. Onshore wind has the lowest capital expenditures and reasonable capacity factors, resulting in the lowest LCOH range of the three sources.

Proton Exchange Membrane

PEM electrolyser systems were first introduced in the 1960s by General Electric. They use compact membrane electrode assembly with solid polymers as both electrolytes and membranes, making them suitable for urban use⁶, and utilise pure water, avoiding the recovery of alkaline solutions⁵⁰. They can rapidly ramp up to 160% of design capacity⁶, which is ideal for integration with intermittent energy^{33, 43}. Despite being more efficient and producing purer hydrogen than AWE, PEM systems face challenges with the oxidative conditions created by the PFSA membrane^{7, 29} which reduces their lifespan, requiring expensive and robust materials like iridium^{6, 43}. Nonetheless, PEM electrolysis is approaching its final stage of technical maturity and is gaining market share, as demonstrated by the increase in PEM installations^{6, 33}.

Hydrogen production using grid electricity found a range between \$6.53-\$16.33/kg H₂, largely influenced by the varying electricity prices. This mirrors the sensitivity of AWE despite PEM's higher efficiency.

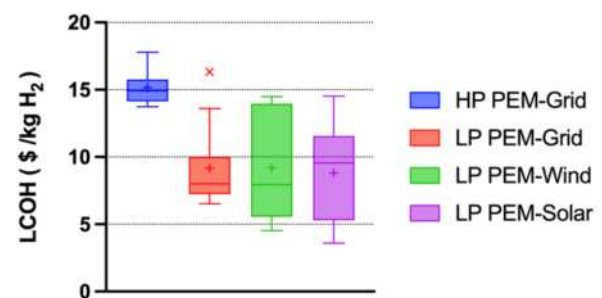


Figure 3. Cost ranges for PEM by energy source

PEM's load flexibility allows it to utilise better the intermittent electricity generated by wind turbines or PV in comparison to AWE. The range of costs found from literature for PV powered production is \$3.59-\$14.53/kg H₂, whilst for offshore and onshore wind powered production, costs range from \$4.53-\$14.49/kg H₂ and \$7.39-\$14.42/kg H₂ respectively. These large ranges can be associated with the electricity cost and capacity factors varying substantially across the literature. Examples of electricity costs identified in the literature include \$0.069/kWh⁷⁷-\$0.136/kWh³² for onshore wind.

Moreover, hydrogen storage and end-use application often require pressures up to 700 bar. While mechanical compression is common, interest in high-pressure PEM electrolyzers is growing due to their simple system configuration^{8, 17, 29, 52} as exemplified by IFE's research⁸ and Honda's 700 bar smart hydrogen station¹⁷. Self-pressurised electrolyzers, despite exhibiting higher energy consumption, offer cost advantages by eliminating mechanical compressor expenses, with optimal operation between 30 to 70 bar⁷. They present a competitive LCOH of \$13.42-\$17.79/kg H₂, compared to \$14.04-\$15.85/kg

H₂ conventionally. Their cost-efficiency nonetheless remain highly dependent on the electricity costs. Research highlight challenges associated with material degradation and gas crossover, which exceeds the lower flammability limit of 4 vol% for H₂ in O₂^{17, 18, 19, 52}, resulting in safety risks and barriers to upscaling. This emphasises the need for mitigating measures such as reinforced membranes or recombining catalysts⁵².

Solid Oxide Electrolyser Cells

Solid Oxide Electrolyser Cells (SOECs) are amongst the least developed technologies for hydrogen production, using electricity and heat⁵³. They are currently in early-stage laboratory development⁶ and are anticipated to reach market maturity within the next decade²⁹. Companies, like Sunfire, are already selling small-scale units³³, with demonstrations reaching up to 1 MW⁷ SOECs use ceramic membranes of yttria-stabilized zirconia (YSZ), enabling high-temperature operation (600–1000 °C), high electrical efficiency³³, and advantageous features such as high current density and reversible operation. As a result, they can provide grid-balancing services alongside hydrogen storage facilities. Co-electrolysis of carbon dioxide and steam is another application of SOECs⁶. However, challenges with material stability due to high operating temperatures often lead to rapid component wear and a shorter lifespan⁵⁴. A recommended strategy presented by Zhang et al.⁵⁵ is hot stand-by mode. This method maintains the SOEC stack temperature above 600°C, enabling fast load variation in response to heat or power availability⁵⁵. Recent developments in proton conducting SOECs could address these challenges by operating at lower temperatures of around 300°C⁵⁴.

SOECs require significant energy to generate and maintain the high temperatures required. This can be supplied from the waste heat of industrial processes, solar energy, nuclear energy, or geothermal systems⁶. The literature primarily explored the potential of waste heat and solar energy, as well as electric heaters.

Integrating waste heat into the electrolysis system is reported by the literature to achieve costs ranging from \$5.90–7.16/kg H₂. This cost-effective technique significantly enhances the efficiency by utilising readily available heat with no additional charge. In contrast, using fossil fuels or grid-connected electric heaters leads to higher LCOH values ranging from \$7.30–10.20/kg H₂. The disparity in the literature values can be attributed to the electricity cost, which doubles across the range. Yet, the enhanced efficiency of WH-SOEC makes it less sensitive to electricity prices.

Alternatively, a key area of focus in the literature is solar energy, notably photovoltaic (PV) and parabolic trough collector (PTC) systems. PTC-powered SOEC typically yields a larger LCOH due to its significant capital investment. Costs are reported at \$4.30/kg H₂³⁰

and \$11.70/kg H₂²⁰ in the UAE and Spain respectively, reflecting differences in regional solar potential. Although PTC-powered SOECs have higher upfront costs, they benefit from high efficiency. Conversely, PV-powered electrolyzers in Spain have a lower efficiency but offer lower capital costs and a decreased LCOH of \$11.49/kg H₂. Consequently, Lin et al.²⁰ examines a hybrid system that combines both PV and PTC, achieving a reduced LCOH of \$9.00/kg H₂. Seitz et al.²¹ discusses the integration of Thermal Energy Storage (TES) to PTC-powered systems, which can increase the production of hydrogen by 50%, and reduce LCOH by 34% with an 11-hour discharge. Zhang et al.⁵⁵ advances this concept further by analysing a hybrid system coupled with TES and batteries, achieving a competitive LCOH of \$5.50/kg H₂. However, constraints associated with the high cost of TES systems makes large scale integration impractical, requiring operation at minimum load⁵⁵.

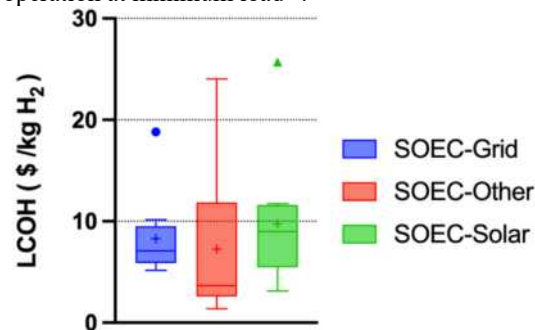


Figure 4. Cost ranges for SOEC by energy source

Nuclear power, for both electricity and heat, could reduce hydrogen production costs to \$1.40–3.00/kg H₂ for SOEC, presenting a competitive option. However, there is limited cost data in the literature.

The significant deviations in SOEC outliers relate to SOEC's greater investment costs range, which is nearly twice as large as the ranges for AWE and PEM electrolyzers⁵⁶.

Overall, the data in grey literature is relatively sparse and typically provides LCOH values for electrolyzers powered by different energy sources, with no distinction between technologies. Nonetheless, green cost data is more readily available compared to blue hydrogen. Grey literature data fits well in the middle of academic data excluding outliers and pressurized systems, reporting renewable hydrogen costs in the range of \$3.62–12/kg H₂^{2,6,7,9}. Endpoints of academic data, around \$1.00–3.00/kg H₂ and \$20.00–25.00/kg H₂, are mostly associated with SOEC systems. However, as it is not yet commercialised, grey literature data mostly represents AWE and PEM technologies.

Costs Synthesis

The LCOH data for each technology are presented in Figure 5. The trends observed reveal average costs of \$8.70/kg H₂, \$10.20/kg H₂, \$7.90/kg H₂, \$3.10/kg H₂,

\$4.00/kg H₂, and \$7.00/kg H₂ for AWE, PEM, SOEC, SMR, SMR+CCS, and SE-SMR, respectively.

Consistent with the previous discussion, SMR and its derivatives are identified as the most cost-effective, while PEM lies on the higher end. SMR and SMR+CCS display low standard deviations of \$1.7/kg H₂ and \$2.6/kg H₂ as opposed to up to \$5.80/kg H₂ for SOEC, emphasizing their cost stability and reliability.

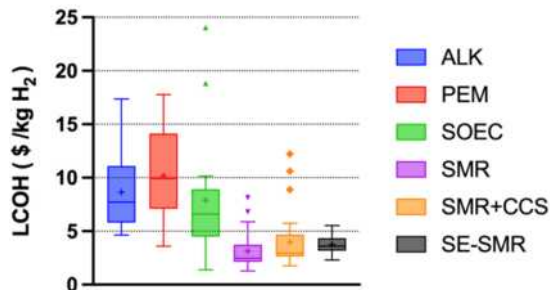


Figure 5. Cost ranges for all technologies

The green hydrogen cost ranges, including outliers, correlate inversely with technology maturity: AWE, PEM, then SOEC. As indicated, PEM's larger levelised costs are associated with its material. SOEC has the smallest interquartile range (IQR), indicating a more consistent cost structure. The lower end of its range aligns closely with the costs of SMR, suggesting potential competitiveness due to its high efficiency, making it a promising technology for future market adoption. Nonetheless, SOEC will need to compete with the maturity and scalability of more established technologies. Furthermore, all technologies, except for PEM, exhibit a right-skewed distribution where the average is almost twice the median, due to the small group of outliers discussed in previous sections.

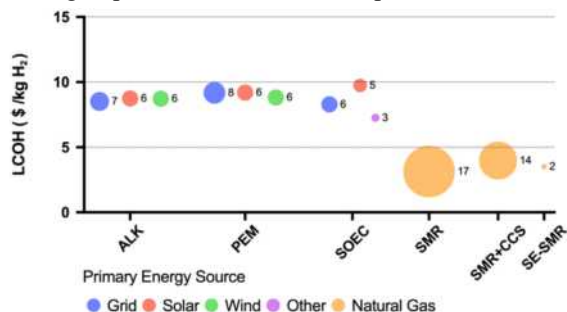


Figure 6. Mean costs of technology, with bubble size indicating the number of references reporting costs for each technology

The cost data were relatively abundant across all technologies, however, varies by energy source. SMR displays the largest number of sources per category, reflecting its prevalence in the hydrogen production market. SMR+CCS has slightly fewer sources, attributed to the complexities associated with CCS integration. The sources for green hydrogen data show a significant number of grid powered electrolyzers, closely followed by renewable sources, highlighting the growing focus on renewable energy integration, due to electricity cost reductions. Particularly for SOEC, the preference leans towards solar energy,

with 5 papers, compared to 3 for other energy sources, likely due to its potential to enhance electrolyser efficiency, especially with PTC solar farms.

III. Harmonisation

The levelised costs of hydrogen production display wide ranges, primarily due to differences in the assumptions associated with the discount rates, energy costs, carbon tax, and capacity factors. These variables are influenced by the region of operation and its economic state, resulting in variations between countries and within a single nation. For instance, the cost of electricity generated by PV panels can range between \$10/MWh¹⁰ in Saudi Arabia to \$42/MWh¹¹ in Finland. These yield differing LCOH values, a sensitivity underscored in the analysis of several papers^{33,34,44,52}. Nonetheless, the fundamental concept of the levelised cost remains consistent across the papers evaluated, with minor variations in its definition. This enables the capturing of the economic performance of green hydrogen on a global scale. However, to effectively compare the costs associated with the present technologies, it is necessary to harmonise extrinsic values, unrelated to the technologies themselves.

Blue Hydrogen

Harmonised LCOH

The harmonised variables were chosen based on the modal values found in the literature and are summarised in Table SI 3.4. The systematic review for blue hydrogen only returned seven papers containing all the relevant cost data for the harmonisation. A significant number of studies consisted of reviews which report referenced LCOHs values and hence do not provide any costing information^{41, 42, 45, 47, 50, 57}. Other papers omitted certain cost parameters such as CAPEX, OPEX or carbon price, preventing the harmonisation being carried out^{40, 58, 59, 60}.

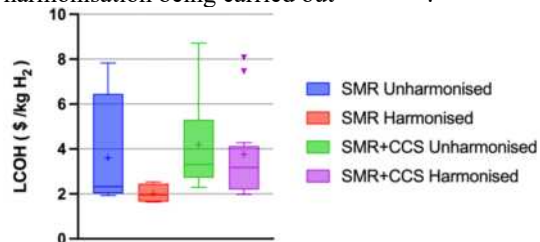


Figure 7. Cost ranges for unharmonized and harmonised cases for grey and blue technologies

Figure 1 clearly demonstrates the effects of the harmonisation. For the grey technologies, an 82.0% decrease in the interquartile range (IQR) of the LCOHs was observed, highlighting the maturity of the SMR technology, as this suggests that the intrinsic costs are similar across all papers. For the blue technologies, a smaller decrease of 24.8% for the IQR was observed, suggesting a variation of the intrinsic costs across the papers. This is partly because the data points are not separated by capture rate, which positively correlate with the LCOH due to the higher CCS costs. Additionally, the data points are not

separated by technology type (SMR+CCS or SE-SMR), hence the spread in intrinsic costs is expected.

In addition, a sensitivity analysis on the harmonisation variables was carried out to gauge their individual influence while keeping all other parameters as reported in their respective sources.

As expected, the natural gas price displayed the largest difference in IQR to the unharmonized case at 21.42%, which reflects its significant contribution to the LCOH, at 45%-75%⁶. This highlights the LCOH's sensitivity to natural gas prices, a consensus supported by the literature. The carbon tax price harmonisation yielded a moderate difference of 7.82%, suggesting a lower LCOH sensitivity. However, this suggest that for green technologies to become more competitive with grey and blue hydrogen, radical increases in carbon taxation are required to incentivise the transition. This aligns with the findings presented by George et al.⁵⁷. The other parameters showed only marginal differences; however, this should not be interpreted as a general insensitivity of LCOH to the parameters. For the discount rate, capacity factor, and plant lifetime, the assumed values in each paper were either identical or closely matched the harmonised value, explaining the minimal differences observed. As for the electricity cost, the marginal difference relative to the base case can be attributed to the negligible consumption of electricity in comparison to that of natural gas⁵⁹.

These results emphasise the significance and necessity for a harmonisation to enable a representative comparison of the LCOHs across different studies.

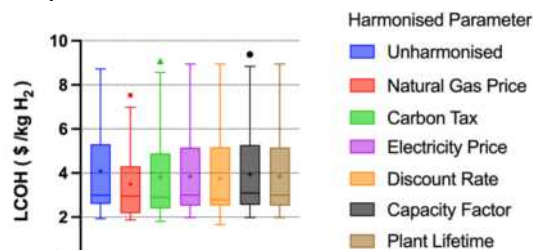


Figure 8. Sensitivity analysis of harmonised parameters on the LCOH

Green Hydrogen

Two issues arise when analysing the data for the green hydrogen harmonisation: on-grid systems and the scarcity of comprehensive data points. Figure 9 illustrates between one and four harmonisable data points within each category, with the bulk found in grid energy sources and other combinations. More specifically, the focus on renewable hydrogen production restricts it to only one data point.

The predominance of on-grid electrolyzers can be associated with the broad search terms, which did not limit green hydrogen strictly to renewable energy sources, and the limitations inherent to the search algorithm, which excludes data based on the title, abstract, and keywords. This results in the grid

hydrogen discussions within the bulk of studies to bypass the filters.

Moreover, further limitations in the harmonisation are linked to the data presented in the papers. While some papers provide comprehensive cost data, they often overlook intrinsic parameters such as system efficiency^{32, 53, 55, 57} or lack information about the lifespan of the electrolyser, plant, and discount rate. Although the latter are generally consistent across the literature and can be reasonably assumed, papers with missing intrinsic data impede this process^{53, 57}. Additional challenges arise when the articles present total costs that cannot be converted into specific costs, rendering them impractical for the selected equation in the methodology^{29, 52, 59, 61}. Alternatively, some papers include energy plant expenditures that cannot be distinctly separated into electrolyser and power plant costs⁶². Finally, several papers are industry reviews with no LCOH calculations or projections and provide minimal data, making them unsuitable for the harmonisation^{38, 50, 63, 44, 65, 66}.

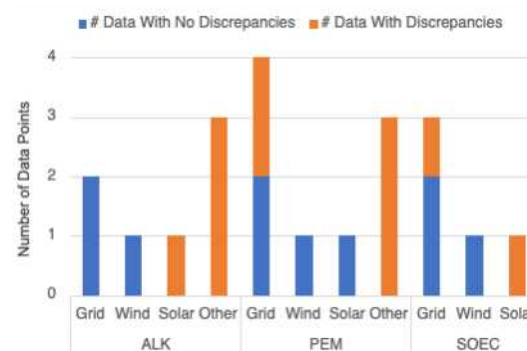


Figure 9. Number of data points with or without discrepancies for each technology and energy source

Furthermore, 52% of the harmonisable data points exhibit discrepancies that could impact the reliability of the collected data. For instance, uncertainties arise from the efficiencies reported by Scheepers et al.²⁸, where its conversion to electrolyser capacity using the hydrogen production rate deviate from the study's reported values. Specifically, capacities of 3.5 MW and 119 MW are obtained instead of 1 MW and 100 MW, respectively. In contrast, applying the same conversion process to the studies by Srettiwat⁴⁶ and Jang et al.³³ yields identical values. As the capacity factor and hydrogen production are set values, this indicates a discrepancy in remaining variables: the efficiency or the quoted capacity. However, the calculation of the LCOH using the study's respective equation and efficiency yields cost of €5.05/kg and €4.88/kg, a 3% deviation from the literature's values of €5.21/kg and €5.04/kg. This suggests that the efficiencies reported are likely true, whereas the electrolyser capacities may be less precise. The discrepancy could be attributed to the use of multiple electrolyser stacks at the reported capacities, however, there are no clear indications in this regard.

In view of these results, the harmonisation of the green hydrogen costs was not pursued. Nonetheless, SI 3 presents all the calculations and data collected from the literature, along with a table detailing the limitations and discrepancies identified in each study.

IV. Cost Projection

The previous sections evaluate the influence of non-intrinsic data on the LCOH, notably the energy costs. However, the literature indicates the system capital cost as the second most significant parameter. Thus, its reductions will play a crucial role in the decarbonisation of the hydrogen production industry. Consequently, a model was developed to investigate the influence of capital cost reduction on the LCOH in 2030. The model makes several assumptions; therefore, its purpose is to offer a qualitative estimate on the market competition based on the anticipated deployment of electrolyzers and influence of capital cost, rather than an accurate cost forecast.

Parametrisation

The hydrogen council's 2023¹² data reveal global electrolyser cumulative capacity of 300 MW by the end of 2020, expected to increase to 232 GW by 2030. However, the report lacks precise data and the plot is not drawn to scale for the 2024-2029 period, making any findings drawn from the global trends highly inaccurate. In the literature, papers often create their own cumulative capacities roadmap based on regional hydrogen ambitions⁶⁵ or country-specific data⁵⁰. Therefore, to assess the global state, the projection is limited to the base and final years. The distribution of the technologies has shown consistency in recent years, with AWEs dominating the market at 60%, followed by PEM at 30% and SOEC at less than 1%². However, this balance is anticipated to change with PEM capturing a larger market share. As several projects have not disclosed their technologies, a precise estimation is unavailable⁷, and the current technology distribution is assumed to remain constant. Consequently, cumulative capacities for AWE are estimated at 0.18 GW in 2020 and 139.2 GW in 2030, while PEM capacities are 0.09 GW and 69.6 GW in 2020 and 2030, respectively.

As blue hydrogen utilises established technologies, their cost is only expected to decrease minimally¹³, a reduction which may also be offset by increased carbon tax in the future. Therefore, the blue hydrogen cost is maintained constant at \$4.00/kg H₂.

Table 1. Learning rates of electrolyzers according to references

Reference	Learning Rate	Technology	Timeframe
[20]	9 %	Alkaline	2020-2030
[20]	13 %	PEM	2020-2030
[22]	18+-6 %	Alkaline	1956-2014
[7]	18+-13 %	Electrolysers	1972-2004
[23]	8 %	Alkaline	-
[24]	28 %	SOEC	1996 - 2008
[25]	15-22%	Electrolysers	-

Numerous studies have examined the learning curve's impact on the electrolyser unit cost, estimating the learning rate over a time span extending from 1956 to 2030. The learning rates of 9% and 13% for AWE and PEM respectively are selected as they provide the most up to date data available and align with the timeframe.

Moreover, Table SI 4.1 displays the techno-economic parameters selected for PEM and AWE, in the base and projected year. Parameters, including CAPEX costs and system efficiency, are drawn from IRENA⁷ as \$750/kW and \$1050/kW, as well as 64 kWh/kg and 66.5 kWh/kg for AWE and PEM electrolyzers, respectively. Techno-economic data for the Solar photovoltaic panel is sourced from the IEA², indicating electricity costs at \$0.071/kWh and a 23.5% capacity factor. Energy Education¹⁴ suggests discount rates for renewable energy installations fall between 3% to 10%, consistent with the systematic review. Hence, the discount rate is selected as 8% to maintain consistency with the literature. Finally, the operating cost (OPEX) is calculated as a percentage of the capital expenditures, a common approach in the literature. A range of OPEX percentages, between 1% to 5%, is typically suggested in studies^{28, 31, 43, 52, 53, 57, 66, 67} with a modal value of around 3%. Therefore, the annual OPEX is defined as 3% of the CAPEX.

Projection

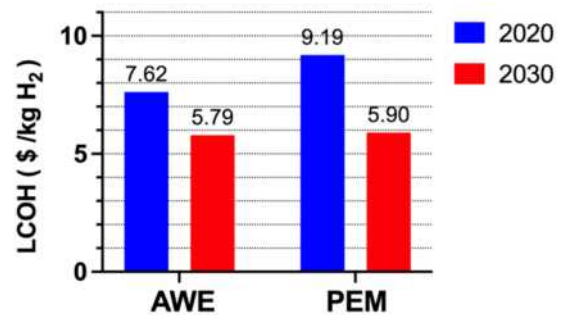


Figure 10. LCOH projections for AWE and PEM

The baseline costs for AWE and PEM electrolyzers in 2020 are \$7.62/kg H₂ and \$9.19/kg H₂, respectively. These costs are projected to decrease by 24% and 36%, respectively. PEM's greater cost reduction is attributed to its higher learning rate, which is linked to its lower technical maturity. This presents PEM as a competitive and potentially more attractive investment option over AWE, behind blue hydrogen. However, long-term cost reductions for PEM may face limitations despite growing capacities, mainly due to constraints with its scarce and expensive materials such as iridium, titanium-based compounds, and platinum, for the porous transport layers^{7, 26, 52}. The supply of these elements is controlled by a few countries, notably South Africa, which contributes about 70% of the global platinum supply and 85% of iridium. The price of the latter is subject to a high price

volatility⁷, with a ratio of 15:1 between the highest and lowest prices in the last two decades. At its peak price of \$1,480/troy ounce⁷, Iridium sets a minimum LCOH threshold of \$1.93/kg. Similarly, platinum yields a minimum cost of \$2.52/kg at its peak. Combining both materials, their cost contribution cannot decrease below \$4.45/kg, posing a potential bottleneck for PEM's large-scale deployment and competitiveness with AWE and blue hydrogen technologies. Minke et al.²⁶ proposes solutions such as more efficient PEM technologies and recycling material.

Conclusion & Outlook

In this paper, the results of an in-depth cost-focused systematic literature review of blue and green hydrogen production technologies are presented. The authors reviewed 63 papers from academic research along with grey literature to construct a comparative cost harmonisation and projection of levelised cost values.

Costs from the academic literature were compiled by technology and energy source and compared to the grey literature where applicable. In general, costs were broadly similar, but the academic literature presented a larger variety of system configurations. Costs for most technologies ranged from \$2-18/kg H₂, resulting in a great amount of uncertainty over their true value. Nonetheless, the averages resulted in levelised costs of hydrogen of \$8.70/kg H₂, \$10.20/kg H₂, \$7.90/kg H₂, \$3.10/kg H₂, \$4.00/kg H₂, and \$7.00/kg H₂ for AWE, PEM, SOEC, SMR, SMR+CCS, and SE-SMR respectively, with SMR and its derivatives leading the market. Thus, SMR+CCS can act as an interim solution until green technologies reach full technical maturity or economic feasibility.

To narrow down the effective cost range, this study has undertaken a cost harmonisation of extrinsic variables. Despite a targeted search of the literature, only about 23.3% and 13% of the reviewed papers contained harmonisable costs for blue and green hydrogen technologies respectively, demonstrating a lack of rigour and uniformity in the literature. For blue hydrogen, the findings demonstrate a substantial decrease in cost ranges for SMR technologies once region-specific costs and assumptions are harmonised. SMR+CCS revealed to be less sensitive due to variations in the technologies and CCS capture rates. A sensitivity analysis indicated that the harmonisation is predominantly impacted by the natural gas costs. Nonetheless, this provides investors with a benchmark to compare technologies. Unlike blue, the green hydrogen harmonisation was unfeasible due to a lack of data and discrepancies, which prevented the identification of the intrinsic cost ranges of each technology. Challenges and areas of improvement of each technology are presented, providing investors with the required insight to make informed decisions.

Finally, a cost projection of the key electrolyser technologies identifies PEM as a promising alternative over AWE, leading with cost reductions of 36% by 2030. However, critical analysis emphasises the potential challenges imposed by its electrode material, which could impose a \$4.45/kg H₂ threshold at peak prices.

Although the production costs for hydrogen have been investigated in detail, depending on the location of the end user, the levelised costs for transportation and storage can exceed those of production¹⁶. Therefore, research into the technologies surrounding them are of paramount importance if large-scale deployment of hydrogen is to be achieved. The use of existing natural gas pipelines for hydrogen transportation and underground salt caverns for storage are possible solutions², however, require substantial research prior to commercial use.

Bibliography

1. IPCC, 2023: *Climate Change 2023: Synthesis Report*. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. IPCC, Geneva, Switzerland, pp. 35-115, doi: 10.59327/IPCC/AR6-9789291691647
2. Global Hydrogen Review 2023. IEA. 2023.
3. *The hydrogen colour spectrum* | National Grid Group. <https://www.nationalgrid.com/stories/energy-explained/hydrogen-colour-spectrum> [Accessed Dec 1, 2023].
4. Chemical Engineering. Chemical Engineering n.d. <http://www.chemengonline.com>.
5. Exchange Rate Archives by Month 2023. https://www.imf.org/external/np/fin/data/param_rms_mth.aspx.
6. Blanco H, Cazzola P, Dulac J, Fukui H, Kim TY, Kurban Z, Levi P, Malischek R, McGlade C, Petrosyan K, Philibert C, Teter J, van Leeuwen J. *The Future of Hydrogen - Seizing today's opportunities. IEA for the G20, Japan*. 2020.
7. Taibi E, Blanco H, Miranda R, Carmo M. *Green Hydrogen Cost Reduction: Scaling up Electrolysers to Meet the 1.5°C Climate Goal*. IRENA; 2020.
8. Water Electrolysis, PEM and Alkaline. <https://ife.no/en/service/water-electrolysis-pem-and-alkaline/> [Accessed Dec 1, 2023].
9. *Hydrogen Insights A perspective on hydrogen investment, market development and cost competitiveness*. Hydrogen Council & McKinsey & Company; 2021.
10. Ahmed A, Hafez, Yasser F, Nassar, Mohamed I, Hammdan & Samer Y. Alsadi. Technical and Economic Feasibility of Utility-Scale Solar Energy Conversion Systems in Saudi Arabia | Iranian Journal of Science and Technology, Transactions of Electrical Engineering. *Iranian Journal of Science and Technology, Transactions of Electrical Engineering*. 2019.
11. Vartiainen E, Masson G, Breyer C, Moser D, Román Medina E. Impact of weighted average cost of capital, capital expenditure, and other parameters on future utility-scale PV levelised cost of electricity. *Progress in photovoltaics*. 2020.
12. *Hydrogen Insights 2023 An update on the state of the global hydrogen economy, with a deep dive into North America*. 2023.
13. *Hydrogen Insights A perspective on hydrogen investment, market development and cost competitiveness*. Hydrogen Council & McKinsey & Company; 2021.
14. *Discounting - Energy Education*. <https://energyeducation.ca/encyclopedia/Discounting> [Accessed Dec 1, 2023].
15. *Path to hydrogen competitiveness - A cost perspective*. Hydrogen Council & McKinsey & Company. 2020.

Effects of Salts on Occurrence Domains of Triglycine Anhydrate and Dihydrate

Claudia Boschín and Rohan Shenoy

Department of Chemical Engineering, Imperial College London

Abstract

As the peptide therapeutics market expands, crystallisation emerges as a sustainable and cost-effective peptide separation technique. To enable critical quality attributes (CQAs), defining occurrence domains of peptide polymorphs is thus of increasing interest. In this study, the effect of salts ($(\text{NH}_4)_2\text{SO}_4$, Li_2SO_4 and MgSO_4 , common industrial precipitants) on the occurrence domains of recently discerned anhydrous and dihydrated triglycine polymorphs is investigated in the 10-30°C temperature range. Triglycine solubility is shown to increase monotonically with concentration of MgSO_4 and up to a threshold with concentrations of $(\text{NH}_4)_2\text{SO}_4$ and Li_2SO_4 , after which specific cation effects appear to reverse. The transition temperature, at which the anhydrate stabilises, is shown to decrease with salt concentration and to the extent: $\text{Li}_2\text{SO}_4 > (\text{NH}_4)_2\text{SO}_4 \gg \text{MgSO}_4$. Experimental results are investigated in terms of molecular mechanisms.

Keywords: peptides, triglycine, crystals, hydrates, solubility, stability, salts, cations

1. Introduction

Peptides present unique therapeutic advantages. Their intermediate size enables desirable properties of both small molecules and biologics. Like small molecules, peptides offer low production costs, low immunogenicity and high bioavailability; like biologics, they are highly specific and can act as inhibitors of peptide-peptide interactions [1]. As a result, their market is large and expanding – estimates for its value in 2022 and its compound annual growth rate over 2023-2032 are USD 42.05 billion and 10%, respectively [2]. Though the average peptide length entering clinical development has increased each decade since the 1980s, the most common range remains 2-10 amino acid residues [3].

Peptide manufacturing encounters a significant bottleneck in the separation process, primarily addressed via chromatography, associated with high solvent usage and high costs of chromatographic adsorbents [4]. Crystallisation has thus received attention as a more sustainable and cost-efficient alternative; additionally, it can offer higher peptide stability and purity [4].

In peptide crystallisation, polymorphic control is critical. Peptides have been shown to exist in polymorphic forms [5] – differing solid forms, including crystalline and amorphous forms, as well as solvates [6]. Polymorphs possess differing physicochemical properties, such as stability and solubility, which affect the bioavailability and efficiency of the therapeutics and downstream operations [7].

The present study investigates co-solutes' effect on the occurrence domains of anhydrous and hydrated peptide crystal forms. Sulphate salts ($(\text{NH}_4)_2\text{SO}_4$, Li_2SO_4 and MgSO_4) have been chosen as model co-solutes for their prominent use in protein crystallisation as precipitants [5]. Triglycine is selected as a model peptide for its simplicity and as it has recently been observed to exist both in folded anhydrous and unfolded hydrated crystal forms (**Figure 1**), stable above and below 30°C, respectively [8] [9].

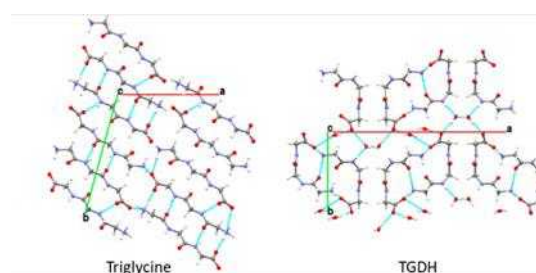


Figure 1: Hydrogen bonding motifs of triglycine anhydrate and dihydrate (TGDH) in the single crystal structure. Reproduced from ref. 8.

2. Background

It has long been appreciated that salts have significant effects on peptide solubility.

In the 1880s, Hofmeister and collaborators observed that salt ions could increase or decrease the solubility of proteins in aqueous solutions – ‘salt-in’ or ‘salt-out’, respectively – and defined

series – later called Hofmeister series – for cations and anions (**Figure 2**) [10]. Specific ion effects (SIEs [11]) were first attributed to specific abilities to adsorb water: small, charge-dense ions – ‘kosmotropes’, adsorbing water molecules better than large, charge-diffuse ions – ‘chaotropes’, dehydrated proteins, favouring their aggregation [10]. However, this explanation already eludes the series for cations: it is not ‘kosmotropic’ ions like Mg^{2+} or Li^+ but ‘chaotropic’ ions like NH_4^+ that are on its salting-out end [10].

Since the 1960s, several studies have attempted to rationalise the Hofmeister series in terms of specific ion-peptide interactions [10]. Salting-in is now understood to arise from ion-peptide interactions, most favoured for strongly hydrated cations at negatively charged side chains and backbone carbonyls and for weakly hydrated anions at backbone amines [10]. Preferential interactions of positively charged side chains with strongly hydrated anions determine a reversal of the anion series for positively charged peptides [12].

In addition, the limitations of treating salt ions separately are emerging, with ion-counterion pairing shown to lead to substantial deviations from the Hofmeister series [11].

The effects of ions on peptide polymorphic outcome have also been studied. In particular, the contributions of cations to peptide folding have been shown to correlate to salting-out and depend on the balance between advantageous cation-mediated peptide dehydration and disadvantageous cation-peptide binding effects [13] [14]. Observed equilibrium shifts toward unfolded conformations have been attributed to the latter [15] [16].

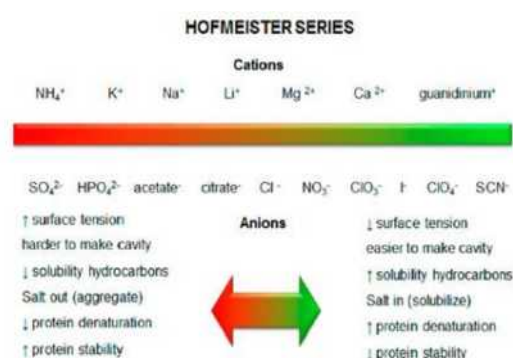


Figure 2: Modern version of the Hofmeister series. Reproduced from ref. 10.

3. Methodology

3.1 Materials

Triglycine (Gly-Gly-Gly, anhydrous, > 98% purity) and salts – ammonium sulphate ($(\text{NH}_4)_2\text{SO}_4$), lithium sulphate (Li_2SO_4), magnesium sulphate (MgSO_4), sodium sulphate (Na_2SO_4), sodium chloride (NaCl) and sodium bromide (NaBr), (all anhydrous, > 98% purity) – were purchased from Sigma-Aldrich and used as received. Deionised water was produced in the laboratory.

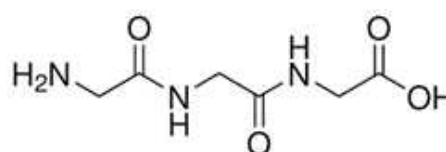


Figure 3: Chemical structure of triglycine.

3.2 Sample Preparation and Equilibration

A stock 2.5M solution of each salt in deionised water was diluted with a liquid handler (Opentrons OT-2) into Eppendorf tubes to obtain 0.5mL samples at 5 concentrations in the range from 0.5 to 2.5M with an increment of 0.5M; a 0.5mL sample of pure deionised water served as control.

After introducing excess amounts of triglycine, the tubes were transferred to a thermostatic mixer (DLAB Scientific HMC100-Pro Thermo Mix) operated at 1500rpm. For $(\text{NH}_4)_2\text{SO}_4$, Li_2SO_4 and MgSO_4 , at least two trials were performed at temperatures of 10, 20 and 30C; due to time constraints, single trials were conducted at intermediate temperatures of 15 and 25C, and results were deemed accurate when fitting with those from adjacent temperature values. For sodium salts, a single trial was performed at 20C.

If, after 30 minutes, a sample appeared clear, additional triglycine was introduced, and the procedure was repeated until the appearance of a suspension.

Phase equilibrium was deemed reached within the following 48 hours (as ref. 9 suggests for an aqueous triglycine solution), after which the samples were left quiescent until all solids appeared to have settled upon visual inspection.

3.3 Measurement of Triglycine Solubility

From the supernatant, 100 μ L were transferred via a micropipette to a second Eppendorf tube. From this, 10 μ L were diluted via the liquid handler in 490 μ L of deionised water in a third Eppendorf tube to prevent further crystallisation and set the triglycine concentration within the limits of calibration. The concentration was thus measured with UV-vis spectroscopy, and its average over five repeats was used to calculate the concentration of the original sample – i.e. the solubility.

A calibration curve was constructed for a microvolume UV-vis spectrophotometer (Thermo Fischer Scientific NanoDrop One C). Absorbance spectra were measured at 230nm for known concentrations of triglycine in deionised water (8 in the range from 0.25 to 2 mg/mL). A linear relation between absorbance and concentration, in accordance with the Beer-Lambert Law, was defined (Figure 4).

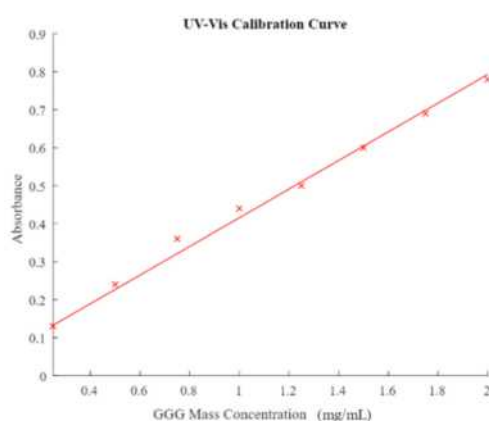


Figure 4: Calibration curve of triglycine (GGG).

3.4 Determination of Stable Triglycine Polymorph

Solids were filtered with filter paper and characterised by optical microscopy (CX-41 Olympus) and PXRD (PANalytical X'Pert PRO X-Ray) to determine their crystal form.

Anhydrous and dihydrated crystals were differentiated by their rod and needle morphologies, respectively, with optical

microscopy (Figure 5.a)) and by characteristic XRD patterns with PXRD (Figure 5.b)) [8].

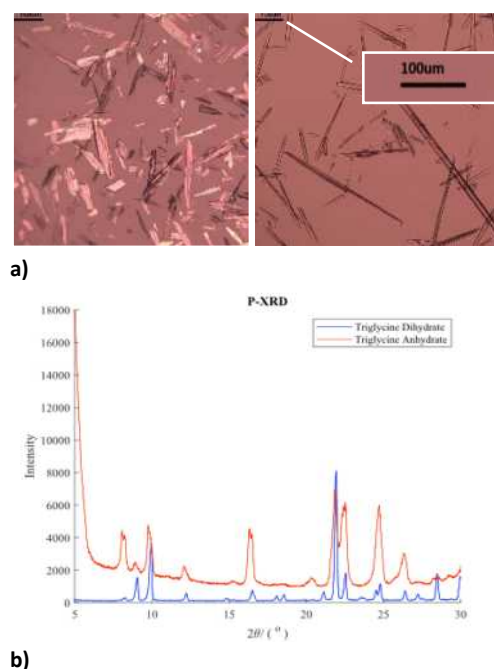


Figure 5.a) Optical microscope images (at 10x magnification) of triglycine anhydrate (left) and dihydrate (right) **b)** The XRD pattern of triglycine anhydrate and dihydrate.

4. Results and Discussion

4.1 Effects of Salts on Triglycine Solubility

Figure 6 illustrates measured triglycine solubilities (normalised over the solubility of the stable form – dihydrate in the cases considered) at salt concentrations 0.5-2.5M and temperatures 10-30C.

All salts are observed to salt-in triglycine in the entire concentration and temperature ranges. Salting-in can be attributed to the screening of peptide dipoles, which would otherwise drive aggregation [17]. Dipoles are expected: all salts used are either neutral or weakly acidic, so their aqueous solutions should have a pH close to the 5.56 triglycine isoelectric point [18]. Ion-peptide interactions are further aided as triglycine is poor in hydrophobic moieties – apolar side chains and backbone methyl groups – which would expel ions [19] [20]. These ion-peptide interactions seem to

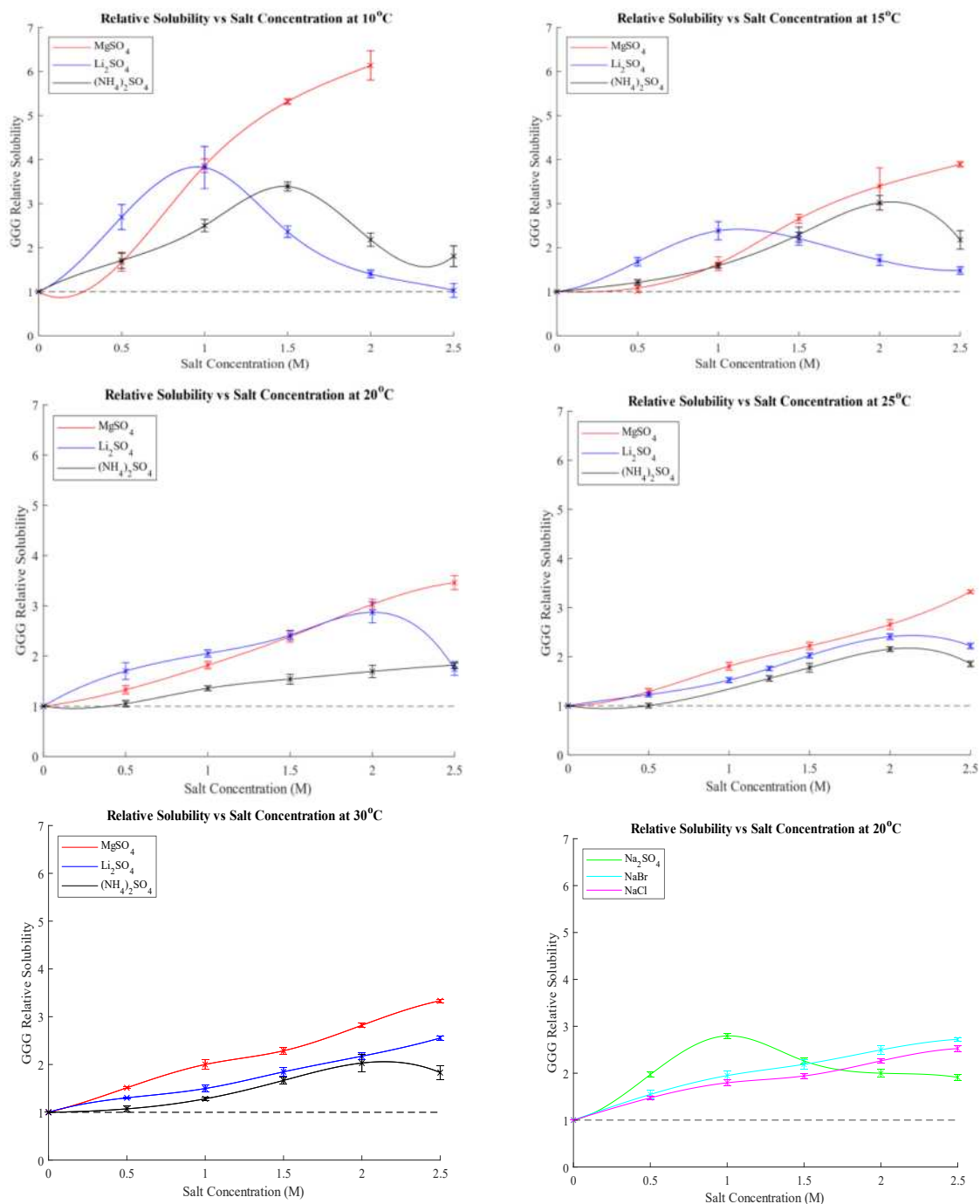


Figure 7: Triglycine relative solubility vs concentration of sulphates (NH₄)₂SO₄, Li₂SO₄ and MgSO₄, at temperatures, from left to right and top to bottom, 10C, 15C, 20C, 25C and 30C and vs concentration of sodium salts (Na₂SO₄, NaCl, NaBr) at 25C at the bottom right. Error bars represent the standard deviation among the five repeats that were averaged to measure solubility.

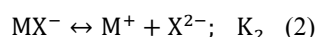
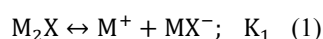
offset the ion-water interactions leading to salting-out (and argued to cause negligible disruption to water structure beyond the first ion hydration shell [21]).

Solubility shows a maximum with (NH₄)₂SO₄ and Li₂SO₄ concentrations and a monotonic

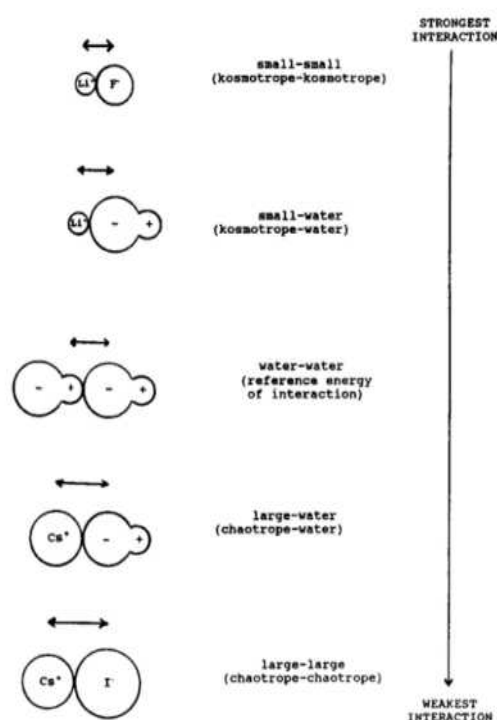
increase with that of MgSO₄. Maxima are understood to result as the primary salt contribution shifts from ion-peptide interactions-driven salting-in to ion-water interactions-driven salting-out at increased concentrations [17]. Before maxima, the salting-in rate is higher for Li⁺ than for NH₄⁺— as is the expected strength of interactions with the

peptide, while after maxima, it is the salting-out rate to be higher for Li^+ than for NH_4^+ – as is the predicted strength of interactions with water.

The different behaviour with MgSO_4 can be attributed to the stronger interactions with C-termini and carbonyls of its cation, which arise from the higher charge density (**Figure 2**). However, differing stoichiometries impede direct comparison of specific cation effects. Indeed, another explanation accounts for the differing stoichiometries themselves. As salt concentration increases, increasing ion-ion interactions favour the formation of ion pairs [22]. For 1:1 MX (MgSO_4) salts, ion pairs are neutral and thus easily partition toward backbones while interacting poorly with water [23]. Meanwhile, as dissociation of 2:1 M_2X salts ($(\text{NH}_4)_2\text{SO}_4$ and Li_2SO_4) occurs via a two-step process:



and, typically, $K_1 \gg K_2$, increased salt concentration favours an increase in MX^- ion pairs, negatively charged [22]. This explanation extends to the observed behaviours of sodium salts (**Figure 6**). Theories such as the Law of Matching Water Affinity [24] (**Figure 7**), capture the extent of ion pairing and predict strong ion pairing for MgSO_4 , observed in past studies [25].



With temperature, maxima appear to shift to higher concentrations, the relative solubility with MgSO_4 to decrease, and the relative solubilities with all salts to converge. At higher temperatures, triglycine dissolution is favoured [9], so the effect of salt addition might then be reduced.

4.2 Effects of Salts on Stable Triglycine Polymorph

The observed stable triglycine crystal forms are illustrated in **Table 1.a) – c)** at salt concentrations 0.5-2.5M and temperatures 10-30C.

The transition temperature – above which anhydrous are favoured over hydrate crystals – appears to decrease with salt concentration. This can be explained as folded configurations are favoured as ions deplete the peptide hydration shell of water molecules [13][14]. Indeed, transition temperature lowers with increased expected water depletion effects. At 1M and 1.5M is highest with MgSO_4 (30C and 25C vs 25C and 15C with $(\text{NH}_4)_2\text{SO}_4$ and Li_2SO_4) and at 2.5M higher with $(\text{NH}_4)_2\text{SO}_4$ than with Li_2SO_4 (20C vs 15C).

In addition, though cation binding is typically understood to favour the unfolded conformation [15][16], this might not be the case when the latter is strongly dependent on buried water networks, such as for triglycine [8]. Cation binding might reduce repulsions between carbonyls and between carboxylates, which has been shown to contribute strongly to buried water networks in peptide hydrates [26], including triglycine dihydrate [8]. The transition temperature decreases more with Li_2SO_4 than with $(\text{NH}_4)_2\text{SO}_4$, as the expected strength of cation-peptide binding increases. However, it decreases the least with MgSO_4 , suggesting that cation-peptide binding, if favourable, is still not as favourable as salt-water binding to anhydrate stabilisation.

Figure 7: Ordering of interactions in aqueous water solutions from strongest to weakest. According to the Law of Matching Water Affinity, small-small ion pairs are energetically favoured as their interactions are strong, whereas large-large ion pairs are favoured as, even though their interactions are weak, their dehydration leads to water-water interactions stronger than large ion-water interactions. In contrast, small-large ion interactions are not energetically favoured, as their weak interactions do not compensate for the work required to dehydrate small ions. Reproduced from ref. 24.

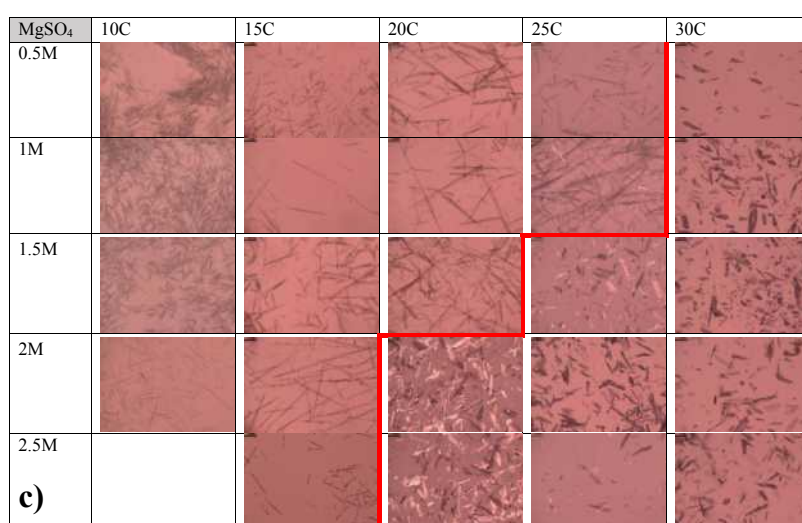
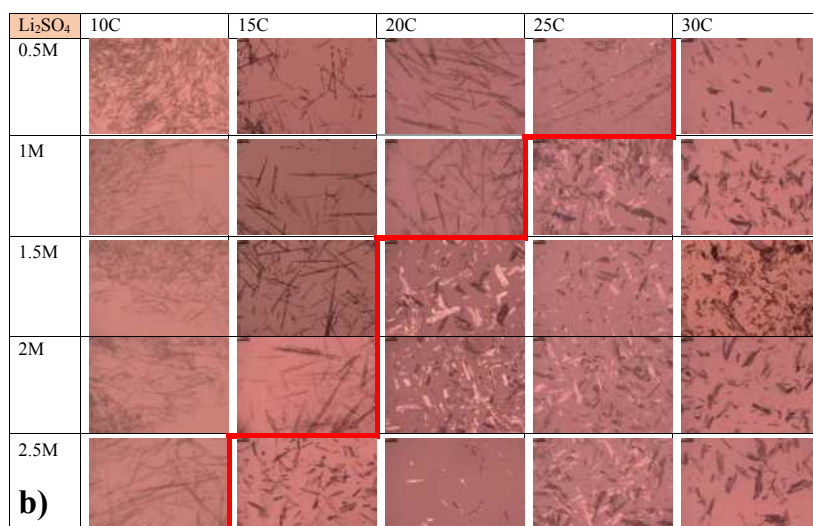
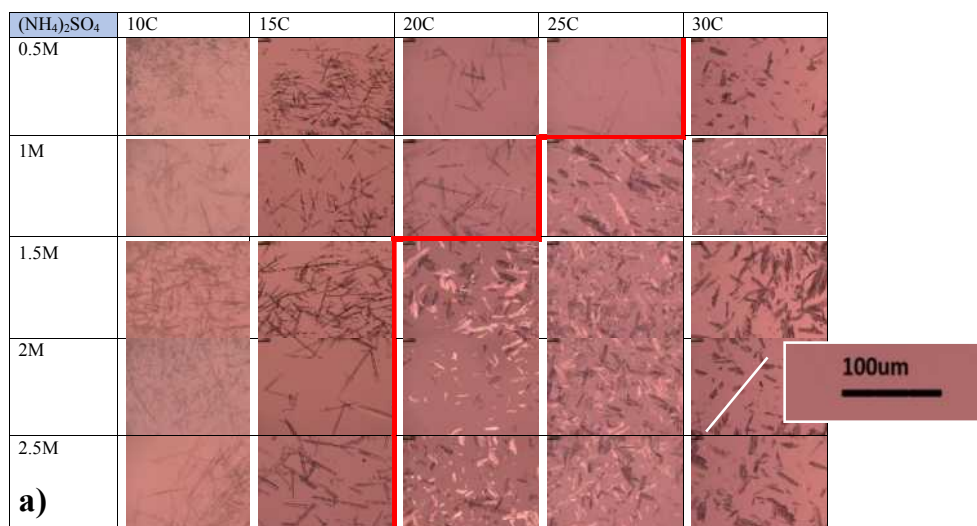


Table 1: Optical microscopy images (at 10x magnification) of the triglycine crystal outcomes at temperatures 10-30C and concentrations 0.5-2.5M of salts **a)** $(\text{NH}_4)_2\text{SO}_4$, **b)** Li_2SO_4 and **c)** MgSO_4 . Red lines mark boundaries between anhydrate (right) and dihydrate (left) occurrence domains.

5. Conclusion and Outlook

Sulphates have been shown to affect triglycine anhydrate and dihydrate occurrence domains significantly.

Triglycine solubility appears to increase to a maximum with $(\text{NH}_4)_2\text{SO}_4$ and Li_2SO_4 concentrations. After maxima, specific cation effects appear to reverse. This is attributed to the dominant effect of salt addition shifting from ion-peptide interactions-driven salting-in to ion-water interactions-driven salting-out with increasing concentration. That this shift is not observed with MgSO_4 is explained in terms of its stronger cation-peptide interactions and its different stoichiometries. Salt addition effects appear to be mitigated as peptide dissolution is favoured at higher temperatures.

The transition temperature, at which the anhydrate becomes the stable crystal form, is shown to lower with salt concentration, to the extent: $\text{Li}_2\text{SO}_4 > (\text{NH}_4)_2\text{SO}_4 \gg \text{MgSO}_4$. This decrease is attributed to salt-water interactions depleting the peptide hydration shell and, to a lesser extent, reduced electrostatic repulsions between charged groups, such as carbonyls and carboxylates, thought to favour buried water networks in the hydrate.

While these results enable predictions about the effect of salts on crystal occurrence domains of triglycine and affine molecules (e.g., longer uncharged peptides), their accuracy and significance could be improved. The individual contribution of salt stoichiometry and salt ion affinity with peptide, water and counterion could be elucidated. This could be done via molecular dynamics (MD) simulations and further experimentation – e.g., over more salt species and with kinetic and more precise and continuous measurements of solubility and transition temperatures via techniques such as Fourier-transform infrared (FTIR) spectroscopy.

6. Acknowledgement

The authors thank the members of the Heng Research Group for their invaluable expertise, guidance and support throughout this research project.

7. References

- [1] Wang, L., Wang, N., Zhang, W., Cheng, X., Yan, Z., Shao, G., Wang, X., Wang, R. and Fu, C. (2022). Therapeutic peptides: current applications and future directions. *Signal transduction and targeted therapy*, 7(1), 48. <https://doi.org/10.1038/s41392-022-00904-4>
- [2] Rossino, G., Marchese, E., Galli, G. and Verde, F., Finizio, M., Serra, M., Linciano, P. and Collina, S. (2023). Peptides as Therapeutic Agents: Challenges and Opportunities in the Green Transition Era. *Molecules*, 28, 7165. <https://doi.org/10.3390/molecules28207165>
- [3] Lau, J. L., and Dunn, M. K. (2018). Therapeutic peptides: Historical perspectives, current development trends, and future directions. *Bioorganic & medicinal chemistry*, 26(10), 2700–2707. <https://doi.org/10.1016/j.bmc.2017.06.052>
- [4] Roque, A. C. A., Pina, A. S., Azevedo, A. M., Aires-Barros, R., Jungbauer, A., Di Profio, G., Heng, J. Y. Y., Haigh, J. and Ottens, M. (2020). Anything but Conventional Chromatography Approaches in Bioseparation. *Biotechnology journal*, 15(8), e1900274. <https://doi.org/10.1002/biot.201900274>
- [5] McPherson, A. and Gavira, J. A. (2014). Introduction to protein crystallization. *Acta crystallographica. Section F, Structural biology communications*, 70(Pt 1), 2–20. <https://doi.org/10.1107/S2053230X13033141>
- [6] Raw, A. S., Furness, M. S., Gill, D. S., Adams, R. C., Holcombe, F. O. Jr. and Yu, L. X. (2004). Regulatory considerations of pharmaceutical solid polymorphism in Abbreviated New Drug Applications (ANDAs). *Advanced drug delivery reviews*, 56(3), 397–414. <https://doi.org/10.1016/j.addr.2003.10.011>
- [7] Halebian, J. and McCrone, W. (1969). Pharmaceutical applications of polymorphism. *Journal of Pharmaceutical Sciences*, 58(8), 911–929. <https://doi.org/10.1002/jps.2600580802>
- [8] Guo, M., Rosbottom, I., Zhou, L., Yong, C.W., Ling, Z., Qiuxiang, Y., Todorov, I.T., Errington, E. and Heng, J.Y.Y. (2021). Triglycine (GGG) Adopts a Polyproline II (pPII) Conformation in Its Hydrated Crystal Form: Revealing the Role of Water in Peptide Crystallization. *The journal of physical chemistry letters*, 12, 8416–8422. <https://doi.org/10.1021/acs.jpclett.1c01622>
- [9] Guo, M., Chang, Z.H., Liang, E., Mitchell, H., Zhou, L., Yin, Q., Guinn, E.J. and Heng, J.Y.Y. (2022). The effect of chain length and side chains on the solubility of peptides in water from 278.15 K to 313.15 K: A case study in glycine homopeptides and dipeptides. *Journal of Molecular Liquids*, 352.

<https://doi.org/10.1016/j.molliq.2022.118681>

[10] Okur, H. I., Hladílková, J., Rembert, K. B., Cho, Y., Heyda, J., Dzubiella, J., Cremer, P. S. and Jungwirth, P. (2017). Beyond the Hofmeister Series: Ion-Specific Effects on Proteins and Their Biological Functions. *The journal of physical chemistry. B*, 121(9), 1997–2014.

<https://doi.org/10.1021/acs.jpcb.6b10797>

[11] Gregory, K. P., Elliott, G. R., Robertson, H., Kumar, A., Wanless, E. J., Webber, G. B., Craig, V. S. J., Andersson, G. G. and Page, A. J. (2022). Understanding specific ion effects and the Hofmeister series. *Phys. Chem. Chem. Phys.*, 24, 12682–12718. doi:10.1039/D2CP00847E [12]

<https://pubs.acs.org/doi/10.1021/acs.jpcb.6b10797>

[12] Hladílková, J., Heyda, J., Rembert, K. B., Okur, H. I., Kurra, Y., Liu, W. R., Hilty, C., Cremer, P. S. and Jungwirth, P. (2013). Effects of End Group Termination on Salting-Out Constants for Triglycine. *The journal of physical chemistry letters*, 4(23), 4069–4073.

<https://doi.org/10.1021/jz4022238>

[13] Asakereh, I., Lee, K., Francisco, O. A. and Khajepour, M. (2022). Hofmeister Effects of Group II Cations as Seen in the Unfolding of Ribonuclease A. *ChemPhysChem*, 23(12), e202100884.

<https://doi.org/10.1002/cphc.202100884>

[14] Wingfield P. (2001). Protein precipitation using ammonium sulfate. *Current protocols in protein science*, Appendix 3, Appendix-3F.

<https://doi.org/10.1002/0471140864.psa03fs13>

[15] Street, T. O., Bolen, D. W. and Rose, G. D. (2006). A molecular mechanism for osmolyte-induced protein stability. *Proceedings of the National Academy of Sciences of the United States of America*, 103(38), 13997–14002.

<https://doi.org/10.1073/pnas.0606236103>

[16] Bykov, S. and Asher, S. (2010). Raman studies of solution polyglycine conformations. *The journal of physical chemistry. B*, 114(19), 6636–6641. <https://doi.org/10.1021/jp100082n>

[17] Dahal, Y. R. and Schmit, J. D. (2018). Ion Specificity and Nonmonotonic Protein Solubility from Salt Entropy. *Biophysical journal*, 114(1), 76–87. <https://doi.org/10.1016/j.bpj.2017.10.040>

[18] Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D. and Bairoch, A. (2003). ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic acids*

research, 31(13), 3784–3788.

<https://doi.org/10.1093/nar/gkg563>

[19] Baldwin R. L. (1996). How Hofmeister ion interactions affect protein stability. *Biophysical journal*, 71(4), 2056–2063.

[https://doi.org/10.1016/S0006-3495\(96\)79404-3](https://doi.org/10.1016/S0006-3495(96)79404-3)

[20] Rankin, B. M. and Ben-Amotz, D. (2013). Expulsion of ions from hydrophobic hydration shells. *Journal of the American Chemical Society*, 135(24), 8818–8821.

<https://doi.org/10.1021/ja4036303>

[21] Collins, K. D., Neilson, G. W. and Enderby, J. E. (2007). Ions in water: characterizing the forces that control chemical processes and biological structure. *Biophysical chemistry*, 128(2-3), 95–104.

<https://doi.org/10.1016/j.bpc.2007.03.009>

[22] Sujanani, R., Nordness, O., Miranda, A., Katz, L. E., Brennecke, J. F. and Freeman, B. D. (2023). Accounting for Ion Pairing Effects on Sulfate Salt Sorption in Cation Exchange Membranes. *The Journal of Physical Chemistry B*, 127(8), 1842–1855. <https://doi.org/10.1021/acs.jpcb.2c07900>

[23] Bruce, E. E., Okur, H. I., Stegmaier, S., Drexler, C. I., Rogers, B. A., van der Vegt, N. F. A., Roke, S. and Cremer, P. S. (2020). Molecular Mechanism for the Interactions of Hofmeister Cations with Macromolecules in Aqueous Solution. *Journal of the American Chemical Society*, 142(45), 19094–19100.

<https://doi.org/10.1021/jacs.0c07214>

[24] Collins K. D. (1997). Charge density-dependent strength of hydration and biological structure. *Biophysical journal*, 72(1), 65–76.

[https://doi.org/10.1016/S0006-3495\(97\)78647-8](https://doi.org/10.1016/S0006-3495(97)78647-8)

[25] Götze, L., Parry, K. M., Hua, W., Verreault, D., Allen, H. C. and Tobias, D. J. (2017). Solvent-Shared Ion Pairs at the Air-Solution Interface of Magnesium Chloride and Sulfate Solutions Revealed by Sum Frequency Spectroscopy and Molecular Dynamics Simulations. *The journal of physical chemistry. A*, 121(34), 6450–6459.

<https://doi.org/10.1021/acs.jpca.7b05600>

[26] Yang, C. H., Brown, J. N. and Kopple, K. D. (1979). Peptide--water association in peptide crystals. *International journal of peptide and protein research*, 14(1), 12–20.

<https://doi.org/10.1111/j.1399-3011.1979.tb01915.x>

A Techno-Economic Analysis of a Novel Process to Treat Pot Ale into Hexanoic Acid

Finlay Barnes Bush and Alexa Dita

Department of Chemical Engineering, Imperial College London, U.K.

Abstract

Pot-ale is one of the most abundant waste products within the global whisky industry. Most commonly used as low-grade animal feed/fertiliser or disposed to land/sea, it is considered to be inefficiently utilised. This paper presents a techno-economic analysis of a novel process to separate pot ales volatile fatty acid content, specifically hexanoic acid for later sale. Two designs were modelled, one where the pot ale is sent straight to the separations system, and one where it is first fermented to concentrate its hexanoic acid content in order to determine if this fermentation is worthwhile. A design of hexanoic acids separation sequence has been built on the process simulation software Aspen Plus v11 using the NRTL-HOC property model under continuous operation. The route the process stream takes is as follows: First as the fermentation is a batch process, the stream is sent to a storage vessel to act as a buffer and allow for a continuous model to be designed. This effluent is then sent to an extraction unit where hexyl acetate, acting as the solvent, removes the high-water content in the stream. The wastewater is sent to a treatment facility and the acid/solvent stream is sent to distillation column where hexanoic acid is isolated in the bottoms and the solvent is recovered in the distillate and recycled back to the extractor. Finally, an economic analysis is presented for both designs over a 25-year lifetime, where typical tax/interest rates and other such charges have been assumed. From this analysis it was determined that fermentation of the pot ale is necessary to build a profitable design, with an estimated rate of return of 68%. Whereas with no fermentation the cost of operation outweighs the revenue seen and thus is not a financially acceptable design.

Keywords: pot ale, hexanoic acid, VFAs, Aspen Plus, property model, solvent extraction, distillation

1. Introduction

The production of whisky dates back to the early 1000s in Scotland and Ireland, with both nations claiming they were the original producers. With limited access to grapes for the production of wine, these early day Europeans decided to ferment grains such as wheat, barley, or rye for alcohol. This alcohol was then distilled to produce the first recorded cases of whisky [1]. Since then the whisky industry has grown exponentially, with improvements in technology and understanding of the process, more complex flavours and production techniques have been refined and perfected. Today whisky is exported and enjoyed all over the globe with the largest producers being the UK, USA, and the EU. The Scotch whisky association reported over £6bn worth of exports in 2022 comprising a quarter of all UK food and drink exports, with roughly 1.6bn bottles being shipped yearly [2]. Although this remains a huge market globally, it has been suggested that the market is currently saturated in the world's major economies [3]. Due to the large numbers of producers in the world there is not much possibility for a unique selling point to separate one business from another. The market is dominated by competition and its growth within the domestic business segment is restricted to the general economic growth of roughly 1~3% per year [3].

One way a business can expand within the industry is to improve on the efficiency of one's process. An effective way to do this could be to make use of an otherwise disregarded side product and turn it into a

throughput for another process. If designed correctly this new process will diversify your revenue streams and lead to a stable increase in economic growth. For the whisky industry the obvious candidate for this is pot ale. Pot ale is the principal effluent by-product of the whisky industry with estimates ranging from 1.4 to 2.7 billion litres produced annually by just Scottish SME distillers [4, 5]. Most commonly pot ale is recycled into animal feed, or it is spread to land as a low-grade fertiliser incurring a disposal cost to the business. In some cases, the pot ale is disposed at sea however this is only available to distilleries located on the coast and a specific discharge licence is required [4]. These are inefficient uses of pot ale and do not provide the industry with any additional income. A newer and potentially better application of pot ale is to concentrate it into a syrup and treat it through anaerobic digestion (AD). This processes the biomass into a biogas which can be used as an alternative energy source to fossil fuels. This is significant as the Scottish Whisky Association (SWA) has issued targets to source 80% of the industries energy from non-fossil fuel sources by 2050 [5] and thus the generation of biogas can help to reach this target. Furthermore there is evidence suggesting that the digestate post AD is a more effective fertiliser than pot ale itself.

As detailed, there are many ways you can process your pot ale with new methods and processes being developed all the time. This paper presents a novel process for the treatment of pot ale whereby it is sent to

a secondary fermenter to produce and concentrate its volatile fatty acid (VFA) content, specifically hexanoic acid. This hexanoic acid is then isolated and purified allowing it to be commercially sold. Hexanoic acid is a valuable feedstock in the chemical and biofuel industry, as well as having application as an antimicrobial agent, animal feed additive and flavour additive [6]. The aim of the report is to decide whether this secondary fermentation is worth the investment or if it would be financially preferable to separate off the lower concentration hexanoic acid directly from the pot ale. With this in mind, two designs shall be built, one where initial acid concentrations reflect no secondary fermentation (case 1) and one where hexanoic acid concentrations have been optimised to reflect the secondary fermentation (case 2).

2. Background

Pot ale is one of the three main side products within whisky production, the others being spent lees and draff. It is comprised of a mixture of Volatile Fatty Acids (VFAs), water, minerals such as Cu, P and K, and a solid fraction mainly consisting of yeast [4]. For Scottish SME malt distillers pot ale has a production rate of roughly 1.4 to 2.7 billion litres per year [4, 5], this is often then concentrated into a syrup through evaporation and used as animal feed selling at roughly £60-200/ton [4,7]. There are several drawbacks to this method such as the presence of copper in pot ale. Ingesting high amounts of copper can lead to copper poisoning causing haemolysis, which is potentially fatal to the animal, the copper content must therefore be regulated before pot ale is allowed to be sold as feed. Furthermore, the syrups high viscosity presents issues with transportation and storage making it an unfavourable choice of feed to farmers [4]. All in all, the low sales price of syrup coupled with energy demanding evaporation and poor transportation and storage make this a bad processing method. Other common practises for the utilisation of pot ale are land and sea disposal. This is also unfavourable as it incurs a disposal fee and is only approved for distilleries located in a suitable location.

The current practises for pot ale utilisation are out-dated and inefficient. Pot ale is constituted of valuable VFAs which if isolated, can be sold commercially. Hexanoic acid sells for roughly £2000/ton [7], 10x that of animal feed and its concentration within pot ale can be greatly improved by fermentation. We are proposing two processes for generating a high purity hexanoic acid stream, both aim to separate off hexanoic acid from pot ale syrup with one having undergone secondary fermentation and the other not. The aim is to determine whether this secondary fermentation is financially favourable. There are several factors to consider here when designing a process like this. The equation of state and activity model used in our calculations and predictions is vitally important as this influences the binary interaction parameters and details how the system will behave and how different components will interact. Due to the high-water content of pot ale syrup LLE extraction is utilised within the separation process to reduce the volume of process fluid.

Therefore, care must be taken in your choice of solvent and a recycle stream must be designed to allow for solvent recovery. Finally, the order of separation must be considered, and it should be determined if there are any other worthwhile components other than hexanoic acid to isolate for commercial sale.

The process proposed within this report details the separation of hexanoic acid from fermented and unfermented pot ale syrup. We do not detail the evaporation of pot ale into said syrup or the actual fermentation process itself. Should you want to design this part of the process then you need to consider the extent of evaporation as a lower water content will reduce the cycle time within the fermenter but too low a water content results in a highly viscous fluid that is difficult to process. You also need to select the bacteria used for fermentation. In our case we are considering a water content of 85% by mass and bacteria that was chosen by the Biorenewables Development Centre (BDC) in York, UK. The identity of this bacteria is protected under a non-disclosure agreement.

3. Methods

All process simulations are done in Aspen Plus v11 under the NRTL-HOC property model.

3.1. Initialising design

A continuous feed basis of 1000L/hr is assumed, as this is a reasonable amount of waste produced by a medium sized distillery [8]. Because fermentation is a batch process, a buffer vessel is installed to allow for continuous feed. The feed is at 35 °C and 1 atm and the mass composition table 3.1.1 was calculated from the acid concentration profile displayed by figure 3.1.1 at time $t=0$ and $t \approx 20$ hours for cases 1 and 2 respectively along with the assumption of 85 wt% water content.

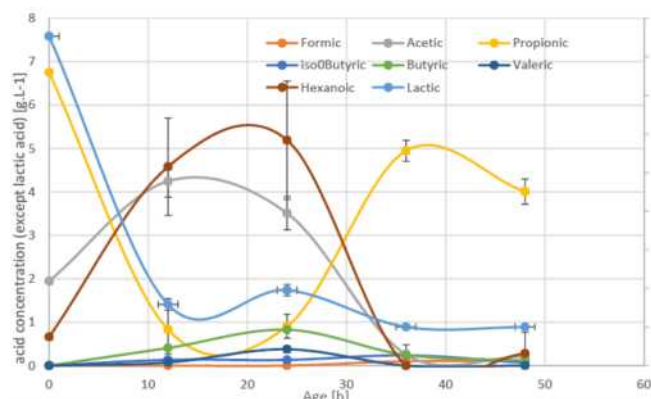


Figure 3.1.1: Acid concentration profile with time spent fermenting [7]

Table 3.1.1: Compositions of feed for both designs by mass

	Water	Hexanoic	Acetic	Lactic	Butyric	Propionic
Case 1 (-)	0.85	0.0062	0.0175	0.0667	-	0.0596

Case 2 (-)	0.85	0.071	0.051	0.019	0.009	-
------------	------	-------	-------	-------	-------	---

A preliminary separation design is started out via distillation, as this is the most used process for this type of mixture [9,10]. First results indicate that the separation is extremely inefficient due to the large water content and high process stream volume.

3.2. Extractor design

Liquid-liquid extraction was found suitable for removing water from a mixture of organic acids as found in literature precedents [11]. An adiabatic extractor using n-hexyl acetate solvent is simulated, and the number of stages is optimized through trial and error with respect to the amount of fresh solvent required to achieve at least 99%wt water removal.

Three different extractants were modelled, two physical extractants, hexyl and nonyl acetate and one reactive extractant, trioctylamine (TOA) with 1-octanol acting as an active diluent. They were modelled in identical extractors and after considering aspects such as its extraction efficiency and ability to be recovered, hexyl acetate was deemed the most suitable.

3.3. Distillation Column Design

Distillation is firstly designed via a RADFRAC block with an arbitrary number of stages. The defining operating specifications are chosen to be the reflux ratio and distillate to feed ratio, both on a molar basis. Both are subsequently manipulated to meet a design specification of 99%wt hexanoic acid purity in the bottoms. Performance metrics are recorded (table 3.3.1): heating/cooling duty per amount of hexanoic acid produced, solvent recovery, hexanoic acid recovery. Afterwards, a sensitivity analysis is performed to find the minimum number of stages that maintains this performance.

Table 3.3.1: Distillation column design and performance metrics

Design	No. of stages	Reflux ratio	Heating duty/mass of product [kWh/kg]	Cooling duty/mass of product [kWh/kg]	Solvent recovery [%]	Hexanoic acid recovery [%]
Case 1	18	1.5	4.31	3.81	96.9	98.5
Case 2	20	2	0.72	0.59	97.4	99.9

3.4. Recycle, Purger & Mixer Design

The recovered solvent stream needs purging to prevent build-up of residual VFAs. Through trial and error, a split ratio of 5% is optimal amount able to converge the simulation whilst minimising loss of solvent. This recycle is mixed with the fresh solvent feed and fed to the extractor.

3.5. Sizing Units

Various methods were employed in sizing the units for our proposed design. First off, when sizing the buffer vessel the only important parameters are its volume, to facilitate a continuous feed of fermentation broth, and its diameter to height ratio, to ensure a stable and space efficient column. As the feed flowrate from the buffer vessel is already specified as 1,000L/hr, it was decided that a column with 10x the volume of this required feed-rate would be suitable to ensure continuous flow. The columns volume was therefore specified as 10,000 L. To design a stable column it was decided to aim for a diameter to height ratio of roughly $1 < D_T/H_T < 1.2$. Now by just specifying either the height or the diameter, the column can be sized. In this case a height of 7ft was selected resulting in a diameter of 8.01 ft.

The extractor column was sized based off of its number of stages and stage efficiency. By assuming a stage efficiency, E_o , of 0.7 the actual number of stages, N_{act} , could be calculated using Eq. 1.

$$N_{act} = \frac{N}{E_o} \quad Eq. 1$$

The columns height was then calculated through Eq. 2 [12].

$$H_T = \frac{1.1N_{act}C_T - C_T}{0.9} \quad Eq. 2$$

Where C_T is the tray spacing and is assumed to be 0.5m (1.64ft), as specified by Aspen, and all length measurements are given in feet. By multiplying by a factor of 1.1, an empty space allowance of 10% for vapour disengagement and liquid sump has been accounted for. For this column of 5 stages, it equates to a height of roughly 18ft and lacking a predicted diameter from Aspen, one was set to be 5ft.

The distillation column was designed in a similar way. Eq. 2 was again used to calculate the height however a factor of 1.15 was used instead of 1.1 to accommodate the increased vapour disengagement and liquid sump seen in a distillation column. Aspen predicted a diameter of 1.5ft, tray spacing of 2ft and again a tray efficiency of 0.7, leading to a column height of 62.43ft.

Finally, the mixing and purging vessel were not specified by Aspen and lacked correlations to size them. All that is known is they both process roughly 1.5L of fluid an hour and therefore must be large enough to accommodate this flowrate. Both vessels are assumed to be negligible in price compared to the other three columns and therefore their exact sizing is not relevant to the economic model.

3.6. Economic Analysis

To build the economic model, several assumptions had to be made. It was first decided that a 'harsh' analysis was to be built with interest rates and other such costs

being overestimated to model a worst-case scenario. If this worst-case proved profitable it would mean that in actual practise the profit margins would be higher.

To start off the model a 25-year lifetime was assumed. An operation time of 8000 hours per year was set. This is around the 10% allowance figure based on a 365-day year and should provide sufficient downtime for the plant. It was then assumed that the project would be geared and fully financed by a loan with 12% interest [13] and a target to pay off the principle and investment payments within 5 years. This target was set arbitrarily and is up to the company to decide. A location factor of 1.3 [14] was used to account for the plant being built in the UK. A linear depreciation of assets was assumed and finally a consumer price index of 5% [15], and tax rate of 25% [16] was used.

To calculate the capital expenditures (CAPEX), correlations from Douglas [17] were utilised:

$$\text{Installed cost of column shell} = \frac{M\&S}{280} 101.9 D_T^{1.066} H_T^{0.802} (2.18 + F_c) \quad \text{Eq. 3}$$

$$\text{Installed cost of trays} = \frac{M\&S}{280} 4.7 D_T^{1.55} H_{stack} F_c \quad \text{Eq. 4}$$

Where M&S is the Marshall Swift index, used to update the correlation to today's prices. A value of 1800 was used [18]. F_c is the cost factor and relates to the materials used, a value of 1 was selected corresponding to a carbon steel column [19]. The capital costs for the distillation columns condenser, reboiler and reflux pump were calculated within Aspen and an estimate of the

mixer's capital was made based off the prices of other similar vessels.

The operating expenditures (OPEX) only stem from the distillation columns utility requirement as well as the cost of solvent. The column used cooling water for the condenser, high pressure steam for the reboiler and required electricity to operate the reflux pump. The respective requirements for these were calculated using equations 5 and 6.

$$\text{Condenser OPEX} = \frac{-Q_c * \text{operating time} * \text{price of cooling water}}{C_{p,water} * \Delta T_{water} * \rho_{water}} \quad \text{Eq. 5}$$

$$\text{Reboiler OPEX} = \frac{Q_r * \text{operating time} * \text{price of steam}}{\Delta H_{vap}} \quad \text{Eq. 6}$$

Where Q_c and Q_r are the heat duties of the condenser and reboiler respectively and were calculated by Aspen. ΔT is the temperature difference of the cooling water, set to be 11.11 K by Aspen. C_p and ρ are the heat capacity and density of water respectively and finally ΔH_{vap} is the heat of vapourisation of water.

Finally, to predict the revenue streams, a price of \$2000/tonne hexanoic acid was used [7] and the cost hexyl acetate (solvent) was set as \$4/kg [20].

4. Results

4.1. Final Design

The final design of both of our proposed processes is displayed by figure 4.1.1, where parameters relating to case 1 are displayed in blue and for case 2 in red. F represents the mass flowrate and x the mass composition.

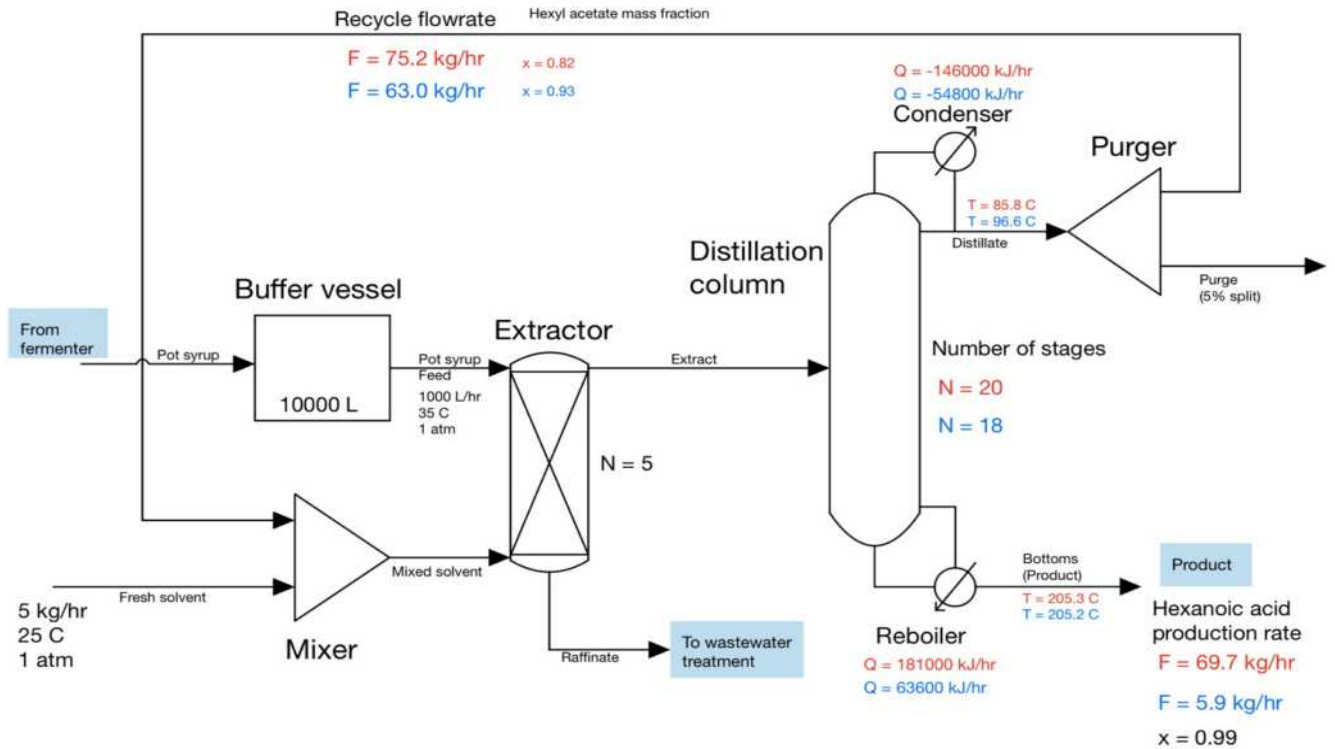


Figure 4.1.1: Process flow diagram for designs 1&2 (blue and red respectively)

4.2. Solvent selection

Table 4.2.1 presents the performance of the modelled solvents, nonyl and hexyl acetate, and TOA with 1-octanol.

Table 4.2.1: Tested solvents performance.

Solvent	Nonyl Acetate (physical)	Hexyl Acetate (physical)	TOA – 1-octanol (reactive)
Distribution coefficient (-)	0.999	0.999	0.975
Solvent : feed ratio	0.06	0.06	0.015
Water leeching (wt%)	1.35	0.58	0.75
Recovery ability	Challenging recovery	Easily recovered	Challenging recovery

4.3. Sizing

The estimated dimensions of each vessel are laid out in table 4.3.1:

Table 4.3.1: Process unit design dimensions

Process Unit	Buffer vessel	Mixer	Extractor	Distillation column		Purger
				Case 1	Case 2	
Volume (L)	10,000	n/a	~10,000	~3000	~3300	n/a
Diameter (ft)	8.01	n/a	5	1.5	1.5	n/a
No. of stages	n/a	n/a	5	18	20	n/a
Tray efficiency	n/a	n/a	0.7	0.7	0.7	n/a
Stage spacing (ft)	n/a	n/a	1.64	2	2	n/a
Tangential height (ft)	7	n/a	18.28	59.14	65.71	n/a

4.4. Economics

With the sizing of each unit complete the CAPEX can be estimated. Table 4.4.1 presents each units predicted capital based on Guthrie's correlation from Douglas' (Eq. 3 & 4), Aspen provided values and predictions based off similar sized vessels.

Table 4.4.1: CAPEX for each unit and its constituents

Unit	Buffer	Mixer	Extractor	Distillation		Purger
				1	2	
Shell cost /\$	119,000	/	119,000	84,600	92,000	/
Trays cost /\$	/	/	6,700	2,900	3,200	/
Condenser cost /\$	/	/	/	55,000	54,800	/
Condenser acc. Cost /\$	/	/	/	110,000	110,000	/
Reboiler cost /\$	/	/	/	68,700	69,200	/
Reflux pump cost /\$	/	/	/	29,500	29,400	/
Total /\$	119,000	8,500 [21]	125,700	350,700	358,600	n.a

Table 4.4.2 provides the material and utility costs associated with the process.

Table 4.4.2: Utility and raw material costs

Cost of cooling water (\$/m ³)	6.46 [22]
Cost of HP steam (\$/kg)	0.0244 [22]
Electricity cost (\$/kWh)	0.0775
Cost of hexyl acetate (\$/kg)	4

Using equations 5 and 6 along with the material prices outlined in table 4.4.2 the OPEX of each plant design could be calculated and is presented in table 4.4.3. For the electricity charge, Aspen provided an electricity usage of 0.09 kW and a 20% ancillary service charge was accounted for.

Table 4.4.3: Summary of OPEX of both designs

OPEX	Yearly charge (\$/yr)	
	Case 1	Case 2
Raw materials (hexyl acetate)	160,000	160,000
Electricity	66.96	66.96
Condenser	61,100	163,000
Reboiler	6,300	17,800
Total	230,000	350,000

Finally, the revenue that can be expected is outlined in table 4.4.4.

Table 4.4.4: Summary of the revenue to be expected for both designs.

Design	Hexanoic acid production rate (kg/hr)	Hexanoic acid sales price (\$/kg)	Yearly revenue (\$/yr)	Ratio of yearly revenue to yearly solvent costs (-)
Case 1	6.1	2	97,500	0.609
Case 2	69.7	2	1,115,000	6.97

With all the cash flows accounted for a full economic analysis over the 25-year lifetime could be produced.

Presented in tables 4.4.5 and 4.4.6 are a summary of the expected cash flows for both case 1 and case 2 respectively.

Table 4.4.5: Summary of cash flows for case 1

Present Cash Flow	-\$1,600,000	Real AT
NPV	-\$2,350,000	Real AT
Present Value Cashflow	-\$2,200,000	Nominal AT
NPV	-\$3,000,000	Nominal AT
IRR	n/a	
ROI	-27%	

Table 4.4.6: Summary of cash flows for case 2

Present Cash Flow	\$4,000,000	Real AT
NPV	\$3,250,000	Real AT
Present Value Cashflow	\$6,300,000	Nominal AT
NPV	\$5,550,000	Nominal AT
IRR	103%	
ROI	68%	

Where AT stands for after tax, NPV for net present value, IRR for internal rate of return and ROI for return of investment. All calculations were done on excel and make use of the 'NPV' and 'IRR' functions within the programme.

Figures 4.4.1 and 4.4.2 present the real cashflows after tax for case 1 and 2 respectively.

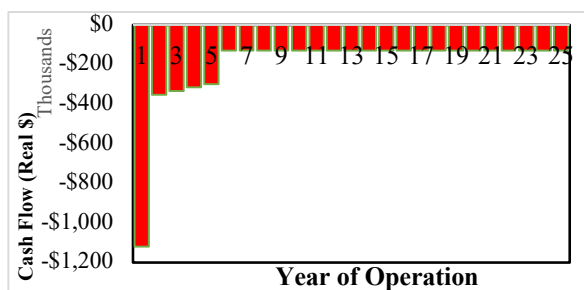


Figure 4.4.1: Real cash flow over plant lifetime for case 1

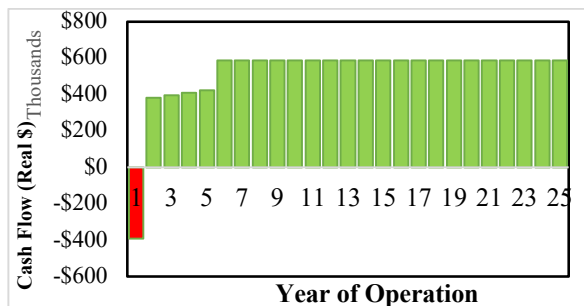


Figure 4.4.2: Real cashflows over plant lifetime for case 2

5. Discussion

There were several areas of research involved in the process design. Figure 4.1.1 presents the final flowsheet but does not show the finer details that went into its creation. First off, when simulating a process on Aspen, an accurate property model and Equation of State (EoS) must be selected to model how the systems components interact with one another. Furthermore, thought must go into the choice of solvent used to ensure an efficient separation and later recovery. Finally, the economic projection of each design must be accurately built to predict the cash flows over both plants' lifetime, whilst also considering its limitations. The areas of research considered are discussed below.

5.1. Property model

The simulation results are fundamentally linked to the choice of property model. For polar or highly non-ideal systems, the "dual approach" [23] is used to choose the property model. This means that an activity model for liquid phases is used in parallel with an EoS that characterizes vapour fugacity. This way, both LLE and VLE applications can be modelled.

After consulting guidelines for choosing property models that are supported by Aspen [24, 25], it was established that NRTL is the most suitable activity model.

NRTL (Non-Random Two-Liquid model) is widely used in process simulations and can be trusted to accurately characterize non-ideality and thus is appropriate to model a system of volatile fatty acids (VFA's) that have complex interactions due to their varying sizes and ability to hydrogen bond. The equation of the NRTL activity model, as developed by Renon and Prausnitz [26] is:

$$\ln \gamma_i = \frac{\sum_j x_j \tau_{ji} G_{ji}}{\sum_k x_k G_{ki}} + \sum_j \frac{x_j G_{ij}}{\sum_k x_k G_{kj}} \left(\tau_{ij} - \frac{\sum_m x_m \tau_{mj} G_{mj}}{\sum_k x_k G_{kj}} \right) \quad \text{Eq. 7}$$

$$G_{ij} = e^{-\alpha_{ij} \tau_{ij}} \quad \text{Eq. 8}$$

Aspen Plus predicts α and τ using a regression (Eq. 9 & 10) with parameters a, b, c, d, e, f, trained on the corresponding EoS databank. Ultimately, the values of these parameters direct the simulation results.

$$\tau_{ij} = a_{ij} + \frac{b_{ij}}{T} + e_{ij} \ln T + f_{ij} T \quad \text{Eq. 9}$$

$$\alpha_{ij} = c_{ij} + d_{ij}(T - 273.15) \quad \text{Eq. 10}$$

As discussed above, it is of great importance to also choose an adequate EoS. Hayden-O'Connell (HOC) equation of state is recommended for mixtures of carboxylic acids, as it can characterize the complex dimerization behaviour of short chain carboxylic acids (C2-C4) in the gas phase [27].

In order to validate the choice of NRTL-HOC and to have a base of comparison, it was decided to investigate other equations of state. At the moment, Aspen does not fully support the implementation of other adequate EoS's. This limitation is mitigated by outsourcing the VLE data via the state-of-the-art Clapeyron.jl package [28]. This open-source package provides access to numerous thermodynamic models, complementing Aspen. However, since it is still developing, Clapeyron.jl is lacking models for lesser-known chemical substances.

Other equations of state have been investigated: SAFT (namely PC-SAFT and SAFT γ -Mie variants) for their reported ability to handle a great variety of non-ideal compounds. Peng Robinson was also investigated, because it is also a widely used adaptation of Van Der Waals, this is less sophisticated but more user-friendly for simulations. SAFT γ -Mie is

also easy to implement thanks to the group contribution method making it applicable to any molecular structure.

Having created NRTL models paired with the above equations of state, the VLE data is manually regressed in Aspen to compute binary parameter values and subsequently plot VLE envelopes. Table 5.1.1 shows the predicted boiling points for each model. Some notable inconsistencies: PC-SAFT prediction for hexanoic acid is an outlier; SAFT gamma Mie model predicts abnormal boiling points for propionic and acetic acid (as well as acetate esters). This can be explained by the fact that the current implementation of this equation of state fails to account for the strong polarity of the methyl and methylene groups adjacent to the COOH group in short chain carboxylic acids [29]. Therefore, SAFT γ Mie was disregarded from further investigation. For the remaining models, there is a general good agreement, and the standard deviations are relatively low, thus the models are accepted.

Table 5.1.1: Predicted boiling points from each equation of state to test their validity.

Component	Pure component boiling points [K]				STD
	PR	PC-SAFT	SAFT γ Mie	HOC	
Lactic acid	485.571	-	476.76	490	5.5115
N-hexanoic acid	477.82	405.34	476.50	478.85	0.9611
N-hexyl acetate	443.82	444.25	444.26	444.65	0.3348
N-butyric acid	435.59	-	435.72	436.42	0.3410
Acetic acid	391.95	392.04	213.28	391.05	0.4560
Water	374.60	373.272	373.62	373.15	0.5805
Propionic acid	413.68	-	202.94	414.32	0.32
N-nonyl acetate	495.77	-	257.70	497.1	0.67

Isobaric (1 atm) VLE envelopes of each model for the relevant binary pairs have been superimposed and compared against NIST experimental data where available (figures 5.1.1 to 5.1.7).

For some pairs, there is agreement between the different models (figures 5.1.1, 5.1.2), the diagram having similar shapes, and some curves even looking identical. For other pairs (figure 5.1.4) the NRTL-HOC model looks much different not only in terms of position in the Txy space but also in terms of convexity of the curves, which could be explained by the fact that only

HOC EoS can characterize acetic acid's dimerization behaviour.

In the cases where the NRTL-HOC model predicted a visibly different VLE envelope, the HOC model had better agreement with experimental data (figures 5.1.3, 5.1.5, 5.1.7). Mean square error analysis yielded relatively low values: 0.018, 0.034, 0.013, respectively, showing that the NRTL-HOC model is a good fit for mixtures containing acetic and butyric acids.

The influence of HOC on short chain acids is validated. Although there is not enough empirical evidence to totally validate NRTL-HOC, it is the most suitable approach currently.

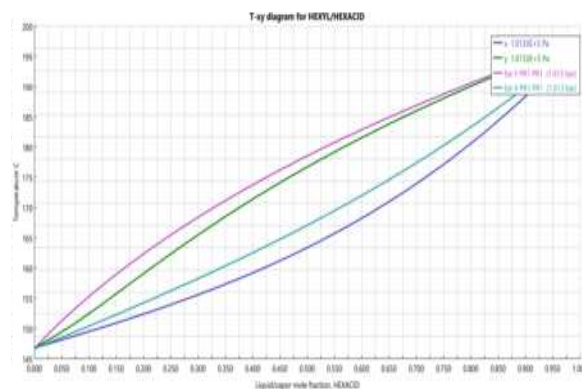


Figure 5.1.1: Hexyl acetate - hexanoic acid binary pair Txy VLE for HOC and PR models superimposed.

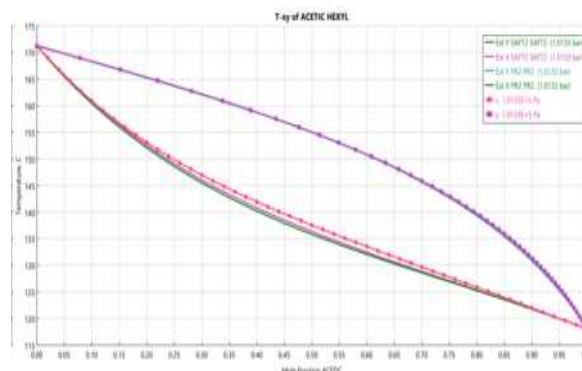


Figure 5.1.2: Hexyl acetate - acetic acid binary pair Txy VLE for HOC, PC-SAFT, PR models superimposed.

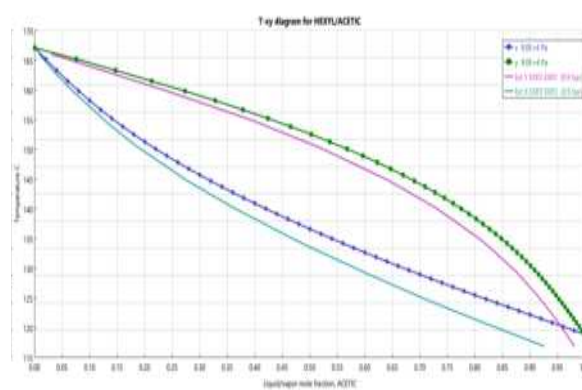


Figure 5.1.3: Hexyl acetate - acetic acid binary pair Txy VLE. Experimental data plotted against NRTL-HOC model at 0.9 bar.

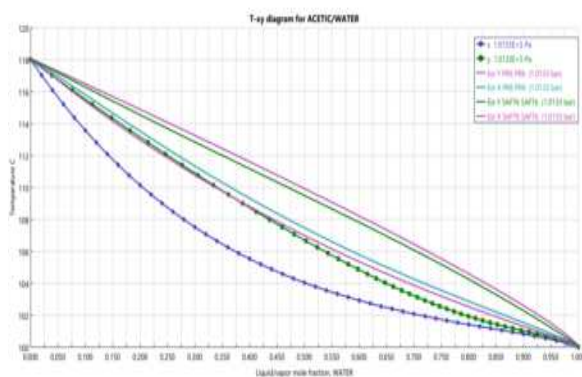


Figure 5.1.4: Acetic acid – water binary pair Txy VLE for HOC, PCSAFT, PR models superimposed.

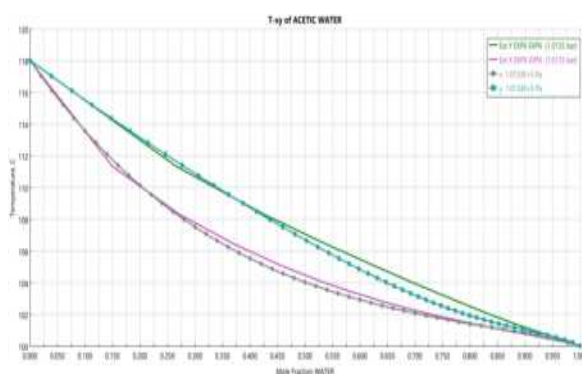


Figure 5.1.5: Acetic acid – water binary pair Txy VLE. Experimental data plotted at atmospheric pressure plotted against NRTL-HOC model.

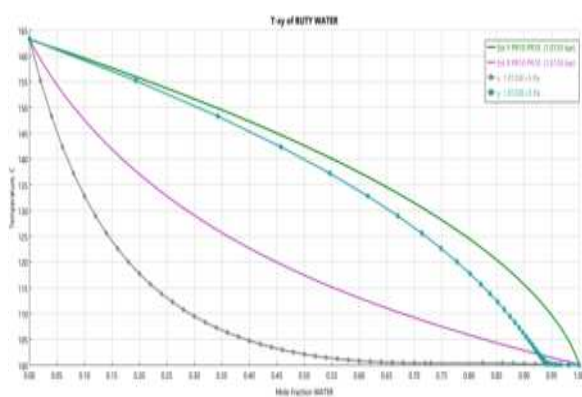


Figure 5.1.6: Butyric acid – water binary pair Txy VLE for HOC and PR models superimposed.

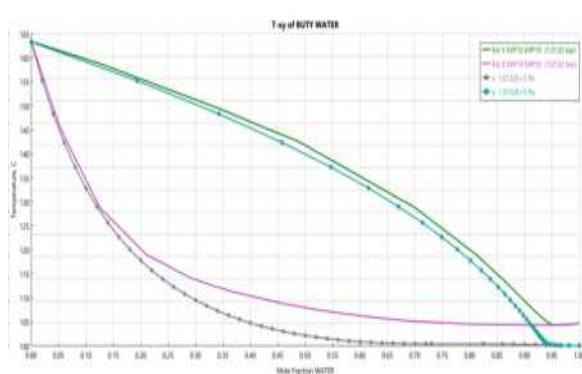


Figure 5.1.7: Butyric acid – water binary pair Txy VLE. Experimental data at atmospheric pressure plotted against NRTL-HOC model.

5.2. Solvent selection

As previously stated, hexyl acetate was selected as the solvent for the LLE extractor. Solvent selection is an important aspect of any design as it determines the recovery of your desired product. Several aspects need to be considered when selecting your solvent including, its distribution coefficients, its extraction efficiency, and its ability to be recovered.

All 3 solvents demonstrated high distribution coefficients allowing for almost 100% recovery of hexanoic acid as shown in table 4.2.1. Although, the TOA-octanol solvent had a slightly lower recovery of hexanoic acid, it displayed far better extraction efficiency due to its much lower solvent to feed ratio. This would suggest that this is the best choice, however when it came to recovering the solvent it displayed poor separation from the hexanoic acid. This is due to the strong complexation that occurs as a part of the reactive extraction [30]. Furthermore, as octanol is considered an active diluent, it is able to hydrogen bond with carboxylic acids which contributes to the improved extraction yet also leads to worse solvent recovery. Poor solvent recovery leads to a lower purity product as well as a higher raw materials cost and thus the TOA-octanol was disregarded.

Nonyl acetate was also disregarded for a similar reason, when the solvent recovery column was designed it appeared that this solvent had a very high affinity for hexanoic acid [11] and therefore presented a challenging separation that would require multiple distillation columns and a high capital investment. The poor recovery could also be a result of the higher levels of water leeching associated with nonyl acetate.

Hexyl acetate however demonstrated efficient extraction and was easily recovered downstream and thus it was chosen as the solvent.

5.3. Economics

As is immediately evident from figures 4.4.1 and 4.4.2 along with tables 4.4.5 and 4.4.6, case 1 is not a financially acceptable endeavour. At no point in the entire lifetime are any profits to be expected. This can be explained by the ratio of yearly revenue to yearly solvent costs as presented in table 4.4.4. This is a ratio of income to expense and thus the only way a profit would be seen is if this value is >1 . Furthermore, this ratio does not account for capital and utility costs and thus the fact that for case 1, with a value of ~ 0.6 , there is no opportunity to make a profit even though it has a lower utility requirement and thus operating costs.

Case 2 however is much more promising and shows that fermentation of the pot ale syrup is necessary for the plant to be profitable. In the first year due to capital expenditures a loss of roughly \$400,000 can be expected but in subsequent years a profit can be expected of roughly \$600,000 a year, after the P&I payments have been repaid in full. Over the whole lifetime an ROI of 68% and IRR of 103% can be

expected for this specific design, and therefore it is considered a worthwhile investment.

As mentioned, case 1 had a lower associated OPEX at roughly \$230,000 a year compared to \$350,000 a year for case 2, as shown in table 4.4.3. This is due to the lower utility requirement of case 1. The distillation column designed was able to reach the same desired recovery and purity of hexanoic acid as case 2 with a lower reflux ratio. This results in lower reboiler and condenser heat duties being observed, and thus less cooling water and high-pressure steam is needed. The distillation column is the only process unit that require utility and is the only source of the varying OPEX. Similarly, it is the only source of a difference in CAPEX between designs. For case 1, the column could be built with slightly fewer stages and thus is smaller and has a lower installed shell and tray costs. Again, as all the other units like extractor and buffer vessel had the same design between cases, their CAPEX remains constant between the two scenarios. This only results in a column roughly \$10,000 cheaper to build but as stated above is inconsequential to the suitability of case 1 due to the ratio of yearly income vs solvent costs being <1 .

There are some limitations to the economic projection presented that result in an overestimate of cashflows. First off, due to a lack of necessary data, a full cash cost of production could not be estimated. Fixed costs of production such as labour or maintenance costs were unknown and thus left out of the model (some fixed costs such as taxes were included). The operating expenditures presented are the variable costs of production such as utility and raw material prices. Furthermore, the cost of the fermentation unit associated with case 2 has not been included as this was designed by the BDC in York, UK and details of its design are protected under a non-disclosure agreement. All of this will result in an actual ROI and IRR lower than what has been presented. However, due to their magnitude it is not expected that these extra expenses will result in a negative return. Typically, annual ROIs of 7% or higher are considered a good investment [31].

6. Conclusion

This report designed two scenarios of a separation route for hexanoic acid from bio-based distillery waste and conducted a techno-economic analysis to support design choices. Research was conducted to validate simulation results and to identify n-hexyl acetate as the best solvent for liquid-liquid extraction due to its efficient capability at extracting hexanoic acid as well as later recovery from said acid. After having planned a preliminary flowsheet, simulations were run numerous times to find and obtain the optimal design specifications: number of stages in the separation units, reflux ratio in RADFRAC, percentage of purge stream, solvent make-up and recycle configuration. The equipment was mapped according to industry standards, either provided by Aspen or discovered in literature, and costing was computed through correlations commonly used in

academia. NRTL-HOC was found as the current most suitable property model available on Aspen.

The results clearly show that case 2 is more financially sensible, producing larger amounts of hexanoic acid at a negligible increase in both capital and operating expenses. This is the profitable design, with an estimated rate of return of 68% over the plant's 25-year lifetime. A negative net cash flow will be seen in the first year due to the initial capital investment however this is offset in subsequent years as the profits substantially outweigh the expenses. Real-life implementation of this process would provide whisky distilleries with a diverse income stream and improve on their financial growth as it converts an underutilised waste product into a valuable throughput.

Going forward the design can be further developed to generate a more accurate economic model. The fixed costs of production such as labour or maintenance should be accounted for as this will give a better representation of cash flows. Furthermore, better estimates are required for the mixing and purging units as currently their capital is either not accounted for or a rough estimate. For the process simulation on Aspen, SAFT type property models should theoretically predict more accurate component interactions. However, before this can be implemented, a wider range of components need to be modelled within its database – this can be done on the property package, Claapeyron.jl. Finally to just improve the process as a whole, it could be redesigned to also isolate the other fatty acids present in the fermentation broth and thus generate a more diverse income stream.

Acknowledgements

The Authors of this report would like to extend their utmost gratitude to Dr. Andrea Bernardi and Mr. Ben Lyons for their continued support and guidance. To Dr. Andrew Haslam for their assistance in understanding property models and Claapeyron.jl and a special thanks to Mr. Malcolm Barraclough and to all those working at the BDC.

References

1. Brad Neathery. *Oak & Eden* (blog), n.d. <https://oakandeden.com/blogs/journal/when-was-whiskey-invented-a-brief-history>.
2. "Facts & Figures." Scotch Whisky Association. Accessed December 27, 2023. <https://www.scotch-whisky.org.uk/insights/facts-figures/>.
3. "Information." Whisky.com. Accessed December 27, 2023. <https://www.whisky.com/whisky-sector.html>.
4. White, Jane S., Kelly L. Stewart, Dawn L. Maskell, Aboubakry Diallo, Julio E. Traub-Moder, and Nik A. Willoughby. "Characterization of Pot Ale from a Scottish Malt Whisky Distillery and Potential Applications." *ACS Omega* 5, no. 12 (2020):

- 6429–40.
<https://doi.org/10.1021/acsomega.9b04023>.
5. Michael Goldsworthy, and Lucie A Pfaltzgraff. Rep. Edited by Adrian Higson. *The Treatment of Pot Ale – a New Process Approach*, n.d.
6. Cavalcante, Willame de, et al. “Anaerobic Fermentation for N-Caproic Acid Production: A Review.” *Process Biochemistry* 54 (2017): 106–19.
<https://doi.org/10.1016/j.procbio.2016.12.024>.
7. Barraclough, Malcolm. *A new process to treat Pot Ale. The principal effluent by-product of the whisky industry*. n.d.
8. Mohana, Sarayu, Bhavik K. Acharya, and Datta Madamwar. “Distillery Spent Wash: Treatment Technologies and Potential Applications.” *Journal of Hazardous Materials* 163, no. 1 (2009): 12–25.
<https://doi.org/10.1016/j.jhazmat.2008.06.079>.
9. López-Garzón, Camilo S., and Adrie J.J. Straathof. “Recovery of Carboxylic Acids Produced by Fermentation.” *Biotechnology Advances* 32, no. 5 (2014): 873–904.
<https://doi.org/10.1016/j.biotechadv.2014.04.002>.
10. Saboe, Patrick O., et al. “In Situ Recovery of Bio-Based Carboxylic Acids.” *Green Chemistry* 20, no. 8 (2018): 1791–1804.
<https://doi.org/10.1039/c7gc03747c>.
11. Woo, H.C., Kim, Y.H. Eco-efficient recovery of bio-based volatile C2–6 fatty acids. *Biotechnol Biofuels* 12, 92 (2019).
<https://doi.org/10.1186/s13068-019-1433-8>
12. Magdah Abdelbasit Nory Salih. “LIQUID-LIQUID EQUILIBRIUM FOR THE DESIGN OF EXTRACTION COLUMN.” *European Journal of Engineering and Technology* 5, no. 4 (2017).
13. Lane, Ryan, and Randa Kriss. “Average Business Loan Rates: December 2023.” NerdWallet, December 1, 2023.
<https://www.nerdwallet.com/article/small-business/small-business-loan-rates-fees>.
14. Towler, Gavin P., and R. K. Sinnott. *Chemical Engineering Design: Principles, practice and economics of plant and Process Design*. Oxford England: Butterworth-Heinemann, 2022.
15. “Inflation and Price Indices.” Inflation and price indices - Office for National Statistics. Accessed December 27, 2023.
<https://www.ons.gov.uk/economy/inflationandpriceindices>.
16. Corporate - Taxes on corporate income. Accessed December 27, 2023.
<https://taxsummaries.pwc.com/united-kingdom/corporate/taxes-on-corporate-income#:~:text=General%20corporation%20tax%20rates,in%20excess%20of%20GBP%205.000>.
17. Douglas, James M. *Conceptual design of Chemical Processes*. New York, NY: McGraw-Hill, 2000.
18. “Inventory Index Factors for 2017.” Marshall and Swift Valuation Service, 2017.
19. Chapter 22 (p558-597), CH EN 4253, Terry A. Ring
20. “Hexanoic Acid Price, Buy Hexanoic Acid.” chemicalbook. Accessed December 27, 2023.
<https://www.chemicalbook.com/Price/Hexanoic-acid.htm>.
21. “Industrial Mixers.” Brookfood. Accessed December 27, 2023.
<https://www.brookfood.co.uk/used-equipment/mixing/industrial-mixers>.
22. “Industrial Steam Cost: Industrial Utilities.” Industrial Steam Cost | Current and Forecast | Intratec.us. Accessed December 27, 2023.
<https://www.intratec.us/products/water-utility-costs/commodity/industrial-steam-cost>.
23. Apen Tech, Inc. “Appendix A.” In HYSYS Simulation Basis, n.d.
24. Carlson, E.C. (1996). Don't Gamble With Physical Properties For Simulations. *Chemical Engineering Progress*, 92, 35-46.
25. Aspen Technology, Inc. “Aspen Plus V11 Help,” 2019.
26. Renon H., Prausnitz J. M., “Local Compositions in Thermodynamic Excess Functions for Liquid Mixtures”, *AIChE J.*, 14(1), S.135–144, 1968
27. Hayden, J. George, and John P. O’Connell. “A Generalized Method for Predicting Second Virial Coefficients.” *Industrial & Engineering Chemistry Process Design and Development* 14, no. 3 (1975): 209–16.
<https://doi.org/10.1021/i260055a003>.
28. Walker, Pierre J., Hon-Wa Yew, and Andrés Riedemann. “Clapeyron.Jl: An Extensible, Open-Source Fluid Thermodynamics Toolkit.” *Industrial & Engineering Chemistry Research* 61, no. 20 (2022): 7130–53.
<https://doi.org/10.1021/acs.iecr.2c00326>.
29. Sadeqzadeh, Majid, et al. “The Development of Unlike Induced Association-Site Models to Study the Phase Behaviour of Aqueous Mixtures Comprising Acetone, Alkanes and Alkyl Carboxylic Acids with the Saft-γ Mie Group Contribution Methodology.” *Fluid Phase Equilibria* 407 (2016): 39–57.
<https://doi.org/10.1016/j.fluid.2015.07.047>.
30. Sprakel, L.M.J., and B. Schuur. “Solvent Developments for Liquid-Liquid Extraction of Carboxylic Acids in Perspective.” *Separation and Purification Technology* 211 (2019): 935–57.
<https://doi.org/10.1016/j.seppur.2018.10.023>.
31. Birken, Emily Guy. “Return on Investment (ROI).” *Forbes*, May 12, 2023.
<https://www.forbes.com/advisor/investing/roi-return-on-investment/>.

Enviro-Economic Assessment of a Scaled-Up Hydrogenolysis Process for the Treatment of Polypropylene Waste

João Costa and Diketsi Karas

Department of Chemical Engineering, Imperial College London, UK

Abstract

The rate at which plastic waste is accumulating in landfills is posing a significant threat to ecosystem and human health. The use of chemical recycling for the treatment of post-consumer plastics has become increasingly popular, including hydrogenolysis, favoured for its mild operating conditions and valuable products. In this report, a hypothetical industrial hydrogenolysis process is designed and analysed. Aspen HYSYS is used to design and model a converged flow-sheet before both a techno-economic analysis (TEA) and a life cycle assessment (LCA) are carried out. Both assessments help determine the feasibility of the scaled-up process and how its performance might compare to current end-of-life pathways, including other chemical recycling technologies. The results demonstrate that the scaled-up hydrogenolysis process is profitable overall with a positive NPV of \$120.5 million. A cost of \$0.46/kg of polypropylene feed and a break-even price of \$0.74/kg of polypropylene were determined. The LCA demonstrates that the proposed design has a significantly lower environmental impact than current recycling processes, particularly concerning human health and resource depletion. However, certain areas of the process require investigation and improvement, from the high capital cost to the low readiness level and uncertainty surrounding the performance at scale. These limitations are discussed in the paper.

Keywords: *Hydrogenolysis, Polypropylene, Chemical Recycling, Circular Economy*

1 Introduction

Annually, approximately 400 million metric tonnes of synthetic polymers are produced worldwide [1] with polyolefins such as polyethylene (LDPE) and polypropylene (PP) accounting for roughly 60% of the global plastics [2]. These polyolefins have low recycling rates due to strong C–C bonds which make them difficult to break down [2]. Primary recycling of plastics typically occurs through mechanical recycling which despite being comparatively cost-effective, results in the contamination and degradation of the plastic (downcycling) making it ultimately unsustainable [3]. In any case, currently only approximately 18% of plastics are recycled, leaving 24% to be incinerated and 58% sent to land-fill or discarded [1]. The continuous production and accumulation of plastic waste results in a sustained loss of resources and poses a serious threat to the environment and human health; for instance through the formation of harmful microplastics which are dangerous both when ingested or upon entering ecosystems[4].

As a result, chemical recycling methods have gained attention for their ability to transform polyolefins into valuable products that can then be re-integrated into a range of industrial processes [5]. Thermal cracking and pyrolysis are currently popular examples of chemical recycling but are limited by their high operating temperatures (400 °C to 900 °C) and poor product selectivity [5]. Hydrogenolysis is promising as it not only produces high-value products but has been shown to operate in a milder temperature range of 200 °C to 300 °C [6].

Hydrogenolysis uses a metal catalyst and high-pressure

hydrogen to cleave the C–C bonds in the polymer chains to break them down into shorter hydrocarbons. In the case of polypropylene, alkanes of varying lengths are formed. Catalyst selection is therefore crucial in process design and determines reaction conditions, kinetics and degradation time [7]. Noble metal catalysts are particularly favourable for hydrogenolysis, particularly platinum (Pt) and ruthenium (Ru) based catalysts [6]. Pt-based catalysts perform very well in thermal cracking reactions but fail to effectively break the C–C bonds at low reaction temperatures without added acid sites [7]. Ruthenium catalysts are currently being investigated as a cheaper alternative, operating at 200 °C to 250 °C whilst still producing promising distributions of useful products. The Rosseinsky catalyst group at the University of Liverpool demonstrated that a Ru/CeO₂ catalyst could produce large yields of useful alkanes from polypropylene waste, with an enhanced selectivity towards liquid alkanes. Thus, suppressing excess methane generation which to date has been a common problem of plastic hydrogenolysis [6].

This report aims to take hydrogenolysis at its low readiness level and to assess its viability as a plastic treatment process. To achieve this, the process is scaled up to meet pre-defined plastic treatment goals by extrapolating experimental data to create a flow-sheet in Aspen HYSYS V11. The approach in this report is “best-case” and uses the highest conversion achieved at laboratory-scale conditions that is realistic at scale. The distribution of products is modelled based on the laboratory data from the Rosseinsky group at the University of Liverpool [6], supplemented by data from a literature review. A comprehensive techno-

economic assessment (TEA) is carried out using data from both literature and the HYSYS model, and a comparative life cycle analysis (LCA) is carried out to analyse the environmental impact of the hypothetical industrial process.

Both analyses aim to evaluate the feasibility and sustainability of the final flow-sheet. In the TEA, capital expenditure (CAPEX) and operating expenditure (OPEX) values are calculated to determine the Net Present Value (NPV) and investigate the profitability of the process. The cost to produce 1kg of polypropylene and the polypropylene break-even price is also calculated in the economic assessment. The CAPEX is then compared to other chemical recycling technologies to assess the initial costs. In the LCA analysis, the ReCiPe 2016 method [8] is used in OpenLCA to compare hydrogenolysis to the common recycling methods of incineration and landfill, assessing the performance of each process against common environmental indicators and identifying key problem areas. This paper aims to lay out a comprehensive model whilst acknowledging the limitations that call for further research; suggestions for intermediate investigations and pilot-scale experiments are laid out throughout and summarised in the outlook.

2 Background

With growing discourse on the advantages of chemical upcycling and its ability to produce various value-added products [3], hydrogenolysis has gained attention as a route to transform plastic waste into valuable chemical feedstocks, contributing to a circular economy. Hydrogenolysis competes with several types of chemical recycling at different stages of development. Some methods are non-catalytic, generally operating at higher temperatures, these include pyrolysis, hydrothermal liquefaction and gasification [9]. Others also involve catalysts, for example, hydrocracking, which operates at lower temperatures and uses hydrogen at high pressure and a noble metal catalyst, making it similar to hydrogenolysis [10]. Crucially, hydrogenolysis can be conducted under relatively mild conditions, minimising energy consumption and associated environmental impacts in comparison to other chemical recycling methods [11].

Recent literature on the hydrogenolysis of plastic waste predominantly involves laboratory-scale experiments exploring catalytic mechanisms and assessing their impacts on the hydrogenolysis process [7]. They also explore the optimisation of laboratory conditions, comparing the effectiveness of the different set-ups [12]. However, there is limited information regarding the scalability of the reaction mechanism and no existing design of an industrial hydrogenolysis process.

Despite its promise, the hydrogenolysis of plastic is complex and challenging. Catalyst development, selectivity of the reaction towards desired products and the behaviour at scale are among the key areas for investigation and improvement. Additionally, understanding the environmental and economic impact of a large-scale hydrogenolysis process are vital for the widespread adoption of this technology and for informing the policy and investment decisions that must be made with the implementation of a novel technology.

One research group working on the development of hydrogenolysis is the Rosseinsky catalyst group at the University of Liverpool which in 2023 conducted a study on hydrogenolysis of polypropylene at a laboratory scale under

batch conditions [6]. Results from the cited paper were supplemented with literature data to form the foundations for this report. The Rosseinsky group carried out a range of tests on polypropylene hydrogenolysis involving different catalysts and varying temperatures to investigate conversion and catalytic properties. Data from these experiments are used in the modelling in this report, taking into account that the process is at a low readiness level.

3 Methodology

3.1 Process Design

3.1.1 Overall Design

The basis for the analysis was the flow-sheet developed on Aspen HYSYS which models an industrial-scale hydrogenolysis process. Some basic parameters were derived from laboratory data, such as the conversion (90%) and the process conditions within the reactor (220°C, 30 bar Hydrogen) [6]. Specific design goals were established, including the ability to process 25 kilo-tonnes (kt) of polypropylene yearly to align with a comparable assessment of chemical recycling technologies [9], 8000 hours of operation based on guidelines for life-cycle cost analysis [13] and purity of useful product of above 95% as a first pass.

A hypothetical solid was modelled in HYSYS to represent polypropylene, defining the molecular weight, the density and the heat of formation from the Polymer Handbook [14]. The products were grouped according to standard crude oil fractions and the midpoint properties of the class were used to define the class, based on a similar methodology found in previous reports [15]. The Rosseinsky group paper contains product yield ranges for the hydrogenolysis reaction [6] which were used to estimate the product distribution, finding average product yields for the conditions selected. Additional data was obtained from the Rosseinsky group to facilitate the estimation of the product splits and this was verified against data from previous hydrogenolysis experiments such as that carried out by Wang *et al.* [2]. The distribution used is summarised in table 1. The Peng-Robinson fluid package was chosen for its general accuracy in determining phase equilibria for a range of substances given the phase transitions in our reactor, as well as its wide usage in the oil and gas industry [15].

Table 1: Chosen product splits for the hydrogenolysis reactor estimated from literature data

Product	Percentage
Light Gases	22.0%
Gasoline	13.2%
Kerosene	19.8%
Diesel	29.8%
Waxes and Lubricants	15.3%

3.1.2 Reactor Design

The reactor is modelled as a simple conversion reactor in HYSYS as in similar studies [9]. The kinetics of the reaction have not been investigated so the reaction equations modelled on HYSYS were designed to represent the distribution

of products found in the literature. Hydrogen consumption was estimated from the equivalent stoichiometric amounts of hydrogen needed for the separation into each product class. By creating a representative balanced stoichiometric reaction in HYSYS the reactor was designed to consume reactants and generate products at ratios proportional to those at the laboratory scale, providing a first-pass prediction for reaction behaviour. For costing purposes, the reactor was sized as a gasification reactor, which has previously been applied in waste treatment [16], as laid out in section 3.2. This provides an order of magnitude estimate for the cost, but would undoubtedly be further refined in advanced reactor design.

Note that several types of reactors could be considered in this process. Agitation, even heat distribution, ability to withstand pressure increases and ability to handle viscous mixtures are key requirements. Spinning basket or screw feeder reactors are commonly used in high viscosity applications [17], whereas a fluidised bed would provide effective contact between the gaseous and solid reactants. Further knowledge of kinetics would inform whether a cascade is desirable. Currently, the data available is insufficient to ascertain which of these options would be ideal. Specific reactor design is therefore not only beyond the scope of this report but it also not necessarily useful given that any design would largely be lacking the core engineering data required, it is, therefore, important to emphasise these knowledge gaps to work towards a robust intermediate scale, this is further addressed in the discussion (Section 4.1.1).

3.1.3 Separation Design

The order of the distillation train is based on standard heuristics to minimise the difficulty of each separation. Light gases are separated off first as the difference in boiling point is the largest. The difference between gasoline to kerosene and diesel to lubricants is small, but gasoline has a molar flow of approximately three times that of the lubricant flow, so this separation is prioritised, subsequent separation order is also based on boiling point differences. Aspen HYSYS is used to design each column, using the shortcut column to estimate the column sizing before fine-tuning it in the flow-sheet. The final separation train is presented in Appendix 1.

A component splitter is used to remove the remaining solid polypropylene after the reactor. By decreasing pressure and temperature after the reactor, most of the hydrogen produced can be flashed out, which decreases the overall energy requirement and makes the separation train easier to converge. A component splitter is also used to model a pressure swing adsorption (PSA) in the HYSYS flow-sheet as HYSYS can not handle non-steady state operations, this PSA is modelled to have 82.5% separation of Hydrogen as the midpoint of the range for typical hydrogen PSA [18].

3.1.4 Recycle and Purge and Conditions

Both the hydrogen and polypropylene were recycled to minimise waste and improve the economic and environmental feasibility of the process. Purge streams were added to both recycle loops to mitigate the build-up of impurities, maintain process efficiency and ensure the quality of products. The hydrogen obtained after the PSA was of high purity, thus a

purge of 1% was deemed sufficient, whereas due to the large uncertainty of the quality and nature of the polypropylene recycle, a purge of 10% was used as an estimate.

3.2 Heat Exchanger Network

Conditions are moderated in the process using compressors, turbines, coolers and heaters to make separations easier and to meet the reaction conditions. These were designed and implemented in HYSYS. Aspen Energy Analyzer is used to analyse net heat and cooling duty and to design a heat exchanger network using Aspen's in-built utilities. Multiple designs are simulated in Aspen Energy Analyzer but the design with the biggest energy savings and lowest cost is chosen. Conventional and readily-available utilities are prioritised.

3.3 Techno-Economic Analysis

3.3.1 Costing and Economic Analysis

A techno-economic analysis (TEA) is used to assess the feasibility of the hydrogenolysis flow-sheet. Aspen Economic Analyzer V11 is used to obtain CAPEX values for all process equipment besides the PSA and the reactor which are sized and costed using methods found in literature. The PSA unit is sized and costed as a packed bed pressure vessel using the Guthrie method [19]. The reactor is costed with a gasification reactor correlation which is scaled based on the dry solid feed to the reactor [20]. This method is chosen because the solid polypropylene reactor feed rate can be modelled more accurately than the residence time, favouring this correlation over others considered. A comprehensive description of the sizing and costing of each unit and associated economic assumptions can be found in Appendix 3.

The catalyst Ru/CeO₂, from the Rosseinsky group paper, is difficult to cost due to the catalyst being prepared in the laboratory and therefore not being directly purchasable. In the experimental set-up of the Rosseinsky paper, the catalyst mass is 5% of the polymer mass [6], it is uncertain how this would scale or how the specialised catalyst could be produced industrially. For estimation, a cost heuristic was used based on a previous TEA [9], assuming that the ratio of the catalyst cost would be comparable given the similarity of the experiment which used ruthenium on platinum/tungstated zirconia as the catalyst in the hydrogenolysis reactor, operating at 250°C and 30 bar. This allows for an estimation of the proportional cost of the catalyst despite the significant uncertainty surrounding the eventual industrial catalyst. Using the cost correlation, the catalyst is calculated to be 18% of the total reactor cost.

All capital costs are updated using the Chemical Engineering Plant Cost Index (PI) [21]. For data acquired from HYSYS and for the PSA unit, the PI from 2019 to 2023 is used, whereas for the reactor the PI from 2014 to 2023 is used to align with the current economic market. OPEX values are also derived from HYSYS, thereby taking into account the heat integration results and utilities. The remaining capital costs and annual costs are calculated as functions of the HYSYS and custom data. These include contingency fees, labour, depreciation, taxes, insurance, general administration, research and development. The labour costs are calculated assuming an average salary

of \$65,000 [22] and that the facility runs on a four-shift schedule with one set of workers per shift. The number of workers required to operate the plant is determined using Perry’s coefficient [23]. An in-depth calculation for labour and other annual costs can be found in Appendix 4.

3.3.2 Profitability Analysis

The profitability of the process is assessed through the calculation of the Net Present Value (NPV), a positive NPV suggests the investment is financially viable and anticipated to yield a higher return than its expenses. A cost per kilogram of polypropylene feed is also calculated, using CAPEX, OPEX and total kilograms of solid polypropylene fed to the system annually. To determine the price at which polypropylene must be purchased to attain a Net Present Value of zero, the break-even point for the chosen design is investigated. The break-even point and price per kilogram are important indicators that enable the comparison of hydrogenolysis to similar technologies for the treatment of polypropylene.

3.3.3 Chemical Recycling Comparison

There is inconclusive information regarding which recycling method has the most overall promise, but some initial cost indicators have been calculated. HYSYS CAPEX values taken from the supplementary material of the Hernandez *et al* report [9] are used to evaluate hydrogenolysis relative to emerging “competitors”. From there, the same capital cost correlations used on the hydrogenolysis data are applied to the literature data to ensure consistent results. Understanding the upfront investment required helps assess whether the project is financially viable and whether it can ultimately generate a satisfactory return on investment whilst also highlighting areas for potential improvement.

3.4 Life-Cycle Assessment

LCA is employed in this study to assess the environmental impact of the process designed, particularly relative to similar plastic recycling processes. The ReCiPe 2016 Endpoint method [8] is used on OpenLCA to quantify the impacts of a given process using 18 indicators which are subsequently grouped into three impact areas: damage to human health, damage to ecosystem quality and damage to resource availability. The functional unit is 1 kg of polypropylene processed and the scope of the study was end-of-life treatment. Data is obtained from the comprehensive Ecoinvent v3.6 database [Ecoinvent]. Results from the LCA are analysed to compare different waste treatment methods to identify ‘hotspots’ of environmental impact and therefore areas for improvement. Using only data from Ecoinvent minimised issues around data integration.

The ReCiPe method can be applied through three ‘cultural perspectives’, the ‘hierarchist’ framework was selected, as it is commonly encountered in similar scientific models mainly due to it aligning closely with the timescale of policy development processes. Ecoinvent does not contain a process flow for hydrogen production such that water electrolysis was simulated using water and electricity inputs. The rest of the inputs and outputs for this process including utilities can be found in the Ecoinvent database. Incineration and landfill of polypropylene are also modelled

in OpenLCA by adapting processes defined on the Ecoinvent database for validation and relative impact analysis.

For comparability, the impacts are converted from the endpoint units in the ReCiPe method into equivalent monetary values using the externalities monetisation method laid out in Dong *et al.* [24]. The Dong *et al.* paper converts the metrics to 2003 Euros, such that the metrics are first converted to 2003 US dollars using data from the OECD [25] and subsequently converted to 2023 US dollars using data from the Bureau of Labour Statistics [26].

Table 2: Conversion factors from ReCiPe environmental impact units to \$₂₀₂₃

Impact Category	Result	€ ₂₀₀₃	\$ ₂₀₂₃
Human Health	DALY	7.40×10^4	1.40×10^5
Ecosystem Quality	Species.yr	9.50×10^6	1.79×10^7
Resource Availability	\$ ₂₀₁₃	N/A	1.32

4 Results and Discussion

4.1 Process Design

4.1.1 Process Overview

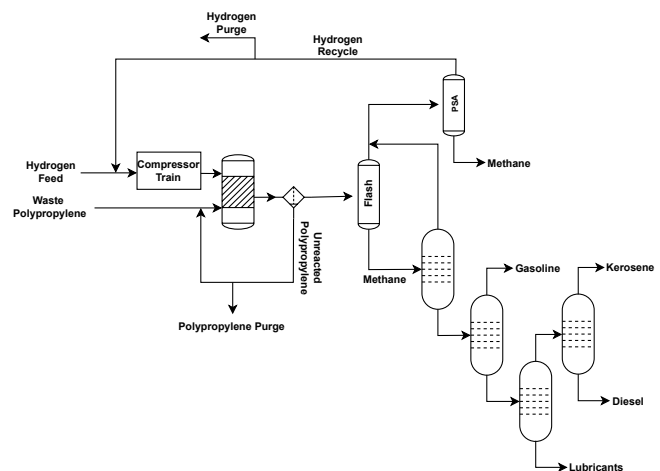


Figure 1: Process flow diagram of designed hydrogenolysis process for treatment of polypropylene waste

The process flow diagram in Figure 1 depicts the main functions of the designed process as described in Section 3, a complete flow-sheet can be found in Appendix 1. Table 2 contains the key process indicators obtained from the HYSYS flow-sheet. The high conversion and carbon efficiency are a result of the small purges, a high reaction conversion and all the products being useful alkanes, however, it is important to emphasise here that this is a best-case scenario and it is likely that in the final process, there will be more impurities, that the splits will not be as ideal and that the catalyst will vary, all of which would lower these indicators and would have to be accounted for in future iterations. At this stage, a significant error margin is expected, the priority of this work is to develop a framework that can be fine-tuned with more inputs as described in the

introduction. Note also that the natural gas purity is lower due to the presence of the hydrogen from the incomplete separation in the pressure-swing adsorption, the further processing of products was beyond the scope of this project but this stream could either be further separated or sold as low-quality natural gas.

Table 3: Key process technical indicators of hydrogenolysis of polypropylene waste

Indicator	Result
Polypropylene Conversion	99.5%
Carbon Efficiency	98.6%
Work Requirements	0.383 MW
Energy Consumption	3.94 MW
Natural Gas Purity	80.2%
Gasoline Purity	99.9%
Kerosene Purity	99.9%
Diesel Purity	99.9%
Lubricants Purity	99.6%

Before implementation at any scale nearing the hypothetical scale presented in this paper, extensive research must be conducted into several aspects of the process. More information on kinetic data and the reaction mechanism is required to predict behaviour at scale. In addition, the catalysts used by the Rosseinsky group are shaped catalysts produced on-site using raw materials obtained from Sigma Aldrich [6], the final catalysts are therefore not available commercially so catalyst preparation would have to be adapted to a large-scale supply chain. Also, more information is required regarding the material properties within the reactor which are especially relevant given that there is a solid-to-gas-and-liquid transition and that the reaction temperature is above the melting temperature of polypropylene [14]. The current residence time at laboratory-scale conditions investigated in this paper is 16 hours [6], although this would change at scale with optimisation and catalyst improvement, it introduces significant uncertainty which should be addressed by kinetic modeling and pilot-scale experiments. Dichloromethane, which has a range of associated health risks [27], is often used to clean the vessel in the laboratory post-reaction and would have to be substituted before implementation at scale. In summary, experiments at laboratory scale and pilot scale are needed to provide a more comprehensive view of the scale-up behaviour which will enable realistic process design.

The data obtained are important as order-of-magnitude predictions are a key objective of this project. However, despite HYSYS allowing flexibility in design, time constraints meant that there was a limit on the number of configurations that could be tested, for example, a full sensitivity analysis of this design or future designs could help identify improvement areas. There are also some limitations to this software, although HYSYS produces a robust high-level model, other modelling methods such as computational fluid mechanics or density functional theory could be used to give a more granular understanding of the molecular-level mechanisms of hydrogenolysis as has been done for comparable processes such as hydrocracking [28], this would supplement large-scale investigations such as this one, giving the modelling a more interdisciplinary perspective, therefore, making it more capable of preempting issues in the process.

4.1.2 Heat Integration

Table 4: Industrial hydrogenolysis process heating and cooling duties

Heat Integration	Heating Duty (MW)	Cooling Duty (MW)
Before	1.65	3.56
After	0.94	3.00

Energy Analyzer calculates a net heating duty of 1.96 MW and a net cooling duty of 3.56 MW. Despite favouring conventional and readily available utilities, fired heat is required to meet the highest reboiler temperatures in the columns which went up to 478.5°C. The cooling duties are supplied by cooling water and the rest of the heating and cooling duties are provided by exchange with other process streams. The largest heat duty was from the heater directly before entering the hydrogenolysis reactor and also the reboiler in column three which account for 32% and 36% of the total heat duty respectively. The largest cooling duty is from the hydrogenolysis reactor which accounts for 43% of the total cooling duty. Heat integration reduced the heating duty to 0.94 MW and the cooling duty to 3.00 MW as shown in Table 4. Appendix 2 provides a detailed explanation of the heat integration used and details the exact changes. The integrated network uses 18 heat exchangers with a total area of 194.7 m². Data from Aspen Energy Analyser suggest that the integrated network will save \$109,500/yr and reduce heat and cooling duty demand by 21%. From this, the amount of cooling water was estimated using the correlation in Turton’s textbook [29] to inform the LCA and costing. Further investigation should be done to minimise the use of fired heat as combustion is not a sustainable way to heat a system and is detrimental to the environmental impact of the process as further discussed in section 4.3. If it is not possible to eliminate fired heat, the light gases from hydrogenolysis could be used for energy recovery similar to other chemical recycling processes such as gasification [9]. This would not only reduce the environmental impact associated with acquiring the fuel for fired heat but also reduce the costs of the process overall.

4.2 Economic Assessment

4.2.1 Capital Investment Cost

Summing the working capital and total fixed capital costs gives a total capital expenditure for the industrial process of \$34,000,000. The total fixed capital costs include process capital, general plant capital (15% of process capital) and contingency costs (25% of the fixed capital cost) [19]. In this model, the working capital only includes adjuvants such as the initial catalyst cost, it does not include accounts receivable or inventory. The total capital cost breakdown is depicted in Figure 2, and in-depth calculations can be found in Appendix 5. As expected, the largest process capital cost comes from the compressors at 42%, given that compressors are one of the most expensive pieces of equipment in a plant operation. Also, since the process requires gas as a reactant, multiple compressors are needed to get the hydrogen to reaction conditions. The reactor cost is also a large percentage of the capital cost (32%), polypropylene hydrogenolysis reactors would undoubtedly be complex as

discussed in section 4.1 and there is no research on an industrial design. Additional column internals would likely be needed to handle the solid plastic similar to a screw-feed gasification reactor device [30]. The hydrogenolysis reactor operates at a comparatively high pressure (30 bar) so it would be costly to find specialized material capable of withstanding these conditions. Also, hydrogenolysis relies on a heterogeneous catalyst, so the reactors must be built to allow for the introduction and regeneration of catalysts, which adds to the total complexity and cost. In summary, the customisation of this reactor is ultimately the reason for the larger capital cost contribution.

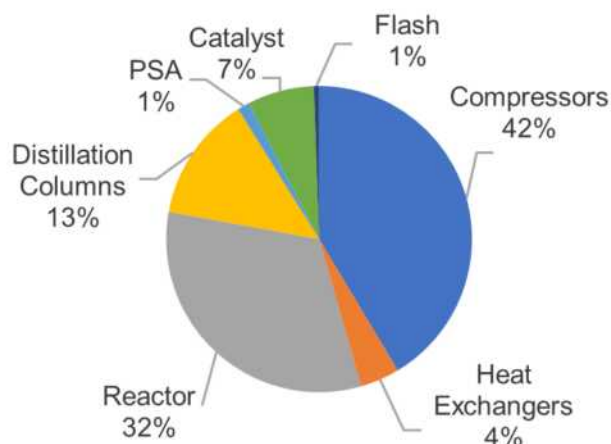


Figure 2: Capital cost breakdown of hydrogenolysis process

4.2.2 Total Manufacturing and Annual Expenses

Adding the total manufacturing costs to additional annual charges gives total annual expenses for the industrial process of approximately \$9,707,000. The total manufacturing cost includes maintenance, depreciation, labour, taxes and insurance, waste disposal and utility costs [19]. Summing these expenses gives a manufacturing cost of approximately \$8,072,000. It was estimated that the maintenance cost is 2.5% of process capital cost and taxes and insurance are 3% of the total capital expenditure. Additionally, depreciation was worked out to be 6.67% of the total capital investment cost. Other annual expenses include general administration and research and development, which are both assumed to be \$10,000/year [19]. The total annual expenses breakdown is shown in Figure 3, this methodology and estimations are based on *Systematic Methods of Chemical Process Design* [19] and full calculations can be found in Appendix 5.

Labour is the highest operating cost due to the number of distillation columns and the fact that Perry's coefficient method [23] used to calculate labour costs gives a larger weighting to distillation columns, necessitating a greater number of people. Also, the other operating costs for the process are relatively small thus labour constitutes a larger proportion; the utility costs for the process make up only 5% of total costs due to the process operating at mild conditions and the use of heat integration which decreases the total heat duty by 21%. Feedstock cost is not large as the only feedstock purchased is hydrogen and the cost of waste polypropylene is considered negligible. It is also assumed that the cost of pre-treatment and separation to get the

polypropylene to reaction conditions is met by the gate fees companies would pay to have their plastic waste recycled. The waste cost is also a small portion due to the process only requiring two small waste streams; one hydrogen purge stream and one reactor polypropylene waste stream.

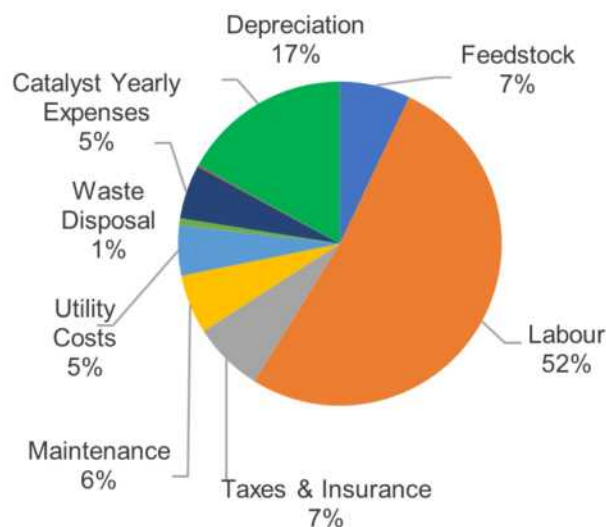


Figure 3: Operating cost breakdown of hydrogenolysis process for the treatment of polypropylene waste

It would be beneficial to compare the calculated OPEX to other recycling methods, but this risks inconsistent results as different reports use different costing methodologies and correlations. Data consistency is important along with data quality, in future, it would be useful to carry out a standardised study of OPEX costs for chemical recycling technologies using a pre-determined methodology for reliable comparison. Consistent data is currently not available in the chemical recycling literature and this investigation was beyond the scope of this report.

4.2.3 Revenue

The sources of revenue are light gases (methane to pentane), diesel, gasoline, kerosene, wax, and lubricants from the various distillation columns, amounting to a total revenue of \$33,600,000. Kerosene produces the most revenue, contributing \$18,000,000 to the total. The detailed calculations for the revenue value can be found in Appendix 5. The light gas stream is costed as natural gas as although the stream contains 18.2% hydrogen, the heating value is 50,000 kJ/kg which is similar to the net heating value of actual natural gas at 41,000 kJ/kg [31]. Note that these revenues assume that the products are of marketable standard when in reality further refining would likely be required constituting additional costs. These costs would increase if the products were used as feedstock for virgin polymer production requiring the breakdown and further processing of the alkanes produced, although these steps are beyond the pre-defined scope of this report they are important for future consideration as they would determine the wider circular economy context of this process. This is further discussed in section 4.3.

4.2.4 Profitability Analysis

Net present value (NPV) is calculated using equation 1, assuming constant revenue and expenses.

$$NPV = -C_i - C_w + (R - X)(1 - t) \frac{1 - (1 + i_t)^{-n}}{i_t} + D_t \frac{1 - (1 + i_t)^{-n_t}}{i_t} + \frac{C_s + C_w}{(1 + i_t)^n} \quad (1)$$

The NPV was calculated to be \$120.5 million. The parameters used to calculate these values can be found in Table 13 in Appendix 5.13. To calculate the NPV, the following assumptions were made: the working capital is equal to the initial catalyst cost as explained in section 3.2; the tax rate was assumed to be 21% [9]; the interest rate was assumed to be 15% [19]; project lifetime was assumed to be 20 years with straight-line depreciation; and salvage value is assumed to be \$0. The large positive NPV suggests that this process is economically favorable.

Despite Net Present Value's (NPV) usefulness as a financial metric for assessing profitability, it has certain drawbacks. When comparing projects of varying sizes or timescales, hydrogenolysis included, NPV might not be appropriate. This is because larger projects that have higher absolute Net Present Values (NPV), do not necessarily have higher percentage profits. In these situations, metrics such as the Internal Rate of Return (IRR) may offer a more comparable measurement. Also, the NPV can be sensitive to assumptions such as interest rate and changes in the market, although a sensitivity analysis partially mitigates this effect, it is nonetheless particularly difficult to predict future cash flows of such a novel process. Using risk-adjusted metrics such as a Risk-Adjusted Return on Investment or a more complex scenario analysis such as a Monte Carlo simulation could complement the use of NPV for highly uncertain processes such as the one presented in this report. These analyses are beyond the scope of this report and in any are currently lacking the reliable inputs needed for their computation. Instead for comparability with other processes, the cost per kilogram of polypropylene and the break-even cost of polypropylene were calculated.

4.2.5 Polypropylene Costs

A cost of \$0.46/kg of polypropylene feed was calculated for the hydrogenolysis process. This is comparable to the cost of recycling plastic in landfills at \$0.77/kg and incineration at \$0.56/kg [32] which partly justifies treating plastic through hydrogenolysis, although this is only indicative as the cited values are for a mixture of plastics as opposed to polypropylene specifically, with the error margin expected it cannot be said with certainty that hydrogenolysis is cheaper than conventional alternatives. However, this only considers economic capital whereas there is a trend in environmental policy discourse towards "natural" capital, that is the economic equivalent value of protecting nature [33]. If the environmental benefits of switching to hydrogenolysis were adequately quantified and considered, the process may be favourable overall despite its potentially lower economic value.

Figure 4 indicates that a polypropylene feedstock price of \$0.74/kg is required to break even. The price to purchase virgin polypropylene is approximately around \$1.46/kg [34].

The lower break-even point of industrial hydrogenolysis provisionally confirms that it has potential for use in plastic production.

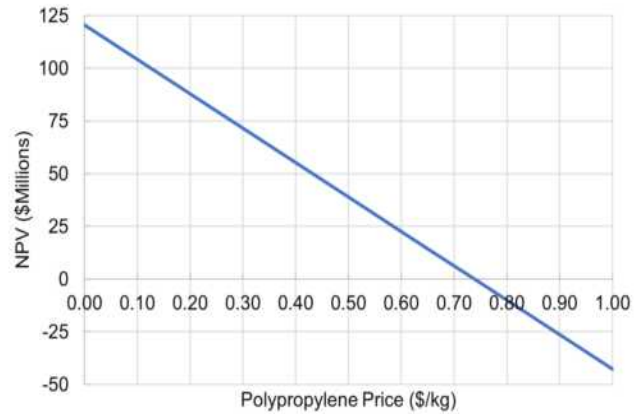


Figure 4: Break-even analysis of NPV with variation in polypropylene price

4.2.6 Sensitivity Analysis

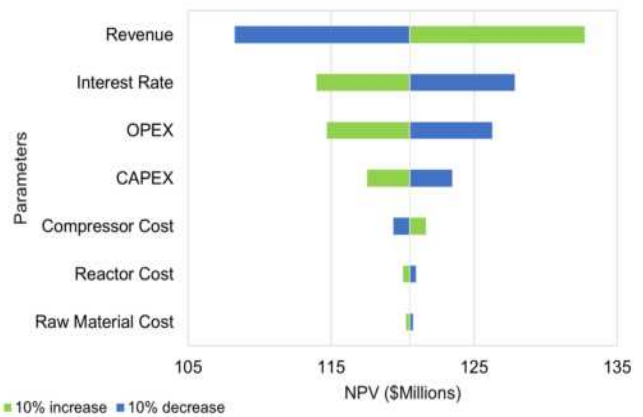


Figure 5: NPV sensitivity analysis of the polypropylene process with 10% parameter change

A sensitivity analysis for the NPV against revenue, interest rate, OPEX, CAPEX, compressor cost, reactor cost, and raw material cost is depicted in Figure 5, computed by varying each variable value by 10%. As expected, revenue has the biggest positive effect on NPV, with a 10% revenue change causing a corresponding 10% NPV change. Thus to increase the NPV, research into reactor conversion and product selectivity towards kerosene and other high-value products could help improve the profitability. The NPV changes by approximately 6% with a 10% interest change, indicating that the interest rate assumption influences the NPV as predicted. NPV assumes a constant interest rate throughout the project because as mentioned it is difficult to predict future scenarios, particularly for new processes, as a result, this may not be an accurate representation of the changing economic market. To make a more well-informed investment decision, NPV should be used in conjunction with other financial indicators such as the Rate of Return (ROI) or the Internal Rate of Return (IRR) as described in section 4.2.4. Changing the OPEX causes a 5% change in NPV, whereas the CAPEX only causes a 2% change. This highlights that to increase the profitability of the

process, focus should be put on reducing the OPEX costs. The OPEX cost is likely to change throughout the project lifetime as it is affected mainly by changing electricity costs, wages and feedstock prices all of which have a degree of variability, this should be considered when laying out an implementation plan.

4.2.7 CAPEX Comparison

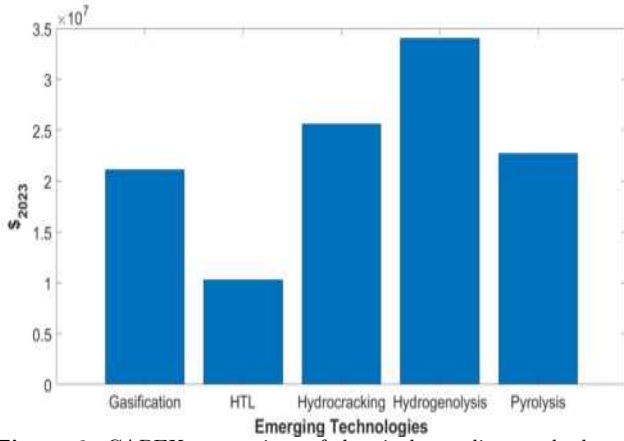


Figure 6: CAPEX comparison of chemical recycling methods

Figure 6 compares the CAPEX for the different chemical recycling technologies. Hydrogenolysis and hydrocracking have high CAPEX values mostly due to the use of expensive catalysts, high hydrogen pressures and complex reactors. To reduce these costs, research could be done to explore alternative catalyst options, particularly favouring milder conditions, increasing the selectivity of desired products, decreasing CAPEX and increasing revenue. The simplicity of the HTL process and the feedstock flexibility are the main reasons for the low capital costs. This flexibility can reduce the need for extensive pre-processing facilities and reduce the number of unit operations required. Pyrolysis and gasification have similar CAPEX and are likely the closest to widespread use compared to the others because the reactors used are low-cost and operate at lower pressures (1 bar). Also, since gasification directly burns leftover gases in a gasification furnace, it does not require any energy recovery equipment. These features could be integrated into future iterations of the hydrogenolysis process provided they satisfy economic and environmental requirements as discussed in section 4.1.2.

4.3 Environmental Assessment

Figure 7 illustrates the endpoint analysis conducted on the process that was calculated using the ReCiPe 2016 method. The negative values represent the negative impact that is avoided, for example by the use of waste polypropylene or the generation of useful products [35]. The generation of valuable products is what acquires most of the environmental "credit" as these products are otherwise generally environmentally damaging to obtain.

The most significant beneficial environmental impact is to human health with \$0.313 of "credit", as opposed to resource availability which was hypothesised to be the leading indicator of hydrogenolysis due to the replacement of environmentally damaging processes such as mining. This

is likely due to the health-threatening impacts of processes such as crude oil extraction but could also be attributed to the hierarchist perspective as its timescale aligns with the timescale for which climate change will have the worst effects on human health [24]. The only notable negative impact is from the utilities. This is expected as the use of fired heat is damaging; its emissions are estimated to be 1.10 kgCO₂/kg_{polypropylene} [31] and it releases toxins, highlighting the need to investigate alternative heat sources.

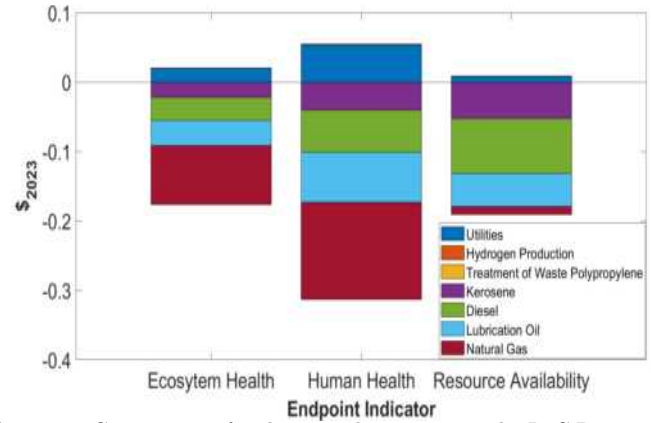


Figure 7: Comparison of endpoint indicators using the ReCiPe 2016 method

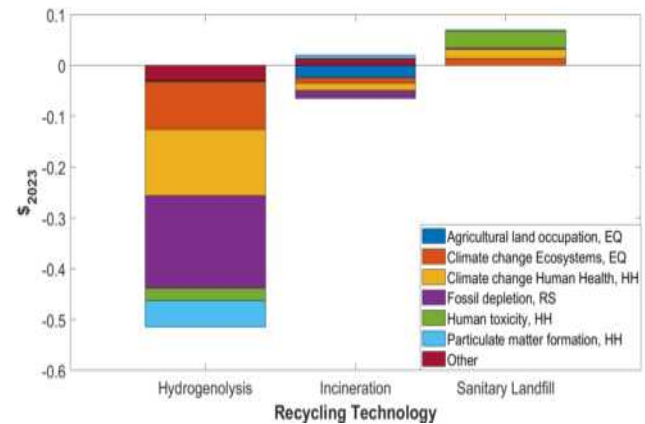


Figure 8: ReCiPe 2016 midpoint indicator breakdown for end-of-life processes

Figure 8 displays the midpoint indicator breakdown of the total externalities costs as represented by the endpoint indicators converted to a monetary basis. For all indicators, the designed hydrogenolysis process is either less environmentally "damaging" than landfill and incineration or comparably "damaging".

The most notable indicators fall into 3 key areas: climate change, because the process avoids climactically damaging end-of-life processes and sustainably produces valuable products; land occupation and resource use metrics, as hydrogenolysis eliminates the need for separate facilities; and the human health indicators, as expected hydrogenolysis produces less health-threatening substances than landfill and incineration. These results are indicative and the large numerical differences in the environmental impacts of the processes should be viewed critically as the LCA is subjective and may be easily skewed by assumptions made. This is not only a best-case scenario, but the

lack of standardised methodologies for the comparison of plastic waste makes it difficult to draw robust comparisons as the outputs are limited by the quality of the data. Despite the reputability of Ecoinvent, variations in process efficiency, regional disparities, socio-economic factors, costs of retrofitting and product and feedstock quality are difficult to model but can affect results.

There are several ways to strengthen the LCA analysis in future iterations: carrying out a sensitivity analysis to identify key parameters, comparing with alternative chemical recycling technologies to validate the preferential investment in hydrogenolysis, comparing different frameworks beyond ReCiPe or carrying out a scenario analysis assessing different scales, future innovations and locations. This model, like the flow-sheet and the TEA, is static and does not take into account variability in parameters such as availabilities and prices and therefore has a limited ability to predict future scenarios. Processes that have been simplified, such as the electrolysis of water, should also be investigated and fully specified as key decisions such as the source of the electricity used can have a significant impact on the sustainability of the process.

Chemical recycling is currently energy-intensive and is therefore unlikely to serve as a direct replacement for simple and cost-effective mechanical recycling, however, mechanical recycling is inherently non-circular as we cannot downcycle indefinitely and that is where hydrogenolysis could be critical if its products are used for the production of virgin polymers and not as combustible fuels.

5 Conclusion

This paper presents a first estimate of the design and environmental assessment of an industrial-scale hydrogenolysis process for the chemical recycling of polypropylene waste. A process flow-sheet is designed based on extrapolations from lab-based data, predominantly those generated by the Rosseinsky group at the University of Liverpool [6]. The flow-sheet is converged with high conversion, carbon efficiency and product purities. A techno-economic assessment is used to summarise the capital and operations cost of the process, suggesting that the process is viable as evidenced by a high NPV and low cost per kilogram, largely owing to the production of useful products. Making it a competitive option amongst emerging chemical recycling processes. The environmental assessment also provisionally suggests that the process is superior to other plastic end-of-life processing methods particularly concerning climate change, resource use and human toxicity.

However, this theoretical analysis, like any model, has inherent limitations. The assumptions made in models often do not align with reality, particularly those concerning complex and under-specified themes from reactor design to policy-making. Any uncertainties in the hydrogenolysis process should be addressed before implementation at any scale nearing the industrial process presented in this report. Despite the use of well-established methods in this report, the results are indicative and not conclusive as several approximations had to be made as a result of the low readiness of the process discussed. Ultimately, the contribution of hydrogenolysis to a circular economy depends on the context of its use, hydrogenolysis is a promising technology but how

and why it will be used will define its wider impacts.

6 Outlook

There are significant areas to be addressed before a decision is made on the investment in hydrogenolysis, beginning with investigations into the kinetics and properties of the hydrogenolysis reaction and how it behaves at scale before subsequent pilot-scale experiments. Additional catalyst design and optimisation could help address current issues surrounding product distribution and residence time whilst allowing for more precise costing. One of the main strengths of this process is the generation of useful and marketable products, the refinement of which could enhance the profitability of the process and further help offset the environmental impact. A weaker area of the process is also the utilities used; investigating other heat sources would also be an important research area.

More information should also be gathered on the process' social, environmental and economic impact. Given the complexity of the process, variability and future changes should be taken into account when assessing its future implementation. Modelling should be holistic and diverse ensuring future studies carry out scenario and sensitivity analyses. Additionally, a more in-depth comparison to alternative chemical recycling technologies could help inform future investment decisions and policy-making regarding plastic waste treatment.

Like any novel technology, hydrogenolysis of plastic waste has its limitations and disadvantages as discussed in this report. It should therefore be subject to standard regulation with careful attention to its integration into the current waste disposal infrastructure, accounting for sources of electricity, delivery of feed stock and perhaps most importantly the usage of its products. With these factors considered, hydrogenolysis could become a core part of sustainable waste handling and contribute to building a circular economy.

7 Acknowledgements

Acknowledgements go to Dr. Andrea Bernardi and Mr. Ben Lyons for continued assistance in modelling practice. Acknowledgments also to Prof. Clemens Brechtelsbauer for advice on scale-up and to the Rosseinsky Group at the University of Liverpool for supplementary data.

References

- [1] Ali Chamas et al. "Degradation Rates of Plastics in the Environment". In: *ACS Sustainable Chemistry & Engineering* 8.9 (2020), pp. 3494–3511. DOI: 10.1021/acssuschemeng.9b06635.
- [2] Cong Wang et al. "Polyethylene Hydrogenolysis at Mild Conditions over Ruthenium on Tungstated Zirconia". In: *JACS Au* 1.9 (2021), pp. 1422–1434. DOI: 10.1021/jacsau.1c00200.
- [3] Tian Tan et al. "Upcycling Plastic Wastes into Value-Added Products by Heterogeneous Catalysis". In: *ChemSusChem* 15.14 (2022), e202200522.
- [4] Winnie W. Y. Lau et al. "Evaluating scenarios toward zero plastic pollution". In: *Science* 369.6510 (2020), pp. 1455–1461. DOI: 10.1126/science.aba9475.

- [5] D. P. Serrano, J. Aguado, and J. M. Escola. "Developing Advanced Catalysts for the Conversion of Polyolefinic Waste Plastics into Fuels and Chemicals". In: *ACS Catalysis* 2.9 (2012), pp. 1924–1941. DOI: 10.1021/cs3003403.
- [6] Ajay Tomer et al. "Enhanced production and control of liquid alkanes in the hydrogenolysis of polypropylene over shaped Ru/CeO₂ catalysts". In: *Applied Catalysis A: General* 666 (2023), p. 119431. ISSN: 0926-860X. DOI: 10.1016/j.apcata.2023.119431.
- [7] Wei-Tse Lee et al. "Catalytic hydrocracking of synthetic polymers into grid-compatible gas streams". In: *Cell Reports Physical Science* 2.2 (2021), p. 100332. ISSN: 2666-3864. DOI: 10.1016/j.xcrp.2021.100332.
- [8] Mark Huijbregts et al. "ReCiPe2016: a harmonised life cycle impact assessment method at midpoint and endpoint level". In: *The International Journal of Life Cycle Assessment* 22 (Dec. 2016). DOI: 10.1007/s11367-016-1246-y.
- [9] Borja Hernández et al. "Techno-Economic and Life Cycle Analyses of Thermochemical Upcycling Technologies of Low-Density Polyethylene Waste". In: *ACS Sustainable Chemistry & Engineering* 11.18 (2023), pp. 7170–7181. DOI: 10.1021/acssuschemeng.3c00636.
- [10] Sibao Liu et al. "Plastic waste to fuels by hydrocracking at mild conditions". In: *Science Advances* 7.17 (2021), eabf8283. DOI: 10.1126/sciadv.abf8283.
- [11] Pavel Kots, Brandon Vance, and Dionisios Vlachos. "Polyolefin plastic waste hydroconversion to fuels, lubricants, and waxes: A comparative study". In: *Reaction Chemistry & Engineering* 7 (Jan. 2022), pp. 41–54. DOI: 10.1039/D1RE00447F.
- [12] Masazumi Tamura et al. "Structure-activity relationship in hydrogenolysis of polyolefins over Ru/support catalysts". In: *Applied Catalysis B: Environmental* 318 (2022), p. 121870. ISSN: 0926-3373. DOI: 10.1016/j.apcatb.2022.121870.
- [13] "Chapter 12 - Life-cycle cost analysis". In: *Maximizing Machinery Uptime*. Ed. by Fred K. Geitner and Heinz P. Bloch. Vol. 5. Practical Machinery Management for Process Plants. Gulf Professional Publishing, 2006, pp. 201–228. DOI: 10.1016/S1874-6942(06)80014-3.
- [14] J. Brandrup, E.H. Immergut, and Eric A. Grulke. *Polymer handbook*. eng. 4th ed. New York ; Wiley, 1999. ISBN: 0471166286.
- [15] Jesse Sarpong-Mensah. *Crude Oil Distillation Using Aspen Hysys*. June 2023.
- [16] Apinya Chanthakett et al. "Performance assessment of gasification reactors for sustainable management of municipal solid waste". In: *Journal of Environmental Management* 291 (2021), p. 112661. ISSN: 0301-4797. DOI: 10.1016/j.jenvman.2021.112661.
- [17] Thierry Meyer. "Scale-Up of Polymerization Process: A Practical Example". In: *Organic Process Research Development* 7 (May 2003). DOI: 10.1021/op025605p.
- [18] Satish Reddy and Sunil Vyas. "Recovery of Carbon Dioxide and Hydrogen from PSA Tail Gas". In: *Energy Procedia* 1.1 (2009). Greenhouse Gas Control Technologies 9, pp. 149–154. ISSN: 1876-6102. DOI: 10.1016/j.egypro.2009.01.022.
- [19] L.T. Biegler, I.E. Grossmann, and A.W. Westerberg. *Systematic Methods of Chemical Process Design*. Physical and Chemical Engineering Sciences. Prentice Hall PTR, 1997. ISBN: 9780134924229.
- [20] Alexander M. Niziolek et al. "Municipal solid waste to liquid transportation fuels – Part II: Process synthesis and global optimization strategies". In: *Computers & Chemical Engineering* 74 (2015), pp. 184–203. ISSN: 0098-1354. DOI: 10.1016/j.compchemeng.2014.10.007.
- [21] Chemical Engineering Essential for the CPI Professional. *Chemical Engineering Plant Index*. <https://www.chemengonline.com/site/plant-cost-index/>. Accessed: (Accessed 30/11/2023). 2023.
- [22] U.S. Bureau of Labor Statistics. *Occupational Employment and Wages*. Accessed: 2023-12-01. URL: <https://www.bls.gov/oes/current/oes518091.htm>.
- [23] R.H. Perry and D.W. Green. *Perry's Chemical Engineers' Handbook, Eighth Edition*. McGraw-Hill chemical engineering series v. 8, pt. 2008. McGraw-Hill Education, 2008. ISBN: 9780071422949.
- [24] Yan Dong et al. "Evaluating the monetary values of greenhouse gases emissions in life cycle impact assessment". In: *Journal of Cleaner Production* 209 (2019), pp. 538–549. ISSN: 0959-6526. DOI: 10.1016/j.jclepro.2018.10.205.
- [25] *Exchange Rates*. Accessed: 2023-12-01. DOI: 10.1787/037ed317-en. URL: <https://data.oecd.org/conversion/exchange-rates.htm>.
- [26] *Inflation Calculator*. Accessed: (Accessed 30/11/2023). URL: https://www.bls.gov/data/inflation_calculator.htm.
- [27] N. Yang. "Dichloromethane". In: *Encyclopedia of Toxicology (Third Edition)*. Ed. by Philip Wexler. Third Edition. Oxford: Academic Press, 2014, pp. 99–101. ISBN: 978-0-12-386455-0. DOI: 10.1016/B978-0-12-386454-3.01218-5.
- [28] Bay Van Tran et al. "Computational fluid dynamics of gas-liquid bubble column with hydrocracking reactions". In: *Computer Aided Chemical Engineering* 44 (2018). Ed. by Mario R. Eden, Marianthi G. Ierapetritou, and Gavin P. Towler, pp. 313–318. ISSN: 1570-7946. DOI: 10.1016/B978-0-444-64241-7.50047-1.
- [29] Richard Turton. *Analysis, synthesis, and design of chemical processes*. eng. 4th edition. India: Prentice Hall, 2012. ISBN: 0132618125.
- [30] G Seely, C Miller, and K Square. *Solids feed to a pressurized reactor*. Accessed: 2023-12-01. URL: <https://patents.google.com/patent/US3841465A/en>.
- [31] The Engineering ToolBox. *Fuel Gases - Heating Values*. https://www.engineeringtoolbox.com/heating-values-fuel-gases-d_23.html. Accessed: (Accessed 30/11/2023). 2005.
- [32] Raymond H.J.M. Gradus et al. "A Cost-effectiveness Analysis for Incineration or Recycling of Dutch Household Plastic Waste". In: *Ecological Economics* 135 (2017), pp. 22–28. ISSN: 0921-8009. DOI: <https://doi.org/10.1016/j.ecolecon.2016.12.021>.
- [33] Arjan Ruijs et al. "Natural capital accounting for better policy." In: *Ambio* 48 (2019), pp. 714–725. DOI: <https://doi.org/10.1007/s13280-018-1107-y>.
- [34] Pınar Polat and Esra Ersöz. *PP, PE producers cut run rates across Asia as high costs hammer margins*. Accessed: 2023-12-01.
- [35] Andrea Ramirez et al. *Guidelines for Life Cycle Assessment of Carbon Capture and Utilisation*. Mar. 2020.
- [36] Se Il Yang et al. "Effects of the residence time in four-bed pressure swing adsorption process". In: *Separation Science and Technology* 44.5 (2009), pp. 1023–1044. ISSN: 0149-6395. DOI: 10.1080/01496390902729122.

Prediction of Thermodynamic Properties and Phase Behaviour of CANDU Nuclear Reactor Fluid Coolant using the SAFT-VR Mie Equation of State

Alkmini Nicolaides, Naser Al-Wsaifer

Department of Chemical Engineering, Imperial College London, U.K.

Abstract Canadian Deuterium Uranium Nuclear reactors (CANDU) form an essential power generation source for Canada and a multitude of European countries. CANDU reactors are characterised by their use of deuterium oxide as coolant as opposed to conventional light water or carbon dioxide coolants used in most nuclear reactors. However, CANDU reactors suffer from deuterium oxide radiolysis - the splitting of coolant to deuterium and oxygen gas. This poses a major threat to the safe and economic operation of CANDU reactors over their lifetime. Hence, clear separation and recombining process of deuterium and oxygen back to deuterium oxide is essential. In light of this, this research paper proposes the first theoretical models of deuterium oxide and the deuterium oxide + oxygen + deuterium mixture present in CANDU reactors. The models of the pure components and mixtures were devised using the SAFTVR – Mie equation of state along with computational techniques to estimate the parameters required by the equation of state to generate the complete pure and mixture models. The models are used to predict crucial physical properties of deuterium oxide and the vapour-liquid mixture under standard operating conditions as well as CANDU operating conditions. The devised models demonstrate excellent accuracy, providing <5%AAD for the pure deuterium oxide, oxygen and deuterium models as well as the mixture system. The aim of devised models is to be used in shortlisting the possible separation and recombining techniques possible.

I. Introduction

Choosing a nuclear reactor coolant is a crucial part of the nuclear reactor design. The power output of any nuclear reactor is determined by the rate of heat removal from the core via the primary coolant loop. In light of this, choosing an effective coolant is crucial for the safe operation of a nuclear power plant.

An effective coolant will potentially have a high isobaric heat capacity and thermal conductivity (rapidly removing plenty of heat from the reactor core), radiolysis resistance (not readily decomposing under the harsh radiation intensive conditions of nuclear reactor core (NRC)), and a low neutron absorption cross section (lower tendency of absorbing neutrons), meaning more of the generated neutrons are reserved for fission reactions. These essential physiochemical properties, amongst many others, (chemical inertness, critical point, cost) form the multi-variate problem of choosing a primary coolant for a nuclear power plant.

With the plethora of industrial coolants present, it is highly desirable to obtain models that can accurately predict their thermophysical properties under reactor conditions without the need for costly and time-consuming experimentation – this is especially true under the severe nuclear reactor core conditions which may entail pressures of up to 120 bar and extreme temperatures radiation.

Canada Deuterium Uranium Nuclear Reactor (CANDU) is a type of nuclear reactor that utilises deuterium oxide as its primary coolant – Figure 1. It currently provides 15% of Canada's electrical power¹. Deuterium oxide is primarily used due to its reduced likelihood of absorbing neutrons (neutron absorption cross section) compared to light water, the primary coolant used in the most common nuclear reactor types, the Pressurised Water Reactor (PWR) and Boiling Water Reactor (BWR). Deuterium oxide's lower neutron absorption cross section allows for the use of much lower uranium fuel enrichment (natural uranium enrichment of 0.7%, instead of 4-5% enrichment used in a typical PWR) as less neutrons are absorbed by the

coolant, leaving more neutrons to be absorbed by the uranium fuel.

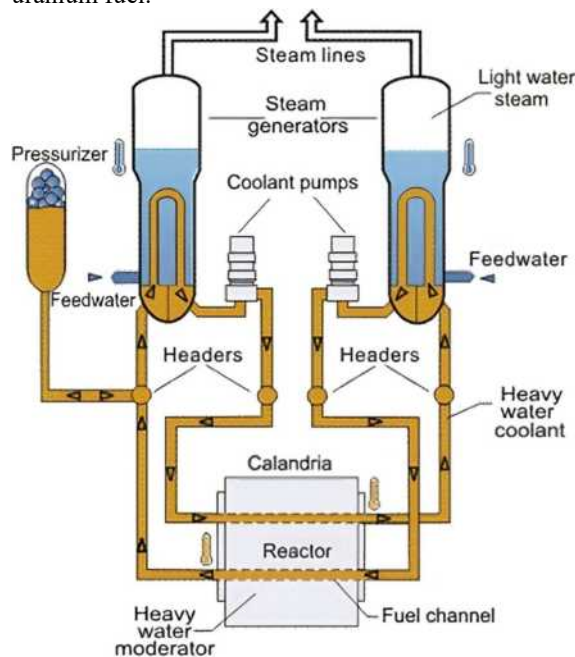
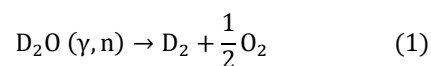


Fig 1. A schematic of a typical CANDU nuclear reactor. Deuterium oxide flows through the calandria, a series of small channels passing through the reactor core. Deuterium oxide also passes through the steam generators (carried by the orange pipes) boiling the secondary coolant water to be sent to the generator. A pressuriser ensures CANDU's high operating pressure while coolant pumps ensure deuterium oxide's circulation around the reactor. (Spinks 2011)

However, deuterium oxide undergoes radiolysis – splitting to form oxygen and deuterium upon gamma radiation absorption. The following reactions are the primary radiolysis reactions taking place in CANDU reactors²:



Deuterium oxide is a costly coolant, amounting up to 20% of the initial nuclear power plant capital cost³. Hence, it is of paramount importance to limit its depletion from radiolysis and ensure its maximum recovery. The recovery of deuterium oxide is achieved by separating the deuterium-oxygen vapour mixture from other vapours present and recombining deuterium and oxygen back to deuterium oxide³. Therefore, the objective of this research is to develop a model for the mixture generated in CANDU's reactor core to predict its physical properties, aiding in the choice of separation and recombination techniques of the deuterium and oxygen generated.

Previous attempts on generating empirical equations of state of pure deuterium oxide, deuterium, and oxygen have been made with great success, predicting physical properties under a wide range of conditions and with excellent accuracy. Interest in deuterium oxide's physical and nuclear properties began in the 1950s. With the advent of the cold war, interest in nuclear weapons programs led to extensive research in deuterium oxide's physical properties and eventually, the development of its correlations. Kirschenbaum, I.⁴ began the discussion on the need for accurate deuterium oxide experimental data and an accompanying empirical correlation. Kesselman, P. M⁵, Mamedov, A. M⁶, Plank et al⁷ and Suvorov⁸ followed by devising the first empirical equation of state for liquid deuterium oxide. A series of improvements to the existing equations of state using the growing experimental data available at wider operating conditions followed by Ikeda et al.⁹, Juza et al.¹⁰ and P.G Hill et al¹¹. The final deuterium oxide equation of state was generated by W. Lemmon et al¹² reducing the number of terms found in the previous equation of state. Oxygen and deuterium follow a similar path with the latest empirical equations of states generated by Weber¹³, Wagner¹⁴ et al, and Lemmon et al¹⁵.

All the aforementioned equations of state are empirical, using experimental data fitting to generate correlations. Although exceptionally accurate, these equations of state do not give much insight into the quantum interactions present. Furthermore, a model for the deuterium oxide + deuterium + oxygen mixture is still absent. Hence, the work to be presented will provide the first theoretical models, using the Statistical Associating Fluid Theory (SAFT) equation of state for pure deuterium oxide and the deuterium oxide + deuterium and deuterium oxide + oxygen mixtures.

The report is structured as follows: in section II we provide the theoretical background of Statistical Associating Fluid Theory and the SAFT-VR Mie's equation of state and we describe the procedure we followed to develop the pure and mixture models. In section III the models' performances are presented and discussed by comparing the model predictions to experimental data graphically and quantitatively using %AADs. Lastly in section IV we summarise our key findings, as well as the implications of our research, and we discuss possible developments and improvements on our current work.

II. Methodology

In this section, we describe the background theory and the procedure we followed for developing the pure component models for deuterium oxide, deuterium, and oxygen, as well as the two binary mixture models of deuterium oxide + deuterium and deuterium oxide + oxygen. Initially, the background theory of Statistical Associating Fluid Theory (SAFT) is provided (subsection II.A), followed by a description of the SAFT-VR Mie equation of state and its molecular parameters (subsection II.B). Finally, the procedure for developing the molecular models using the SAFT-VR Mie equation of state is outlined - subsection II.C.

II.A Statistical Associating Fluid Theory

The molecular framework underlying the Statistical Associating Fluid Theory (SAFT) is a chain of fused spherical segments that represent a molecule. The segments interact with each other through an interatomic potential. Numerous SAFT equations of state have been developed^{16, 22, 25}, each using varying types of potentials to describe the segment-segment interactions. Association interactions – strong, directional intermolecular bonds – are modelled as interactions through a square-well potential between association sites on the segments¹⁶.

The SAFT equations of state are expressed in terms of the fluid's total Helmholtz free energy A . The Helmholtz free energy is given as a sum of contributions¹⁶:

$$A = A_{\text{ideal}} + A_{\text{mono}} + A_{\text{chain}} + A_{\text{assoc}} \quad (3)$$

A_{ideal} is the free energy of an ideal gas mixture of the molecules in the fluid. It incorporates the contributions from the translational, rotational, and vibrational modes of motion. A_{mono} is the residual free energy of each segment. It incorporates the segment-segment interactions taking place within the molecule. A_{chain} is the contribution from the fusing of the segments forming a chain. A_{assoc} is the free energy from the strong intermolecular association interactions, like hydrogen bonding, in the fluid¹⁶.

Wertheim's Thermodynamic Perturbation Theory (TPT) provides the basis for describing the relation between site-site interactions and bulk fluid properties of associating molecules¹⁷. In his theory, molecules are represented as single spheres. To model the strong attractive interactions of associating fluids, he defines acentrically positioned attractive sites. These sites interact through a short-range square-well potential. Wertheim takes advantage of the short range of the interactions to introduce steric hindrance effects and limit the number of intermolecular bonds each association site participates in. This led to the development of his Thermodynamic Perturbation Theory. For particles with two attractive sites, Wertheim derives the first-order TPT1 equation of state¹⁸. Chapman et al. restates the TPT1 into a form which can be used to describe mixtures of species of different sizes,

with a non-spherical molecular shape, as well as with different numbers of association sites¹⁹. This is incorporated in the SAFT equation of state¹⁸.

II.B SAFT-VR Mie equation of state

SAFT Variable Range (VR) Mie is one of the many equations of state belonging to the SAFT equation of state family. The underlying molecular framework is a chain of homonuclear (identical) segments. SAFT-VR Mie enables the manipulation of the range of the segment-segment dispersion interactions, rendering it a “variable range” equation of state. The SAFT-VR Mie equation of state incorporates the Barker and Henderson high-temperature perturbation expansion of the $\mathcal{A}_{\text{mono}}$ term. In this perturbation expansion, the segment-segment interactions are described by the Mie potential (Figure 2) – a generalised form of the Leonard-Jones potential²⁰:

$$\phi_{\text{Mie}}(r) = \frac{\lambda_r}{\lambda_r - \lambda_a} \left(\frac{\lambda_r}{\lambda_a} \right)^{\frac{\lambda_r}{\lambda_r - \lambda_a}} \varepsilon \left[\left(\frac{\sigma}{r} \right)^{\lambda_r} - \left(\frac{\sigma}{r} \right)^{\lambda_a} \right] \quad (4)$$

where ε is the depth of the potential well, σ is the segment diameter, λ_a and λ_r are the attractive and repulsive exponents, respectively²⁰.

SAFT-VR Mie’s improvements from its predecessor equation of state make it an attractive option to model the species of interest of this study. In general, a good description of phase behaviour is easily achievable through any potential with fixed attractive and repulsive exponents, but most equations of state struggle around the critical region. The thermodynamic property prediction around the critical region is one point that SAFT-VR Mie excels in. Additionally, most equations of state struggle with the prediction of second-derivative properties like isobaric heat capacity. Second derivative properties are sensitive to the nature of the repulsive interactions, captured by the repulsive exponent, λ_r . Thus, a more versatile potential, like the Mie potential, which allows adjustment of the repulsive exponent, can offer a better description of second derivative properties²¹. The SAFT-VR Mie equation of state serves the purposes of this research well, as it enables accurate phase behaviour prediction up to the critical point. It also provides the means to achieve better predictions of heat capacity, which is of interest in a nuclear reactor setting.

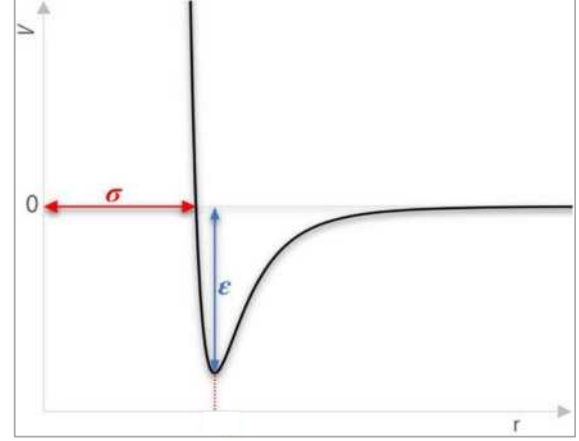


Fig. 2 Mie intermolecular potential, where ε is the depth of the potential well, σ is the segment diameter. (Sikorska 2020)

A set of molecular parameters need to be defined to describe a pure component using the SAFT-VR Mie equation of state – Table 1. Four of these parameters can be seen in equation 4. These are the energetic parameter ε/k_B (K), the size parameter σ (Å), and the attractive and repulsive exponents, λ_a and λ_r , respectively. Additionally, the molecular weight and the number of segments comprising the molecule need to be specified. For associating species, like deuterium oxide, the association energy $\varepsilon_{\text{assoc}}/k_B$ (K), the bonding volume parameter K (Å³), and the number of attractive (hydrogen sites) NST_a , and repulsive (electron sites) NST_e association sites, also need to be specified^{20, 22}.

Table 1 SAFT-VR Mie molecular parameters

Parameter	Units	Description
m	-	Number of spherical segments
ε/k_B	K	Depth of Mie potential-well
σ	Å	Segment diameter
λ_a, λ_r	-	Attractive and repulsive exponents of the Mie potential
$\varepsilon_{\text{assoc}}/k_B$	K	Depth of association square-well potential
K	Å ³	Bond volume parameter
NST_e	-	Number of site types corresponding to lone a lone electron pair
NST_a	-	Number of site types corresponding to H atom

II.C Parameter estimation procedure

To develop equations of states to describe the target pure components and binary mixtures, we used experimental data to regress the molecular parameters of the SAFT-VR Mie equation of state. This section describes in detail the parameter estimation procedure for the pure components and for the mixtures.

Firstly, we modelled deuterium oxide – the key component found in the primary coolant of CANDU nuclear reactors. We collected pseudo-experimental data from NIST²³ for the saturation pressure (P_{sat}), saturated liquid density (ρ^{sat})¹⁷, isobaric heat capacity (C_p) at 100 μbar, 1 bar, and 100 bar.

Important to any equation of state is the ideal contribution. There are various ideal gas models available, which can be implemented in a SAFT equation of state. We used the Reid ideal model, which accounts for the translational, rotational, and vibrational modes of motion of the ideal gas²⁴. The Reid coefficients were obtained by fitting the Reid heat capacity polynomial to the vapour phase heat capacity data at 100 μ bar data, as the heat capacity is approximately that of an ideal gas at this pressure. The molecular interaction parameters were then estimated by regressing the saturation and heat capacity data at 1 bar. The saturation pressure and saturated liquid density were given a weighting of 1.0 while the heat capacity was given a weighting of 0.1 in the objective function. The parameters estimated were the potential-well depth ϵ , the segment diameter σ , the repulsive exponent λ_r , and association energy ϵ_{assoc} . The existing SAFT-VR Mie light-water parameters²² were used as initial guesses for the parameter estimation.

SAFT models of deuterium and oxygen have already been developed. Deuterium was modelled using another SAFT equation of state – the SAFT-VRQ Mie equation of state²⁵ – while oxygen was modelled using the SAFT-VR Mie equation of state²¹. The SAFT-VRQ Mie parameters of deuterium were implemented in the SAFT-VR Mie equation of state. To improve their performance in predicting the heat capacity, we implemented the Reid ideal contribution in the SAFT-VR Mie models of the two components.

Additional data was collected from NIST²³ for the evaluation of the pure component models. For deuterium oxide, the saturated molar volume and the liquid density and heat capacity at 1 bar and 100 bar, were collected. For deuterium and oxygen, data of vapour density and heat capacity at 1 bar were collected. The models were evaluated quantitatively with the percent average absolute deviation %AAD and qualitatively by examining the graphs of the properties of interest.

We developed the binary mixture models of deuterium oxide + deuterium and deuterium oxide + oxygen using the finalised pure component models. Experimental data for the mole fraction of deuterium and oxygen dissolved in deuterium oxide at varying temperatures were obtained from Scharlin et al.²⁶ and Setthanan et al.²⁷. In addition to defining the pure component parameters when modelling a mixture, it is required to estimate unlike molecular interaction parameters to describe the interactions between the different species present in the mixture. In this study, we defined the ϵ and λ_r unlike parameters. There are two approaches of doing so. The first approach uses combining rules – an augmented geometric-mean rule²⁰ and the Berthelot rule¹⁶ to estimate the unlike ϵ and λ_r respectively. The second method involves regression of the parameters using experimental data. Initially, combining rules were used. The combining rules underpredicted all experimental data sets. Hence, they were used as lower bounds in the parameter estimation, which was subsequently conducted. The parameter estimation for the mixtures was done by experimental data regression, similar to the pure components. Finally, the mixture models’

performance was evaluated using average absolute deviation (AAD) and %AAD values and graphs.

III. Results and discussion

In this section, we present the molecular parameters of the SAFT-VR Mie equations of state for the pure component (deuterium oxide, deuterium, and oxygen) and binary mixture (deuterium oxide + deuterium, deuterium oxide + oxygen) models. Additionally, we evaluate the models’ performance in predicting thermophysical properties. We first evaluate the pure components in subsection III.A and then proceed to discussing the mixture models in subsection III.B. For each model we report the average absolute deviation (AAD) or the percentage average absolute deviation (%AAD) between the experimental data points and the model predictions:

$$\%AAD = 100 \times \left(\frac{1}{N_{\text{data}}} \sum_{i=1}^{N_{\text{data}}} \left| \frac{Z_i^{\text{exp}} - Z_i^{\text{model}}}{Z_i^{\text{exp}}} \right| \right) \quad (5)$$

$$AAD = \frac{1}{N_{\text{data}}} \sum_{i=1}^{N_{\text{data}}} |Z_i^{\text{exp}} - Z_i^{\text{model}}| \quad (6)$$

where N_{data} is the number of data points and Z^{exp} and Z^{model} are the experimental and predicted values of property Z , respectively. Along with the models developed, we present the existing model and experimental data of light-water in order to compare it with deuterium oxide to demonstrate that the deuterium oxide model can capture the slight differences in thermophysical properties between the two isotopes.

III.A Pure component models

This section discusses all the pure component models with a focus on the deuterium oxide model and its performance in predicting liquid properties at 1 bar and at 100 bar – CANDU reactor operating conditions. The SAFT-VR Mie molecular parameters for each component are reported in Table 2. The %AAD of the pure component models’ predictions from pseudo-experimental data are reported in Table 3. It should be noted that the uncertainty of the NIST pseudo-experimental data is reported to be below 0.1% for the saturation and density data used, and 1% for the heat capacity data²³.

III.A.1 Deuterium oxide model performance for saturation properties

The deuterium oxide model provides highly accurate predictions of the saturation properties. The molar volume in the temperature-molar volume saturation envelope (Figure 3a) is predicted with a 2.1% AAD. Small deviations are noticeable near the critical region and in the low temperature region of the gas phase branch. The saturation pressure (Figure 3b) is also predicted with outstanding accuracy up to the critical region – with 1.0% AAD. In both the saturation envelope and the saturation pressure curve, the deuterium oxide model is able to capture the slight differences of deuterium oxide’s physical properties when compared with light-water’s.

III.A.2 Deuterium oxide model performance for liquid properties at 1 bar

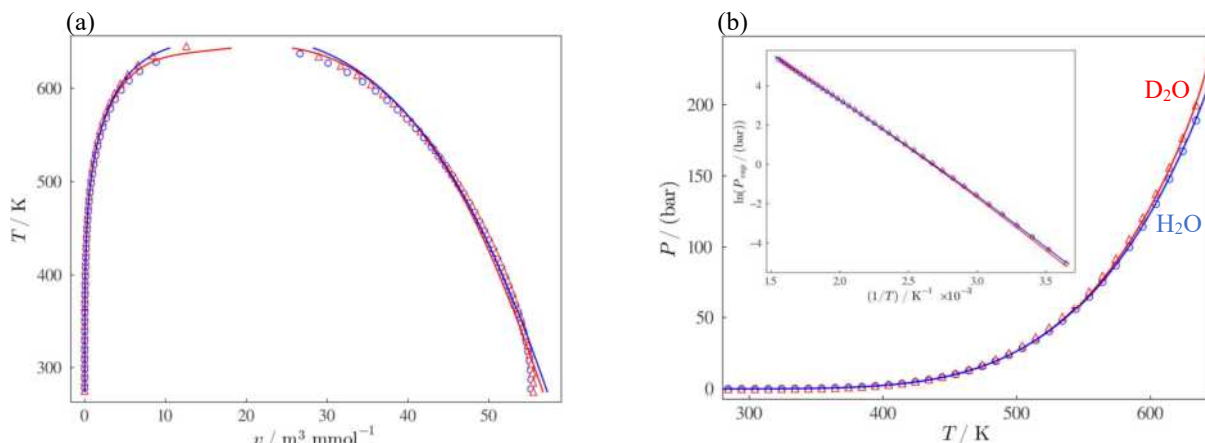


Fig. 3. SAFT-VR Mie's model predictions for the saturation envelope (a) and vapour pressure (b) variation with temperature of light-water (blue solid) and deuterium oxide (red solid) along with the respective experimental data – blue circle and red triangle. The model accurately predicts the pressures and saturation molar volumes even in near the critical point – a region notoriously difficult to accurately predict physical properties for.

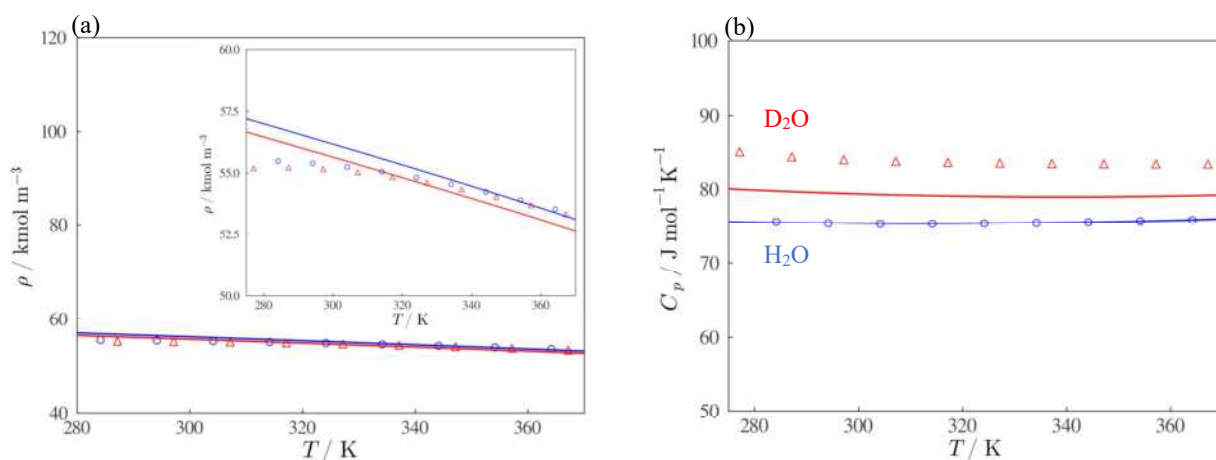


Fig. 4. SAFT-VR Mie's model predictions of the molar density (a) and heat capacity (b) variation with temperature for light-water (blue solid) and deuterium oxide (red solid) at 1 bar in the liquid phase along with the respective experimental data - blue circle and red triangle. The model predicts light water and deuterium oxide's molar density with exceptional accuracy while it slightly underestimates deuterium oxide's heat capacity for the same temperature range indicating the presence of nuclear quantum effects unaccounted for by the model.

The liquid phase of deuterium oxide is the most interesting, yet challenging to model, as complex intermolecular and intramolecular interactions play a deterministic role in its thermophysical properties. In this section we examine deuterium oxide's model predictions of the heat capacity and density of the liquid phase at 1 bar. We examine limitations that the SAFT-VR Mie equation of state faces when predicting these liquid properties, as well as the implications this has for the estimated SAFT-VR Mie molecular parameters.

The deuterium oxide model predicts with exceptional accuracy the pseudo-experimental liquid density at 1 bar, (1.1% AAD), capturing the subtle differences between deuterium oxide and light-water – Figure 4a. However, it is also observed that the model does not capture the liquid density maximum, which is an inherent limitation of the SAFT-VR Mie theory. By closely examining the liquid density data of the two components, one can make an interesting observation – the heavier deuterium oxide has a lower density than its light-water counterpart. This anomaly in the density

indicates that the arrangement of the molecules in the liquid is what has a determining effect on its density. Deuterium oxide forms on average more hydrogen bonds per molecule (3.76) than light-water (3.62)²⁸, giving it a more organised, tetrahedral structure (Figure 5) when compared to light water. Deuterium oxide's tetrahedral structure has a greater intermolecular void, as opposed to light water's more disordered structure, causing deuterium oxide to have a lower density²⁹. The density anomaly is analogous to the more familiar phenomenon observed in light-water and ice; upon freezing of light-water to form ice, all molecules participate in four, tetrahedrally-oriented hydrogen bonds³⁰. The molecules become arranged with a low packing efficiency, resulting in a less-dense solid phase compared to the liquid phase²⁸. To capture this effect when modelling the liquid density, the four-body intermolecular interactions would have to be considered.

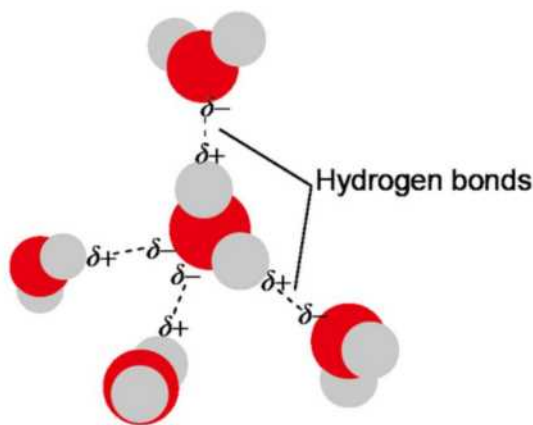


Fig.5. Tetrahedral hydrogen bond network of liquid deuterium oxide, where each deuterium oxide molecule participates in four-body interactions (Physics Open Lab)³¹.

SAFT-VR Mie faces a limitation when it comes to modelling fluids when many-body interactions are present. These many-body interactions have a prominent effect on the physical properties of the components they are affecting, as observed in deuterium oxide.

SAFT uses an effective-pair-density function to describe associating fluids. It assumes that only two-body intermolecular, association interactions are taking place in the hydrogen-bonding liquid. As a result, the hydrogen-bonding molecules in the liquid phase are modelled as dimers, instead of tetrahedrals^{17, 18}. Since SAFT-VR Mie does not incorporate many-body interactions, it implicitly accounts for the density in the association energy parameter, ϵ_{assoc} , which relates to the hydrogen bond strength. We speculate that when regressing the SAFT-VR Mie molecular parameters, the association energetic parameter, ϵ_{assoc} , is estimated to have a lower value than it would have if it were modelled taking the tetrahedral arrangement into consideration. This is further supported by regressing the experimental data for the pure model using a lower relative weighting on the density data – the quantum effects introduced by the density data, including the strength of the hydrogen bonds, will have less of an effect on the model parameters³². The resulting model parameters predict a stronger hydrogen bond strength than the previous model, reducing the effect of the tetrahedral structure formation on the hydrogen bond strength as expected.

The isobaric heat capacity is predominantly defined by the ideal contribution, with the remaining residual contribution being affected by the hydrogen bond strength³³ – ϵ_{assoc} . The model consistently underpredicts the pseudo-experimental liquid heat capacity data at 1 bar (Figure 4b) with a 5.4% AAD. The underprediction of the heat capacity is speculated to be a byproduct from the model's interpretation of the intermolecular interactions present when regressing the density data – interpreting the lower deuterium oxide density compared to light water as weaker hydrogen bonds. This, in turn, causes the underprediction of the heat capacity by the model.

Studying the performance of other classical thermodynamic models of water, it becomes evident that

without quantum treatment and consideration of many-body interactions, it proves challenging to reproduce the heat capacity with greater accuracy^{34, 35} – the light water model proposed by Graham et al²² uses alternative, more sophisticated parameter estimation techniques to reach the excellent light-water model performance demonstrated (Figure 4b). The SAFT-VR Mie equation of state does not explicitly account for quantum effects. They are to an extent implicitly incorporated in the molecular parameters, as heat capacity was used to regress the model parameters. To predict the heat capacity with greater accuracy, the heat capacity data should be given a greater weighting in the optimisation function used to optimise the deuterium oxide model parameters. Quantum effects would then be implicitly accounted for in the molecular parameters. This, however, results in poor modelling of the phase behaviour and loss of physical meaning of the molecular parameters.

Comparing deuterium oxide's model with the light-water model developed by Graham et al²² (Table 2), it is observed that deuterium oxide's association parameter, ϵ_{assoc} , is slightly lower than light water's association parameter. This suggests weaker hydrogen bonding in deuterium oxide when compared with light water. This is contrary to literature belief – X-ray spectroscopies and observations of physical properties, like the higher melting point of deuterium oxide³².

III.A.3 Deuterium oxide model performance at operating conditions

Once we have tested and validated the performance of the deuterium oxide model for saturation and liquid properties at 1 bar, we then examined how the model performs under CANDU reactor operating conditions – 100 bar and 500 K – 550 K³. The model predicts the liquid density remarkably well across the liquid temperature range. The model's predictions of heat capacity become increasingly more accurate as temperature increases, with a perfect prediction of the heat capacity within CANDU's operating temperature. Pressure, as expected, does not have a significant effect on the liquid properties, as the values and curvature of the experimental data and model prediction curves at 100 bar closely resemble the model predictions at 1 bar over the same temperature range (figures 6a and 6b).

Temperature, by contrast, has a significant effect on the physical properties and accuracy of the deuterium oxide model. The heat capacity remains relatively steady and begins to gradually increase around 450 K. This gradual increase in heat capacity can be attributed to the nuclear quantum vibrational modes of motion (stretching and bending) beginning to activate as the temperature increases²³. The vibrational modes of motion introduce their contribution to the heat capacity through the ideal gas heat capacity contribution.

Another interesting effect is the SAFT-VR Mie model's increasingly improving performance with the increase in temperature. The model initially underpredicts the heat capacity identically to the model predictions at 1 bar – the effect of the tetrahedral structure of deuterium oxide on the heat capacity prediction is still present. Yet, at higher temperatures,

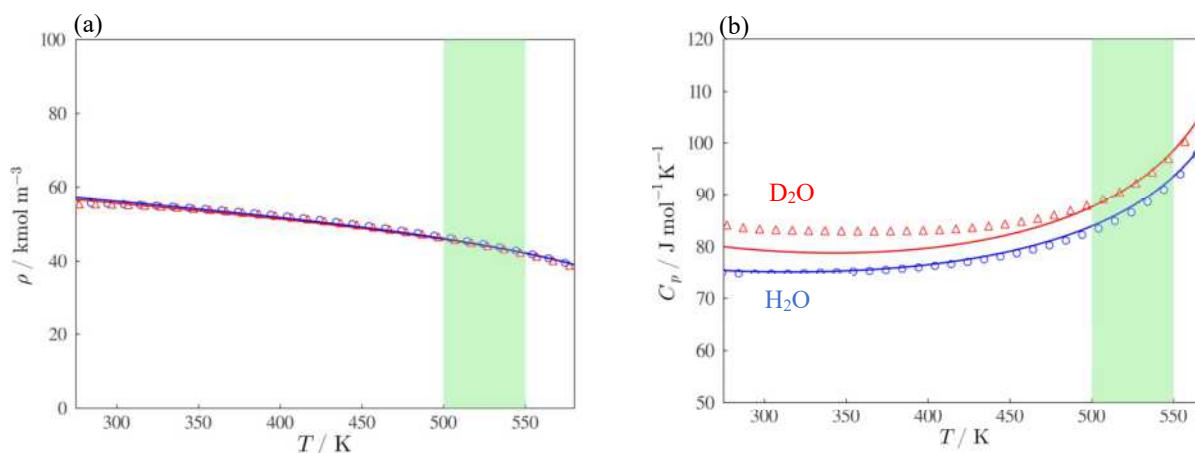


Fig. 6. SAFT-VR Mie's model predictions of the molar density (a) and heat capacity (b) variation with temperature for light-water (blue solid) and deuterium oxide (red solid) at 100 bar (CANDU operating pressure) in the liquid phase along with the respective experimental data – blue circle and red triangle. The model predicts light-water and deuterium oxide's density with exceptional accuracy across the liquid phase temperature range. Deuterium oxide's heat capacity is predicted by the model with an increasing improvement, starting with a slight underestimation and progressively improving with a perfect prediction at CANDU's operating temperature (500-550 K and depicted in a green band) indicating the presence of nuclear quantum effects at lower temperatures unaccounted for by the model.

where the liquid is more disordered, the intermolecular attractions more closely resemble two-body interactions³². Thus, at higher temperatures the model's predictions become more accurate, as SAFT's interactions more closely resemble two-body interactions.

III.A.4 Deuterium and oxygen models performance for vapour-phase properties at 1 bar

To complete the CANDU reactor primary coolant mixture that occurs due to deuterium oxide radiolysis, we test the deuterium and oxygen model performances. The two models demonstrate perfect performance in predicting the vapour phase density and heat capacity – Figures 7a and 7b. The gases display close to ideal behaviour, which is why in the data points have almost identical values.

In the total Helmholtz free energy expression of a component, the ideal contribution, A_{ideal} , is the most well understood and well-defined. The gases at 1 bar behave almost ideally. Thus, the model's predictions perfectly overlap with the gaseous density data points. The same applies to the heat capacity³³. Since the Reid ideal heat capacity coefficients were fitted specifically for oxygen and deuterium, as expected, the models match the data exactly. Comparing the deuterium and the hydrogen heat capacity data, it is clear that there are differences in their values, which are captured by the deuterium model.

III.B Binary mixtures models performance

This section presents and discusses the two binary mixture models (deuterium oxide + deuterium, deuterium oxide + oxygen) and their performance in predicting the mole fraction of deuterium and oxygen in deuterium oxide. The additional unlike molecular interaction parameters describing the mixtures and the corresponding AAD values are presented in Table 3.

The model predicts the amount of deuterium and oxygen dissolved in deuterium oxide with %AADs of

5.1% and 17%, respectively. It is worth noting that the %AADs are misleadingly high, due to the very small order of magnitude of the mole fraction values of dissolved deuterium and oxygen in deuterium oxide.

The observed disparity between the theory and the experimental data (Figure 8) could firstly be attributed to the structure and dynamics of deuterium oxide. The limited insight into the hydrogen bond network in deuterium oxide impedes the accurate modelling of the interactions of the dissolved gases in deuterium oxide²⁹. Additionally, information regarding the accuracy of the experimental data is lacking. As outlined in Scharlin et al²⁷, ideality is assumed initially, when determining the amount of gas dissolved in the solvent. The mole fraction values are later corrected for non-ideality. However, the accuracy of this approximation is not stated. Instead, only the error associated with the experimental procedure is given.

The optimal unlike parameters for the mixture were obtained by regression of the experimental data. Initially, though, combining rules were used to estimate the unlike parameters. The combining rules tended to underpredict the experimental data. When the combining rules were used, the predictions generated 70% %AAD for the mole fraction of deuterium dissolved in deuterium oxide and 29% for the mole fraction of oxygen dissolved in deuterium oxide. The unlike parameters obtained from the combining rules were not chosen as the optimal parameters, since experimental data and a better parameter estimation tool are available. However, combining rules demonstrate that even if there are no experimental data available to regress the unlike parameters, the theory can provide reliable predictions. It is not uncommon that in the absence of experimental data, combining rules would be relied on for unlike parameter estimations¹⁶.

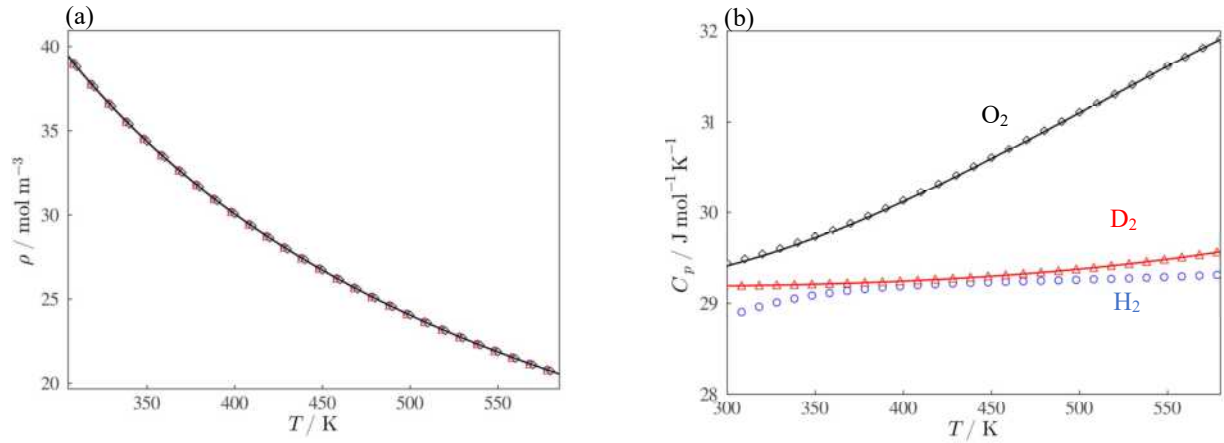


Fig. 7. SAFT-VR Mie's model predictions of the molar density (a) and heat capacity (b) variation with temperature for deuterium (red solid) and oxygen (black solid) at 1 bar in the vapour phase along with the respective experimental data – red triangle, black diamond and blue circle for hydrogen. The model predicts the molar density and heat capacity with exceptional accuracy for all the species across the whole tested temperature range – a tribute to the ideal behaviour of the vapours.

Table 2. SAFT-VR Mie parameters for the pure components: deuterium oxide, deuterium, oxygen and light-water

Model	Parameters								
	m	$(\epsilon/k_B)/K$	$\sigma/\text{\AA}$	λ_r	λ_a	$(\epsilon_{\text{assoc}}/k_B)/K$	$K/\text{\AA}^3$	$N ST_c$	$N ST_a$
D ₂ O	1.257	382.0	2.824	27.98	6.000	1590	177.6	2	2
D ₂	1.000	21.20	3.154	8.000	6.000	-	-	-	-
O ₂	1.000	81.48	2.967	8.922	6.000	-	-	-	-
H ₂ O *	1.257	351.2	2.802	25.13	6.000	1630	177.6	2	2

* H₂O model by Graham et al²² for comparison with D₂O model

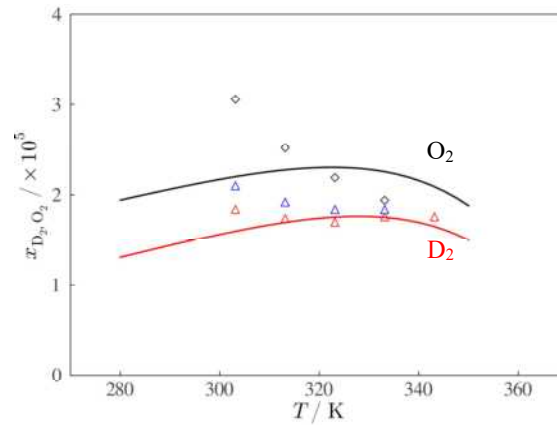


Fig. 8. SAFT-VR Mie's solubility prediction of the solubility variation with temperature for deuterium (red solid) and oxygen (black solid) vapour mixture in liquid deuterium oxide along with the respective experimental data – red and blue triangle for deuterium and black diamond for oxygen. The different triangle colours represent the different deuterium solubility data sources with the red and blue triangle being from Setthanan²⁶, and Scharlin²⁷. The two SAFT-VR Mie models (D₂O + D₂ and D₂O + D₂) exhibit a concave shape while the experimental data for both components exhibit a convex shape. Although the difference in shape between the experimental data and the model is noticeable, the scale at which the solubility data is exhibited is extremely small, demonstrating a significantly accurate mixture model.

Table 3. Mixture model unlike parameters and absolute average deviations (AAD) and %AAD from mole fraction, x , experimental data

Model	Unlike Parameters		AAD × 10 ⁷	
	$(\epsilon/k_B)/K$	λ_r	x	AAD(%)
D ₂ O + D ₂	136.9	20.97	9.2	5.1
D ₂ O + O ₂	183.5	15.16	45	17

IV. Conclusions

To conclude, this research sought to generate theoretical models capable of predicting thermophysical properties of the nuclear coolant mixture found in CANDU reactor cores. We established a deuterium oxide SAFT-VR Mie model and improved upon existing deuterium and oxygen SAFT models. The pure models predict heat capacities and densities with excellent accuracies, both at 1 bar and at CANDU reactor operating conditions – 100 bar. The mixture models predict the deuterium + deuterium oxide and oxygen + deuterium oxide binary mixture solubilities with slightly less accuracy. Nevertheless, the mixture model performance is more than adequate for an engineering purpose.

The research's significance lies in its aid in the selection of separation techniques of the generated vapours in CANDU reactors and the added insight into the quantum effects present in the coolant mixture. The models will reduce the need for experimental data under CANDU operating conditions for deuterium and oxygen to determine their physical properties and hence decide on the suitable recombination methods. Furthermore, deuterium oxide's SAFT-VR Mie model parameters and their comparison to light water's parameters, provide insight on the quantum effects present and their respective strengths, indicating that hydrogen bonding is weaker in liquid deuterium oxide compared to light-water hydrogen bonding. Furthermore, deuterium oxide's parameters validate the observed differences in physical properties between deuterium oxide and light water – the differences in liquid densities and heat capacities.

In the greater picture, this research further validates SAFT-VR Mie's theoretical model of molecules and their interactions by demonstrating highly accurate predictions of physical properties in both pure and mixture systems. Furthermore, this research introduces a starting point for theoretical analysis of fluids in nuclear reactor settings.

Nuclear reactors contain a complex mixture of many fluids. Hence, the mixture models devised can be greatly improved by introducing the missing components found in CANDU liquid coolant mixtures. The introduction of tritium oxide, T_2O , and deuterium tritium oxide, DTO, into the liquid phase can be an area of further study, probing into their effect on the mixture physical properties. Additionally, including interactions between deuterium and oxygen in the vapour phase may prove beneficial to the prediction performance of the model. Finally, compiling further experimental data for the current mixture system will prove valuable in improving the current mixture model as well as be used to validate the model performance in a wider range of operating conditions.

Acknowledgements

The authors would like to thank Pierre Walker, Thomas Bernet, Karl Tischler, Klaus Hesch, Hisako Niko and Kareem Hijazi for their contributions and continued support in the duration of this research project.

References

- [1] Winfield, M. "Nuclear Power in Canada." (2006).
- [2] Simnad, M. T. "Nuclear reactors: shielding materials." *Encyclopedia of Materials: Science and Technology*: 6377-6384 (2001).
- [3] Nanis, R. "Heavy water cycle in the CANDU reactor." (2000).
- [4] Kirschenbaum, I. "Physical Properties and Analysis of Heavy Water", (div III, vol 4A), (1951).
- [5] Kesselman, P. M. "The Equation of State for Liquid Heavy Water." *Teploenergetika* 7.4: 72 (1960).
- [6] Mamedov, A. M. "The Equation of State for Heavy Water According to Experimentally Determined p-V-T," *Teploenergetika* 7, (9), 71 (1960).
- [7] Plank, R. B. W. K. 13, 257 (1961).
- [8] Suvorov, N. P., Zhur. *Fiz. Khim.* 36, 216 (1962).
- [9] Ikeda, Munetaka, Kageyama and Nagashima. "Equation of State for D₂O in the Liquid Region of up to 1000 bar." *Bulletin of JSME* 20.149: 1492-1498 (1977).
- [10] Juza, J., Mares, R. "Equation of state for saturated and superheated steam D₂O up to 500°C." *Acta Technica CSAV*, 23(1), 1-10 (1978).
- [11] Hill, P. G., MacMillan, RD. and Lee, V. "A fundamental equation of state for heavy water." *Journal of Physical and Chemical Reference Data* 11.1: 1-14 (1982).
- [12] Herrig, S., Thol, M., Harvey, A. H., and Lemmon, E. W. "A reference equation of state for heavy water." *Journal of Physical and Chemical Reference Data*, 47(4) (2018).
- [13] Weber, L. A. "Extrapolation of Thermophysical Properties Data for Oxygen to High Pressures (5,000 to 10,000 PSIA) at Low Temperatures (100 - 600°R)" *Nat. Bur. Stand. (U.S.), Internal Report No. 10727* (1971).
- [14] Wagner, W. "Thermodynamic Properties of Oxygen from the Triple Point." *J. Phys. Chem. Ref. Data*, 20(5) (1991).
- [15] Richardson, I. A., Leachman, J. W., and Lemmon, E. W. "Fundamental equation of state for deuterium." *Journal of Physical and Chemical Reference Data*, 43(1) (2014).
- [16] Haslam, A. J., et al. "Expanding the applications of the SAFT- γ Mie group-contribution equation of state: Prediction of thermodynamic properties and phase behavior of mixtures." *Journal of Chemical & Engineering Data* 65.12: 5862-5890 (2020).
- [17] Chapman, Walter G., et al. "SAFT: Equation-of-state solution model for associating fluids." *Fluid Phase Equilibria* 52: 31-38 (1989).
- [18] Zmpitas, W., and Gross, J. "Detailed pedagogical review and analysis of Wertheim's thermodynamic perturbation theory." *Fluid Phase Equilibria* 428: 121-152 (2016).
- [19] Zmpitas, W. "Analysis and extension of Wertheim's thermodynamic perturbation theory." (2019).
- [20] Lafitte, T., et al. "Accurate statistical associating fluid theory for chain molecules formed from Mie segments." *The Journal of chemical physics* 139.15 (2013).
- [21] Dufal, S., et al. "Developing intermolecular-potential models for use with the SAFT-VR Mie

equation of state." *AIChE Journal* 61.9: 2891-2912 (2015).

[22] Graham, E. J., et al. "Multi-objective optimization of equation of state molecular parameters: SAFT-VR Mie models for water." *Computers & Chemical Engineering* 167: 108015 (2022).

[23] Marzouk, O., "Thermo Physical Chemical Properties of Fluids Using The Free NIST Chemistry Web Book Database." *Fluid Mechanics Research International Journal* (MedCrave Publishing Group) 1.1: 14-18 (2017).

[24] Walker, P. J., Hon-Wa Yew, and Andrés Riedemann. "Clapeyron.jl: An extensible, open-source fluid thermodynamics toolkit." *Industrial & Engineering Chemistry Research* 61.20: 7130-7153 (2022).

[25] Aasen, A., et al. "Equation of state and force fields for Feynman–Hibbs-corrected Mie fluids. I. Application to pure helium, neon, hydrogen, and deuterium." *The Journal of Chemical Physics* 151.6 (2019).

[26] Setthanan, U., David R. Morris, and Derek H. Lister. "Solubilities of H₂ in H₂O and D₂ in D₂O with Dissolved Boric Acid and Lithium Hydroxide." *Canadian journal of chemistry* 84.1: 65-68 (2006).

[27] Scharlin, P., and Battino, R. "Solubility of 13 nonpolar gases in deuterium oxide at 15–45° C and 101.325 kPa. Thermodynamics of transfer of nonpolar gases from H₂O to D₂O." *Journal of solution chemistry* 21: 67-91 (1992).

[28] Clark, T., Heske, J., and Thomas D. Kühne. "Opposing electronic and nuclear quantum effects on hydrogen bonds in H₂O and D₂O." *ChemPhysChem* 20.19: 2461-2465 (2019).

[29] Paesani, F. "Hydrogen bond dynamics in heavy water studied with quantum dynamical simulations." *Physical Chemistry Chemical Physics* 13.44: 19865-19875 (2011).

[30] Chaplin, M. "Explanation of the Density Anomalies of Water" *Water Structure and Science* (2022).

[31] Lappetito, L. "Water Molecule Vibrations with Raman Spectroscopy." *PhysicsOpenLab* (2022).

[32] Ceriotti, M., et al. "Nuclear quantum effects in water and aqueous systems: Experiment, theory, and current challenges." *Chemical reviews* 116.13: 7529-7550 (2016).

[33] Walker, P., J., and Haslam, A., J. "A New Predictive Group-Contribution Ideal-Heat-Capacity Model and Its Influence on Second-Derivative Properties Calculated Using a Free-Energy Equation of State." *Journal of Chemical & Engineering Data* 65.12: 5809-5829 (2020).

[34] Vega, C., et al. "Heat capacity of water: A signature of nuclear quantum effects." *The Journal of chemical physics* 132.4 (2010).

[35] Motoyuki, S. and Shinoda, W. "Calculation of heat capacities of light and heavy water by path-integral molecular dynamics." *The Journal of chemical physics* 123.13 (2005).

Electrochemical Reduction of CO₂: Insights into Cobalt Single-Atom Catalysts via a Decoupled Two-Step Synthesis

Hashem Ghandour and Mencia Izaga

Department of Chemical Engineering, Imperial College London, U.K.

Abstract Electrochemical CO₂ reduction offers a sustainable solution to environmental challenges by converting CO₂ emissions into valuable chemicals and fuels, using excess renewable energy to close the anthropogenic carbon cycle. This study explores single metal atom catalysts embedded in nitrogen-doped carbon (MNC), with emphasis on cobalt due to its potential high activity for the reduction of CO₂ to CO. Employing a two-step decoupled synthesis method, the aim was to successfully integrate cobalt into the highly porous nitrogen-doped carbon matrix, while preventing nanoparticle formation that typically limits SACs efficiency. Characterisation of the catalyst material using X-ray Photoelectron Spectroscopy (XPS) suggests high metal and oxygen content (8.39 wt% cobalt and 9.76 wt% oxygen respectively), which often entails moderate catalytic performance due to nanoparticle formation. However, electrochemical testing exhibited high CO faradaic efficiency (FE), reaching a maximum value of 97% at -0.8V vs. RHE, with a total current density of 17.6 mAcm⁻². These results positioned TAP900@Co as a competitive alternative amongst other state-of-the-art Co-SACs. Future work should include advanced analytical techniques to confirm cobalt aggregation into the formation of other species and elucidate their role in the catalyst's performance.

Keywords: electrochemical CO₂ reduction, single atoms catalysts, cobalt, transmetalation

1. Introduction

In 2015, leaders from 55 countries worldwide agreed on an international treaty aimed at combatting global warming, famously known as the Paris Agreement. This pact aspires to limit the increase of the global average temperature to no more than 1.5°C by the end of this century, primarily through the reduction of greenhouse gases (GHGs) emissions [1]. Fast forward to 2021, the AR6 Climate Change report predicted a concerning increase of 0.45°C for every 1000Gt of cumulative CO₂ emissions [2]. Furthermore, the International Energy Agency (IEA) recorded a significant 36.8Gt of CO₂ emissions in 2022, originating from global energy-related sources like power plants, automobiles and airplanes [3]. This data serves to reveal the role of CO₂ as one of the primary anthropogenic contributors to GHG emissions and emphasizes the necessity of long-term approaches for full decarbonisation, such as CO₂ capture and utilisation, in addition to emission reduction technologies [4].

address these environmental concerns and to ultimately close the anthropogenic carbon cycle [5]. As seen in Figure 1, by utilizing electricity derived from renewable sources and protons from an aqueous electrolyte, the eCO₂RR can transform CO₂ into a wide variety of feedstock chemicals, valuable for processes like fuel production, thereby standing as a promising way to eliminate our dependence on fossil fuels at a commercial scale. Nevertheless, the scale-up of this technology faces several barriers, including the thermodynamics and kinetics of the reaction, catalytic efficiency and stability, product separation and the purity and sourcing of CO₂ [5].

One of the main challenges arises from the competition of eCO₂RR in aqueous electrolytes with the hydrogen evolution reaction (HER), due to the overlapping of their thermodynamic potentials [6]. In addition, the reduction of CO₂ itself can result in multiple products (Figure 1) which share similar thermodynamic potentials [7], causing them to compete during the reduction process. Over time, researchers have explored different catalysts, reaction conditions and electrochemical set-ups with the aim of achieving high activity and stability catalysts that can selectively produce desired products, at low overpotentials.

Despite the difficulties in the development of these catalysts, the economic value of eCO₂RR products remains an incentive to continue the pursuit of more effective alternatives [8]. As illustrated in Figure 2, formic acid and propanol emerge as the most profitable eCO₂RR products. Metals such as Sn, Bi, Hg, In and Pb have been shown to produce formic acid as their main product [7],[9], due to their weak adsorption energy with the CO₂⁻ intermediate radical. Despite the high profitability of formic acid and propanol, their current low production levels (0.2MtC/yr and 0.1MtC/yr respectively) [8], are insufficient to meet the targets of the Paris Agreement. In contrast, C₂⁺ products like ethylene (120MtC/yr) and ethanol (40MtC/yr) offer a balance between profitability and the potential for an impactful eCO₂RR [8].

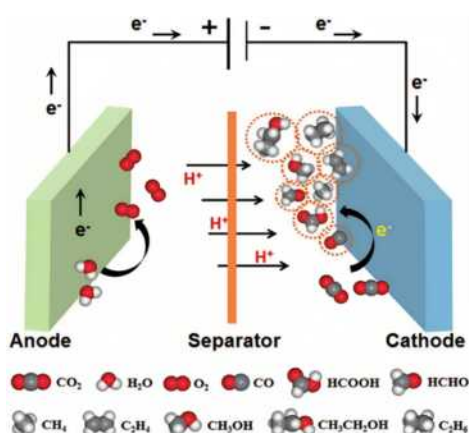


Figure 1. Schematic of the electrochemical set-up for the eCO₂RR. Reprinted from Ref [1].

The electrochemical CO₂ reduction reaction, or eCO₂RR, has emerged as a promising technology to

The state-of-the-art catalyst for hydrocarbons containing more than one carbon is copper (Cu), since it has demonstrated a unique ability to facilitate C-C coupling^[10]. It is the only metal that has weak binding to H₂ and moderate binding to CO₂, facilitating the further reduction of *CO into C₂₊ products, rather than the desorption and subsequent formation of CO₂ or the HER. However, Cu is still far from an ideal solution. The main issue lies in its lack of selectivity due to its involvement in several reaction pathways, resulting in a wide range of products which diminishes its effectiveness in selectively yielding a single desired product. This was illustrated in the work of Kuhl et al., who identified a total of 16 products when studying Cu across a range of potentials. The highest faradaic efficiency (FE%) for a C₂₊ product was the one for ethylene, 23% approximately^[11]. Additionally, the production of longer chain hydrocarbons includes multiple proton and electron transfer steps, which require high overpotentials (> than 1V), leading to energy inefficiencies within the process^[12].

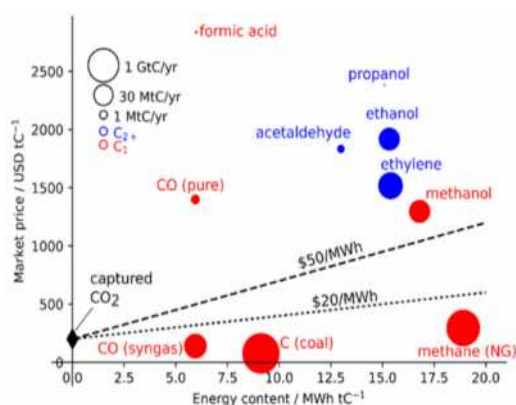


Figure 2. Market price of select CO₂ recycling products as a function of energy content. Reprinted from Ref^[8].

The formation of CO is a simpler electrochemical process involving the transfer of two protons and two electrons, making it a more practical option for study and application. Furthermore, it is usually accompanied by the HER resulting in syngas production, which can be used as feedstock in synthetic fuels production via the catalytic Fischer-Tropsch process^[13]. Currently, precious metals like gold (Au) and silver (Ag) are the state-of-the-art catalysts for CO formation, achieving high selectivity towards CO at low overpotentials^[14]. For example, ultrathin Au nanowires have shown the ability to reduce CO₂ to CO with a selectivity of 94%, at a potential of -0.35V vs. RHE^[15]. Silver nanoparticles have achieved greater than 95% selectivity towards CO₂, at a potential of -0.8V vs. RHE^[16]. Despite their high performance, earth-abundant catalysts are being studied as an alternative to reduce the reliance on these expensive and limited metals.

Single-metal-atom catalysts (SACs) offer a promising option. Isolated single-metal-atoms are evenly distributed on a conductive support, typically made of carbon-based materials such as carbon nanotubes, graphene or amorphous carbon^[17]. The

uniqueness of these catalysts lies in their configuration. Each metal atom is individually anchored to the support, ideally allowing every metal atom to act as an active site, thus maximising the utilisation of the material. This minimises metal wastage that is typically present in conventional, aggregated metal catalysts, where a significant number of metal atoms are rendered inactive due to lack of accessibility^[18].

Building upon the promise of SACs as optimal catalysts for eCO₂RR, efforts have been directed towards metal-nitrogen-doped-carbon (MNC) catalysts. These MNC materials can be sustainably prepared from sources like biomass and feature porous structures that ensure active sites are electrochemically exposed and conductive, as they are supported on a carbon substrate^[19]. A range of metal centres in MNC catalysts, including Mn, Fe, Ni, Co and Cu have been investigated to demonstrate their effect on activity and selectivity^[20]. The comparison between these metals is depicted in Figures 3a-d. Fe is the most active catalyst and has the ability to initiate eCO₂RR at low overpotential, due to the strong binding affinity to *COOH. Ni presents the highest selectivity towards CO, attributed to the weak bond with *H, hindering the HER. As a result, Ni and Fe containing catalysts are considered some of the most promising materials

In the synthesis of SACs, the final treatment often involves heating the catalyst material, which already contains the metal centre integrated in the nitrogen-doped carbon support, to temperatures ranging from 600°C to 1000°C. These high temperatures are known to enhance the conductivity of the material^[21]. However, they can also lead to the formation of metal oxides, nitrides and carbides, creating a diverse range of active sites, and thereby complicating the ability to draw clear conclusions regarding catalyst performance. Furthermore, such elevated temperatures can also facilitate the carbothermal reduction of metal ions into pure metal, which lacks functionality as an active site. At temperatures exceeding 800°C, it has been shown that a thin carbon layer can form on the elemental metal^[22]. This protects the metal atoms from being removed during acid washing and leads to their retention in the final catalyst, reducing the efficiency and utilisation of the catalyst material.

A new synthesis method (Figure 4) was proposed recently to prevent these challenges and limit the undesired aggregation of active sites, which occurs due to the system's inclination towards stability and the high surface energy of free-standing atoms^[22]. This approach decouples the synthesis of the support material from the low temperature metal coordination step through transmetalation, avoiding the production of metal oxides, nitrides and carbides. One recent study following this method showed promising outcomes for Ni and Fe metal centres, leveraging exceptional porosity to both inhibit aggregation of single atoms and maximise active site utilisation. However, such high active site utilisation comes at the compromise of site density, which inhibits the catalytic activity of the material^[23]. Future work should aim to enhance metal loading without sacrificing performance to establish the

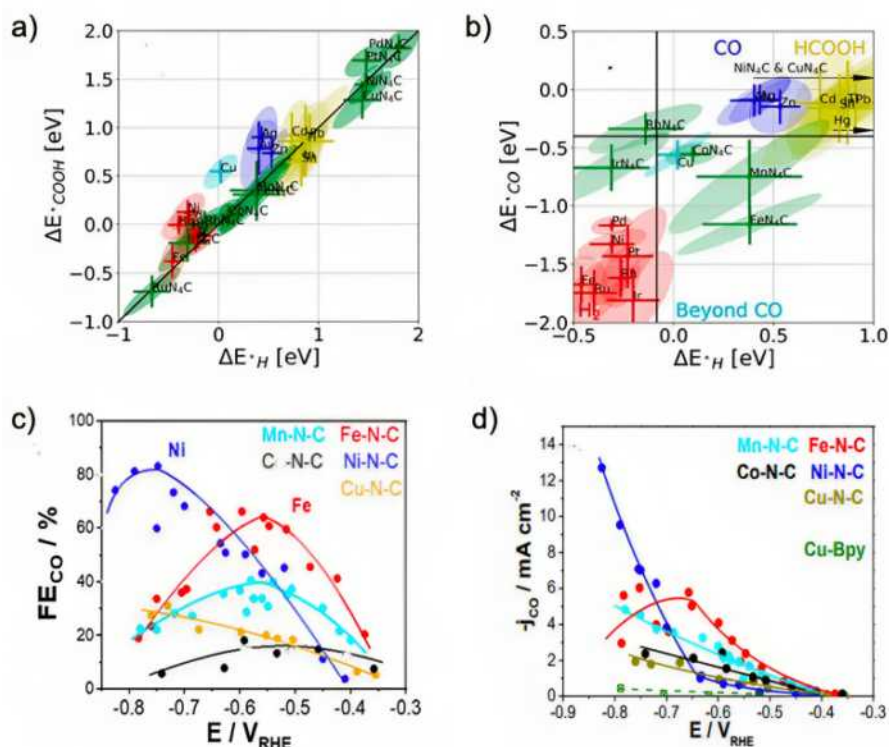


Figure 3. a-b Binding energies of CO₂ for elemental and MNC reduction electrocatalysts, towards *COOH, *CO and *H. c Faradaic Efficiencies (FE) vs. applied, IR-corrected electrode potential of CO. d Catalyst surface area-normalized CO partial currents vs. applied potential for the 6 catalysts. Reproduced from Ref^{[14],[19]}.

method as a commercially viable option for catalyst production.

The synthesis mentioned involves pyrolysis of the organic precursor 2,4,6-triaminopyridine (TAP) and a MgCl₂·6H₂O templating agent at 900°C to provide an adequate balance between nitrogen content, electrical conductivity, and porosity^[24]. This is followed by transmetalation, where a desired metal atom is coordinated through low temperature wet impregnation in methanol reflux, replacing Mg. This process led to 59±6% and 45±14% electrochemical utilisation for TAP900@Ni and TAP900@Fe respectively, which are unprecedented figures for MNC catalysts^[25]. TAP900@Fe displayed a slightly lower utilisation, which can be attributed to the inherent contamination of inactive Fe sites within the uncoordinated TAP900 as shown in Figure 5, where the metal content was calculated from inductively coupled plasma mass spectrometry (ICP-MS) measurements.

The notable utilisation is a result of the high mesoporosity of the catalyst, which arises from the interaction between TAP and Mg²⁺ salt. Upon heating, TAP organises around the salt's water molecules

through hydrogen bonds, melting together to form a homogeneous liquid that polymerizes without forming grain boundaries. This process yields a material with high porosity once acid washed, hindering aggregation and facilitating a relatively high density of electrochemically active sites on the nitrogen-doped carbon support (~3295m²g⁻¹)^[25]. The selectivity improvements are apparent when comparing with the previous data displayed in Figure 3, as TAP900@Fe presents a FE_{CO} of 93.5±3.7% at -0.55V, while TAP900@Ni displays a FE_{CO} of 95.3±4.7% at -0.59V, both at stable current densities of approximately 15mAcm⁻²^[25]

The acute sensitivity of cobalt to different coordination environments has been a subject of interest, with recent studies highlighting its potential for high FE in CO production. Although the CoNC material in Figure 3c displays a poor performance over the potential range applied, research investigating alternate support structures has achieved efficiencies towards CO exceeding 99% between -0.73V and -0.77V vs. RHE^[26]. The inconsistent performance between materials underscores the need for further investigation into the

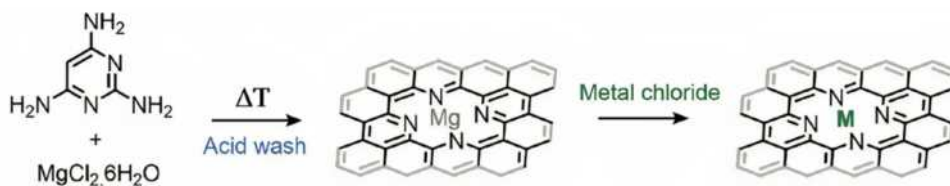


Figure 4. Schematic of the synthesis for the preparation of pyrolyzed TAP 900 and the subsequent low temperature metal coordination M. Reproduced from Ref^[25].

intrinsic catalytic abilities of cobalt for the eCO₂RR. The synthetic method used in this paper that coordinates cobalt into TAP900 through transmetalation, seeks to clarify these ambiguities. By leveraging the material's inherent high porosity, the intrinsic activity of cobalt can be accurately quantified, due to the direct interaction of e CO₂RR intermediates with fully exposed Co single atom sites. Research conducted into the d-band centres of Ni, Fe and Co supports the premise that cobalt's intrinsic activity is high^[27], as opposed to showing good performance only in specific tailored environments. This is evidenced by the proximity of cobalt's d-band centre to the fermi level, which facilitates the electron transfer from the metal d-band to the adsorbate. Characterisation showed that the d-band of Ni was closest to the fermi level, followed by Co and Fe. The CO₂ adsorption energy on the Co active site tested was also significantly stronger than that of the Ni active site, further substantiating the appeal of cobalt as a high performing electrocatalyst for CO₂ reduction.

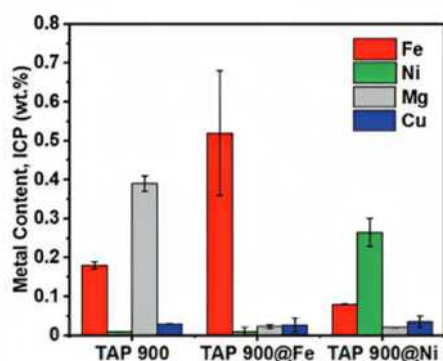


Figure 5. Metal loadings calculated through ICP-MS of different TAP900-derived catalysts. Reproduced from Ref^[25].

2. Experimental procedure

Synthesis. TAP (97% Sigma Aldrich) and magnesium chloride hexahydrate (99% Sigma Aldrich) were measured in a weight ratio of 1:8, then ground to a homogenous powder with a pestle and mortar. The mixture was then pyrolysed in a ceramic crucible that was filled to 1/3 capacity. The pyrolysis was conducted at 900°C for 3 hours in a N₂ atmosphere (>99.998%, BOC). The N₂ flowrate was maintained at 300mlmin⁻¹ and the heating rate was set at 5Cmin⁻¹. The material was collected, ground to a fine powder, and then acid washed overnight in 2M HCl, (prepared by dilution of fuming 37%, Merck), to eliminate any residual MgCl₂ and MgO. After acid washing, the powders were filtered extensively with DI water, then dried at 80°C under vacuum conditions. The final product was designated as “TAP900”.

Cobalt Coordination. To coordinate cobalt into the material, 60 mg of TAP900 were added to a 250ml round-bottom flask containing 75ml of MeOH (AnalaR NORMAPUR Reag. Ph. Eur., ACS, VWR). This mixture was stirred until a uniform dispersion was achieved. Following this, a 75ml solution of 25 × 10⁻³ M CoCl₂·6H₂O (98% Sigma Aldrich) in MeOH was prepared. The flask was connected to a reflux condenser

and subjected to a low temperature wet impregnation method, where it was continuously stirred at 90°C for 24 hours. Next, the material was filtered and rinsed with MeOH. It was then washed overnight with 0.5M H₂SO₄ (95-98% Sigma Aldrich) to remove any aggregated Co species. The cobalt coordinated catalyst, (TAP900@Co), was filtered thoroughly with DI water to remove the acid and dried at 80°C under vacuum.

Electrode preparation. To prepare the cathode ink, 12mg of TAP900@Co, and 40 mg of ball milled polytetrafluoroethylene, (PTFE), were ground to a homogenous powder with a pestle and mortar. PTFE was employed to act as a binder, due to its adhesive characteristics, and provide a hydrophobic layer between the electrolyte and catalyst. The hydrophobicity of PTFE prevents the electrolyte flooding onto the electrode, therefore allowing for the formation of a three-phase interface between the CO₂ passing through the flow field, the cathode, and the aqueous electrolyte^[25]. Since the solubility of CO₂ presents mass-transport limitations, it was crucial to form this interface as a way of ensuring consistent supply of gaseous CO₂ to maintain the concentration gradient that allows for diffusion from the bulk of the electrolyte to the catalyst surface. The mixture was dispersed in 10mL of isopropanol and spent 10 minutes in a sonication bath, followed by 10 minutes under probe sonication for further homogenisation. The probe sonicator followed a pattern of 5s on 5s off to avoid excess heat generation that could possibly damage the sample. The cathode ink was then air-brushed onto the hydrophobic side of 4x4 cm carbon paper, (GDL – Sigracet 39BB), which was then cut into 16 cathodes, each measuring 1 cm². The dilution with IPA and subsequent homogenisation served to reduce the probability of the air-brush clogging, and effectively disperse the particles in the ink for even spraying onto the electrode. For the anode ink, 100 mg of 40wt% Pt/C was dispersed in a 12mL solution of 80% DI water and 20% ethanol by volume. The ensuing steps taken were identical to the cathode preparation.

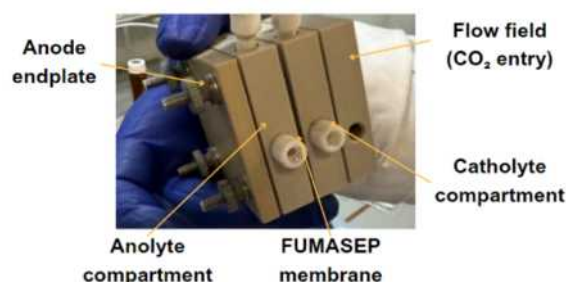


Figure 6. Image of the electrochemical setup.

Cell assembly. The electrolyte was prepared from 30ml of 0.5M KOH. It was placed in a centrifuge and saturated with CO₂ for 30 minutes to ensure sufficient supply of CO₂ at the electrode surface. The electrochemical cell used had a gas diffusion configuration and consisted of a flow field, catholyte compartment, anolyte compartment, 4 gaskets, a FUMASEP membrane that has been wet with DI water

in advance, and an anode end plate. Copper tape was added around the flow field and end plate to ensure efficient current distribution, improving contact with the electrodes. The compartments were then assembled as shown in Figure 6. Each compartment was separated by a gasket, which provided an additional seal to the cell, preventing any liquids or gases from escaping or entering. Once the cell was put together and tightened, the Ag/AgCl reference electrode was screwed into the catholyte compartment, and 1.5 ml of KOH electrolyte was added to the catholyte and anolyte compartments.

Electrochemical tests. Gas chromatography (GC) was paired with chronoamperometry (CA) at varying potentials utilizing an AUTOLAB PGSTAT302N potentiostat. This configuration enabled the simultaneous analysis of gaseous products and assessment of the electrocatalyst's performance. The GC is equipped with a Flame Ionisation Detector (FID), which is highly sensitive, for the detection of carbon-based products, alongside a Thermal Conductivity Detector (TCD) for sensing H₂. Before electrochemical measurements were taken, the GC was conditioned to remove any residual contamination from prior experiments. This required raising its temperature to 210°C and letting in Ar, H₂ and air at 10psi, 20psi and 5psi respectively.

Prior to CA, Frequency response analysis (FRA) and cyclic voltammetry (CV) measurements were conducted to correct the resistance of the electrolyte between the electrodes and stabilise the catalyst surface by forming an interface between the electrode and the electrolyte. All three tests were run using NOVA 2.1.4 software. Post-resistance adjustment, CV was conducted between 0 and -0.5V vs Ag/AgCl for 10 cycles with a 50% internal resistance (iR) compensation. Subsequently, CA was performed at -1.2, -1.3, and -1.4V vs. Ag/AgCl, each for 2400 s. The overpotential conversion to vs. RHE, is defined as:

$$E_{RHE} = E_{Ag/AgCl} + 0.59V \quad (1)$$

The gas chromatography (GC) system performed three injections during each CA run, each lasting 16 minutes, (13.25 min injection and 2.75 min cooling down), to monitor the formation of gaseous products, with peaks and currents noted at 300, 1260, and 2220 s into the CA. Each peak is proportional to the concentration of the detected gaseous species. Partial currents j_i for each product i were calculated as follows:

$$j_i = \frac{A \times F \times Q_{flow}}{V_m} \quad (2)$$

where A is the integrated area under the peak for each gas i from the GC data, F is the calibration constant specific to each gas i , V_m is the total volume of gas i that passes through the GC and Q_{flow} is the rate at which gas i is passed through the GC. Partial current densities were found by dividing j_i by the surface area of the working electrode S (1cm²). The Faradaic efficiency (%) of each product i was then found with:

$$FE_i = \frac{j_i}{j_{total}} \times 100 \quad (3)$$

where j_{total} is the total current density measured during each CA.

Material Characterisation. X-ray Photoelectron Spectroscopy (XPS) was performed in a Thermo Fisher K-Alpha XPS system and analysed using Avantage software. All spectra were calibrated relative to the carbon C1s peak at 284.8 eV for correcting for charging effects.

3. Results and Discussion

Characterisation Results. XPS confirms the incorporation of Co into the TAP900 framework, quantifying elemental composition in TAP900@Co by both weight (wt) and atomic (at) percent, as detailed in Table 1. The at% is derived by normalizing the wt% with the atomic mass of each element. This shows the primary constituents of TAP900@Co: carbon, followed by nitrogen, oxygen and cobalt.

Table 1: Compositions by weight % and atomic % of TAP900@Co, calculated from XPS data.

Element	Weight %	Atomic %
C1s	72.64	81.1
O1s	9.76	8.18
N1s	9.21	8.82
Co2p	8.39	1.91

TAP900@Fe synthesized in another work from TAP900 with 0.18 wt% Fe (from contamination of the TAP precursor) following the same method, allowed the coordination of 0.520 wt% of Fe single atoms [25]. This low metal content, characteristic of SACs, indicates the formation of highly dispersed single metal atoms on the surface of the catalyst. In contrast, TAP900@Co's high metal content of 8.39 wt% implies significant cobalt aggregation and the consequent formation of nanoparticles. This is further supported by the pronounced peak for Co2p in the XPS spectrum in Figure 7, where the peak's magnitude correlates with elemental concentration [28]. The O1s spectrum additionally indicates a high oxygen content, suggesting substantial oxidative characteristics of the sample.

The synthesis of TAP900@Co includes several critical steps where any slight deviation could introduce undesired high metal content. A crucial aspect is achieving the proper nitrogen content, around 4.5 at% [24], which provides the necessary number of lone electron pairs for the coordination of cobalt within the TAP framework. This step requires precise pyrolysis conditions. Should the furnace fail to reach or maintain the necessary temperature, or if the sample tube is not sealed properly, it could detrimentally impact the structural properties of TAP900. Other analytical techniques should be used to confirm the findings of XPS. Inductively coupled plasma mass spectroscopy (ICP-MS) could be applied to verify the chemical composition of the catalyst, since it is highly sensitive

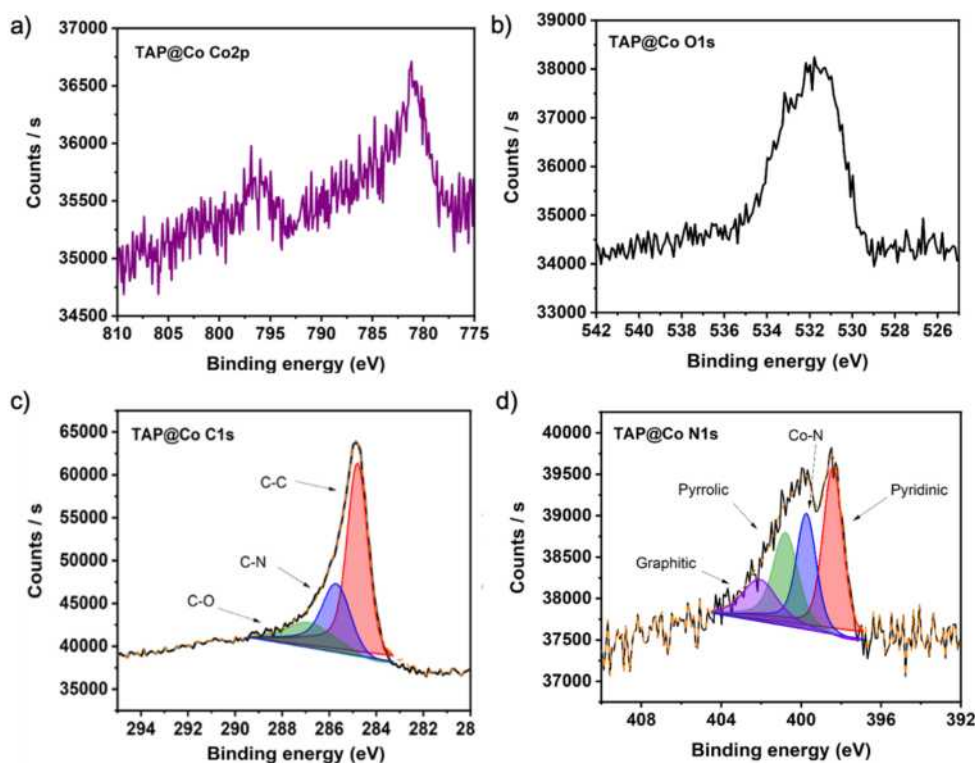


Figure 7. X-ray photoelectron spectra of TAP900@Co **a** Co 2p spectrum, **b** O 1s spectrum, **c** TAP900 C 1s spectrum and **d** TAP900@Co N 1s spectrum.

and can detect trace metals within the bulk composition of the measured material, as opposed to XPS, which is a surface analytical technique. The oxidation state and coordination environment around TAP900@Co could be further assessed by X-ray absorption spectroscopy (XAS) and extended X-ray absorption fine structure (EXAFS). To show the distribution of single atoms and verify the existence of metallic nanoparticles, high angle annular dark field scanning transmission electron microscopy (HAADF-STEM) should be employed, as well as X-ray diffraction (XRD). It would also be beneficial to quantify the porosity of TAP900@Co through Brunauer–Emmett–Teller (BET) analysis.

TAP900@Co shows three different contributions that correspond to C-C, C-N and C-O bonding. In the N 1s spectrum, a peak arises at 399 eV that corresponds to N-Co coordination. The N 1s spectrum also presents three other contributions that stand for pyridinic, pyrrolic and graphitic. Compared to TAP900 in Figure 8, the total contribution of the pyrrolic moieties decreases and the one corresponding to nitrogen coordinated to metals increases (Mg in the case of TAP900 and Co in the case of TAP900@Co), while the other components remain similar. This fact suggests that Co is coordinated via pyrrolic N rather than pyridinic.

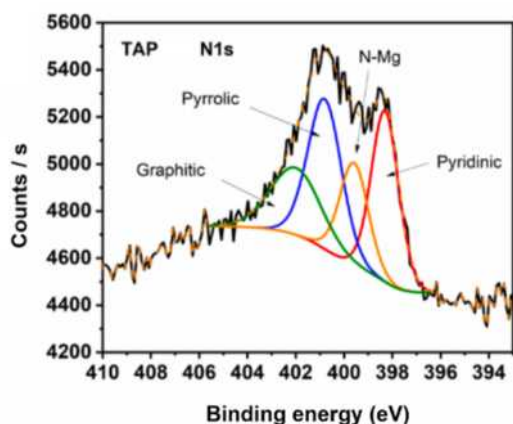


Figure 8. TAP900 N 1s spectrum. Reproduced from Ref^[24].

XPS spectra in Figure 7c and 8d confirm the metal loading after metal coordination, and the common features of MNC materials. The C 1s spectrum of

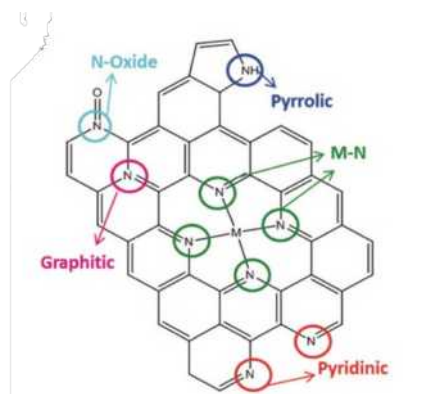


Figure 9. General structure of TAP900, reproduced from Ref^[12].

Electrochemical testing results. In eCO₂RR, current density is a key metric that indicates the rate of flow of electrons with respect to the working electrode.

Applying a negative potential to the electrode increases the energy of electrons, which can reach such a high energy that they transfer into the vacant states in the electrolyte [29], causing the reduction to take place. As the applied potential is increased, the reactants receive additional energy which facilitates their ability to overcome the activation energy barrier. Electrons are transferred more energetically and the mass transport of reactants to and products away from the electrode surface is enhanced, thereby increasing the reaction rate and, as a result, the current density.

Current density reaches a relatively steady state after 300s approximately, implying that the electrochemical environment has stabilised. The observed fluctuations in current density, specially at higher potentials suggest concentration overpotential, where the consumption rate of reactants exceeds their supply rate to the catalyst [30]. This is evident in Figure 9, where current oscillations could be attributed to the dynamic formation and detachment of gas bubbles that intermittently block and unblock electrode active sites. At -0.71V vs. RHE, the frequency of these fluctuations is increased. This implies that smaller gas bubbles are formed, which quickly detach. The quick detachment helps to avoid larger bubbles that can disrupt the process. Therefore, -0.71V vs. RHE is high enough to accelerate the reaction, while avoiding the limitations imposed by mass transport when reactants cannot reach the electrode fast enough, allowing the reaction to run with less disruption.

Partial current density represents the current associated with the formation of a specific product, serving as an indicator of a catalyst's activity. FE, on the other hand, measures the proportion of the total amount of current that contributes to the formation of a product, thus serving as an indicator of selectivity. It is also a significant parameter when it comes to industrial scalability. Although low faradaic efficiency towards CO can indicate the production of more profitable hydrocarbons, such as ethylene or ethane, the implications can include much greater separation and purification costs. To assess the performance of TAP900@Co, both factors were considered in this analysis, as shown in Figure 10.

With rising potentials, an increase in partial current densities was observed for both H₂ and CO, aligning with the general trend of current density increasing as a result of elevated energy received by electrons and reactants. The partial current of CO grows more sharply with higher potential, reaching a value of 17.1 mAcm⁻² at -0.8V vs. RHE, indicating a dominance over H₂ formation and therefore, the inhabitation of the HER. In contrast, the marginal increase in H₂ partial currents up to 1.7-2.5 mAcm⁻² implies that HER's dependency on applied potential is minimal, suggesting its occurrence regardless of potential adjustments.

At -0.6V vs. RHE, TAP900@Co displayed a relatively low FE_{CO} of 66%, without the presence of any other profitable products. One possible reason for this, in addition to the inadequate energy received by reactants mentioned earlier, could be insufficient structural changes in the catalyst surface. These structural changes, which lower the activation energy

for CO₂ reduction to CO by exposing desirable active sites, could become more pronounced as the applied potential is increased. This is proven as the chemical environment surrounding Co active sites in TAP900@Co can promote CO₂ reduction towards CO with approximately 82% and 97% selectivity at the potentials of -0.7V and -0.8V vs. RHE respectively.

The potential of -0.6V vs. RHE displays

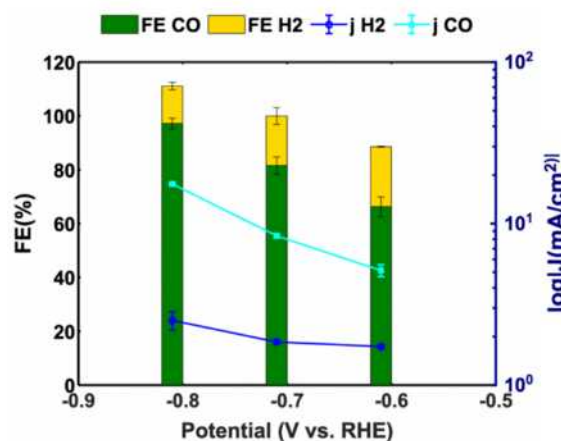


Figure 10. Catalytic performance of TAP900@Co as a function of applied potential (vs. RHE).

approximately 90% total FE. The discrepancy could be due to the formation of liquid byproducts, such as formate, which due to their anionic character [31] are not entirely constrained by the anion exchange membrane (AEM). At higher current densities and CO₂ flowrates, product concentration at the electrolyte increases. This can cause product crossover through the AEM and consequent oxidation at the anode, due to a greater diffusion rate. This suggests that a portion of liquid products may have formed because of the low CO₂ flowrate used. The fraction of liquid products caused by this electromigration effect was not significant, presumably as a result of the relatively lower current densities measured. Typically, electromigration is more impactful at current densities greater than 50 mAcm⁻² [31]. Considering the summation of total FE at -0.7V and -0.8V vs. RHE, it is possible that the structural changes that enhance CO formation begin to dominate over the electromigration effect at higher potentials. This finding can be supported by additional evidence, where similar TAP900 materials coordinated with Fe and Ni also experienced discrepancies in the summation of FE at only low potentials [25].

Assessing the specific surface area and total pore volume of TAP900@Co through BET analysis would have been pivotal in elucidating the notable results achieved at -0.7V and -0.8V vs. RHE. It is possible that, despite the complications encountered during synthesis, TAP900@Co was able to maintain a sizable fraction of the porosity achieved by TAP900@Fe. This is because the excessive presence of O₂, and subsequent aggregation of Co, likely evolved from imperfect vacuum conditions during the drying process. Although this would have diminished the porosity of the catalyst, TAP900@Co could have retained reasonable active site exposure as the porous

framework originated from the bubble templating effect during pyrolysis [24]. Therefore, the favourable results could be attributed to the inherent conductivity of the TAP precursor, with reasonable porosity and the hypothesised intrinsic activity of cobalt. However, it is difficult to draw definite conclusions without additional characterisation to fully understand TAP900@Co.

Table 2: Performance comparisons of TAP900@Co to other Co-SACs in literature. The compiled data corresponds to measurements at maximum FE_{CO} , reproduced from ref [26], [32], [33].

Catalyst	Potential (vs. RHE)	FE_{CO} (%)	J ($mAcm^{-2}$)
TAP@Co	-0.81	97	17.6
Single atom Co-N5 (HNPCSs)	-0.79	99	10.2
CoNC	-0.48	45	1.0
Co-Tpy-c	-0.70	97	~7.5

Table 2 shows cobalt catalysts tested in literature, covering a broad spectrum of performance characteristics. These variations in performance occur because of structural and chemical differences within the material that come about due to alternate paths of synthesis, giving rise to distinct coordination environments. The performance of TAP900@Co fares surprisingly well against top performing cobalt coordinated eCO_2RR catalysts, despite the complications that may have emerged during synthesis. The selectivity towards CO is among the top found in literature, although there is potential for minor overestimation due to the total FE summing to over 100%, likely through inherent experimental error with the GC. This error could have risen from inaccurate parameter calibration during set-up of the GC.

A Co-SAC with atomically dispersed Co sites anchored on hollow N-doped porous carbon spheres (HNPCSs) displayed an excellent FE_{CO} of 99%, that seems to arise due to the Co-N₅ active centre in the HNPCS environment [26]. As the coordination number drops, so does the selectivity towards CO. Remarkably, in this specific environment, commonly successful coordination metals such as Fe, Ni and Cu perform substantially worse. Additional study is evidently necessary into cobalt, as demonstrated by a Co-doped zeolitic imidazolate framework (ZIF-8) precursor that formed Co-N₄ atomic structures. This precursor achieved a maximum FE_{CO} of only 45%, despite Fe coordination in the same structure, exhibiting a FE_{CO} of 93% [32]. The superior performance of Fe over Co is generally expected. However, the precedent set by the HNPCSs presents a unique case where the FE_{CO} remained high when the coordination number changed from 5 to 4, only dropping from 99% to 91% [26]. It is possible that the microporosity of the ZIF-8 support structure does not promote eCO_2RR for cobalt-based catalysts, causing the poor selectivity and lack of activity towards CO₂ reduction.

Both the Co-HNPCS and Co-Tpy-c catalysts display significant increases in current density as the potential was raised past the displayed results in Table 2. This came at the compromise of CO selectivity. The greater of which, Co-Tpy-c, was able to reach a maximum current density of 46.6 $mAcm^{-2}$ at -1.2V vs. RHE, producing CO with a selectivity of approximately 72% [32]. The Co-HNPCS was only measured up to a potential of -0.88V vs. RHE and exhibited a current density of 17.5 $mAcm^{-2}$. However, it was still able to form CO with a selectivity near 90% [26]. This indicates testing TAP900@Co over a wider potential range would have been useful in determining its full capabilities, especially to establish the range of potentials that were able to produce CO with high selectivity. Additionally, although TAP900@Co has shown high current densities, long term testing would have given more meaning to the results, as catalyst stability is a crucial factor to consider when demonstrating what could make an ideal industrial catalyst.

4. Conclusion

TAP900@Co was prepared via a decoupled synthesis approach, which involved the coordination of cobalt sites into a highly porous nitrogen-doped carbon structure. The material achieved a high FE_{CO} of 97% at -0.81V vs. RHE, with a total current density of 17.6 $mAcm^{-2}$, making it a competitive catalyst amongst state-of-the-art cobalt single-atom catalysts in literature. However, an unexpected high cobalt and oxygen content was found, as evidenced by the 8.39 wt% of Co and 9.76 wt% of O₂ calculated, and large Co2p and O1s peaks. Further analytical techniques should be employed to confirm cobalt aggregation and oxidation of species during the synthesis, and to ascertain their nature and impact on catalyst performance. Analysing the electrolyte post-electrocatalysis via HPLC could confirm the formation of liquid products, validating the explanations for total FE remaining below 100% at -0.6V vs. RHE. Assessing the D/G band intensity ratio through Raman spectroscopy would have identified possible defects and grain boundaries within the material, as well as the degree of graphitisation of TAP900@Co. Gaining this information would have given insights into the conductivity of the material. Additional experiments with varying electrolytic concentrations or solutions, such as NaHCO₃ and KHCO₃, may shed light on the CO₂ reduction conditions favourable to TAP900@Co, as these would have offered greater stability than KOH (at the compromise of conductivity, due to the higher pH of KOH). Further research is vital to meet the industrial benchmarks for catalyst performance, which consist of current densities and selectivity exceeding 200 $mAcm^{-2}$ and 90% respectively, with long term stability greater than 80,000 hours, at minimised overpotentials [34]. Despite the shortcomings Co-SACs face to achieve this criterion, cobalt has clearly displayed interesting and encouraging capabilities that this paper set out to find, adding to the foundation that is currently being set up in the novel field of electrochemical CO₂ reduction.

Acknowledgements

The authors of this report would like to express thanks to Dr Jesus B.Hermida for guidance and conduction of XPS, as well as Jinjie Zhu and Helena Perez del Pulgar Villena for their experimental support.

References

- [1] UNFCCC (2015). *The Paris Agreement*. [online] United Nations Climate Change. Available at: <https://unfccc.int/process-and-meetings/the-paris-agreement>.
- [2] Stephens, L. et al. (2022). 2022 roadmap on low temperature eCO₂RR. *JPhys energy*, 4(4), pp.10–12. doi:<https://doi.org/10.1088/2515-7655/ac7823>.
- [3] International Energy Agency (2023). *CO₂ Emissions in 2022 – Analysis*. [online] IEA. Available at: <https://www.iea.org/reports/co2-emissions-in-2022>.
- [4] Sharifian, R., et al. (2021). CO₂ capture to close the carbon cycle. *Energy & Environmental Science*, [online] 14(2), pp.781–814. doi:<https://doi.org/10.1039/D0EE03382K>.
- [5] Chen, C., et al (2018). Progress toward Commercial Application of Electrochemical Carbon Dioxide Reduction. *Chem*, 4(11), pp.2571–2586. doi:<https://doi.org/10.1016/j.chempr.2018.08.019>.
- [6] Ooka, H., et al. (2017). Competition HER and eCO₂RR on Cu Electrodes. *Langmuir*, 33(37), p.9307.
- [7] Zhang, et al. (2020). eCO₂RR: from fundamental principles to catalyst design. *Materials Today Advances*, 7, pp.1–4. doi:<https://doi.org/10.1016/j.mtadv.2020.100074>.
- [8] Ruiz-López et al. (2022). Electrocatalytic CO₂ conversion to C₂ products. *Renewable and Sustainable Energy Reviews*, 161(112329), pp.1–2. doi:<https://doi.org/10.1016/j.rser.2022.112329>.
- [9] Zhu, D.D., Liu, J.L. and Qiao, S.Z. (2016). Recent Advances in Inorganic Heterogeneous Electrocatalysts for eCO₂RR. *Advanced Materials*, 28(18), pp.3423–3452. doi:<https://doi.org/10.1002/adma.201504766>.
- [10] Nitopi, S., et al. (2019). Progress and Perspectives of eCO₂RR on Copper in Aqueous Electrolyte. *Chemical Reviews*, 119(12), pp.7610–7672. doi:<https://doi.org/10.1021/acs.chemrev.8b00705>.
- [11] Kuhl, et al. (2012). New insights into eCO₂RR on metallic Cu surfaces. *Energy & Environmental Science*, [online] 5(5), p.7050. doi:<https://doi.org/10.1039/c2ee21234j>.
- [12] Zheng, Y, et al. (2019). Understanding the Roadmap for eCO₂RR to Multi-Carbon Oxygenates and Hydrocarbons on Cu-Based Catalysts. *Journal of the American Chemical Society*, 141(19), pp.7646–7659. doi:<https://doi.org/10.1021/jacs.9b02124>.
- [13] Dry, M.E. (2001). High quality diesel via the Fischer-Tropsch process - a review. *Journal of Chemical Technology & Biotechnology*, 77(1), pp.43–50. doi:<https://doi.org/10.1002/jctb.527>.
- [14] Bagger, A., et al. (2017). eCO₂RR: A Classification Problem. *ChemPhysChem*, [online] 18(22), pp.3266–3273. doi:<https://doi.org/10.1002/cphc.201700736>.
- [15] Zhu, W., et al. (2014). Active and Selective Conversion of CO₂ to CO on Ultrathin Au Nanowires. *Journal of the American Chemical Society*, 136(46), pp.16132–16135. doi:<https://doi.org/10.1021/ja5095099>.
- [16] Ma, S., et al. (2014). Silver Supported on Titania as an Active Catalyst for eCO₂RR. *ChemSusChem*, 7(3), pp.866–874. doi:<https://doi.org/10.1002/cssc.201300934>.
- [17] Nguyen, et al. (2020). Fundamentals of eCO₂RR on SACs. *ACS Catalysis*, 10(17), pp.10068–10095. doi:<https://doi.org/10.1021/acscatal.0c02643>.
- [18] Ren, S., et al. (2023). Catalyst Aggregation Matters for Immobilized Molecular CO₂RR. *Journal of the American Chemical Society*, 145(8), pp.4414–4420. doi:<https://doi.org/10.1021/jacs.2c08380>.
- [19] Varela, et al. (2018). Molecular NC Catalysts, Solid MOF Catalysts, and MNC Catalysts for the eCO₂RR. *Advanced Energy Materials*, 8(30), pp.17–20. doi:<https://doi.org/10.1002/aenm.201703614>.
- [20] Ju, W., et al. (2017). Understanding activity and selectivity of MNCs for eCO₂RR. *Nature Communications*, 8(1), pp.3–4. doi:<https://doi.org/10.1038/s41467-017-01035-z>.
- [21] Mehmood, et al. (2017). Facile Metal Coordination of Active Site Imprinted Nitrogen Doped Carbons. *Advanced Energy Materials*, 8(9), p.1701771. doi:<https://doi.org/10.1002/aenm.201701771>.
- [22] Glatzel, S., et al. (2013). From Paper to Structured Carbon Electrodes by Inkjet Printing. *Angewandte Chemie International Edition*, 52(8), pp.2355–2358. doi:<https://doi.org/10.1002/anie.201207693>.
- [23] Fang, J., et al. (2023). The synthesis of SACs for heterogeneous catalysis. *Chemical Communications*, [online] 59(20), pp.2854–2868. doi:<https://doi.org/10.1039/D2CC06406E>.
- [24] Barrio, J., et al. (2023). FeNC Oxygen Reduction Electrocatalyst with High Utilization Penta-Coordinated Sites. *Advanced Materials*, 35(14), p.2211022. doi:<https://doi.org/10.1002/adma.202211022>.
- [25] Saurav Ch. Sarma, et al. (2023). Reaching the Fundamental Limitation in CO₂ Reduction to CO with SACs. *Advanced Functional Materials*, 33(41), p.2302468. doi:<https://doi.org/10.1002/adfm.202302468>.
- [26] Pan, Y., et al. (2018). Design of Single-Atom Co–N₅ Catalytic Site. *Journal of the American Chemical Society*, 140(12), pp.4218–4221. doi:<https://doi.org/10.1021/jacs.8b00814>.
- [27] Cui, X., et al. (2018). Turning Au Nanoclusters Catalytically Active for Visible-Light-Driven CO₂ Reduction. *Journal of the American Chemical Society*, 140(48), pp.16514–16520. doi:<https://doi.org/10.1021/jacs.8b06723>.
- [28] Stevie, F. et al. (2020). Introduction to XPS. *J. Vac. Sci. Technol. A*, [online] 38, p.63204. Available at: <https://mmrc.caltech.edu/XPS%20Info/Practical%20Guides%20to%20XPS/Intro%20to%20XPS.pdf>.
- [29] Bard, et al. (2000). *Electrochemical Methods: Fundamentals and Applications*, 2nd Edition. Wiley Global Education.
- [30] Gabardo, et al. (2019). Continuous eCO₂RR to Concentrated Multi-carbon Products Using a Membrane Electrode Assembly. *Joule*, 3(11), pp.2777–2791. doi:<https://doi.org/10.1016/j.joule.2019.07.021>.

- [31] Zhang, J., et al. (2020). Crossover of liquid products from eCO₂RR through GDE and AEM. *Journal of Catalysis*, 385, pp.140–145. doi:<https://doi.org/10.1016/j.jcat.2020.03.013>.
- [32] Pan, F., et al. (2018a). Unveiling Active Sites on NC and Atomically Dispersed Iron and Cobalt Catalysts. *ACS Catalysis*, 8(4), pp.3116–3122. doi:<https://doi.org/10.1021/acscatal.8b00398>.
- [33] Hou, P., et al. (2020). Well-Defined Single-Atom Cobalt Catalyst. *Small*, 16(24). doi:<https://doi.org/10.1002/sml.202001896>.
- [34] Xu, D., et al. (2022). Electrocatalytic eCO₂RR towards industrial applications. *Carbon Energy*, 5(1). doi:<https://doi.org/10.1002/cey2.230>.

Data-driven Modelling and Prediction of Complex Systems Using Neural ODEs

Eran Emirzadeoglulari and Mario Triguero Munoz
Department of Chemical Engineering, Imperial College London, U.K.

Abstract Dynamical systems, i.e., systems which involve observable quantities that evolve over time, are omnipresent in our ever-changing world. From pandemic evolution prediction to weather forecasting, being able to accurately predict future changes in our environment is crucial for the greater good. Therefore, scientists and engineers are constantly working on methods to model future outcomes using past data. In fields such as Chemical Engineering, the ability to precisely model a complex system and its time dependent behaviour leads to a fundamental understanding of said system, which yields the benefit of facilitating the design of robust and highly optimal operations. Conventional methods often fail to find solutions to complex engineering problems. Hence, as part of a rapidly growing field, engineers can harness the power of Artificial Intelligence, particularly that of Machine Learning, for effective, data-driven approaches to complex problems. One recently developed model of deep neural networks, Neural Ordinary Differential Equations (Neural ODEs), is used to predict the dynamics of three prototypical models belonging to distinct generic classes: the Lotka-Volterra system in population biology, the SIR model in epidemiology and the Lorenz system in chaotic dynamics. This paper serves as a proof of concept that aims to explore the limitations and capabilities of Neural ODEs. Different validation techniques were used for each of the three systems, in consideration of their distinct nature. Ultimately, the Neural ODE model was successful in capturing the underlying dynamics of the Lotka-Volterra and SIR models but shows limitation in predicting the chaotic dynamics of the Lorenz system.

Keywords Neural Ordinary Differential Equations, SIR Model, Lotka-Volterra System, Lorenz System

1. Introduction

Nowadays, the society faces many dynamic challenges such as pandemics and climate change. For the most part of the last 4 years, the Covid-19 pandemic has had devastating impacts on our daily lives, and it has been necessary to take a plethora of precautions, such as lockdowns, to control and prevent the spread of the virus. At the same time, Earth's climate is changing, resulting in extreme temperature fluctuations, which in turn increase the frequency of severe weather phenomena. Both of these natural phenomena have something in common: they are complex dynamical systems with numerous intricate interactions and dependencies.

Dynamical systems are time-dependent mathematical models that illustrate the behaviour of an artificial or natural system [1]. Fundamental understanding of dynamical systems is crucial if the aim is to predict its behaviour, and thus engineer a highly optimal and robust system in order to respond to the challenges imposed by said behaviour.

There are numerous systems that interact with one another and change as a result of their interactions. The vast majority of these systems are characterised by non-linearity and high dimensionality. Thus, the task of modelling and predicting them with traditional methods is strenuous. Scientists and engineers continuously try to find patterns based on past data from these systems, to understand their behaviour and make predictions by using data-driven techniques. Artificial Intelligence (AI) and Machine Learning (ML) in particular, is a current trend and a rapidly growing field with numerous applications regarding the modelling of complex engineering problems which cannot be solved with conventional methods. It allows for rapid and

accurate decision-making, comparable to human analysis, especially when processing large amounts of data. Specifically, ML can be applied on a wide spectrum of problems, ranging from modelling, pattern recognition and classification to forecasting and estimation, with outstanding performance. While achieving precise predictions with absolute certainty might be impossible for some systems, e.g., chaotic systems, ML techniques still enable a basic understanding of the system's governing principles.

For the scope of this research paper, a novel type of neural network models, called Neural Ordinary Differential Equations (Neural ODEs), was employed for the modelling and prediction of prototypical systems, each representative of a generic field of science: population biology, epidemiology, and chaotic dynamics. This proof-of-concept approach is performed to illustrate the capabilities and limitations of Neural ODEs. All the systems that were modelled come in the form of a system of ODEs. The solutions to said systems of ODEs yield the datasets that are then used for the validation of the effectiveness of their corresponding Neural ODEs in reconstructing the underlying dynamics. Furthermore, these solutions represent a time-dependent trajectory, and the data is therefore time series data. Hence, appropriate validation methods were selected in order to take into account the time dependency of the data.

2. Background

2.1 Neural Networks

Neural Networks are a subset of ML algorithms whose origin can be traced back to the development

of the perceptron in the 1950s [2]. They were originally modelled to loosely represent the human brain [3], as they are composed of interconnected neurons, which vaguely represent the neurons in our brains, and the connections between them represent the synapses. Said neurons and connections can be structured in many ways, that is to say that neural networks have many different architectures.

The most basic example of a neural network architecture is that of a Feedforward Neural Network, or FNN, which consists of a series of layers, where the first and last layers are those corresponding to the input and output layers, respectively. In between said layers are additional layers, called hidden layers. Each layer contains a series of neurons, where each neuron is connected to all neurons in the previous and next layer. A characteristic of FNNs, is that the flow of information is unidirectional, i.e., the inputs sequentially go through the layers of the network, where a series of transformations are applied to them, until they reach the output layer. Said transformations are linear, such as those resulting from the weights and biases, which are the parameters of a neural network, and nonlinear, such as those applied by an activation function at each neuron.

In the context of this paper, neural networks can be thought of as a means of non-linear regression, and their role is to accurately predict the underlying dynamics of some data. Neural networks are capable of doing so given that, under certain conditions, they have been proven to be universal function approximators [4]. Therefore, in theory, a neural network should be able to model and fit any possible trajectory to a given dataset.

Neural networks achieve this by learning as a result of training. The first step of training is the forward pass, where the data passes through the network from the input to the output layer. Once the information has reached the output layer, the network makes a prediction, which is then compared to the true data of the input layer. The comparison yields some error, typically quantified by a loss function. Said loss function is then minimised using optimisation algorithms such as stochastic gradient descent, in order to obtain the values of the parameters at the minimum of the loss function. Next, the parameters of the network are updated via a process called backpropagation. Lastly, the process is repeated in an iterative fashion for a given number of steps, called epochs, or until the network predictions are within some specified tolerance.

Oftentimes, especially in fields such as Chemical Engineering, the raw time-series input data stems from highly complex, nonlinear systems of differential equations. Therefore, in order to uncover the relationships present in these highly complex datasets, the number of hidden layers in a network is

increased, with the purpose of introducing a larger number of transformations and increase the dimensionality, so that the network will be able to interpret the more intricate dynamics of the system. The procedure of creating and training networks with more than one hidden layer is called deep learning, which has gained significant momentum in recent decades [2].

2.2 Neural Ordinary Differential Equations

At the 2018 conference on Neural Information Processing System (NIPS), Chen et al. introduced a new family of deep neural network models, called Neural Ordinary Differential Equations [5]. The main characteristic of these models is that the networks are of continuous depth, rather than composed of discrete layers, as all standard neural networks are. This is illustrated in Figure 1. The idea originates from an observation that there exists an inherent similarity in the structure of ODEs and Residual Networks (ResNets).

ResNets are a type of FNN first introduced in 2015 [6], which utilise skip connections such that the output of a layer is added to the output of the next layer, therefore skipping the transformation of this next layer. Hence, a discrete step in a ResNet network is described by:

$$\mathbf{z}_{i+1} = \mathbf{z}_i + f(\mathbf{z}_i, t_i, \theta) \quad (1)$$

where \mathbf{z}_i is the input to the current layer, \mathbf{z}_{i+1} is the layer's output and f the transformation applied by some network's layer i , and θ represents said layer's weights. If we multiply a small constant h to the function f , this can be seen as a single step of Explicit Euler method:

$$\mathbf{z}_{i+1} = \mathbf{z}_i + hf(\mathbf{z}_i, t_i, \theta) \quad (2)$$

$$\frac{\mathbf{z}_{i+1} - \mathbf{z}_i}{h} = f(\mathbf{z}_i, t_i, \theta), \quad (3)$$

where:

$$\frac{d\mathbf{z}_i(t)}{dt} = f(\mathbf{z}_i, t_i, \theta). \quad (4)$$

Based on this observation, in the limit, the continuous derivative can be parameterised by a neural network by an ODE:

$$\frac{d\mathbf{z}(t)}{dt} = f(\mathbf{z}(t), t, \theta). \quad (5)$$

A parametrised derivative is a derivative function where the exact form of the function is not known a priori but is instead determined (parametrised) by a set of parameters that are learned through a training process. Therefore, in a Neural ODE, the derivative of the state of a system with respect to time is represented by a function, which is not explicitly defined as in traditional theoretical models, but

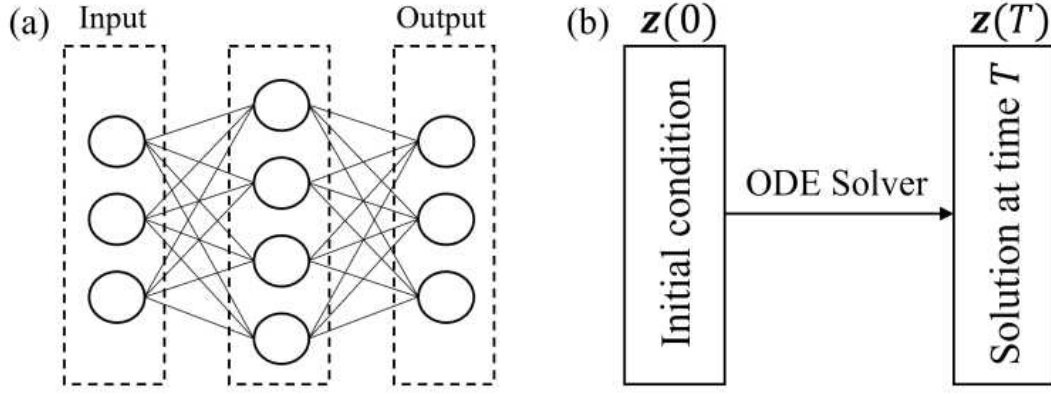


Figure 1: (a) Feedforward neural network with one hidden layer. (b) Neural ODE.

rather represented by a neural network. Therefore, Neural ODEs are called augmented ODEs, since they can uncover dynamics even when dealing with the noisy data that one collects in practice.

Given an initial condition $\mathbf{z}(0)$, the output $\mathbf{z}(T)$ is specified as the solution of the ODE at time T . The output can be computed with any desired accuracy using an ODE solver as:

$$\begin{aligned}\mathbf{z}(T) &= \mathbf{z}(0) + \int_0^T \frac{d\mathbf{z}(t)}{dt} dt \\ &= \mathbf{z}(0) + \int_0^T f(\mathbf{z}(t), t, \theta) dt.\end{aligned}\quad (6)$$

Neural ODEs come with several benefits. Since the network can be defined and evaluated using an ODE solver, it can take advantage of the error control and adaptive strategies of modern ODE solvers. Additionally, these models have constant memory cost and can explicitly trade numerical precision for speed [5]. Most importantly, due to their continuous nature, Neural ODEs can naturally incorporate data which arrives at arbitrary times, i.e., irregular interval time-series data, as opposed to traditional discrete neural networks, which impose a predefined, fixed time step. Neural ODEs are therefore ideal for complex systems in engineering as they can handle noisy, irregularly sampled data whose dynamics are complex, highly non-linear and previously unknown or not accurately modelled.

3. Application in Population Dynamics: Lotka-Volterra system

The Lotka-Volterra system, often referred to as the predator-prey equations, is a model in mathematical biology, specifically the field of population dynamics, developed in the 1920s, which consists of a pair of first-order nonlinear differential equations. This system describes the population dynamics of two interacting species, where one acts as a predator and the other acts as prey, i.e., the model describes how the populations of said species evolve over time under the influence of their mutual interactions.

If we denote the populations of the prey and predator by x and y respectively, the Lotka-Volterra equations are:

$$\frac{dx}{dt} = \alpha x - \beta xy \quad (7)$$

$$\frac{dy}{dt} = -\gamma y + \delta x, \quad (8)$$

where α and γ describe, respectively, the maximum prey per capita growth rate, and the effect of the presence of predators on the prey growth rate, and β and δ represent the predator's per capita death rate, and the effect of the presence of prey on the predator's growth rate, respectively. All parameters are positive and real.

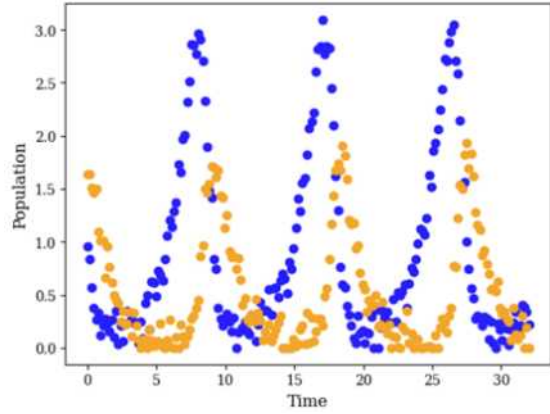


Figure 2: Noisy data generated for the Lotka-Volterra system. Prey and predator data shown in blue and orange, respectively.

The goal is to employ a Neural ODE to model the predator prey dynamics. Firstly, the system is solved in order to generate the dataset, and then Gaussian noise, with a standard deviation of 0.1, was added to the population data, in a way that no datapoint can take negative values, as that would be unphysical. This was done by forcing any negative values to be zero. Then, the resulting data is converted into tensors and divided into two categories: the first one

contains the first 80% of the data, which will be used for training, whilst the second one contains the rest of the data, which will be used for validation. This is done so that the validation is performed on unseen data, therefore evaluating the model's ability to generalise. Generalisation is a key aim of neural network training as it ensures that the model will function in a live environment.

The training-validation strategy chosen for the Lotka-Volterra system was random time series batching. This method involves dividing the training and validation datasets into smaller batches, such that each batch contains a sequence of datapoints from the time series data. The continuous and sequential nature of the original dataset is conserved within each batch. Next, in order to avoid overfitting to a sequence of data, a batch is selected randomly from the training data. Overfitting occurs when a model memorises the data and fits exactly to said data, such that it also memorises the noise in the data rather than the actual dynamics of the system. As a result of overfitting, the model will perform extremely well on the training set but poorly on the validation set.

Training is then performed on the randomly selected batch, followed by validation on a different random batch, this time selected from the validation set. This constitutes a single epoch, and the process continues until a specified number of epochs have passed. Random time series batching essentially guarantees that the model does not memorise the order of the data in the dataset, resulting in a more robust model.

Given that the Lotka-Volterra system is ultimately a simple system, and since this work is a proof of concept, the network only has a single layer of 100 neurons. The activation function was chosen to be ReLU, simply due to the fact that it is computationally inexpensive and, when it comes to deep learning, it is the most used in practice [7]. The weights of the network were initialised with a normal distribution centred around zero, with a standard deviation of 0.1.

The biases were initialised as a constant value of 0. The optimiser used was Adams and the solver used was dopri5, formally known as Dormand-Price method. Dopri5 was chosen as it is an explicit, adaptive solver of high order with good error control. Weight decay was added to discourage overfitting, and early dropout was implemented to reduce training time by stopping the training process if the model performed well enough before the training loop finished.

4. Application in Epidemiology: SIR Model

Until today, the world has experienced many types of infectious diseases which were caused by viruses and bacteria. Since the onset of the COVID-19 pandemic, there has been a global focus on

preventing the virus' spread for over two years. Understanding how populations respond to such outbreaks is critical and has direct implications for our daily lives. The time evolution of these population dynamics can be modelled using non-linear dynamical systems. In this section, an epidemic model proposed by Kermack and McKendrick was investigated and modelled using Neural ODEs.

Epidemics are sudden outbreaks of diseases, in a specific region, where the number of disease cases grows quickly [8]. In 1927, Kermack and McKendrick introduced a mathematical model that accurately captures the dynamic patterns observed in epidemiological studies. Their model, which aligns closely with the trends observed in multiple epidemics, reduces the complex interactions into a simple system consisting of three distinct population groups: susceptible, infected, and removed which are denoted by S , I and R respectively. The SIR model is an epidemiological model based on simple assumptions on the rates of flow between different classes of members of the population. In this model, it was assumed that the population size remains constant (meaning no entry or departure from the population) and that, when individuals recover, they gain immunity against the re-infection, and are thus removed from the susceptible population [9].

Usually, the SIR system can accurately model viral diseases' behaviour [9]. $S(t)$ shows the number of individuals who are susceptible to the disease but who are not infected yet. $I(t)$ number individuals who are infected and have the chance to spread the disease through various ways such as contact. $R(t)$ number of individuals who were infected but then removed from the possibility of spreading the disease or being infected again [9].

The SIR model used in this paper consists of a set of three ordinary differential equations:

$$\frac{dS}{dt} = -\beta SI \quad (8)$$

$$\frac{dI}{dt} = \beta SI - \alpha I \quad (9)$$

$$\frac{dR}{dt} = \alpha I, \quad (10)$$

where α is a positive real parameter which represents the rate of infected individuals leaving the infective class per unit time and β is a parameter which denotes the contact rate among the individuals who are in susceptible and infective class. With these parameters and initial conditions, the basic reproduction number (R_0) can be calculated by $R_0 = \frac{\beta S_0}{\alpha}$ where $S_0 = S(0)$. If $R_0 > 1$, the infection spreads and if $R_0 < 1$ the infection dies out [9]. Neural ODEs were trained and tested on this simple epidemic model to observe the predictions of the neural network. Firstly, before the training process,

SIR data were generated with an ODE solver with the assumptions that the time span is 50 days, $\alpha=0.1$ and $\beta=0.001$. Moreover, it was assumed that the total number of individuals (N) in a population is 1000 where $N = S + I + R$ and $S_0 = 997, I_0 = 3$ to make this specific case an epidemic, $R_0 > 1$.

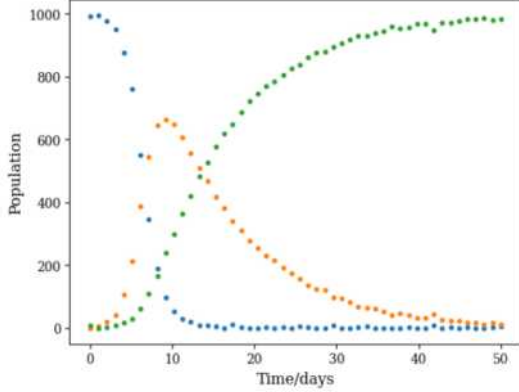


Figure 3: Generated true SIR model dataset with Gaussian noise, blue for Susceptible, orange for Infected and green for Removed.

Finally, Gaussian noise was added with a standard deviation of 5 to emulate a real-world example and to make network predictions more robust. This was implemented in a way that there are no negative values, as that would be unphysical. This was done by implementing a constraint similar to that used in Lotka-Volterra. Additionally, mass is conserved by equating the summation of all standard deviations at

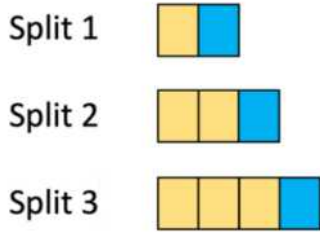


Figure 4: Yellow blocks representing training data whereas blue blocks representing validation data.

each time step to zero.

Unlike Lotka-Volterra model, random time-series batch technique was not utilised and instead a “extending window” strategy was employed to improve the learning. Since the trajectories of the Lotka-Volterra system are inherently cyclical, the neural network can capture the trend despite being trained with random batches of data. This is not true for the SIR model, and it cannot predict accurately when the batches are selected randomly. In the extending window strategy, a window of data is divided into training set and validation set. This means that, right at the beginning, the training window consists only of the first datapoint for training, and the datapoint immediately after for validation. Figure 4 depicts that at each timestep, the

number of training data increases by 1 and the validation data moves one step along the window, expanding the window size. This way, network uses the weights and biases from the previous data to predict the unseen validation data.

This network involves a single layer of 100 neurons and the activation function was chosen to be ReLU, due to the same reasons stated in Lotka-Volterra. The weights and biases of the network were initialised with a normal distribution centred around zero, with a standard deviation of 0.1 and 0.06 respectively. The biases were initialised as a constant value of 0.06. The optimiser used was Adams and the solver used was Dopri5. Finally, weight decay was also added to discourage overfitting.

5. Application in Chaotic Dynamics: Lorenz System

The Lorenz system is a system of three nonlinear ODEs developed by mathematician and meteorologist Edward Lorenz in 1963 [10]. It is a simplified mathematical model for atmospheric convection, in which the equations relate the properties of a two-dimensional fluid layer which is uniformly warmed from below and cooled from above. The variable x is proportional to the rate of convection, and y and z are proportional to the horizontal and vertical temperature variation, respectively:

$$\frac{dx}{dt} = \sigma(y - x) \quad (11)$$

$$\frac{dy}{dt} = x(r - z) - y \quad (11)$$

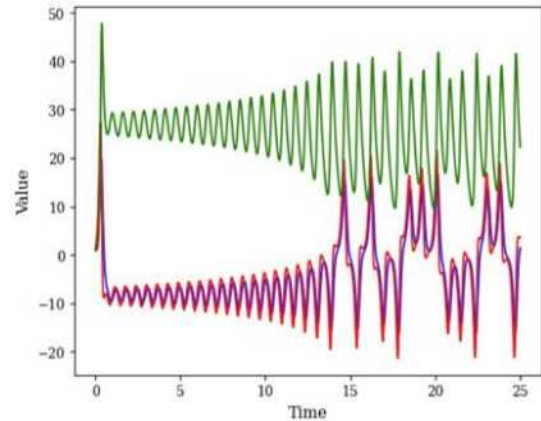


Figure 5: Trajectories resulting from the data generated for the Lorenz system. X , Y and Z components shown in blue, red and green, respectively.

$$\frac{dz}{dt} = xy - \beta z. \quad (12)$$

The constants σ , r , and β are system parameters proportional to the Prandtl number, Rayleigh number, and certain physical dimensions of the layer itself. There exists some critical value for the control parameter r :

$$r > r_c = \sigma \frac{\sigma + \beta + 3}{\sigma - \beta - 1}. \quad (13)$$

When $r > r_c$, the model becomes chaotic. Chaos, or state of disorder, is a very common phenomenon in fields such as chemistry and physics [11]. Chaotic systems are very complex systems that appear to be random in nature. Such systems, including the Lorenz system, are actually deterministic. Therefore, for a given set of initial conditions, a solution to the system can be found. However, these systems are extremely sensitive to initial conditions, and are very difficult to predict, as small numerical errors in the prediction of a trajectory will result in a totally different outcome to that expected for the initial conditions used. Lorenz himself coined the term butterfly effect to describe this extreme sensitivity to initial conditions when he noticed that tiny, butterfly-scale changes to the initial conditions of his weather model simulations resulted in anything from sunny skies to violent storms. Additionally, when the dynamics become chaotic, the Lorenz system is characterised by two strange attractors, around which the phase trajectories will oscillate but never converge to.

Given that chaotic systems are often present in real world systems, it was decided to attempt the prediction of the Lorenz system with a vanilla Neural ODE, in order to test the limitations and capabilities of this neural network model. In order to see if the plain vanilla Neural ODE model can learn chaotic dynamics, the control parameter r was chosen to have a value of 28, which together with the values of $\sigma = 10$ and $\beta = 8/3$, ensures that the system is chaotic as the critical control parameter value for the aforementioned parameter values is 24.74.

The data was then generated, and Gaussian noise was added. The training was performed using the same extending window strategy that was used for the training of the SIR model. Given that the Lorenz system dynamics have much higher rates of change than any of the two models covered in this paper, the dataset generated was much larger. A dataset consisting of 1000 datapoints was generated, it was clear that the training was going to be computationally expensive.

Therefore, the model only has a single layer of 100 neurons, and the activation function was chosen to be ReLU. The weights of the network were initialised with a normal distribution centred around zero, with a standard deviation of 0.1. The biases were initialised as a constant value of 0. The

optimiser used was Adam, and the solver used was midpoint, which was chosen over dopri5 since it is second order, and therefore less computationally demanding. Lastly, weight decay was added to discourage overfitting.

6. Results and Discussion

Regarding the Lotka-Volterra system, the network proved to be very good at uncovering the underlying dynamics in the data despite the noise, whilst showing no major signs of overfitting, as shown in Figure 6. However, a few steps were taken to arrive at such a result. At first, the network refused to capture the full height of the peaks of both the prey and predator populations. Upon plotting the generated data to explore this issue, it became clear that, relative to the rest of the trajectory, the peaks were sparsely populated with datapoints. Hence, the data resolution was doubled, i.e., twice as many datapoints were generated for the same time horizon, such that the network would be penalised much more for ignoring the datapoints at the peaks.

Doubling the batch size from 10 to 20 also increased the quality of the network prediction, which makes intuitive sense since the network will train on twice the amount of data per epoch. However, further increases in batch size did not yield any improvement. Hyper parameter tuning was not performed as trial and error proved to be sufficient for such a simple system. Should the data have been noisier, and the timespan explored longer, hyper parameter tuning might have been necessary, although perhaps not achievable due to the long computation times.

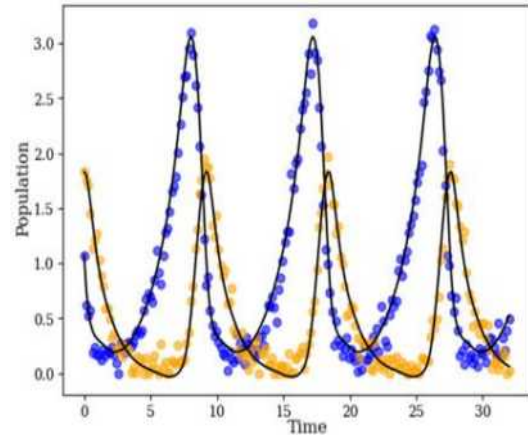


Figure 6: True prey data (blue) and true predator data (yellow) of the Lotka-Volterra system, plotted with their corresponding predicted trajectories (solid black lines).

When considering our epidemic case, as depicted in Figure 8, network was able to capture the behaviour of the SIR model very accurately. At first the model was attempted to be done with random time series batch validation technique, but after plotting due to its non-periodic nature it has failed to make as accurate predictions as it did with Lotka-

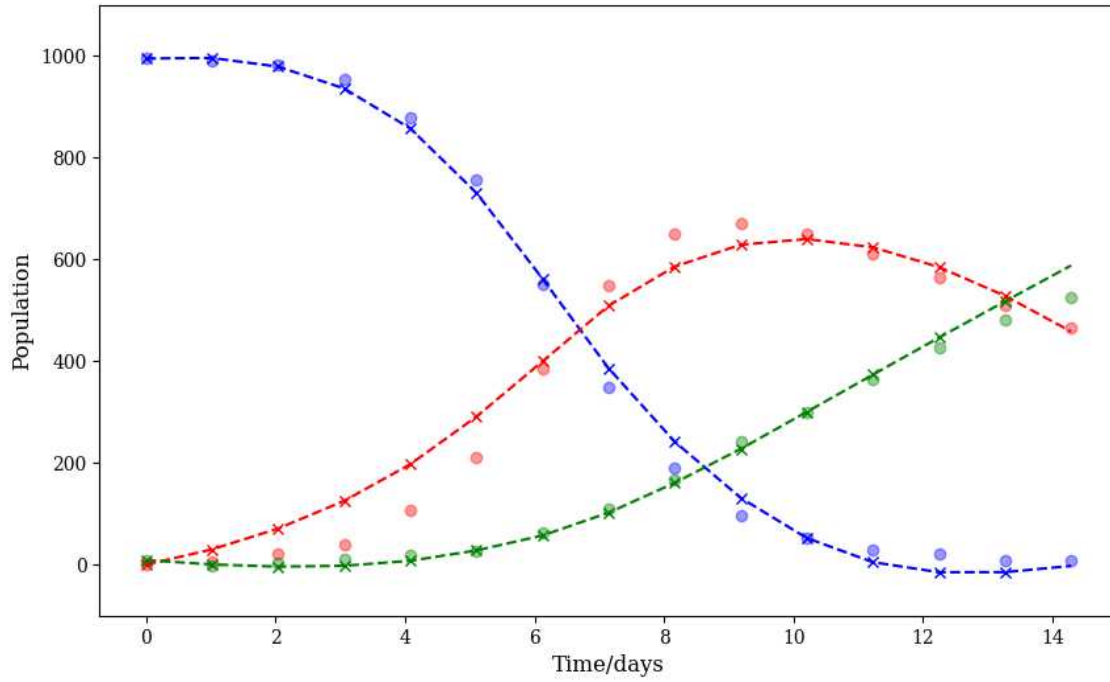


Figure 7: True data – dots, Network predictions – crosses, Test data predictions for validation - dashed lines, Blue (Susceptible), Red (Infected), Green (Removed) plotted to observe the training process with “extending window”.

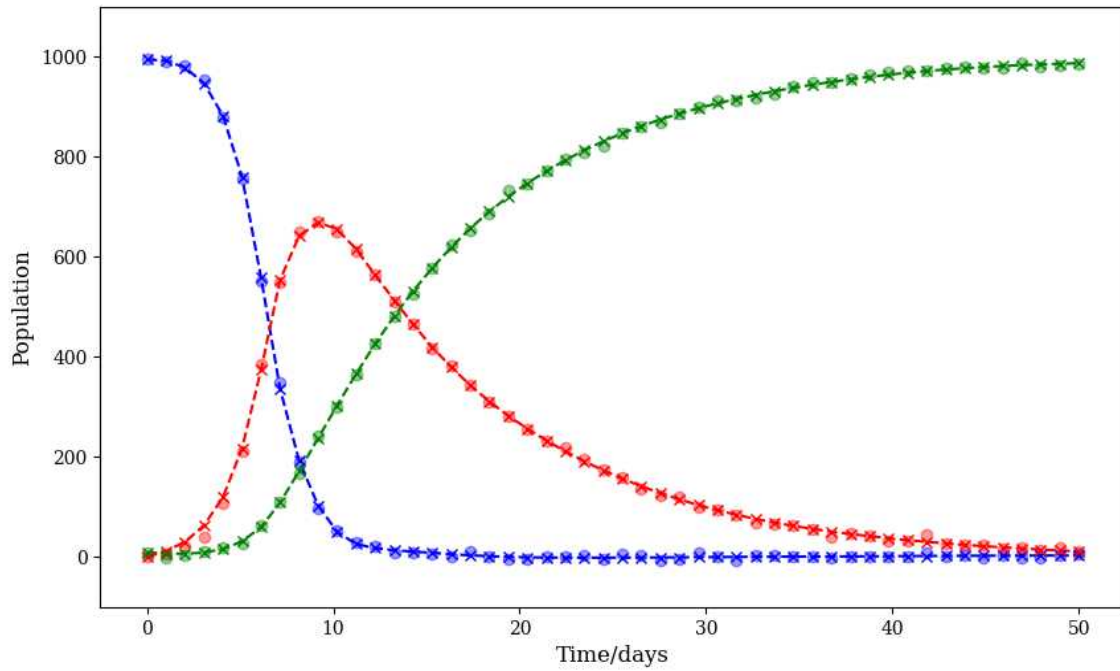


Figure 8: True data – dots, Network predictions – crosses, Test data predictions for validation - dashed lines, Blue (Susceptible), Red (Infected), Green (Removed)

Volterra. Because of this, extending window strategy was employed. The neural network was able to learn from the previous parameters and predict the next unseen data. This is an iterative process, where if the datapoint at $t = 13$ is for validation, it will be used as a training data when the dataset is expanded to first 14 datapoints. To achieve the final result,

parameters were tuned manually. Although decreasing the number of data points had a negative impact on data resolution, it was reduced from 200 to 50 and the network still managed to capture the dynamics at the peak with a shorter computational time. For future applications, if the standard deviation of the Gaussian noise and the number of

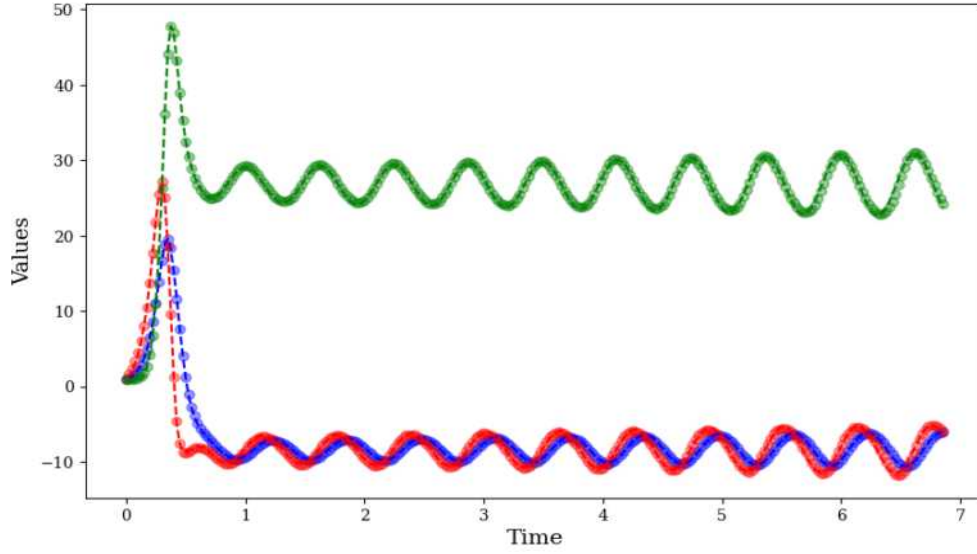


Figure 9: True data of the x , y and z components (blue, red and green data points respectively) of the chaotic Lorenz system, plotted with their corresponding predicted trajectories (dashed lines) for the training window extending up to time $t = 7$.

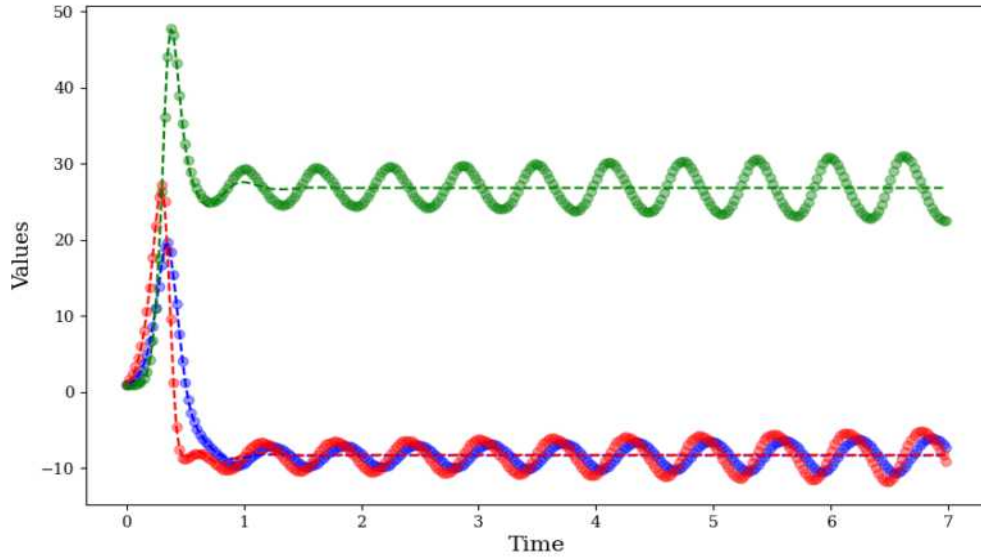


Figure 10: True data of the x , y and z components (blue, red and green data points respectively) of the chaotic Lorenz system, plotted with their corresponding predicted trajectories (dashed lines) for the training window extending up to time $t = 7.125$.

validation points in the extending window strategy is increased, it will be harder to train the network for the same time-period.

With our final system, the network was ultimately incapable of accurately fitting to the chaotic dynamics of the Lorenz system. However, many attempts were made to overcome the challenges imposed by said system. Since the Lorenz system, unlike the Lotka-Volterra system, is not cyclical in nature but instead ever changing, random time series batching was not an appropriate training and validation method. Hence, the training was initially performed using K-fold blocked cross validation. K-fold blocked cross validation is a variant of K-fold cross-validation. The latter consists of randomly shuffling the dataset and splitting it into K equally-sized folds. Said folds are then further divided into

train and validation partitions [12]. This is done to prevent overfitting. However, since the data for the Lorenz system is of a sequential nature, it cannot be randomly shuffled since the network might then try to predict past dynamics from future dynamics. Blocked K-fold cross validation omits the shuffling step to avoid this [13]. The 5-fold cross validation strategy, which was used in the initial training attempts, failed due to exploding gradients. Gradient explosion occurs when the gradients keep on getting larger and larger as the backpropagation algorithm progresses. This, in turn, causes very large weight updates and causes the gradient descent algorithm to diverge [14]. Several measures were taken to avoid the issue. Firstly, proper weight initialisation was implemented to alleviate the gradient explosion issue by ensuring the proper flow of information

through the network by ensuring that: the variance of outputs of each layer is equal to the variance of its inputs and that the gradients have equal variance before and after flowing through a layer in the reverse direction [15]. This is called Glorot initialisation. A variant of Glorot initialisation, Kaiming He initialisation, which is tailored specifically to the ReLU activation [16], was used. However, this method yielded results that were no better than those obtained when initialising the weights using normal initialisation. Batch normalisation, introduced in 2015 to address the gradient explosion issue [17], was not successful either. This did not come as a surprise, as this technique is not usually applied to Neural ODEs [18]. The gradient explosion issue was solved by clipping the norm of the gradients if the threshold of 1 was exceeded.

Unfortunately, despite solving this issue, training would be interrupted very early on due to an underflow in dt , i.e., the time step of the ODE solver. This could be due to the fact that the Lorenz system exhibits high sensitivity to initial conditions, and as a result the solver required extremely small steps to capture the dynamics. Loosening the absolute and relative error tolerances of the solver in order to allow for larger timesteps did not work either. As a last resort, it was decided to use the extending window strategy that was used for the SIR model.

Additionally, `dopri5` was substituted by the midpoint solver, since this extending window training method involves longer training times. Interestingly, this new model resulted in better results than any of the previous model. It was capable of accurately predicting the Lorenz system up to time $t = 7$, after which the model completely failed to capture the initial oscillations of the system and would instead predict a horizontal line that passes through the midpoint of the highs and lows of the oscillations.

This result was not surprising, as that exact behaviour has been previously observed when attempting to capture the Lorenz system's chaotic dynamics using a plain vanilla Neural ODE [19]. Perhaps if the network was deeper and the hyperparameters were tuned, better results would be obtained. But since the backpropagation step of the Neural ODE model entails solving an augmented ODE backwards in time [18], deeper networks and larger datasets would result in extremely long computation times which are simply not feasible.

7. Conclusion

The objective of the paper was to demonstrate Neural ODEs' ability to reconstruct the hidden dynamics of prototypical systems by means of empirical, noisy datasets. The Neural ODEs achieved the desired objective for the Lotka-Volterra and SIR models. The predictions were accurate, and the trained models proved to be robust

to unseen, noisy data. Both of these models were simple, but despite their simplicity, the dynamics of these systems is highly nontrivial. The Lorenz system in particular, a set of three ODEs with a quadratic nonlinearity, the simplest nonlinearity one can think of, exhibits chaotic behaviour, and thus is, as Lorenz himself concluded, inherently unpredictable. This was confirmed by the results obtained, which show that the data-driven Neural ODE approach presents some limitations for such chaotic systems.

Moreover, it is clear this plain vanilla Neural ODE approach is computationally demanding, and therefore training times are a severe limitation. Ultimately, the results in this paper show that Neural ODEs hold substantial promise for real-world applications, as the increased complexity of real-world system dynamics could be modelled using deeper, finely tuned Neural ODEs, which could be successfully trained given enough computational time and resources. Furthermore, this approach requires very small amounts of data for training, as the Lotka Volterra and SIR datasets were 200 and 100 datapoints in size, respectively, whereas traditional neural network approaches are much more data hungry.

Given that the time available for the completion of this work, the scope was severely limited. Should there be an opportunity to continue this work, there are many closely related, interesting areas of research. An example would be to delve into Physics Informed Neural Networks (PINNs), which are a type of universal function approximator that can embed the knowledge of the physical laws that govern a given dataset in the learning process. They are a popular line of work because they overcome the low data availability of some biological and engineering systems that makes most state-of-the-art ML techniques lack robustness [18]. It would also be interesting to explore other types of Neural Differential Equations, such as Controlled Differential Equations, which are often used to model systems where you have some degree of control over the system's dynamics and could therefore be applied to any practical Chemical Engineering process. Furthermore, the work could be extended to cover Stochastic Differential Equations. They are a natural extension of ODEs for modelling systems that evolve in continuous time subject to uncertainty, and they have seen widespread use for modelling real-world random phenomena [18].

Lastly, since the computation time was the main obstacle when training the Neural ODEs, different but faster methods for backpropagation could be explored in order to reduce training times and allow for deeper networks.

8. Acknowledgements

The authors of this work are grateful to our supervisors Prof. Serafim Kalliadasis, Dr. Miguel Angel Durán-Olivencia and Mr. Antonio Malpica-Morales, as well as group members Mr. Frank-Ioannis Papadakis-Wood and Mr. Tushar Verma for their continuous support and guidance throughout this research project.

9. Bibliography

- [1] www.collimator.ai. (2022). Understanding Dynamical Systems in Real-world Situations. [online] Available at: <https://www.collimator.ai/post/what-are-dynamical-systems#:~:text=A%20dynamical%20system%20is%20a> [Accessed 12 Dec. 2023].
- [2] Nielsen, M.A. (2015). Neural Networks and Deep Learning. Determination Press, p.3.
- [3] Berndt Müller, Joachim Reinhardt and Michael Thomas Strickland (1995). Neural Networks : An Introduction. 2nd ed. Berlin: Springer, p.14.
- [4] Hornik, K., Stinchcombe, M. and White, H. (1989). Multilayer feedforward networks are universal approximators. Neural Networks, 2(5), p.363.
- [5] Chen, R.T.Q., Rubanova, Y., Bettencourt, J. and Duvenaud, D.K. (2018). Neural Ordinary Differential Equations. NeurIPS2018, 31, pp.1–2.
- [6] He, K., Zhang, X., Ren, S. and Sun, J. (2015a). Deep residual learning for image recognition. p.1.
- [7] Shen, G. and Yuan, Y. (2019). On theoretical analysis of single hidden layer Feedforward Neural Networks with ReLU activations. 2019 YAC, 34th, p.1.
- [8] Columbia University (2021). Epidemic, endemic, pandemic: What are the differences? [online] Columbia University's Mailman School of Public Health. Available at: <https://www.publichealth.columbia.edu/news/epidemic-endemic-pandemic-what-are-differences> [Accessed 12 Dec. 2023].
- [9] Kermack, W.O. and McKendrick, A.G. (1927). A contribution to the mathematical theory of epidemics. Proc. R. Soc. A: Math. Phys. Eng. Sci. 115, pp. 700-721.
- [10] Lorenz, E.N. (1963). Deterministic nonperiodic flow. Journal of the Atmospheric Sciences, 20(2), pp.130–141.
- [11] Ditto, W. and Munakata, T. (1995). Principles and applications of chaotic systems. Comm. of the ACM, 38(11), p.97.
- [12] Chollet, F. (2018). Deep Learning with Python. (Manning Publications Co., Shelter Island), p.87.
- [13] Shrivastava, S. (2020). Cross validation in time series. [online] Medium. Available at: <https://medium.com/@soumyachess1496/cross-validation-in-time-series-566ae4981ce4> [Accessed 12 Dec. 2023].
- [14] Bohra, Y. (2021). Vanishing and Exploding Gradients in Deep Neural Networks. [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2021/06/the-challenge-of-vanishing-exploding-gradients-in-deep-neural-networks/> [Accessed 12 Dec. 2023].
- [15] Glorot, X., Bordes, A. and Bengio, Y. (2011). Deep sparse rectifier Neural Networks. Int. Conf. on Artif. Intel. and Stat., 15, pp.315–323.
- [16] He, K., Zhang, X., Ren, S. and Sun, J. (2015b). Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. Proc. of the IEEE ICCV, p.1026.
- [17] Ioffe, S. and Szegedy, C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. p.1. arXiv.1502.03167.
- [18] Kidger, P. (2021). On Neural Differential Equations. [Thesis] p.1-117. Available at: <https://arxiv.org/pdf/2202.02435.pdf> [Accessed 12 Dec. 2023].
- [19] Ko, J.H., Koh, H., Park, N. and Jhe, W. (2022). Homotopy-based training of NeuralODEs for accurate dynamics discovery. p.8. arxiv.2210.01407.

Tabular and Deep Q-Learning for Optimal Control of a Commercial HVAC System

David Shamash and Jeremy Yee Tong Wah

Department of Chemical Engineering, Imperial College London, U.K.

Abstract

The use of Reinforcement Learning for the control of HVAC systems in buildings is of growing interest – its ability to adapt to different environments, without the need to exhaustively fine-tune key parameters for each individual environment, has sparked the interest of many. This paper investigates the use of Tabular and Deep Q-Learning for the control of the HVAC system of a supermarket store. A Tabular Q-Learning controller and a Deep Q-Learning controller are developed and tested over an existing simulation environment of the store. Both controllers use 21 states, 11 actions, a learning rate (α) of 0.001, a discount factor (γ) of 0.99 and an exploration probability (ϵ) of 0.4, and are trained using an equivalent of 18 years of historical data. The Deep Q-Learning algorithm comprises of a neural network with 2 hidden layers. It is found that the Tabular Q-Learning controller outperforms the existing PI controller by 4% in energy efficiency and 12% in user comfort. However, the Deep Q-Learning does not present any improvement over the baseline PI control, and requires further fine tuning. Overall, this paper demonstrates the potential for Tabular Q-Learning for the control of HVAC systems in buildings, with potential improvements in both energy and comfort metrics.

1. Introduction

Climate change and global warming have been the subject of many political and public concerns in recent years. In 2019, the United Kingdom became the first major economy to legislate for net zero by 2050 [1]. Buildings are currently responsible for 30% of global final energy consumption and 26% of global energy-related emissions [2]. In UK supermarkets, Heating, Ventilation and Air Conditioning (HVAC) systems account for 20 to 30% of total energy consumption [3].

Sainsbury's Supermarkets Ltd, one of the largest food retailers in the UK, is committed to reach net zero emissions by 2035 [4], and optimising energy consumption in their stores is key in reaching this goal. It is therefore of high interest to investigate the potential benefits of transitioning their existing HVAC systems to smart energy management controls to enhance the control process.

1.1 Reinforcement Learning

Reinforcement Learning (RL) is a Machine Learning technique where an *agent* is trained within a set *environment* using a trial-and-error approach. This is achieved by allocating *rewards* based on the agent's actions [5]. The fundamental principles of Reinforcement Learning can be attributed to a *Markov Decision Process* (MDP), where a system exists in *states* (quantified conditions a system can be in), while *actions* taken by the agent can move the system from one state to another. The *policy* defines how the agent behaves in a particular state and aims to maximise the reward – it can be a function, a table or a neural network. For an MDP, actions and states are discrete, or need to be discretised. Moreover, the future state of the system only depends on the current state and action; it is independent of the previous states and actions [6].

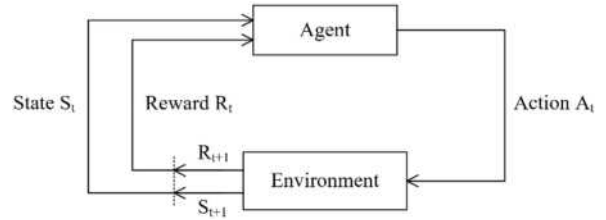


Figure 1: A schematic of Reinforcement Learning. The loop represents the agent continuously learning from the environment by taking actions and receiving feedback in the form of rewards.

1.1.1. Value-based Reinforcement Learning

Value-based RL methods focus on learning a *value function* which estimates the expected cumulative reward of the agent being in a particular state, or taking a specific action in a particular state. The agent makes decisions based on this value function. The value $q_{\pi}(s, a)$ of taking action a in state s under policy π is given by:

$$q_{\pi}(s, a) \triangleq \mathbb{E}_{\pi} \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s, A_t = a \right] \quad (1) [7]$$

where γ denotes the discount factor and S_t , A_t and R_t denote the state, action and reward at timestep t respectively. The algorithm ultimately seeks to find the optimal value function $q_{\pi^*}(s, a)$ corresponding to an optimal policy π^* . Examples of value-based methods include Tabular Q-Learning and Deep Q-Learning.

1.1.2. Tabular Q-Learning (TQL)

Q-learning methods aim to find the optimal policy by finding the optimal *Q-value* for each state-action pair. The Q-value $Q(s, a)$ measures the desirability of taking action a in state s , and is updated as follows [8]:

$$\begin{aligned} \text{New } Q(s, a) \\ = Q(s, a) + \alpha [R + \gamma \max_{a'} Q'(s', a') - Q(s, a)] \end{aligned} \quad (2)$$

where α and γ denote the learning rate and discount factor respectively, R represents the reward for the current run and $\max Q'(s', a')$ corresponds to the future Q-value with the highest expected reward.

In Tabular Q-Learning (TQL), the Q-values are stored in an $N_s \times N_a$ table (known as the *Q-Table* or *look up table*), where N_s and N_a are the number of states and actions respectively.

1.1.3. Deep Q-Learning (DQL)

Deep Q-Learning (DQL) uses a neural network (known as the *Q-Network*) instead of a table to approximate the Q-value function. The use of neural networks allows the algorithm to process data more effectively and increase the efficiency of multi-dimensional processes.

DQL uses a *replay buffer* which stores the agent's past experiences $e_t = (S_t, A_t, R_t, S_{t+1})$ in a data set $D_t = \{e_1, \dots, e_t\}$, where S_t , A_t and R_t denote the respective state, action and reward at timestep t . Instead of training the model with experiences in the order that they occur, the DQL algorithm randomly selects experiences from the replay buffer during the learning process [9]. This decoupling of temporal correlations in the data reduces the likelihood of overfitting, allows for more stable and efficient learning, enhances the controller's ability to handle non-stationary environments and improves overall convergence [10]. DQL works by adjusting the parameters θ of the Q-network to minimise the *loss function*, defined for a sample $(s, a, R, s') \sim U(D)$ and iteration i as follows [9]:

$$L_i(\theta_i) = \mathbb{E}_{(s,a,R,s') \sim U(D)} [(R + \gamma \max_{a'} Q'(s', a'; \theta_i^-) - Q(s, a; \theta_i))^2] \quad (3)$$

where γ denotes the discount factor and S , A and R are the respective state, action and reward. θ_i^- and θ_i^+ represent the network parameters and the target network parameters respectively.

1.2. Literature Review

The interest in Reinforcement Learning for energy systems has increased exponentially in recent years, with approximately 20 research papers published on the topic in 2010 to almost 400 in 2020 [11]. Research has shown that RL can easily adapt to dynamic environments (changing weather conditions, for example) and might therefore prove beneficial in controlling the heating rates of HVAC systems to achieve energy-efficient temperature control in buildings [12].

Value-based RL methods can become computationally less efficient for large action-state spaces [13] (*large* usually regarded as more than 50). As a result, this field has received less research interest compared to other RL methods [14]. However, research is actively ongoing for relatively simple systems. In 2018, S. Baghaee and I. Ulusoy [15], implemented a Tabular Q-Learning method for the ventilation control system of a building and justified its use due to the relative simplicity of the model. It was reported that the RL agent only consumed marginally less energy than the existing PID controller. In 2015, E. Barrett and S. Linder [16] combined Tabular Q-Learning with an occupancy prediction method for the control of a building's heating systems. This was achieved by simplifying the model and discretising external temperatures ranges. This controller

achieved a reduction of approximately 10% in operating costs without compromising comfort standards.

Recent developments in computational capabilities have proven beneficial for Deep Reinforcement Learning (DRL) methods, allowing them to process larger sets of data within more expansive environments [17]. DRL algorithms are commonly being tested in self-driving vehicles and open-world games amongst other applications [18]. They have also been reviewed for complex control systems with continuous environments [19].

S. Brandi. et al. [17] investigated the use of a Deep Q-Learning algorithm in buildings to maximise user comfort and minimise energy consumption, and reported energy savings of between 5% and 13% based on the occupancy and season amongst other factors. In 2021, Z. Jiang et al. [20] implemented a DQL controller in an office building space using 4 months of historical temperature and energy data, and found that the energy efficiency of the DQL controller exceeded that of a baseline PI controller by 6% to 8%. Further research into DQL for energy efficiency in office spaces was conducted in 2022 by X. Zhong et al. [21], who noted an increase in energy efficiency of 12.8% compared to the existing fixed-schedule control strategy.

It is important to note that models from literature use different technologies – results are not standardised. As such, a higher increase in efficiency does not necessarily imply a better model [12]. In general, research on Tabular Q-Learning is limited, with the method generally dismissed for being too simple for complex control systems. While the performances of different RL algorithms were compared, no direct quantitative comparisons between Tabular and Deep Q-Learning have been made.

Reinforcement Learning represents a significant potential for the control of HVAC systems as it allows for relatively simple and fast implementation of a control system across stores. This can be attributed to its ability to adapt to different environments, without the need to exhaustively fine-tune key parameters for each individual environment [11]. It is therefore of value to explore the potential of Tabular Q-Learning in HVAC systems as they are generally less dynamic than other processes and do not require complex algorithms to account for various sensitivities [22]. Moreover, the simplicity of Tabular Q-Learning would offer a more computationally efficient method, allowing for less powerful computers to run simulations. It is also of interest to train a Deep Q-Learning agent to determine how a more complex model compares.

1.3. Project Scope

This paper aims to train, tune and test a Tabular Q-Learning controller and a Deep Q-Learning controller for the HVAC system of a supermarket store to qualitatively and quantitatively assess their performances against a PI controller.

2. Methods

2.1. System Specifications

2.1.1. Simulation Environment

An existing ‘living-lab’ pilot [23] is used as simulation environment. This simulation environment replicates the temperature dynamics of a supermarket store building with 4 walls and a floor space surface area of 7600m², located 100 miles north of London. A resistance capacitance (RC) network approach is used for this purpose. The network uses 3R2C models for the external walls and the roof, and an additional 2R2C model to represent the building’s internal heat capacity. The model collects internal temperature data from 16 temperature sensors from different parts of the store. However, a uniform temperature distribution is assumed, with every part of the store having the same internal temperature.

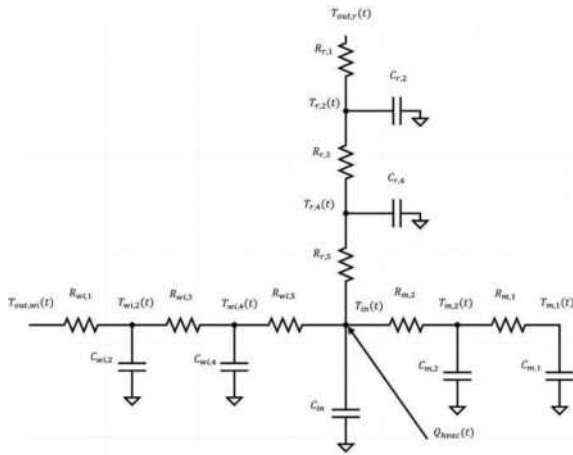


Figure 2: Thermal model of the supermarket store [23]. R and C represent resistance and capacitance values respectively, T is the temperature, and subscripts w , r , m and i denote the i^{th} wall, the roof, internal mass of the building and internal air of the building respectively.

2.1.2. The PI controller

To assess the performance of the RL algorithms, it is necessary to compare their performance against a baseline controller. An existing Proportional-Integral (PI) controller [24] is used for this purpose.

A PI controller aims to minimise the differences between a system’s output and the desired setpoint by employing two parameters: a proportional-gain (P) parameter, which responds to the current error, and an integral (I) parameter, which addresses past errors accumulated over time. These parameters contribute to the controller’s output to regulate the system’s behaviour and are described by constants [20].

The PI controller used in this paper was tuned to a P value of 1×10^6 and an I value of 5×10^4 for optimal performance in the simulation environment described in section 2.1.1.

2.1.3. Environment Dynamics

An effective approach to assess the environment dynamics involves investigating its time constant τ , defined as the time taken to move 63% closer to the new

value following a step change. This is achieved using the existing PI controller and a setpoint change of 16°C to 19°C. Based on the graphical method described by C. Kontoravdi [25], the value of τ for the thermal model is determined as approximately 3 hours.

Due to the slow dynamics of the environment, changing the HVAC heat load at small intervals (every minute, for example) resulted in overfitting and unstable temperature oscillations [26]. It is determined that a heating interval of 1 hour performed best, as this maintained the stability of the system while adjusting the heating regularly enough.

2.1.4. External Temperatures and Setpoints

Eleven months of hourly measured external temperature data for the supermarket store (ranging from May 2022 to March 2023) [23] were available for use. The store aims to maintain an internal temperature of 19°C from 7 a.m. to 12 a.m. (operating hours) and at 16°C between 12 a.m. and 7 a.m. (non-operating hours).

2.1.5. States and Actions

States were described as deviations of the internal temperature from the setpoint (*setpoint deviations*). Setpoint deviations directly relate to the amount of heat that should be added to the system, unlike the internal temperatures themselves. Since the setpoint changes across hours of the day, the measured temperature does not provide a good indication of how close the internal store temperature is to the temperature setpoint [27]. A total of 21 states were used, corresponding to temperature deviations from the setpoint ranging from -10°C to $+10^\circ\text{C}$, with temperature increments of 1°C between states. Using a state range of this magnitude allowed the controller to account for various scenarios, including cold winter months where constant heating is required, and hot summer months where no heating is necessary. Temperatures above or below the maximum setpoint deviations were approximated as the largest possible deviation states ($+10^\circ\text{C}$ and -10°C respectively). This is because the best actions at these temperatures correspond to the best actions as at the largest possible deviation states: the maximum possible amount of heating is required for a setpoint deviation of -10°C or more negative values, while no heating is required at a setpoint deviation state of $+10^\circ\text{C}$ or more.

Actions were described as changes in the power output of the HVAC system. J. da Silva and A. Secchi [27] established that describing the actions in such a way is effective to maintain a stable setpoint in the context of a process production plant. Designating the actions as changes in the heat loads instead of the heat load values themselves improved the stability of the control system, as shown in the *Supplementary Information* section. A total of 11 actions were used, corresponding to changes in the power output of the HVAC system ranging from -250kW to $+250\text{kW}$, with increments of 50kW between actions. Using changes of magnitude 250kW prevented the HVAC system from directly switching its power output between 0 and 500kW , over the concern that this could result in significant instability, similar to cases presented in [25].

Table 1: Effect of action on HVAC power output

HVAC Power Output (Timestep = t-1)	Action, a_t	HVAC Power Output (Timestep = t)
\dot{Q}_{HVAC}	$\Delta \dot{Q}_{HVAC}$	$\dot{Q}_{HVAC} + \Delta \dot{Q}_{HVAC}$

2.1.6. Reward function

The reward function ensures that the actions taken by the controller are as close as possible to the desired outcome. An iterative approach is taken to fine-tune the reward function for optimal results. It is common to set the rewards as negative values, favouring rewards closest to zero [7]. The absolute value of each term in the reward function represents a penalty which needs to be minimised.

The reward function aims to simultaneously penalise two separate metrics: the temperature deviation from the setpoint and the HVAC heat load relative to the maximum. These two metrics were specifically chosen as they gauge the user comfort and thermal energy usage respectively. Both metrics were standardised with adequate weights.

Table 2: Temperature and energy rewards for different deviations from the setpoint ΔT , where $\Delta T = T_{\text{internal}} - T_{\text{setpoint}}$

ΔT (°C)	$\Delta T > 0$	$-2 < \Delta T < 0$	$\Delta T < -2$
Temperature reward	$-w_1 \Delta T $	$-w_3 \Delta T $	$- \Delta T ^3$
Energy reward	$-w_2 \cdot \frac{\dot{Q}_{HVAC}}{\dot{Q}_{\max}}$	$-w_4 \cdot \frac{\dot{Q}_{HVAC}}{\dot{Q}_{\max}}$	$-5 \cdot \frac{\dot{Q}_{HVAC}}{\dot{Q}_{\max}}$

As explained in section 2.1.4., setpoint deviations directly relate to the amount of heat that should be added to the system. Accounting for setpoint deviations instead of internal temperatures values in the reward function prevents the internal temperature from converging at the weighted average value of the two setpoints.

The constraint for $\Delta T < -2$ is implemented to prevent the model from converging around the 16°C setpoint. A setpoint change from 16°C to 19°C would activate the cubic temperature factor. This would in turn prompt the controller to increase the internal temperature to the new setpoint and prevent the algorithm from taking enormous penalties. This constraint proves particularly useful in cold temperature conditions, during which the controller would otherwise tend to minimise the energy usage and incidentally prioritise the 16°C setpoint over the 19°C setpoint.

2.1.7. Exploration Probability (ϵ)

A significant aspect of RL involves the distribution of training between *exploration* (investigating actions with unknown outcomes) and *exploitation* (taking the best-known action, based on the current estimated Q-values [12]). A balance between these must be achieved to optimise performance and computational cost.

The *Epsilon Greedy* method is the most common approach to strike this balance [20]. In this method, the exploration probability parameter ϵ is set between 0 and 1. As its name suggests, the probability of the agent exploring new state-action pairs stands at ϵ , while the agent exploits the known state-action space $(1-\epsilon)$ of the time.

2.1.8. Learning Rate (α)

The learning rate α is also set to a value within the range 0 to 1. A learning rate close to 0 corresponds to a slow rate of learning during which the system tends to stabilise its Q-estimates; Q-values only undergo small changes when updated. However, this may result in the algorithm taking a significant amount of time before converging to the optimal solution. A learning rate close to 1 results in faster learning, with the Q-values undergoing more significant updates. This can however increase the likelihood of unstable and oscillatory learning, and the possibility of overshooting the optimal solution [7].

2.1.9. Discount Factor (γ)

The discount factor γ measures the importance of future rewards for the agent. It is also designated a value between 0 and 1. A discount factor close to 1 implicates giving more importance to long-term consequences. This encourages the agent to consider future rewards in decision-making. A discount factor close to 0 corresponds the algorithm prioritising short-term gains and mostly ignoring long-term consequences. In most of the literature [12], γ values between 0.9 and 0.99 were used to prioritising future rewards. It would therefore be sensible to use γ values of similar magnitudes as a guideline [20].

2.2. Training the Tabular Q-Learning Controller

A Q-Table is first initiated with Q-values for all state-action combinations being zero. This ensures a neutral starting point and prevents biases in the learning process [7].

The controller is then trained on the eleven months of available external temperature data ran 20 times, equating to 18 years' worth of training data. This is consistent with how RL agents should be trained, often requiring numerous time steps to provide meaningful results [6]. Opinions in literature vary regarding how much data suffices for training, ranging from 3 months to 100 years. However, it is determined that an equivalent of 18 years of data is enough.

The TQL controller is then tested for different scenarios, including periods of extreme heat and extreme cold to assess its robustness. Plots for these scenarios can be found in the *Supplementary Information* section.

2.3. Training the Deep Q-Learning Controller

A Deep Q-Learning algorithm is trained with the same system specifications as described in Section 2.1., with the same 18 years equivalent of training data used to train the TQL controller. The controller is modelled after N. Joshi's Deep Q-Network [28] and makes use of two hidden layers in its neural network.

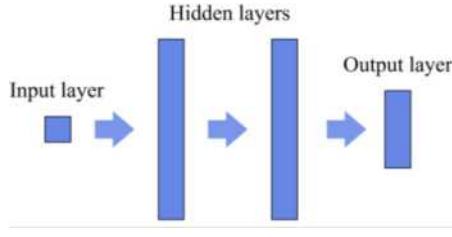


Figure 3: A neural network with two hidden layers used as Q function. The input layer is the state; the output layer is the Q -values corresponding to every possible action

2.4. Key Performance Indicators (KPIs)

2.2.1. Comfort Violation

The Comfort Violation is calculated using Equation 4, based on M. Bird's definition of comfort violations [23]. Only deviations below the setpoints were considered. In Equation 4, N represents the number of timesteps of duration 1 hour each and ΔT_i is the deviation from the setpoint at timestep i . The Comfort Violation is normalised per year, and 1Kh/yr corresponds to deviations from the temperature setpoint of 1 Kelvin for 1 hour within 1 year [23].

Comfort Violation =

$$\begin{cases} \frac{24 \cdot 365.25}{N} \sum_{i=0}^N |\Delta T_i|, & T_{\text{internal}} < T_{\text{setpoint}} \\ 0, & T_{\text{internal}} > T_{\text{setpoint}} \end{cases} \quad (4)$$

2.2.2. Total Energy Usage

The total thermal energy produced by the HVAC system Q_{HVAC} (kWh_{thermal}/yr) over N timesteps of length 1 hour each is given by Equation 5, where $\dot{Q}_{\text{HVAC},i}$ (kW_{thermal}) is the thermal power into the building during timestep i . The total thermal energy value is normalised to obtain the energy produced per year.

$$Q_{\text{HVAC}} = \frac{24 \cdot 365.25}{N} \sum_{i=0}^N \dot{Q}_{\text{HVAC},i} \quad (5)$$

The total electrical energy consumed by the HVAC system (kWh_{electrical}/yr) can then be determined from Equation 6, where the Coefficient of Performance (COP) of a typical HVAC system is 3 kWh_{thermal}/kWh_{electrical} [3].

$$Q_{\text{Electrical}} = \frac{Q_{\text{HVAC}}}{\text{COP}} \quad (6)$$

2.2.3. Total Energy Cost

The total electricity cost (£) is calculated using:

$$\text{Total Electricity Cost} = Q_{\text{Electrical}} \cdot P_{\text{Electricity}} \quad (7)$$

where the price of electricity $P_{\text{Electricity}}$ can be approximated as £0.20/ kWh_{electrical}. [3]. A fixed price is used rather than real time data (from the N2EX database for example) as energy consumption and user comfort constitute the primary objectives of the model. If the model was constructed to also include an electricity cost minimisation target, the comfort target would have risked being compromised in scenarios where electricity prices surge (during the Russia-Ukraine war in 2022, for example).

2.2.4. CO₂e emissions

The UK Government Department for Energy Security and Net Zero [29] reports that 0.207 kg of carbon dioxide equivalent is released on average per kWh of electrical energy used. As a result, the total mass of CO₂e released (kg/yr) can be calculated using Equation 8 below.

$$m_{\text{CO}_2\text{e}} = 0.207 \cdot Q_{\text{Electrical}} \quad (8)$$

2.5. Sensitivity analysis

After tuning a controller, it is of value to perform a sensitivity analysis on its different model specifications. This helps to analyse the most sensitive hyperparameters and offers an opportunity to further understand the dynamics of the reward function. Sensitivity analyses are independently performed on w_1 , w_2 , w_3 , w_4 , α , γ and ϵ for the Tabular Q-Learning controller.

3. Results

3.1. System Specifications

In most cases, it is necessary to take an iterative approach and manually tune the reward function term weights and the model hyperparameters. Table 3 shows the temperature and energy rewards with optimal w_1 to w_4 , while Table 4 displays the optimal hyperparameters, which are used in for both TQL and DQL controllers.

Table 3: Temperature and energy reward terms, with optimal $w_1=0$, $w_2=0.3$, $w_3=2$ and $w_4=1$

ΔT (°C)	$\Delta T > 0$	$-2 < \Delta T < 0$	$\Delta T < -2$
Temperature reward	0	$-2 \Delta T $	$- \Delta T ^3$
Energy reward	$-0.3 \cdot \frac{\dot{Q}_{\text{HVAC}}}{\dot{Q}_{\text{max}}}$	$-\frac{\dot{Q}_{\text{HVAC}}}{\dot{Q}_{\text{max}}}$	$-5 \frac{\dot{Q}_{\text{HVAC}}}{\dot{Q}_{\text{max}}}$

Table 4: Optimal values for hyperparameters α , γ and ϵ .

α	γ	ϵ
0.001	0.99	0.4

Figure 4 depicts how the total reward for a training run varies, where each run corresponds to one training loop where the 11 months of training data described in Section 2.1.4. is used. The total reward plateaus before 20 training loops.

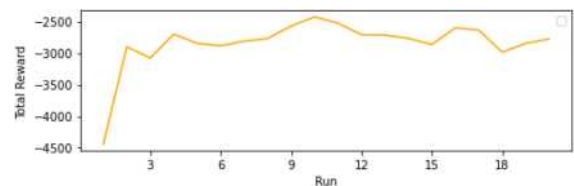


Figure 4: Total reward plotted against training run number for 20 runs for TQL controller

3.2. Tabular Q-Learning

The Q-Table contains 231 entries (21 states \times 11 actions), which is considered a reasonable Q-Table size [7]. Figure 5 depicts the temperature profile of the baseline PI controller and the trained TQL Controller tested in May 2023, with Figure 6 presenting a more detailed

analysis over a 48-hour period. It can be observed that the TQL controller maintains the internal temperature closer to the 19°C setpoint compared to the baseline PI controller.

The Tabular Q-Learning controller is also proven versatile, performing efficiently in both the warm summer 2022 and the cold winter 2023, as shown in the *Supplementary Information* section.

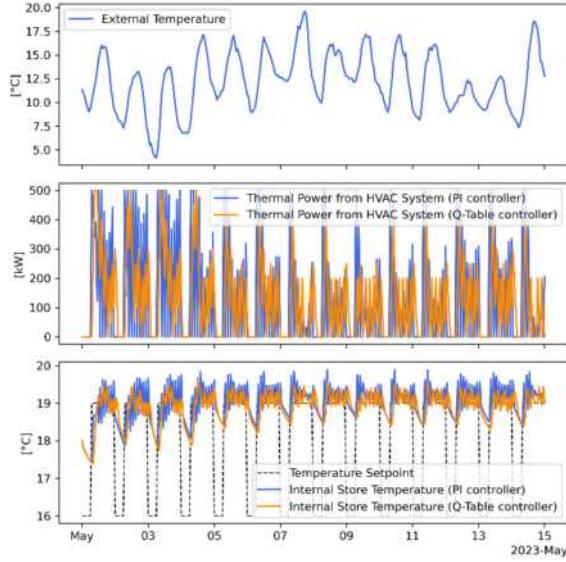


Figure 5: Testing plots for the TQL controller in the period 1 to 15 May 2023. In the two bottom plots, the orange and blue lines illustrate the performance of the TQL and PI controllers respectively. The black dashed line in the bottom plot represents the temperature setpoint.

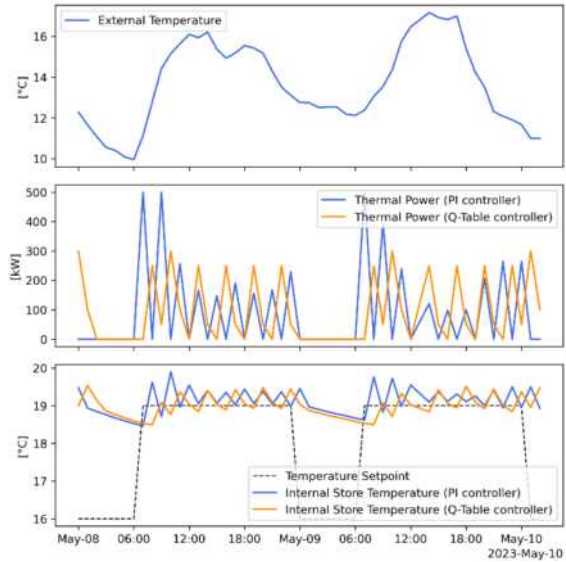


Figure 6: Testing plots for the TQL controller in the period 8 to 10 May 2023. In the two bottom plots, the orange and blue lines illustrate the performance of the TQL and PI controllers respectively. The black dashed line in the bottom plot represents the temperature setpoint.

Quantitative results shown in Table 5 demonstrate that the Tabular Q-Learning agent outperforms the PI controller in every metric. As the total energy cost and CO₂e emissions are directly proportional to the total thermal energy produced by the HVAC system, these indicators will all experience the same percentage change across models.

Table 5: TQL quantitatively compared to PI control

	PI	TQL	% Reduction
Comfort Violation (Kh/year)	710	626	12%
Total Thermal Energy (kWh/year)	1.09×10^6	1.04×10^6	4%
Total Energy Cost (£/year)	7.25×10^4	6.96×10^4	4%
CO ₂ e Emissions (kg CO ₂ e/year)	7.50×10^4	7.20×10^4	4%

3.3. Deep Q-Learning

Figures 7 and 8 depict the performance of the DQL controller compared to the baseline PI controller, tested over the first two weeks of May 2023. It is found that training a DQL controller takes on average 7 times longer than a TQL controller.

The DQL controller displays significantly less dynamic behaviour compared to the TQL controller. However, it did not perform as well as the TQL agent, improving energy savings by only less than 1%, while increasing the comfort violation by 22% compared to the baseline PI controller.

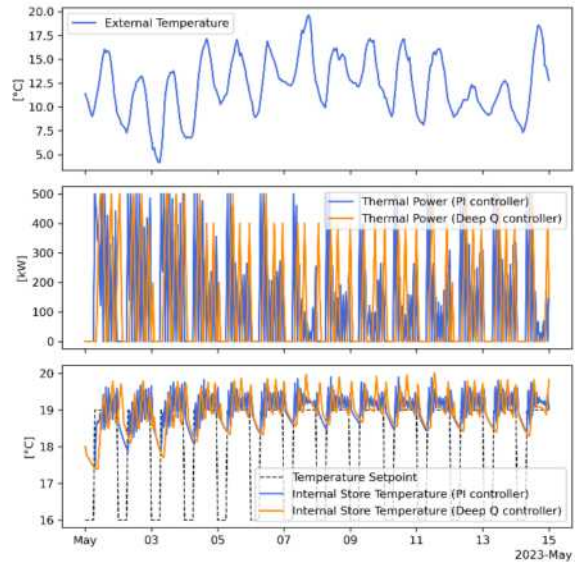


Figure 7: Testing plots for the DQL controller in the period 1 to 15 May 2023. In the two bottom plots, the orange and blue lines illustrate the performance of the DQL and PI controllers respectively. The black dashed line in the bottom plot represents the temperature setpoint.

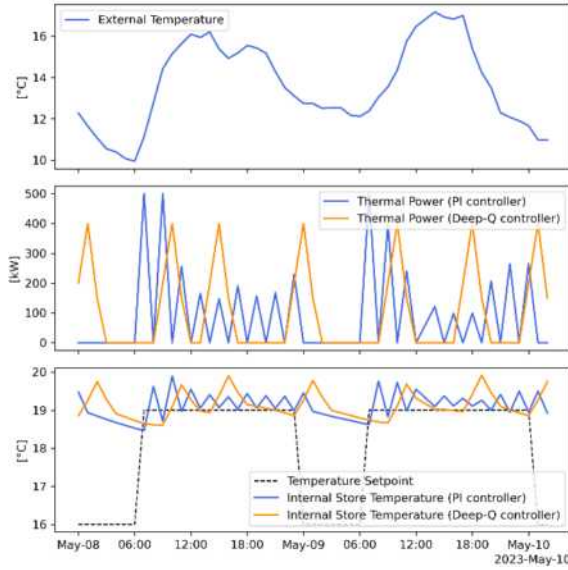


Figure 8: Testing plots for the DQL controller in the period 8 to 10 May 2023. In the two bottom plots, the orange and blue lines illustrate the performance of the DQL and PI controllers respectively. The black dashed line in the bottom plot represents the temperature setpoint.

Table 6 shows a direct comparison between the PI, TQL and DQL controllers. The TQL controller outperforms the DQL controller in every metric.

Table 6: DQL quantitatively compared to TQL and PI control

	PI	TQL	DQL
Comfort Violation (Kh/year)	710	626	865
Total Thermal Energy (kWh/year)	1.09×10^6	1.04×10^6	1.07×10^6
Total Energy Cost (£/year)	7.25×10^4	6.96×10^4	7.15×10^4
CO ₂ e Emissions (kg CO ₂ e/year)	7.50×10^4	7.20×10^4	7.40×10^4

3.4. Sensitivity analysis

The tuned values for the system parameters are presented in Section 3.1. It is found that the w_1 , α and ϵ parameters play a critical role in the dynamics and robustness of the system. Factors including stability and values used in literature are used to determine optimal values for the hyperparameters.

The temperature penalty weight w_1 is identified as the most sensitive of all the weights. Increasing its value by small increments significantly impacted the controller performance, as shown in Figure 10

Learning rate (α) values larger than 1×10^{-3} result in internal temperatures deviating from the setpoints, as shown in Figure 9, while smaller values result in overfitting, leading to poor control in extreme winter and summer conditions.

The exploration probability ϵ is varied between 0.01 and 0.6 to experiment with training scenarios with almost no exploration, and scenarios where the majority training consists of exploration. The system does not consistently converge to the setpoints at ϵ values close to 0.01. The RL agent performs better at larger ϵ values. However, it is found that ϵ values greater than 0.5 result in system

instability. An ϵ value of 0.4 is determined optimal for the environment.

The *Supplementary Information* section includes sensitivity analysis plots for the remaining weights and hyperparameters, which do not impact the system at the same level.

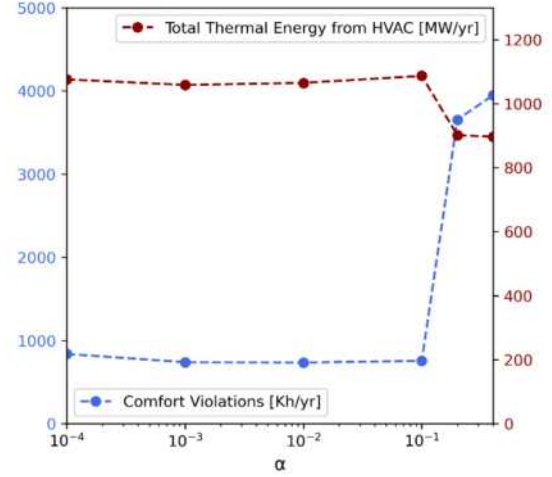


Figure 9: Plots of Comfort Violation and Total Thermal Energy against the learning rate (α) from testing in May 2023. Red data points represent the total thermal energy from the HVAC system in MW/yr. Blue data points indicate the Comfort Violation in Kh/yr.

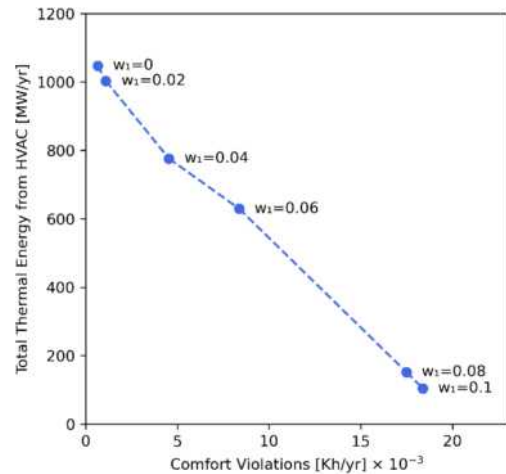


Figure 10: Plot of Total Thermal Energy against Comfort Violations for different w_1 values and a constant weight of $w_2=0.3$ from testing in May 2023. Blue data points represent the total thermal energy from the HVAC system in MW/yr and comfort violations in Kh/yr multiplied by a factor of 10^{-3} .

4. Discussion

4.1. System Specifications

From Figure 4, it is seen that the total reward plateaus before reaching the 20th training loop. This indicates that 20 training runs (corresponding to 18 years' worth of data) prove well sufficient to fully optimise the Q-Table and provide adequate training. A similar trend is also seen with the DQL controller.

Increments of 1°C between states might initially appear too large. However, using such increments does not hinder the control system. To confirm this hypothesis, both the DQL agent and the TQL agent are trained using state increments of 0.5°C and 0.2°C; no

visible impact on the controller performance is observed. Instead, the model takes significantly longer to train due to the larger number of states, indicating that increments of 1°C prove more efficient. Although using state increments of 2°C further improves the computation time, such a large temperature difference between states leads to a decline in performance. These results prove surprising for the DQL agent, as DQL usually performs better when state values are more granular [21]. This suggests that other factors may lead to its underperformance.

The reward function owes its effectiveness to the weights in its respective terms. From an iterative approach, it is found that for $\Delta T > 0$, w_1 and w_2 should be tuned such that energy penalties outweigh temperature deviation penalties. For $-2 < \Delta T < 0$, the temperature deviations should be penalised more heavily than energy use. As such, values of $w_2 > w_1$ and $w_3 > w_4$ should be used to ensure stable control.

4.2. Controller Performance Analysis

Section 3.2. establishes that the TQL agent outperforms the PI controller in every metric. This is indicative of the superior decision-making capabilities of the Tabular Q-Learning agent compared to the baseline PI controller. Superior energy efficiency not only aligns with Sainsbury's sustainability goals, but also represent savings in operational costs.

Both TQL and DQL models exhibit more stable behaviours compared to the PI controller. It can be seen from Figure 5 and Figure 7 that both the TQL and DQL controllers display more consistent HVAC power output profiles compared to the PI controller. This is due to the fundamental differences in the control systems. The HVAC heat load for the PI controller is proportional to the setpoint error (as discussed in section 2.1.2), while the RL agents can choose between a limited number of actions (as discussed in section 1.1.1.). This more consistent heating pattern reduces the wear and tear effect in the HVAC system and helps to improve its lifespan. It also leads to higher user satisfaction [19].

A look at Figure 8 indicates that the subpar performance of the DQL in terms of comfort can be attributed to a time lag between the setpoint change and the start of heating. This time lag might be due to an inadequate neural network structure, which might be too simple for this particular problem. A neural network that lacks capacity can struggle to identify the intricate patterns, knowledge of which is required for effective control [7].

The DQL algorithm is further tested with changes in the reward function and individual hyperparameters. However, no attempted combination of these results in a superior outcome. As such, it is deduced that the issues most probably originate from the structure of the neural network, which usually represents the greatest challenge in DRL methods [7]. Finding an optimal solution for this problem can prove expensive and time consuming, as no universal structure exists for a neural network. Neural networks are constructed for specific problems and are thus dependent on the characteristics of the problem in question [20]. Due to time constraints, attempting to

restructure the neural network was not possible. It was not considered worthwhile to conduct a sensitivity analysis on the constraints and hyperparameters used for the DQL agent as their ideal values were still unknown.

The much longer time to train the DQL compared to the TQL highlights the greater complexity of the former: the forward and backward passes in DQL prove computationally more intensive and time consuming [17] than the more straightforward table updates in TQL. Experience replays also constitute an additional step and requires management of the replay buffer, which leads to further increases in computational overhead.

By allowing the model to learn from a more diverse set of experiences, incorporating experience replays generally results in an improved controller performance in complex and dynamic environments. However, this does not prove true in this particular case due to the relative simplicity of the environment dynamics [17]. Moreover, Deep Q-Learning algorithms can struggle with relatively simpler environments because of the approximations made to the state-action spaces. Tabular Q-Learning, on the other hand, reviews every state-action combination independently, leading to more optimal results. Although TQL might prove computationally demanding for models comprising a large number of states and actions, it can demonstrate superior performance for cases involving a relatively small number of actions and states.

It is important to note that the *Results* section only illustrates some typical trained models. Results may slightly vary each time the agent is trained due to the mildly stochastic nature of Q-Learning, as explained by R. Sutton and A. Barto [7].

4.3. Sensitivity Analysis

4.3.1. Temperature Penalty Weight w_1

The system behaves counterintuitively as w_1 is varied. Using a larger w_1 value implies more importance placed on the comfort metric compared to the energy term, usually corresponding to smaller comfort violations and more energy used. However, the opposite is observed. A value of $w_1 = 0$ is deemed optimal as larger values w_1 result in monumental deviations from the setpoints and comfort violation. Although time constraints did not allow for a full investigation of this temperature penalty weight, this unusual behaviour could benefit from further research.

4.3.2. Learning Rate (α)

As described in section 2.1.8, a higher learning rate leads to very poor convergence of Q-values in the Q-Table. In the TQL model, it is found that in such a scenario, Q-values for different actions in a given state remain almost identical, and the action with the highest Q-value does not always correspond to the best action. This effectively results in fully stochastic control behaviour, independent of the amount of training data used. This explains why past research often considers very small α values as optimal. N. Ali and T. Tahir, for example, used a learning rate of 3×10^{-5} as they found that a value of this magnitude

significantly improved the performance of their model [30].

4.3.3. Exploration Rate (ϵ)

For ϵ values close to 0.01, the system would not consistently converge to the setpoints as a result of the agent not sufficiently exploring the environment during the training set, and prevents the agent from successfully finding the optimal actions as a result. The controller performance improves at larger ϵ values, as the model requires less training data to converge. S. Ghode and M. Digalwar [31] advise against using ϵ values greater than 0.5 value as this would risk erratic behaviour in the system – they were proven right. The ϵ value of 0.4 leads to optimal system performance as it most successfully balances exploitation and exploration within the environment. Instead of using a fixed exploration rate, T. Wei et. al. [19] experimented with an ϵ value which gradually decreased after each training loop. This approach was tested but did not impact the controller performance; it is therefore disregarded.

5. Conclusion

This paper demonstrates a tangible approach of Tabular Q-Learning (TQL) and Deep Q-Learning (DQL) to optimise the control of the HVAC system of a supermarket store. It is found that the TQL controller outperforms the existing PI controller by 4% in energy efficiency and 12% in user comfort. However, the DQL controller did not present any improvement over the baseline PI controller, and performs 2% worse in energy efficiency compared to the TQL controller. This suggests that further fine tuning of the DQL controller is required to unlock its full potential. Overall, this study demonstrates the potential of TQL for the control of energy systems in buildings. It would therefore be sensible to further investigate TQL as it offers a relatively simple solution to reduce improve the energy efficiency of HVAC systems.

The work from this paper could be extended by fine-tuning the neural network of the DRL agent and assessing whether it could exceed the performance of the TQL controller. Further research could investigate the potential of the TQL agent trained in this paper on similar processes, such as refrigeration, which uses significantly more energy compared to heating systems, indicating a greater energy saving potential. It can also prove beneficial to train the TQL agent on a Model Predictive Control (MPC) controller. MPC controllers often require a significant amount of time to fine tune, and RL can offer a solution to this problem. Finally, the TQL controller should be tested in real world environments, as the simulation environment used is heavily idealised. A real-life store may behave differently when accounting for a non-uniform temperature distribution or large unforeseen fluctuations in internal temperature and such factors need to be considered carefully before implementing the model in the real world.

Acknowledgements

We would like to express our sincere gratitude to Mr. Max Bird MEng, who assisted us in understanding the theory behind some of the more complicated concepts of reinforcement learning, as well as sharing his existing research knowledge with us.

References

- [1] UK Government - Department for Business, Energy & Industrial Strategy, "UK becomes first major economy to pass net zero emissions law, available online at <https://www.gov.uk/government/news/uk-becomes-first-major-economy-to-pass-net-zero-emissions-law>, accessed on 4 December 2023," 2019.
- [2] International Energy Agency, "Tracking Clean Energy Progress 2023, available online at <https://www.iea.org/reports/tracking-clean-energy-progress-2023>, accessed on 4 December 2023," 2023.
- [3] M. Bird, 2023.
- [4] Sainsbury's, "Available online at <https://about.sainsburys.co.uk/sustainability/better-for-the-planet/carbon>, accessed on 4 December 2023".
- [5] C. Shyalika, "A Beginners Guide to Q-Learning, available online at <https://towardsdatascience.com/a-beginners-guide-to-q-learning-c3e2a30a653c>, accessed on 5 December 2023," 2019.
- [6] D. Azuatalam et al., "Reinforcement learning for whole-building HVAC control and demand response," 2020.
- [7] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction (Second Edition), 2018.
- [8] udit, The Q in Q-learning: A Comprehensive Guide to this Powerful Reinforcement Learning Algorithm, 2023.
- [9] V. Mnih et al., Human-level control through deep reinforcement learning, 2015.
- [10] Y. Ye et al., "Model-Free Real-Time Autonomous Control for a Residential Multi-Energy System Using Deep Reinforcement Learning," 2020.
- [11] A.T.D. Perera, P. Kamalaruban, "Applications of reinforcement learning in energy systems," 2021.
- [12] K. Mason, S. Grijalva, "A review of reinforcement learning for autonomous building energy management," 2019.
- [13] M. Han et al., The reinforcement learning method for occupant behavior in building control: A review, 2021.
- [14] A. Del Rio Chanona, Theme III: Machine Learning for Optimisation, Data-driven sequential decision making and reinforcement learning, 2023.

- [15] S. Baghaee and I. Ulusoy, User Comfort and Energy Efficiency in HVAC systems by Q-learning, 2018.
- [16] E. Barrett and S. Linder , Autonomous HVAC Control, A Reinforcement Learning Approach, 2015.
- [17] S. Brandi, M. Piscitelli, M. Martellacci, A. Capozzoli , "Deep reinforcement learning to optimise indoor temperature control and heating energy consumption in buildings," 2020.
- [18] J. Lubars et al., Combining Reinforcement Learning with Model Predictive Control for On-Ramp Merging, 2021.
- [19] T. Wei, Y. Wang, Q. Zhu, "Deep Reinforcement Learning for Building HVAC Control," 2017.
- [20] Z. Jiang et al., "Building HVAC control with reinforcement learning for reduction of energy cost and demand charge," 2021.
- [21] X. Zhong et al., End-to-End Deep Reinforcement Learning Control for HVAC Systems in Office Buildings, 2022.
- [22] V. Pong et al., "Temporal Difference Models: Model-Free Deep RL for Model-Based Control," 2020.
- [23] M. Bird et al., Real-world implementation and cost of a cloud-based MPC retrofit for HVAC control systems in commercial buildings, 2022.
- [24] M. Bird, Centre for Process System Engineering, Imperial College London, 2021.
- [25] C. Kontoravdi, Process Dynamics and Control – Lecture 9: Process Modelling, 2021.
- [26] A. Metelli, Control Frequency Adaptation via Action Persistence in Batch Reinforcement Learning, 2020.
- [27] J. I. S. da Silva and A. R. Secchi, Model Predictive Control for Production of Ultra-Low Sulfur Diesel in a Hydrotreating Process, 2018.
- [28] N. Joshi, Part 1 — Building a deep Q-network to play Gridworld — DeepMind’s deep Q-networks, 2021.
- [29] UK Government - Department for Energy Security and Net Zero, UK Government GHG Conversion Factors for Company Reporting, 2023.
- [30] N. Ali and T. Tahir, Deep Reinforcement Learning Algorithms to Optimize Supply Chain Processes with Uncertain Demand, 2022.
- [31] S. Ghode and M. Digalwar, A Novel Model based Energy Management Strategy for Plug-in Hybrid Electric Vehicles using Deep Reinforcement Learning, 2023.
- [32] S. Mousavi, M. Schukat, E. Howley, "Traffic light control using deep policy-gradient and value-function-based reinforcement learning," 2017.
- [33] G. P. Henze, "Evaluation of Reinforcement Learning Control for Thermal Energy Storage Systems," 2003.
- [34] S. Brandi, M. Fiorentini, A. Capozzoli, "Comparison of online and offline deep reinforcement learning with model predictive control for thermal energy management," 2022.
- [35] J. Liu, Y. Li, Y. Ma, R. Qin, X. Meng, J. Wu, "Coordinated energy management for integrated energy system incorporating multiple flexibility measures of supply and demand sides: A deep reinforcement learning approach," 2023.
- [36] Y. Zhou, Z. Ma, J. Zhang, S. Zou, "Data-driven stochastic energy management of multi energy system using deep reinforcement learning," 2022.
- [37] Cornell University Ergonomics Web, DEA3500: Ambient Environment: Thermal Environment, available online at <https://ergo.human.cornell.edu/studentdownloads/DEA3500notes/Thermal/thcomnotes2.html>, accessed on 11 December 2023.

A Flexible Calcium Ion Holographic Sensor for Wound Monitoring via Smartphone Readout

Shihabuddeen Waqar and Ali Fadlelmawla

Abstract Chronic wounds pose significant challenges to patients and healthcare systems alike. Continuously monitoring the changing calcium ion (Ca^{2+}) concentration in the wound milieu, using a holographic calcium ion sensor integrated in a bandage via smartphone readout, can allow patients and doctors to monitor the progress of wound healing. The smartphone application is user friendly and limits the requirement for the intervention of medical professionals to obtain accurate results, thereby having the potential to reduce the burden on the healthcare system. Herein, a double photopolymerisation method is used to fabricate a holographic calcium ion sensor, which is the hydrogel poly (HEMA-co-PEGDA-co-MAA). The fabricated free standing holographic calcium ion sensor is attached to the flexible substrate, Polydimethylsiloxane (PDMS). The sensor replays a total Bragg peak shift of -18.53 ± 0.51 nm upon varying the calcium ion concentration in the biological contaminant free buffer solution between $0.0 - 4.0$ mmol L^{-1} at a pH of 8.0 and wound model diameter of 70mm. Individual selectivity tests in the presence of various biological contaminants at physiological concentrations in the buffer solutions show that the sensor maintains a notable selectivity towards calcium ions. The sensor also preserves its sensitivity when subjected to various levels of bending as the diameter of the wound models is varied. The smartphone application is capable of detecting and processing the holographic reflection spectrum, using region of interest detection (ROID) and thresholding algorithms, to quantify the calcium ion concentration in buffer solutions.

Keywords Continuous monitoring, Region of interest detection (ROID), Poly (HEMA-co-PEGDA-co-MAA), holographic sensor, Polydimethylsiloxane (PDMS), thresholding algorithm

1. Background and Introduction

1.1 Wound healing

Wound healing is a complex physiological process that results in tissue reconstitution in response to injury. Wound healing initially restores the protective epithelial barrier, which is the body's primary defence against infection from external pathogens and prevents fluid loss. The wound healing process typically consists of four distinct, but overlapping stages: haemostasis, inflammation, proliferation, and remodelling.¹ Wounds that take longer than 3 months to heal are classified as chronic wounds.² Chronic wounds stagnate in the inflammatory phase without further healing which predisposes patients to further complications such as disfigurement, loss of function and amputation.¹ In the US alone, chronic wounds affect over 6.7 million people annually, incurring an annual cost in excess of 50 billion US dollars, underscoring the need for effective and accessible wound monitoring and treatment.³ The current standard for wound assessment relies heavily on the clinician's observations, which are prone to subjective errors based on the clinician's experience and judgement. Biomarkers provide an indication of a patient's biological state and are potentially useful for understanding, or

predicting, the healing trajectory of a wound. Conventional biomarker quantification in wound fluids requires time-consuming and costly laboratory testing, such as enzyme-linked immunosorbent assays (ELISAs), which do not enable real-time monitoring or on-demand quantification of wound healing progress.⁴ This presents the need for objective wound healing assessments that can be quantified without the need for complex laboratory equipment and the limited intervention of medical professionals. The accurate monitoring of biomarkers is critical for deploying effective treatment and patient recovery.

1.2 Importance of calcium ions (Ca^{2+})

Calcium ions (Ca^{2+}) play a critical role in a range of physiological processes, including blood coagulation, enzyme activity, the release of neurotransmitters and hormones. A deficiency in calcium can lead to various health complications, such as rickets in children and osteomalacia in adults.⁵ Calcium has also shown to be effective in promoting wound healing when used in wound dressings.⁶ Research indicates that calcium concentration in the wound area varies with biochemical activities during the healing process. The extracellular

calcium ion concentration increases upon injury, persisting through the inflammatory and proliferative phases and declining during the remodelling phase.⁷ In the haemostasis phase, calcium facilitates blood clotting, while in the inflammatory phase, it modulates the function of immune cells like neutrophils. Extracellular calcium is essential for the proliferation and differentiation of skin cells involved in the wound healing process.⁸ The change in concentration that occurs during wound healing provides the opportunity for calcium ions to act as effective biomarkers to monitor the progress of healing.

1.3 Current methods to monitor calcium ions (Ca²⁺)

Current methods for monitoring the calcium ion levels include the use of fluorescent, electrochemical and colorimetric sensors. A coumarin based fluorescent sensor for calcium ion has been proven to work within living cells.⁹ Whilst it has proven to detect calcium ions, the fluorescence spectroscopy requires specialised equipment and trained personnel for analysis, in addition to high cost, which makes it a widely inaccessible form of monitoring biomarkers like calcium ions in patients.¹⁰ Moreover, fluorescent sensors are vulnerable to photobleaching over time due to the destruction of fluorophores.¹¹ Electrochemical sensors are also effective in calcium ion detection and monitoring. Electrochemical sensors offer many advantages over other classes of calcium ion sensors owing to their affordability, high sensitivity, selectivity, rapid response time and low power consumption.^{12,13} However, electrochemical sensors have drawbacks, including a lower shelf life and the occurrence of signal drift, which requires regular recalibration.¹⁴ A colorimetric sensing system measures the calcium ion concentration, where the calcium ions react with Arsenazo III and form a coloured complex which absorbs light in the visible spectrum. Through the use of a 650nm LED as a light source and a high-speed photodiode as a detector, the absorbance can be measured using Beer Lambert's law.¹⁵ Whilst the calorimetric sensors are generally promising, the lack of reversibility renders this class of sensors unsuitable for real time monitoring. Moreover, calorimetric sensors require the use of valuable laboratory equipment to obtain results, reducing its suitability for continuous monitoring.

1.4 Holographic sensors

While holograms are mostly known for their data storage and artistic capabilities, they have managed to find other uses in modern science such as holographic sensors. Holographic sensors utilise two coherent laser beams interfering in a photosensitive substance, triggering a

photochemical reaction that generates interference fringes that store the original light information.¹⁶ The colour observed is determined by the fringe spacing between changes in refractive index of these scales, as given by Bragg's law (Equation 1).

$$n\lambda_{max} = 2d\sin\theta \quad \text{Equation (1)}$$

Where λ_{max} is the Bragg peak wavelength, n is the diffraction order, d is the grating spacing, and θ is the angle of incident light from the normal. By varying the grating spacing according to the environment, a hologram can function as a sensor and convey information about the environment. The observed colour is redshifted if the grating spacing increases, giving a visual indication of the environment's state. This approach has been previously used for a wide variety of holographic sensors including pH, alcohol, temperature and humidity, glucose, strain, drug detection and ion sensors such as Cu²⁺ and Fe²⁺.¹⁷ Holographic sensors provide label free sensing that removes several complex processing steps required by conventional chemical analysis, therefore reducing the level of training required and removing the need for specialist laboratory equipment for users to obtain reliable data. Holographic sensors have many advantages over the variety of sensors discussed earlier due to their low cost and reversibility. Additionally, holographic sensors do not require electrical currents or large metallic components, allowing them to have a reduced impact on the environment. Holographic sensors are also free of the need to regularly re-calibrate like electrochemical sensors and are not impacted by photobleaching. In this work, a double photopolymerisation process was used to fabricate a hydrogel based holographic calcium ion sensor. Double photopolymerisation removes the need for nanoparticles in holographic gratings, reducing the complexity of the process, thus making mass production more achievable.¹⁷ Additionally, nanoparticle-free systems avoid the potential health risks of nanoparticle leaching.¹⁸ This work investigates the reliability of the sensor by measuring the influence of factors such as the pH environment, the level of bending the sensor experiences and the presence of biological contaminants on the sensor's sensitivity.

1.5 Hydrogel based holographic sensors

The holographic sensor in this work is hydrogel based. Hydrogels are three-dimensional networks of hydrophilic polymers that swell in water and retain water whilst maintaining their structural integrity because of chemical or physical cross linking of individual polymer chains. Hydrogels undergo reversible volume phase transition in

response to physical and chemical stimuli including temperature, solvent composition, pH and ions.¹⁹ In chemically cross-linked hydrogels, the polymer chains are covalently linked, providing excellent mechanical strength. Photoinitiated chemical cross-linking occurs through radiation exposure, where photo initiators absorb the photons and form free radicals, which in turn can react with vinyl bonds in monomers and form a polymer network that preserves its structure in an aqueous medium. This cross-linking technique was applied in the fabrication of poly (HEMA-co-PEGDA-co-MAA) hydrogels in this work. A unique property of hydrogels includes non-solubility in water while remaining hydrophilic due to the presence of hydroxyl, carboxylic and amidic groups.²⁰ Non-solubility in water is desirable as it allows for the sensor to monitor the calcium ion concentration in an aqueous environment. Additionally, poly (HEMA-co-PEGDA-co-MAA) hydrogels are flexible, biocompatible and transparent making them suitable for applications to wounds and skin. The free-standing poly (HEMA-co-PEGDA-co-MAA) hydrogel is then attached to a flexible substrate Polydimethylsiloxane (PDMS) that provides it with mechanical stability. Moreover, Polydimethylsiloxane (PDMS) is flexible, biocompatible and closely resembles the skin, making it suitable for skin adhesion and decreases discomfort as wound sites are particularly sensitive regions.

2. Methods

2.1 Materials

All chemicals were analytical grade. 3-(trimethoxy silyl)-propyl methacrylate, acetone, hydroxy ethyl methacrylate (HEMA), ethylene glycol dimethacrylate (EGDMA), 2-(dimethyl amino) ethyl acrylate (DMAEA), 2-hydroxy-2-methylpropiophenone (HMPP), isopropanol, tris(hydroxymethyl)amino methane (TRIS), TRIS hydrochloride (TRIS HCl), potassium chloride, calcium chloride, magnesium chloride, sodium bicarbonate, urea, uric acid, sodium lactate, and hydrochloric acid (1M) were purchased from Sigma Aldrich. Methanol, glucose, and sodium chloride were purchased from VWR. Aluminised polyester films were purchased from HiFi Industrial Film Ltd. SYLGARD 184 Silicone Elastomer Kit 1.1Kg was purchased from Dow. Cotton layers were purchased from Synergy Health (UK) Ltd.

2.2 Equipment

Microscope slides were purchased from Fisher Scientific, cover glass slides from VWR and UVP crosslinker from Analytik Jena. Nd: YAG frequency tripled Quantel Q-smart (5 ns, 355 nm) solid-state laser purchased from Lumibird, France. The drying oven UN30 used for hydrogel drying was purchased from Memmert. The plano-concave lens (−75.0 mm, Ø1" UV fused silica plano-concave lens, uncoated), optical posts (Ø12.7 mm, L = 20 mm), pedestal post holder (L = 20 mm, Ø12.7

mm), power and energy meter interface (PM100USB), UV extended Si photodiode, motorized precision rotation stage (Ø1"), collimated laser-diode-dumped DPSS laser module (532 nm, 4.5 mW), laser diode module mounting Kit (Ø11 mm), and iris (mounted standard iris, Ø12 mm max aperture, TR3 Post) were purchased from Thorlabs, United States. A 25 mm dielectric 355 nm Nd: YAG laser line mirror was purchased from Edmund Optics, UK. Fisherbrand classic vortex mixer was purchased from Fisher Scientific, Bishop Meadow, UK. The Mettler Toledo FiveEasy Plus™ pH benchtop meter was purchased from Mettler-Toledo Ltd. Orion Star A212 Benchtop Conductivity Meter from ThermoFisher. Holographic responses were analyzed using UV-VIS bifurcated fibre optical cables, a 20 W tungsten halogen broadband light source, Flame-S-VIS-NIR-ES spectrophotometer, and Oceanview software (2.0.8) purchased from Ocean Insight.

2.3 Free-standing holographic calcium ion sensor fabrication.

The holographic calcium ion sensor was fabricated using two different monomer solutions. The primary monomer solution (P1) contains 63 mol% 2-hydroxyethyl methacrylate (HEMA), 6 mol% polyethylene glycol diacrylate (PEGDA 700), 30 mol% methacrylic acid (MAA), and 1 mol% 2-hydroxy-2-methylpropiophenone (HMPP). The P1 solution is then diluted at a ratio of 1:1 v/v with isopropanol. The secondary monomer solution contains 55 mol% HEMA, 40 mol% ethylene glycol dimethylacrylate (EGDMA), and 5 mol% HMPP. The secondary monomer (P2) solution monomers are then diluted with 90 vol% methanol at a ratio of 1:1 v/v.

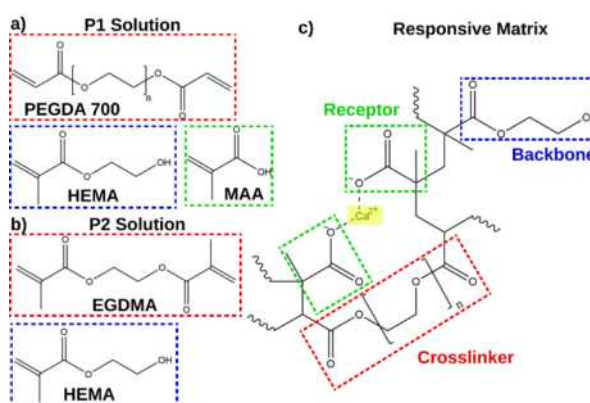


Figure 1. Schematic representation of the structure of a holographic calcium ion sensor. a) Composition of P1 solution. b) Composition of P2 solution. c) Composition of the responsive matrix.

Figure 1c depicts the formation mechanism of the responsive matrix. Exposing the HMPP to incoherent UV light leads to its rapid decomposition into highly active free radicals, which initiates a polymerisation chain reaction of monomers in the P1 solution (Figure 1a) and MAA in the responsive matrix. EGDMA was used as the

crosslinker in secondary monomer (P2) solution (Figure 1b) due to its mechanical stiffness, which prevents any volumetric change. While the surface of healthy human skin is acidic, the wound's nature is alkaline due to the underlying tissues of the body being exposed, which typically have a pH of 7.4.^{21,22} This alkaline environment leads to the rapid deprotonation of MAA. Calcium ions that penetrate the hydrogel are able to coordinate with the carboxylate ions to form ionic bonds. This ionic bonding results in the hydrogel shrinking, which leads to the interference layer (IL) stripes narrowing and the observed wavelength blue shifting, as shown in Figure 2. As the ionic bonding results in a volumetric change in the hydrogel, the shift in the observed peak wavelength can be used to quantify the calcium ion concentration in the wound.

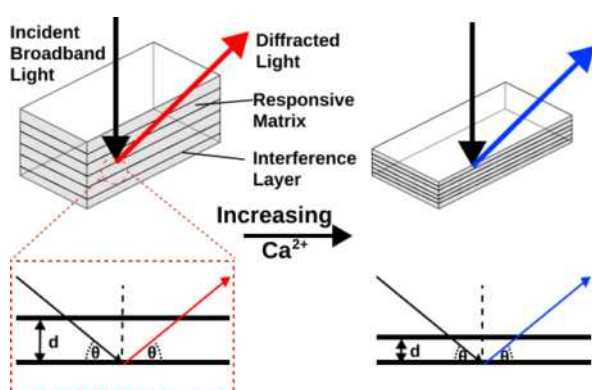


Figure 2. Working principle of hydrogel shrinking due to the ionic bonding of MAA and calcium ions, leading to a blue shift in the peak wavelength.

On the polyester side of a flat aluminized polyester film, 40 μL of the primary monomer solution was pipetted, and a clean untreated glass slide was placed onto the droplet. It was then exposed under 4 UV (A) strip lights for 30 minutes. The polymerized primary hydrogel layer was then removed from the polyester film and washed in methanol and water at a ratio of 1:1 (v/v) for 30 minutes to remove any by-products and unpolymerized monomers from the hydrogel layer. The hydrogel was then wiped dry, and 150 μL of the P2 solution was pipetted onto a clean, flat, aluminized polyester film.

Hydrogel side facing downward, the slide was then placed on the droplet to allow the hydrogel to be soaked in the P2 solution for 5 minutes. The hydrogel surface was then given a single wipe to remove excessive secondary monomer solution. The hydrogel was then dried in an oven at 55 $^{\circ}\text{C}$ for 4.5 minutes. The hydrogel was then left to cool down to 25 $^{\circ}\text{C}$. Since the thickness of hydrogel at laser exposure is a critical factor affecting replay wavelengths of holographic sensors, these steps must be maintained at the same temperature and humidity. Slides were kept under safe lighting and exposed to a single UV pulse from an Nd: YAG laser (355 nm, 5 ns, 30 μs delay)

on the planar mirror with the hydrogel side facing downwards with a tilted angle of 5 $^{\circ}$ from the mirror. While maintaining safe lighting, hydrogels were washed in methanol and DI water (1:1 v/v) overnight. This setup is shown in Figure 3.

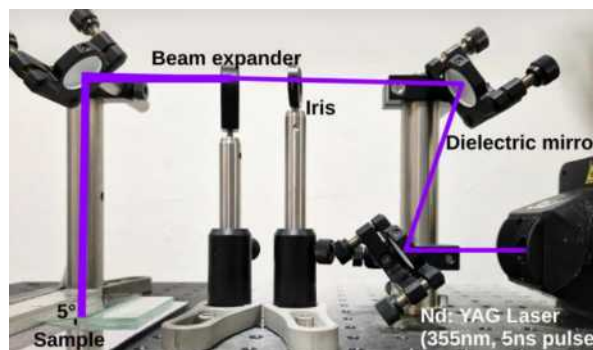


Figure 3. A photograph of the laser arrangement showcasing the sample positioned at a 5 $^{\circ}$ elevation from one side and undergoing exposure to a single UV pulse (355 nm, 5 nanoseconds, with a 30-microsecond delay).

2.4 PDMS substrate fabrication

The siloxane was mixed with the curing agent at a mass ratio of 10:1. The mixture was degassed using a vacuum desiccator, poured into a mould, and covered using clean microscope glass slides to fabricate a PDMS sheet with a thickness of 80 μm . The mould was then placed in an oven at 80 $^{\circ}\text{C}$ for 1 hour. The fabricated PDMS sheet was then removed from the mould, treated using the plasma cleaner, and silanised using a 3-(trimethoxysilyl) propyl methacrylate in acetone at 1:50 (v/v) overnight.

2.5 Attach free-standing holographic sensor on PDMS substrates

Both free-standing holographic calcium ion sensors and PDMS substrates were rinsed with methanol to clean surfaces. An adhesive monomer solution was prepared by mixing PEGDA 700 and 3 vol% HMPP in DI water at a ratio of 2:1 v/v. The diluted PEGDA 700 solution was pipetted on the free-standing holographic calcium ion sensor surface. The PDMS substrate was then placed on top of the PEGDA 700 solution droplet. A pressure of 1 kPa was applied to facilitate the attachment. The sample was then placed under 4 UV (A) strip lights for 15 minutes. The PDMS substrate holographic calcium ion sensor was then carefully removed from glass slides and washed with in methanol and DI water (1:1 v/v) for 30 minutes. The two distinct layers of the dried holographic calcium ion sensor and PDMS can be seen in Figure 4.

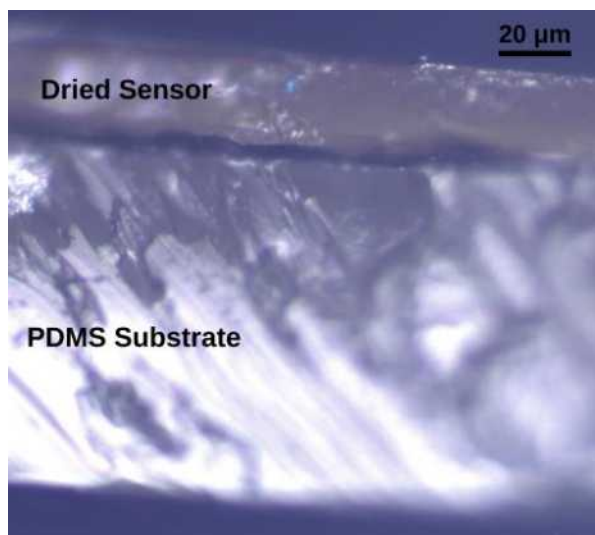


Figure 4. A photograph of the hydrogel based holographic calcium ion sensor attached to PDMS Substrate, under a microscope.

2.6 Buffer solution formulation

The buffer solutions used to perform all experiments contain calcium chloride ($0.0 - 4.0 \text{ mmol L}^{-1}$) dissolved in deionized water (DI). TRIS, TRIS HCl and potassium chloride (4.40 mmol L^{-1}) were added to the buffer solutions, maintaining TRIS concentration at 50 mmol L^{-1} at a constant pH of 8.00 to stabilise the pH of the buffer solutions. To stabilise the calcium concentration, 3 mmol L^{-1} citric acid was added. The pH value of buffer solutions was corrected using 1.0 mol L^{-1} hydrochloric acid and 3.0 mol L^{-1} potassium hydroxide confirmed using the pH meter. To perform the pH test, the buffer solutions' pH values were corrected using the same method to prepare solutions in the pH range of 7.0-9.0. Buffer solutions used to perform the selectivity test contain physiological concentrations of common electrolytes and biological contaminants. The buffer solutions were prepared by dissolving the following contaminants separately into buffer solutions with the same composition used in the sensitivity and pH tests; sodium chloride ($128.33 \text{ mmol L}^{-1}$), magnesium chloride (0.94 mmol L^{-1}), potassium chloride (4.40 mmol L^{-1}), urea (8.90 mmol L^{-1}), uric acid (0.35 mmol L^{-1}) and sodium lactate ($10.90 \text{ mmol L}^{-1}$) in DI water. The calcium ion concentration was tested at 0.0 mmol L^{-1} and 4.0 mmol L^{-1} . On each change of calcium ion concentrations, the artificial wound model was rinsed using the next buffer solution three times to ensure a reliable calcium concentration in the artificial wound model.

2.7 Statistical analysis

Spectra taken by the spectrophotometer were firstly processed by the Savitzky-Golay filter and then subtracted by broadband light spectra taken before the

measurement. The processed spectra were then normalized to $[0,1]$. All replay wavelength and Bragg peak shift data were expressed as mean \pm standard error. All the above-mentioned data processing was carried out using Origin 2020. The data processing was carried out on Python 3.9.7 using NumPy, OpenCV and Pillow libraries.

3. Results & Discussion

3.1 Sensitivity test

To quantify the calcium ion concentration at the site of the wound, a calibration curve was generated using experimental results so that the Bragg shift values can be converted into the calcium ion concentration. The blood calcium ion concentration is maintained within a narrow range of $2.2 - 2.7 \text{ mmol L}^{-1}$.²³ The calcium ion concentration increases during the inflammatory and proliferative phases, the two phases that this work seeks to monitor and address. The inflammatory phase is when extracellular calcium enters neutrophils, increasing intracellular calcium concentration which modulates the neutrophil function.²⁴ The proliferative phase is characterised by the resurfacing of the wound with a new epithelium, where the increase in the concentration of calcium ions at the site of the wound initiates epithelial healing.^{25,26} Due to the variation in calcium ion concentration, the sensitivity test was performed across a concentration range of $0.0 - 4.0 \text{ mmol L}^{-1}$ of free calcium ions in buffer solutions where 3 mmol L^{-1} citric acid was added to stabilise the calcium ion concentrations in the buffer solutions. The pH at the site of chronic wounds have been recorded to be in the range of 7.15-8.90.²⁷ In order to replicate the alkaline environment at the site of wounds, the 50 mmol L^{-1} TRIS buffer solutions were prepared to be at a pH of 8.0. A Bragg shift

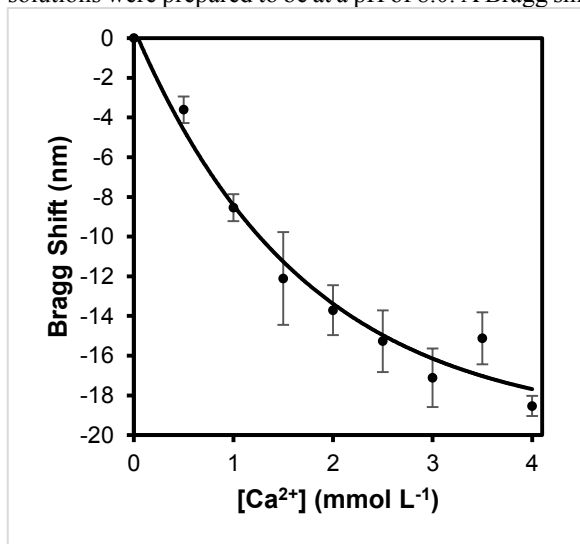


Figure 5. Calibration curve of $0.0 - 4.0 \text{ mmol L}^{-1}$ calcium ion concentration by observed Bragg shift in nm. The error bars represent the standard error.

of -18.53 ± 0.51 nm was measured across the range of calcium ion concentrations as shown by Figure 5. The negative Bragg shift (blue shift) is expected as an increase in calcium ion concentration causes the hydrogel networks to decrease in volume, reducing the spacing between the IL stripes. Two main trends can be observed from Figure 5. Firstly, the Bragg shift exhibits an approximately linear trend between the calcium ion concentrations of $0.0 - 1.0$ mmol L⁻¹. The second trend is one resembling exponential decay across the entire range of calcium ion concentrations, where the size of the change in Bragg shift appears to decrease as the calcium ion concentration increases. The first trend can be explained by an abundance of deprotonated MAA within the hydrogel matrix that can form ionic bonds with the free calcium ions, therefore increasing the calcium ion concentration leads to a proportional Bragg shift. However, a factor limiting the Bragg shift achievable is the availability of deprotonated MAA that can bond to the free calcium ions. As more calcium ions form ionic bonds with the deprotonated MAA, less sites are available for calcium ions to bond to, hence the change in Bragg shift becomes less significant, as not all the calcium ions are able to participate in ionic bonding. This limits the volumetric change in the hydrogel structure, and this is reflected in an exponential decay in the impact increasing calcium concentration has on the observed Bragg shift.

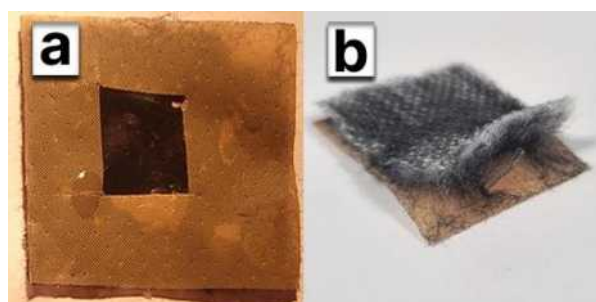


Figure 6. Holographic calcium ion sensor integrated into a bandage. (a) Sensor fitted into the bandage cut-out. (b) Black cotton on the underside of the bandage prevents light scattering.

3.2 Prototype and bending test

As the location of the wound can vary across the body, so will the bending the sensor integrated into the bandage (Figure 6a) experiences. For example, wounds on regions with a greater amount of curvature, including fingers and the brow ridge, will cause the hydrogel to bend more than flatter surfaces like the legs and torso. The prototype of the sensor integrated into a bandage has black cotton on the underside of the bandage. As the wound dressing is worn on the skin, light is reflected off the skin which causes light scattering that is much greater than the holographic signal. Using black cotton prevents light scattering, which simplifies the detection of the holographic signal. To investigate any changes in the hydrogel sensitivity with the extent of bending, a bending

test was conducted. The diameters of the artificial wound site ranged from 30mm to 70mm and were tested at 10mm increments. The results in Figure 7 show that as the diameter of the wound model is reduced, the magnitude of the Bragg shift also reduces. There is an average decrease of 2.86 nm in sensitivity across the range of non-zero calcium ion concentrations. The reason for this is because bending increases the distance between adjacent holographic gratings. Increasing bending, increases the resistance to the opposing hydrogel shrinkage when the calcium ion concentration is increased. As a result, the holographic signal is altered so that the peak wavelength of light in the reflection spectrum is greater. This effectively reduces the magnitude of the Bragg shift observed at higher levels of bending (smaller diameters).

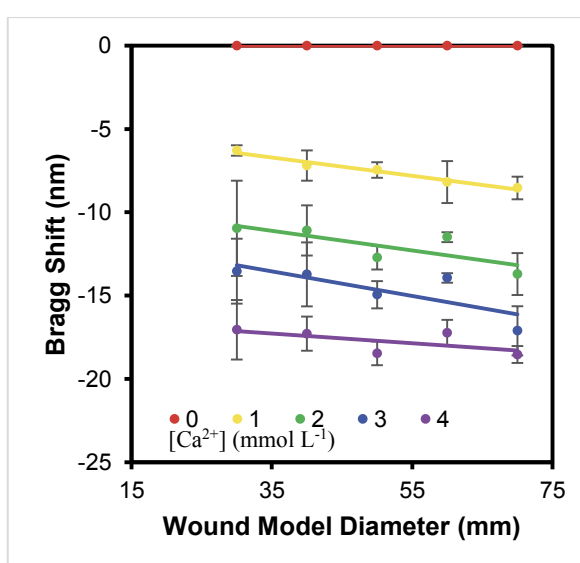


Figure 7. The Bragg shift achieved at different calcium ion concentrations ($0.0 - 4.0$ mmol L⁻¹) at different diameters of the artificial wound model. The error bars represent the standard error.

3.3 Selectivity test

Calcium ions are one of many cations in the blood that can form ionic bonds with deprotonated MAA. Various components aside from calcium ions in blood can impact the Bragg shift values obtained as calcium ions will not exclusively be responsible for any volumetric change in the hydrogel matrix. Some of the biological contaminants present in blood that were tested in the form of a selectivity test at physiological concentrations alongside calcium ions at concentrations of 0.0 mmol L⁻¹ and 4.0 mmol L⁻¹ in buffer solutions including sodium ions (Na⁺), magnesium ions (Mg²⁺), Urea, Uric acid and Lactate.²⁸⁻³⁰

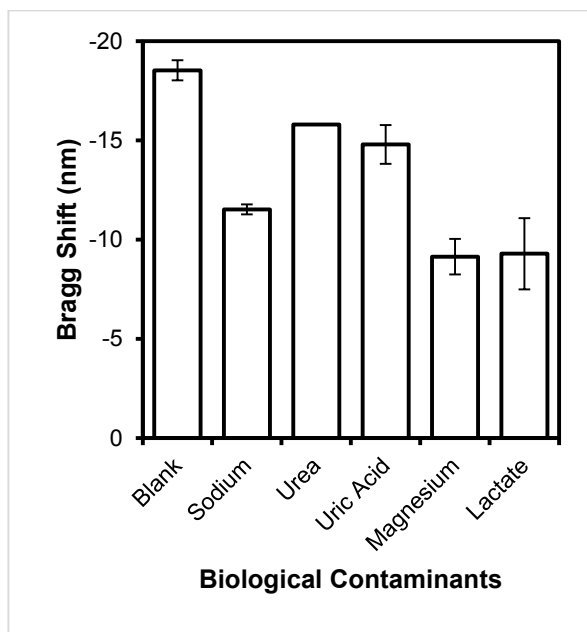


Figure 8. The Bragg shift measured between calcium ion concentrations of $0.0 - 4.0 \text{ mmol L}^{-1}$ in the presence of various biological contaminants in the buffer solutions. The error bars represent the standard error.

Figure 8 shows the total observed Bragg shifts observed in buffer solutions with the biological contaminants present, and a blank sample which provides a benchmark for the expected Bragg shift with no biological contaminants present, which was $-18.53 \pm 0.51 \text{ nm}$. The results in Figure 8 show that the calcium ion sensor consistently achieves a Bragg shift of a smaller magnitude with the biological contaminants present compared to the blank sample. The samples with magnesium ions present showed the greatest change in Bragg shift, $-9.14 \pm 0.90 \text{ nm}$. Magnesium ions possess the same positive charge as the competing calcium ions, and readily form ionic bonds with the deprotonated MAA's carboxylate group too. This explains the reduction in sensitivity in the sensor. Similarly, the monovalent sodium cations reduce the sensor's sensitivity for the same reason as magnesium ions, resulting in a Bragg shift of $-11.52 \pm 0.25 \text{ nm}$. The presence of Urea and Uric acid did not result in a significant reduction in the sensitivity of the sensor. Finally, the presence of sodium lactate decreased the sensitivity of the sensor, exhibiting a Bragg shift of $-9.29 \pm 1.80 \text{ nm}$. This can be explained by the combination of the presence of sodium ions in the compound and, to a lesser extent, the lactate anions. The sodium ions compete with calcium ions when forming ionic bonds with the deprotonated MAA. Also, the lactate ions possess a carboxylate group that the calcium ions can also form ionic bonds with instead of with the deprotonated MAA.

3.4 pH test

The pH of a wound varies during healing, initially intact skin has a pH of 5.4-5.6, upon injury the pH increases resulting in an increasingly alkaline environment. The pH of chronic wounds is in the range of 7.15-8.9 when wounds are still undergoing the inflammatory and proliferative phases of healing where wounds have yet to restore epithelial tissue.¹ As epithelial tissue is restored, the wound environment shifts to a slightly acidic pH of 6.0. As this project focuses on addressing chronic wounds delayed in the inflammatory and proliferative stages, a pH test was conducted in the pH range of 7.0-9.0 to investigate any changes in the sensor's sensitivity.

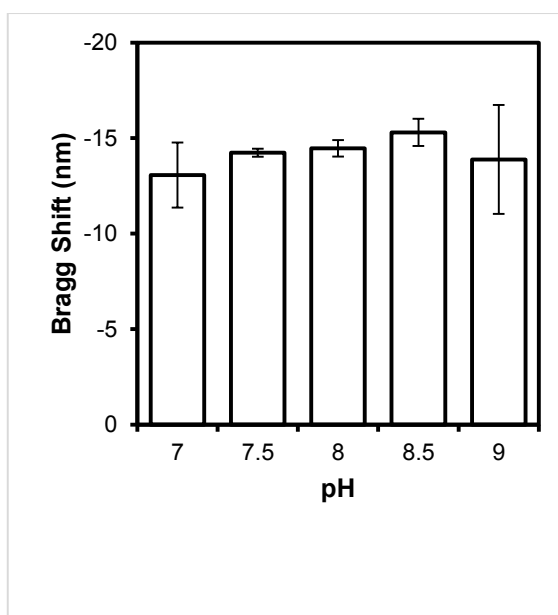


Figure 9. The Bragg shift achieved between calcium ion concentrations of $0.0 - 4.0 \text{ mmol L}^{-1}$ in pH environments ranging from 7.0-9.0 in the buffer solutions. The error bars represent the standard error.

The results in Figure 9 show that overall, the sensor is moderately resistant to changes in the pH environment. As the pH increases the Bragg shift achieved increased in magnitude between pH values of 7.0-8.5. The Bragg shift at a pH of 7.0 was measured to be an average of $-13.07 \pm 1.70 \text{ nm}$, and steadily blue shifts to $-15.03 \pm 0.71 \text{ nm}$ at a pH of 8.5. This is because the deprotonation of carboxylic acid groups in MAA can be affected by the pH value.³¹ As the buffer solution environment becomes more basic, the MAA deprotonates to a greater extent, allowing for a greater proportion of carboxylate ions to be available to form ionic bonds with the free calcium ions. The increase in coordination between the calcium ions and deprotonated MAA results in a more significant volumetric change in the sensor's hydrogel matrix.

3.5 Smartphone application

A smartphone application was developed to detect and interpret the holographic signal from the sensor to quantify the calcium ion concentration at the site of a wound. The user interface is shown in Figure 10. The smartphone application demonstrated proficiency in conducting Region of Interest Detection (ROID) to effectively isolate the hologram from the surroundings. Initially, the image underwent a conversion from Red, Green, and Blue (RGB) values to Hue, Saturation and Brightness (HSB) values as a pre-processing step to facilitate ROID. Given that the hologram typically exhibits elevated saturation and brightness compared to the overall image, thresholding algorithms were employed. An OTSU threshold was first applied to the saturation channel to identify pixels with significantly high saturation. The remaining pixels were filtered using a basic thresholding algorithm to eliminate dark spots with brightness values falling below a predetermined threshold. The assessment of ROID effectiveness involved the utilisation of 10 images, revealing a success

rate of 90%. Some examples of the ROID can be seen in Figure 10j, Figure 10k and Figure 10l.

The changes in hue values resemble changes in wavelength, so it can be theoretically used to quantify the Bragg shift in the isolated region. However, the practical application of this approach encountered substantial challenges. The application faced challenges in consistently determining calcium concentration, attributed to various factors. Firstly, a maximum Bragg shift of -18.53nm , achieved at a concentration of 4.0 mmol L^{-1} , was determined to be too small. Such a shift is susceptible to errors due to viewing angle and lighting variations. This issue was complicated further as several different colours are reflected by the hologram, as seen in Figure 10g. Finally, due to the non-linear correlation between hue and wavelength, the hue changes tend to be lower when the light is redshifted, a prevalent characteristic in the holograms used.³²

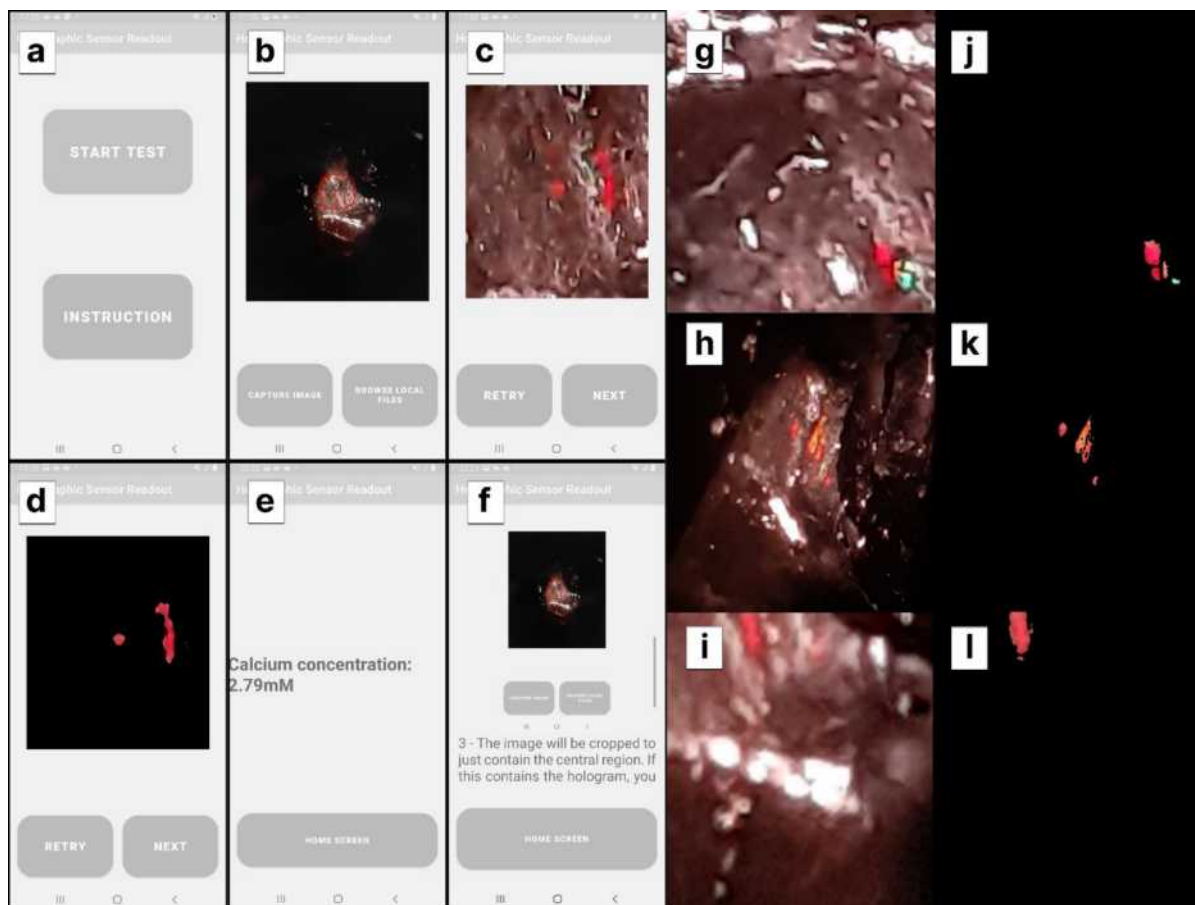


Figure 10. Images taken on the smartphone showcasing the mobile application and the ROID. (a) Main Page. (b) Second Page to capture image. (c) The chosen image is cropped. (d) The cropped image is processed. (e) The calcium concentration is displayed. (f) Instructions Page. (g, h, i) Examples of images taken before ROID. (j, k, l) Examples of images taken after ROID.

4. Conclusion

In this work a flexible hydrogel-based holographic calcium ion sensor was successfully fabricated and tested alongside a developed smartphone app designed to interpret the holographic output and relay results. Overall, the sensor proved to be moderately resistant to various levels of bending, experiencing an average decrease of 2.86 nm in sensitivity across the range of calcium ion concentrations. The sensor also showed a limited change in sensitivity when tested in various pH environments. In the individual selectivity tests the sensor showed good selectivity towards calcium ions apart from sodium, magnesium and lactate ions, where the sensor demonstrated limited selectivity towards calcium ions, dropping by approximately 50% in sensitivity. A small decline in sensitivity due to various factors is an issue in this project as the sensitivity of the sensor is low to begin with, resulting in a Bragg shift of -18.53nm at a pH of 8.0 and artificial wound site diameter of 70mm (little to no bending). This ultimately limits the use case of the sensor in its current state as the Bragg shift achieved is not large enough to be reliably and accurately detected and processed by the smartphone application. Accuracy of data is highly important when providing medical data to patients and doctors, inaccuracy may lead to unnecessary cause for concern or a failure to alert patients with deteriorating wound health. The low sensitivity can be addressed in further research by investigating alternative flexible and bio-compatible substrates to PDMS and altering the hydrogel composition. Additional factors that complicate the use of the sensor that should be noted and resolved is the limited viewing angle to observe the reflection spectrum from the holographic sensor, which requires time consuming alignment to obtain results. The limited selectivity of the sensor towards calcium ions can also be improved by investigating alternatives receptor chemicals to MAA. Prospective receptor chemicals would ideally be strongly selective towards calcium ions by reacting based on properties unique to calcium ions, like specific charge or ionic radius. The responsive sensor and capable smartphone app developed lay the foundation to a flexible, accessible, and reusable wound monitoring system that can reduce the burden on the healthcare system if the aforementioned shortcomings are addressed in further research.

Acknowledgements

Ali Fadlelmawla and Shihabuddeen Waqar express their sincere gratitude to Mr. Zhang Yihan for his guidance throughout the project and to everyone in the Yetisen group for their support and encouragement.

References

1. Gethin, G. The Significance of Surface PH in Chronic Wounds. *Wounds UK* **2007**, 3 (3).
2. Iqbal, A.; Jan, A.; Wajid, M.; Tariq, S. Management of Chronic Non-Healing Wounds by Hirudotherapy. *World Journal of Plastic Surgery* **2017**, 6 (1).
3. Businesswire. *Wound Care Awareness Week Highlights Chronic Wound Epidemic in U.S.* [www.businesswire.com](https://www.businesswire.com/news/home/20160607006326/en/Wound-Care-Awareness-Week-Highlights-Chronic-Wound).
<https://www.businesswire.com/news/home/20160607006326/en/Wound-Care-Awareness-Week-Highlights-Chronic-Wound>.
4. Hoyo, J.; Bassegoda, A.; Ferreres, G.; Hinojosa-Caballero, D.; Gutiérrez-Capitán, M.; Baldi, A.; Fernández-Sánchez, C.; Tzanov, T. Rapid Colorimetric Detection of Wound Infection with a Fluidic Paper Device. *International Journal of Molecular Sciences* **2022**, 23 (16). <https://doi.org/10.3390/ijms23169129>.
5. M. Biga, L.; Bronson, S.; Dawson, S.; Harwell, A.; Hopkins, R.; Kaufmann, J.; LeMaster, M.; Matern, P.; Morrison-Graham, K.; Oja, K.; Quick, D.; Runyeon, J.; OSU OERU; OpenStax. *Anatomy & Physiology*; OpenStax, 2019.
6. Oda, Y.; Tu, C.-L.; Menendez, A.; Nguyen, T.; Bikle, D. D. Vitamin D and Calcium Regulation of Epidermal Wound Healing. *The Journal of Steroid Biochemistry and Molecular Biology* **2016**, 164. <https://doi.org/10.1016/j.jsbmb.2015.08.011>.
7. LANSDOWN, A. B.; SAMPSON, B.; ROWE, A. Sequential Changes in Trace Metal, Metallothionein and Calmodulin Concentrations in Healing Skin Wounds. *Journal of Anatomy* **1999**, 195 (3). <https://doi.org/10.1046/j.1469-7580.1999.19530375.x>.
8. Subramaniam, T.; Fauzi, M. B.; Lokanathan, Y.; Law, J. X. The Role of Calcium in Wound Healing. *International Journal of Molecular Sciences* **2021**, 22 (12). <https://doi.org/10.3390/ijms22126486>.
9. Yao, K.; Chang, Y.; Li, B.; Yang, H.; Xu, K. A Novel Coumarin-Based Fluorescent Sensor for Ca²⁺ and Sequential Detection of F⁻ and Its Live Cell Imaging. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **2019**, 216, 385–394. <https://doi.org/10.1016/j.saa.2019.03.035>.
10. Suvetha, S.; Muruganandam, G.; Hariharan, G.; Nesakumar, N.; Jayalatha Kulandaisamy, A.; Bosco Balaguru Rayappan, J.; Mahendran Gunasekaran, B. Electrochemical Investigation on Calcium Ion Sensing and Steady-State Diffusion Analysis Using Zinc Oxide Modified Glassy Carbon Electrode. *Measurement* **2023**, 221, 113511. <https://doi.org/10.1016/j.measurement.2023.113511>.

11. Hashizume, R.; Fujii, H.; Mehta, S.; Ota, K.; Qian, Y.; Zhu, W.; Drobizhev, M.; Nasu, Y.; Zhang, J.; Bito, H.; Campbell, R. E. A Genetically Encoded Far-Red Fluorescent Calcium Ion Biosensor Derived from a Biliverdin-Binding Protein. *Protein Science* **2022**, *31* (10). <https://doi.org/10.1002/pro.4440>.
12. Almeida, J. M. S.; Dornellas, R. M.; Yotsumoto-Neto, S.; Ghisi, M.; Furtado, J. G. C.; Marques, E. P.; Aucélio, R. Q.; Marques, A. L. B. A Simple Electroanalytical Procedure for the Determination of Calcium in Biodiesel. *Fuel* **2014**, *115*, 658–665. <https://doi.org/10.1016/j.fuel.2013.07.088>.
13. Alarfaj, N. A.; El-Tohamy, M. F.; Oraby, H. F. A Reduced Graphene Oxide/Gold Nanoparticles Composite Modified Glassy Carbon Electrode as Electrochemical Sensor for Calcium Ions Detection in Bottled Water. *International Journal of Electrochemical Science* **2020**, *15* (6). <https://doi.org/10.20964/2020.06.49>.
14. Menon, S.; Mathew, M. R.; Sam, S.; Keerthi, K.; Kumar, K. G. Recent Advances and Challenges in Electrochemical Biosensors for Emerging and Re-Emerging Infectious Diseases. *Journal of Electroanalytical Chemistry* **2020**, *878*, 114596. <https://doi.org/10.1016/j.jelechem.2020.114596>.
15. Kajalkar, R.; Gaikwad, A. Colorimetry Based Calcium Measurement. *International Journal of Engineering Research and Development* **2013**, *7* (8), 8–11.
16. Jiang, N.; Davies, S.; Jiao, Y.; Blyth, J.; Butt, H.; Montelongo, Y.; Yetisen, A. K. Doubly Photopolymerized Holographic Sensors. *ACS Sensors* **2021**, *6* (3). <https://doi.org/10.1021/acssensors.0c02109>.
17. Davies, S.; Hu, Y.; Jiang, N.; Blyth, J.; Kaminska, M.; Liu, Y.; Yetisen, A. K. Holographic Sensors in Biotechnology. *Advanced Functional Materials* **2021**, *31* (47). <https://doi.org/10.1002/adfm.202105645>.
18. Kawata, K.; Osawa, M.; Okabe, S. In Vitro Toxicity of Silver Nanoparticles at Noncytotoxic Doses to HepG2 Human Hepatoma Cells. *Environmental Science & Technology* **2009**, *43* (15). <https://doi.org/10.1021/es900754q>.
19. Bahram, M.; Mohseni, N.; Moghtader, M. An Introduction to Hydrogels and Some Recent Applications. In *Emerging concepts in Analysis and Applications of Hydrogels*; Majee, S. B., Ed.; IntechOpen, 2016.
20. Choi, J. R.; Yong, K. W.; Choi, J. Y.; Cowie, A. C. Recent Advances in Photo-Crosslinkable Hydrogels for Biomedical Applications. *BioTechniques* **2019**, *66* (1). <https://doi.org/10.2144/btn-2018-0083>.
21. Rippke, F.; Schreiner, V.; Schwanitz, H.-J. The Acidic Milieu of the Horny Layer. *American Journal of Clinical Dermatology* **2002**, *3* (4). <https://doi.org/10.2165/00128071-200203040-00004>.
22. O'Meara, S.; Cullum, N.; Majid, M.; Sheldon, T. Systematic Reviews of Wound Care Management: (3) Antimicrobial Agents for Chronic Wounds; (4) Diabetic Foot Ulceration. *Health Technology Assessment* **2000**, *4* 23.
23. Goldstein, D. A. Serum Calcium. In *Clinical Methods: The History, Physical, and Laboratory Examinations*. 3rd edition; Walker, H. K., Hall, W. D., Hurst, J. W., Eds.; Butterworths, 1990.
24. Immler, R.; Simon, S. I.; Sperandio, M. Calcium Signalling and Related Ion Channels in Neutrophil Recruitment and Function. *European Journal of Clinical Investigation* **2018**, *48*, e12964. <https://doi.org/10.1111/eci.12964>.
25. Jeffcoate, W. J.; Vileikyte, L.; Boyko, E. J.; Armstrong, D. G.; Boulton, A. J. M. Current Challenges and Opportunities in the Prevention and Management of Diabetic Foot Ulcers. *Diabetes Care* **2018**, *41* (4). <https://doi.org/10.2337/dc17-1836>.
26. Cordeiro, J. V.; Jacinto, A. The Role of Transcription-Independent Damage Signals in the Initiation of Epithelial Wound Healing. *Nature Reviews Molecular Cell Biology* **2013**, *14* (4). <https://doi.org/10.1038/nrm3541>.
27. Tsukada, K.; Tokunaga, K.; Iwama, T.; Mishima, Y. The PH Changes of Pressure Ulcers Related to the Healing Process of Wounds. *Wounds* **1992**, *4* (1), 16–20.
28. Otsuka Pharmaceutical Co., Ltd. *What are electrolytes (ions)?* [otsuka.co.jp](https://www.otsuka.co.jp/en/nutraceutical/about/rehydrat/ion/water/electrolytes/). [https://www.otsuka.co.jp/en/nutraceutical/about/rehydrat ion/water/electrolytes/](https://www.otsuka.co.jp/en/nutraceutical/about/rehydrat/ion/water/electrolytes/).
29. Kathleen Deska Pagana; Timothy James Pagana; Theresa Noel Pagana. *Mosby's Diagnostic and Laboratory Test Reference*, 14th ed.; Elsevier: St. Louis, Missouri, 2019.
30. Goodwin, M. L.; Harris, J. E.; Hernández, A.; Gladden, L. B. Blood Lactate Measurements and Analysis during Exercise: A Guide for Clinicians. *Journal of Diabetes Science and Technology* **2007**, *1* (4). <https://doi.org/10.1177/193229680700100414>.
31. Marshall, A. J.; Blyth, J.; Davidson, C. A. B.; Lowe, C. R. PH-Sensitive Holographic Sensors. *Analytical Chemistry* **2003**, *75* (17). <https://doi.org/10.1021/ac020730k>.
32. Kuehni, R. G. On the Relationship between Wavelength and Perceived Hue. *Color Research & Application* **2011**, *37* (6). <https://doi.org/10.1002/col.20701>.

Effect of 2D thickness on the performance of 2D/3D organic-inorganic metal halide perovskite solar cells

Pavan Ayyar, Tom Parkinson, Matyas Daboczi, Salvador Eslava
Department of Chemical Engineering, Imperial College London, U.K.

Abstract Perovskite solar cells (PSCs) are an emerging technology in the field of renewable energy. One of the primary challenges encountered with organic-inorganic metal halide perovskites is their instability, especially when exposed to humid air or water. Layering a 2D perovskite on top of a thicker 3D perovskite crystal has been shown to improve moisture stability in addition to passivating defects in the crystal structure resulting in less recombination and higher performance. The primary objective of this project was to investigate the effects of the thickness of the 2D layer on the overall performance as both photoelectrochemical (PEC) and photovoltaic (PV) devices. This was done using a $(\text{FAPbI}_3)_{0.99}(\text{MAPbBr}_3)_{0.01}$ perovskite that underwent chemical dipping in an FEA1 solution. The results showed that there is an improvement in the performance up to the thickest 2D layer that was investigated, with power conversion efficiencies of up to 15.9% and onset potentials (V_{on}) of 0.614 V_{RHE} . The main conclusion is that dipping time of 30 seconds showed the best performance.

Introduction

The most urgent and defining crisis of the 21st century is climate change. With urban population growth and socio-economic development comes a surge in energy demand, resulting in the release of billions of tons of carbon dioxide from burning fossil fuels. Developing renewable energy technologies plays a crucial role in replacing fossil fuels and stabilising climate change. Solar power shows promise, owing to the large amount of sunlight that reaches the Earth's surface annually [1]. However, effective energy storage solutions must be employed to address its intermittent energy generation. One notable solution is photoelectrochemical (PEC) water-splitting, whereby semiconductor electrode materials drive the electrolysis of water to produce green hydrogen. Hydrogen stands in contrast to conventional lithium-ion batteries, which rely on materials with limited resources and may not be environmentally friendly.

In recent years, perovskite solar cells (PSC) have received much attention, with a lot of the research centred around 3D organic-inorganic metal halide (OIMH) perovskites. Due to their outstanding optoelectronic properties, these OIMH PSCs have achieved remarkable power conversion efficiencies (PCE) of up to 25.7% [2]. They also have enormous potential for commercialisation due to their solution-processibility at low temperatures, allowing for the production of PSCs through cost-effective and scalable methods.

Despite their rapid development in photovoltaics, they remain relatively unexplored in other solar applications, such as photoelectrochemistry. This is because OIMH perovskites are very susceptible to water degradation, significantly impacting performance [3]. These effects would only be exacerbated when in direct contact with water for PEC applications.

There is also the presence of defects within 3D perovskite and on its surface that encourage the

recombination of charge carriers, lowering performance.

A novel concept involves creating a 2D/3D perovskite structure composed of a thick 3D perovskite layer and a thin 2D perovskite layer. It has been demonstrated that 2D treatment of 3D perovskites can passivate defects at the surface and within the 3D structure [4]. The thin 2D layer also protects the device from water degradation, improving stability [5]. As such, 2D/3D devices could pave the way for efficient PEC water-splitting. However, research is still in its experimental stages, with much trial and error. One variable that needs to be investigated is the effect of the 2D layer thickness on device performance. While 2D treatment can enhance performance via passivation, if the 2D layer is too thick, there will instead be a drop-off in performance caused by a decrease in charge carrier mobility [6].

Hypothetically, there is an optimal 2D layer thickness that balances these opposing effects. Therefore, the aim of this project is to improve the performance of 2D/3D perovskites by studying the effects of 2D layer thickness on PEC and PV performance parameters in order to find an optimal thickness.

Background

3D Perovskites

3D perovskites have the general formula ABX_3 where (1) the A-site is occupied by a monocation such as caesium (Cs^+), methylammonium (MA^+), or formamidinium (FA^+), (2) the B-site is a divalent metal cation, usually lead (Pb^{2+}), or tin (Sn^{2+}), and (3) the X anion is often a halide ion, such as iodide (I^-) or bromide (Br^-).

OIMH perovskites in particular have shown exceptional performance in photovoltaic (PV) applications. They possess high absorption coefficients, making them highly efficient at generating electron-hole pairs (excitons) [7].

Additionally, they exhibit long carrier diffusion lengths, allowing photogenerated charge carriers to travel considerable distances without recombination [8]. Their low exciton binding energies facilitate efficient charge separation and minimize the likelihood of recombination [9,10]. Other than that, many of the constituent materials of perovskites devices are relatively abundant and inexpensive. The bandgap of OIMH perovskites can be tuned by varying the organic cation composition. Lead-free compositions also mitigate the environmental and health concerns of lead-based materials.

Nevertheless, when it comes to commercialisation, one of the major concerns of OIMH perovskite solar cells is the vulnerability of their organic constituents. FA, MA, and mixed FAMA configurations have all been shown to degrade when exposed to light, humidity, and elevated temperatures. For instance, MAPbI₃ PSCs show significant degradation when varying temperatures up to 85°C and humidity up to 80% [11]. On the other hand, FAPbI₃ has the potential for higher thermal stability while still improving performance due to its narrower bandgap and a higher absorption coefficient. Despite this, FAPbI₃ still suffers due to the phase transition from the photoactive α -perovskite polymorph to the yellow, photoinactive δ -phase under humid air, making water-based degradation one of the biggest challenges for OIMH perovskites [3,12].

There are also defects that occur throughout the 3D perovskite crystal structure, such as mismatches/dislocations between grain boundaries and ionic vacancies [13]. The surface is also highly defective, mainly because of the interruption in the crystal structure that causes dangling bonds (unpaired electrons).

These defects introduce electronic trap states within the bandgap of the perovskite, which can capture charge carriers, leading to non-radiative recombination—i.e., the electron-hole pairs generated from sunlight are not converted into electricity. A higher defect density would increase the recombination rate, leading to a decrease in voltage, photogenerated current, and hence power output.

2D Perovskites

The basic structure of 2D perovskites consists of alternating organic spacer layers and 3D perovskite layers [14]. The organic spacer layers are formed from large organic spacer ions and are intercalated in between each octahedral slab. The octahedral layer can be pure 2D ($n=1$) or quasi-2D ($n>1$) where n is the number of octahedral layers in each slab, representing thickness and dimensionality [15].

2D/3D Perovskites

One way of forming the 2D/3D perovskite is by dipping pre-synthesized 3D perovskite crystals into

a solution containing the organic spacer cations, allowing them to intercalate into the 3D structure [16]. The extent to which this occurs (2D layer thickness) depends on the dipping time.

The 2D layer formed protects the 3D perovskite from water degradation, improving stability [5]. This is due to the large organic cations that provide both steric hindrance and hydrophobic resistance, preventing water molecules from penetrating deep into the device.

This stability increase is not without its drawbacks. 2D perovskites tend to exhibit decreased charge carrier mobility, for instance. The intercalated organic cations act as insulating barrier hinder movement across the spacer layer.

Methods

Preparation of Substrates

To begin with, FTO-coated glass was cut into 25x27 mm rectangles and placed into staining dishes for cleaning. These staining dishes were filled with deionised water and a drop of detergent (Hellman) and sonicated for 10 minutes. After sonication, they were rinsed until there were no bubbles. The sonication process was repeated with acetone and finally isopropanol before the substrates were dried with compressed air.

Electron Transport Layer

The electron transport layer (ETL) included two layers: a thin compact TiO₂ layer (c-TiO₂) and a thicker mesoporous TiO₂ layer (m-TiO₂).

The c-TiO₂ precursor solution was produced by first adding 284 mg of titanium isopropoxide (Sigma-Aldrich), which is kept in a glovebox to prevent oxidation, and 105 mg of diethanol amine (Alfar-Aesar) into 2 mL of ethanol (Fisher Science), which was then mixed for 20 minutes, filtered and used within a day. Before spin coating, the top 2 mm of the substrate was covered with CAPTON tape to allow electrical contact with the FTO during testing. 65 μ L of the precursor solution was dropped onto the substrate before it was spin-coated at 7000 rpm with an acceleration of 7000 rpm/s for 30 seconds. It was ensured that the precursor was mixed between each spin coat to prevent comets [17]. This layer was annealed starting at 100°C, then increasing the temperature by 100°C every 5 minutes until it reached 500°C, at which point it was left to anneal for 2 hours.

The precursor solution for the m-TiO₂ was made by mixing 300 mg of 30 N-RD Dyesol TiO₂ paste (30 NR-D, Greatcell Solar) with 2 mL ethanol. This solution was stirred overnight to ensure a uniform concentration of nanocrystals. Before spin-coating, 2 mm of the substrate was once again covered by the CAPTON tape and 65 μ L was dropped on before being spin-coated at 4000 rpm with an acceleration of 2000 rpm/s for 10 seconds. This layer was annealed using the same temperature sintering as in

the c-TiO₂ layer, however, instead of 2 hours of annealing at the end, the m-TiO₂ layer was annealed for 30 minutes.

A further lithium-based dopant was applied to the ETL. This dopant was produced by adding 10 mg of bis(trifluoromethane)sulfonimide lithium salt (LiTFSI) to 1 mL of acetonitrile [18]. This was then spin-coated onto the m-TiO₂ at 3000 rpm with 1000 acceleration for 10 seconds. The annealing process for this is identical to the m-TiO₂ layer.

FAMA Perovskite layer

A 1.81 M FAMA precursor solution was made using two precursor solutions, all within a glovebox. The first solution (FAPbI₃ with MACl) consists of 384 mg of PbI₂ (TCI), 567 mg of FAI (GreatCell Solar) and 78.3 mg of MACl dissolved in 2 mL of an 8:1 v/v DMF:DMSO (dimethylformamide:dimethylsulfonimide) solution. The second solution (MAPbBr₃) consists of 122 mg of MABr (GreatCell Solar) and 367 mg of PbBr₂ (TCI) dissolved in 1 mL of the same 8:1 DMF:DMSO solution. Both solutions were vigorously stirred for 1 hour before 28 µL of the MAPbBr₃ solution was added to the FAPbI₃ solution and the resulting mixture was filtered.

70 µL of the FAMA precursor was distributed across the substrates and was then spin-coated with the method shown in Table 1 [16].

In order to get a uniform crystalline layer, the antisolvent deposition was optimised to remove the solvent. The antisolvent chosen was chlorobenzene (CB), which was applied via an automatic solvent dispenser built into the spin coater. A syringe pressure of 0.7 psi dispensing for 0.1 seconds was found to produce a continuous drop without any defects in the perovskite.

Table 1 – Program used to disperse antisolvent and spin-coat FAMA perovskite

Step	Duration	rpm	Acceleration
1	10s	5000	2000
2	0.1s (antisolvent dispensed)	5000	2000
3	30s	5000	2000

These substrates were then annealed at 150°C for 15 minutes.

To prevent defects within the perovskite, a solvent filter regeneration was done overnight. Furthermore, during the spin-coating, the glovebox was purged every 2 spin-coats, and 2-minute breaks were taken between each spin-coat during which a handheld battery-powered fan was used to evaporate any leftover solvents and allow the circulation reactor to remove them from the system. This prevented any crystallisation of the perovskite prior to spin coating because of the ambient chlorobenzene within the confined space of the glovebox.

2D Perovskite Layer

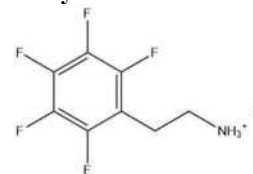


Figure 1: 2-(perfluorophenyl)-ethanaminium iodide (FEAI)

To create the 2D/3D layer, a chemical bath of 2-(perfluorophenyl)-ethanaminium iodide (FEAI) (shown in Figure 1) dissolved in isopropanol was used [5]. This was made by dissolving 152 mg of the FEAI (synthesized in-house) in 15 mL of isopropanol to create a 30 mM solution before being left to stir overnight.

As the independent variable of this experiment, a range of 2D layer thicknesses were required. This was achieved by varying the length of time that the chemical bath was applied to the devices within each batch. The 3D FAMA perovskites were dipped for 5 seconds, 10 seconds and 30 seconds. Some of the substrates were left as 3D perovskite as a reference.

To dry the substrate of any leftover FEAI solution, the substrates were spin-coated at 2000 rpm with acceleration of 2000 rpm/s for 20 s before being annealed at 120°C for 10 minutes.

Hole Transport Layer

The chosen HTL was oxygen doped Spiro-MeTAD. The oxygen doped Spiro-OMeTAD solution consists of two precursor solutions. The first solution requires 72.3 mg of Spiro-OMeTAD (Luminescence technologies) to be dissolved in 1 mL of acetonitrile, before being doped with 35 µL of 4-tertbutylpyridine. The second solution consists of 13 mg of LiTFSI dissolved in 250 µL of acetonitrile. Both solutions are stirred before 25 µL of the LiTFSI solution is added to the Spiro-OMeTAD solution. This was then left stirring overnight. The deposition of the layer was done by spreading 30 µL onto the substrate, before spin-coating at 5000 rpm at 5000 rpm/s for 30 seconds, with no annealing required. These substrates were then placed in a desiccator outside of the glovebox to allow exposure to oxygen for doping.

Gold Layer

A thin 80 nm gold layer was plated on top of the 2D layer by thermal evaporation.

Functionalisation of Graphite

Two graphite layers were used to protect the perovskite. The first being a 30 µm (G3) thick layer applied directly onto the gold, followed by a 150 µm (G15) thick layer which was functionalised by a NiFeOOH catalyst by electrodepositing onto the surface [19]. Both graphite layers had an adhesive layer to make application easier. The electrodeposition was completed by preparing a

solution of 139 mg of $\text{FeSO}_4 \cdot 7\text{H}_2\text{O}$ (Sigma-Aldrich) and 525 mg of $\text{NiSO}_4 \cdot 7\text{H}_2\text{O}$ (Sigma-Aldrich) dissolved in 50 mL of deionised water. This was then bubbled through with nitrogen for 20 minutes to remove any dissolved oxygen as well as mix it thoroughly. The G15 graphite sheets were cut into 3x3 cm squares, and the bottom 2 cm were marked for deposition. The marked area was then held in the solution while a linear sweep from 1 to 0.4 V(Ag/AgCl) was performed at a scan rate of 50 mV/s. The excess solution was washed off with deionised water, and they were allowed to dry.

Preparation for Testing

For the devices to be tested, electrical contact to the FTO must be established and any potential holes should be removed. This was done by first scratching off the top 2 mm of the device that were covered during the ETL deposition with a scalpel until the FTO was exposed, before scratching off the perovskite around the edges of the device, resulting in a square in the centre. The resistance between the gold layer and the FTO was then measured to ensure the device was acting as a semiconductor with a resistance in the $\text{k}\Omega$ - $\text{M}\Omega$ range. If not, a visual inspection of the device was performed, and any holes and visual imperfections identified were scratched off. Once a reasonable resistance was measured, the G3 layer could be cut and applied directly onto the gold surface, making sure not to accidentally remove any gold with the adhesive. The functionalised G15 sheet was then cut and applied on top of the G3 layer, although this was not required for PV testing.

Photovoltaic (PV) Testing

In order to prevent water degradation of the perovskite, photovoltaic testing was performed first. The device was treated as a two-electrode setup, with the working electrode wire connected to the FTO and the reference electrode wire connected to the G3. A linear sweep was then performed from -0.2 to 1.2 V at a scan rate of 10 and 50 mV/s with a further reverse scan at 50 mV/s. The power density of the light for both PEC and PV was 100 mW/cm^2 .

Photoelectrochemical (PEC) Testing

A three-electrode setup was chosen to test the photoelectrochemical performance of the devices. This consisted of a platinum counter electrode, an Ag/AgCl reference electrode saturated in 3.5 M KCl and a pH 14 aqueous NaOH supporting electrolyte. The electrolyte was made by dissolving 20 g of NaOH in 500 mL of deionised water and testing with a calibrated pH probe. Linear sweep voltammetry was then performed between -0.4 to 1.2 V with a scan rate of 50 mV/s. All PEC data was converted from Ag/AgCl to V_{RHE} using the Nernst equation:

$$V_{\text{RHE}} = V_{\text{Ag/AgCl}} + 0.0592 \cdot \text{pH} + 0.1976.$$

Stability measurements were performed at an applied bias of 1.23 V and the measurements were again converted to V_{RHE} . A stirrer bar was added to disrupt bubbles forming and keep a uniform electrolyte concentration.

All structural and energy characterisation was done on samples of structure: FTO/c-TiO₂/m-TiO₂/FAMA/2D layer.

Scanning Electron Microscope (SEM)

For characterisation of the device structure, SEM (Zeiss) imaging was used at 5 kV accelerating voltage to measure samples of different 2D layer thickness.

UV-Vis Spectroscopy

UV-Vis (Shimadzu) absorbance spectra were obtained on the different thickness 2D films.

Photoluminescence (PL)

Steady state PL measurements were obtained on the samples with an FLS1000 'Edinburgh Instruments' photoluminescence spectrometer, which used a 450 W xenon arc lamp with a 405 nm excitation wavelength. Time correlated single photon counting (TCSPC) measurements were also obtained using a long-pass filter with a 45.5 nm cutoff and a 405 nm pulsed diode laser.

Energy Level Measurements

Ambient pressure photoemission spectroscopy (APS) was done on the samples with UV light energies of 4.8-6.0 eV. Fermi levels were measured using a vibrating tip Kelvin probe.

X-ray Diffraction (XRD)

X-ray diffraction was performed on the samples with an automatic 'Malvern Panalytical' machine.

Results and Discussion

Photoelectrochemical Performance

To investigate the impact of the 2D layer and its thickness, the 2D/3D devices were configured as photoanodes for PEC water-splitting. Key parameters to assess performance can be extracted from linear sweep voltammetry J-V curves. These include the onset potential, V_{on} , and the photocurrent produced at 1.23 V_{RHE} , J_{ph} , which are used to calculate the fill factor, FF.

The onset potential represents the voltage at which the photoanode starts to generate a photocurrent under illumination. On its own, the devices do not generate enough voltage to drive the oxygen evolution reaction (OER) at 1.23 V_{RHE} . Thus, additional voltage must be supplied to start the reaction. Achieving a lower onset potential is desirable in PEC water-splitting because it reduces the energy input required to drive the OER. The J_{ph} represents the rate of reaction when the additional voltage supplied is equal to 1.23 V_{RHE} , so a higher

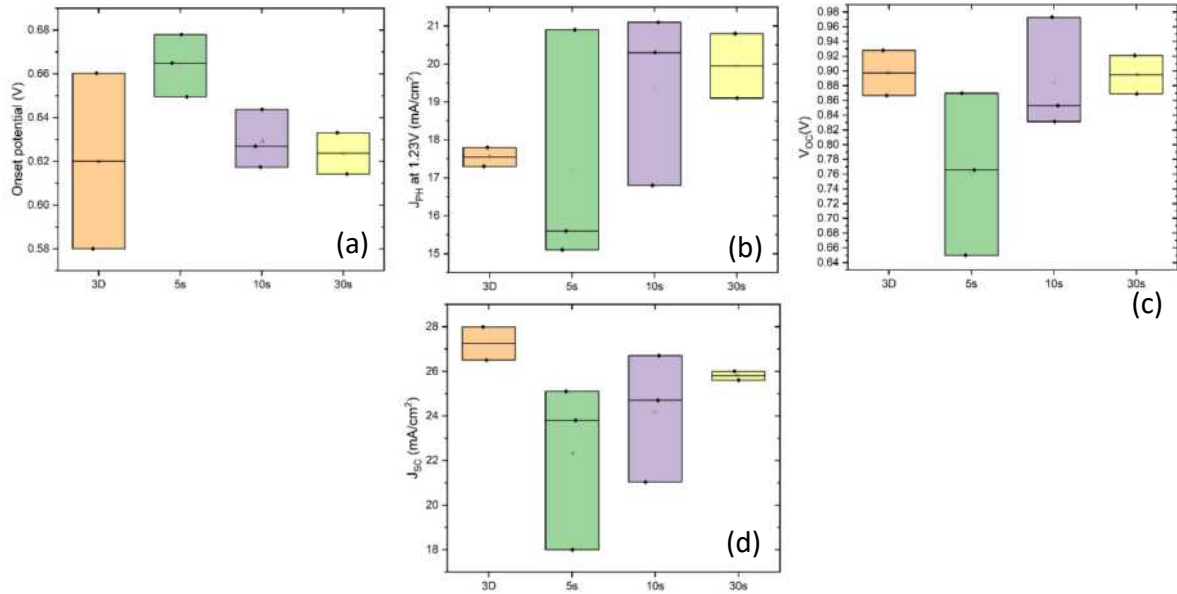


Figure 2 – Box plots of PEC and PV measurements of 3D FAMA and 2D/3D FAMA+FEAI. (a-b) Effect of 2D treatment time on the V_{on} and J_{ph} , (c-d) Effect of 2D treatment time on V_{oc} and J_{sc}

J_{ph} would indicate a better performing device. The FF represents the squareness of the J-V curve, correlating with the efficiency of the device.

The V_{on} and J_{ph} was recorded for each working device and any data points that were more than one standard deviation from the median were determined to be anomalies. The statistical analysis of the onset potential is shown in Figure 2.a and the mean onset potentials and photocurrents are summarised in Table 2.

Table 2 – Mean values of PEC parameters measured from 3D and 2D/3D PSCs

PSC	$V_{on}(V_{RHE})$	$J_{ph}(mA/cm^2)$
3D	0.6201	17.61
2D/3D (5s)	0.6641	17.19
2D/3D (10s)	0.6293	19.41
2D/3D (30s)	0.6236	19.97

The 3D perovskite had the lowest mean onset potential of 0.620 V, meaning that it had the best average performance compared to all the 2D treated devices. However, this is unlikely to be the case based on the literature surrounding 2D/3D devices. For example, under 10 seconds of 2D treatment, it was demonstrated that a similar composition of 3D FAMA benefited from 10 seconds of 2D treatment, with an increase in PCE of 1.54% [5]. Furthermore, there were only two data points for the 3D reference devices, and it had a significant range of 0.080 V_{RHE} . This suggests that the onset potential of 0.580 V_{RHE} is most likely anomalous as it is far lower than all the other devices tested. Therefore, more data needs to be collected for the 3D perovskite devices to accurately assess the effect of 2D treatment relative to the 3D perovskite.

Nonetheless, when focusing on the average performance of the 2D devices, there is a far clearer

trend produced. The mean V_{on} falls by 0.036 V when changing the dipping time from 5 to 10 seconds. Whereas from 10 to 30 seconds, there is a relatively smaller decrease of 0.006 V in the mean V_{on} .

The overall improvement shown from increasing the 2D treatment time can be attributed to the passivation of defects which improves with 2D layer thickness. As the defect density decreases, the rate of non-radiative recombination falls, meaning that less energy is lost from the solar cell. Instead, more electron-hole pairs contribute to the voltage and photocurrent, leading to improved performance. This is especially clear from 5 to 10 seconds. The smaller increase in performance from 10 to 30 seconds, however, reveals the diminishing improvement associated with a longer 2D treatment. After 10 seconds of 2D treatment, the 2D layer is beginning to grow too thick, leading to decreased charge mobility and higher rates of recombination, counteracting the benefits derived from 2D treatment passivation.

It can be speculated that, since the increase in performance at 30 seconds was very small, the optimum dipping time would be somewhere around 30 seconds. Acquiring more data at 20 seconds and 40 seconds may confirm this.

For the J_{ph} , as shown in Figure 2.b and Table 2, the mean of the 2D treated devices follows a similar trend to the V_{on} with an increase of 2.22 mA/cm^2 from 5 to 10 seconds and a smaller increase of 0.56 mA/cm^2 from 10 to 30 seconds. However, the trend is less likely to be accurate because of the wide spread of data for the 5 and 10 second 2D treated devices.

This could be due to the functionalised graphite layer. Some PEC measurements were conducted a day after functionalising the G15 graphite, which

may have caused the catalyst to degrade. Moreover, the graphite layers are not of the highest quality as they were meant for commercial use, which could have introduced random errors that could be eliminated if more data was collected. As such, whether the 2D treatment had any effect on the photocurrent generated is inconclusive.

Photovoltaic Performance

The devices were also configured as photovoltaic solar cells and J-V curves were produced to determine the performance of the devices. The parameters used to assess performance are the open-circuit voltage, V_{oc} , and the short-circuit current, J_{sc} . The V_{oc} is defined as the voltage across the terminals of a solar cell when there is no external load connected to it, representing the maximum voltage that the solar cell can produce. The J_{sc} is the current that flows through a solar cell when its output terminals are short-circuited. It represents the maximum photocurrent that the device can produce. Both a higher V_{oc} and J_{sc} is indicative of better performance.

Box plots were also produced for each of these parameters as shown in Figure 2.c and 2.d and the mean values are presented in Table 3.

Table 3 – Mean values of PV parameters measured from 3D and 2D/3D PSCs

PSC	V_{oc} (V)	J_{sc} (mA/cm ²)
3D	0.8973	27.23
2D/3D (5s)	0.7618	22.30
2D/3D (10s)	0.8858	24.15
2D/3D (30s)	0.8949	25.80

Note that the same devices were used to produce both the PEC data and the PV data, to ensure a fair comparison. Unfortunately, this meant that the 3D perovskite devices were anomalous for the PV data as well and only the trend of the 2D treated devices could be analysed.

For these devices, the V_{oc} and V_{on} should follow a similar trend, which is the case. Again, the V_{oc} improves substantially by 0.124 V from 5 to 10 seconds, but only 0.009 V from 10 to 30 seconds. Therefore, the data supports the original hypothesis that there is a balance between the passivating effect of defects and the reduced charge carrier mobility with a thicker 2D layer. The optimum dipping time should also be approximately 30 seconds, but more data is needed to confirm this.

The J_{sc} box plot is shown in Figure 2.d. The results also show that a longer 2D treatment time increases the J_{sc} of the device by 3.50 mA/cm² from 5 to 30 seconds. Again, the trend does not correspond to V_{oc} and V_{on} as the photocurrent increases somewhat linearly. While the functionalised G15 graphite layer was not used for PV measurements, the G3 layer could have still affected performance, leading to the

wider range in data shown for the 5 and 10 seconds 2D treatments.

Other characteristics measured by the PV measurements include the PCE and the FF. The best FF was shown to be 0.698 and the best PCE was 15.91%. Both the FF and the PCE are significantly lower than results in literature with similar device structures, indicating that this experiment could be far more impactful with a larger dataset that's closer to the optimum.

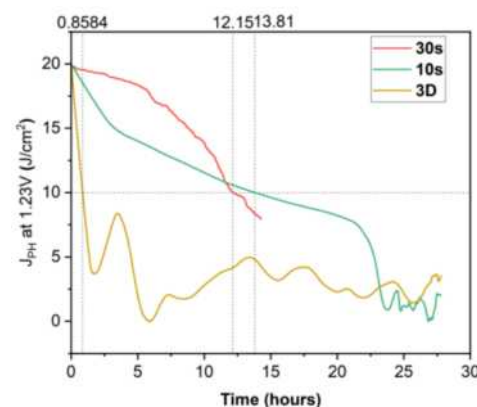


Figure 3 – Normalised stability data for the best samples tested, with the t_{50} indicated.

Stability

As a secondary objective for the experiment, the effect of the 2D layer on stability was investigated. These results, shown in Figure 3, were primarily characterised by the t_{50} , which is the time taken for the current to reach half of the initial current. Despite there being a minor trend, the best result for the t_{50} on a 3D sample was less than one hour, which is significantly lower than similar cells in literature, which reached up to 83 hours in some cases [17]. The 2D samples were slightly better, with 12.15 hours being the best for the 30s sample and 13.81 hours for the 10s sample. Due to the large disparity between this data and the data obtained from literature, a trend cannot be established confidently.

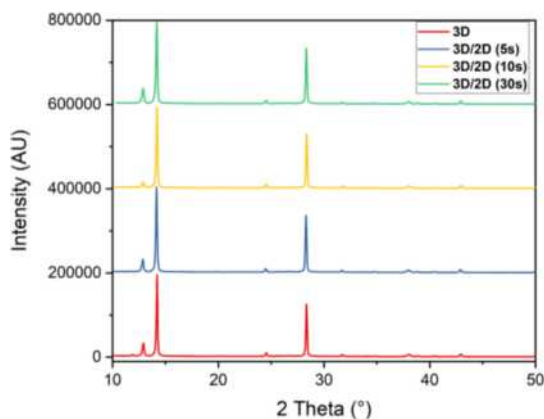


Figure 4 – X-ray diffraction data, showing diffraction of perovskites with different dipping times

Characterisation

In order to analyse the phase composition and crystal structure of the 2D/3D layers, X-ray diffraction (XRD) was performed on a representative sample of devices, as shown in Figure 4. The primary difference found between them is the far smaller peak for the 10 second 2D treated sample at 13 degrees. This is likely to be a peak for PbI_2 , which is a degradation product of the FAPI perovskite, indicating that the 10 second sample degraded less than the others. Disregarding this difference, the spectra are almost identical, with similar large peaks, suggesting a uniform perovskite layer.

This trend is further confirmed in the SEM images, shown in Figure 5, zoomed in with a magnification factor of 50000. All the samples show small imperfections in the grains except for the 10 second sample, in which almost no visible imperfections are present, indicating the presence of degradation products in the other results.

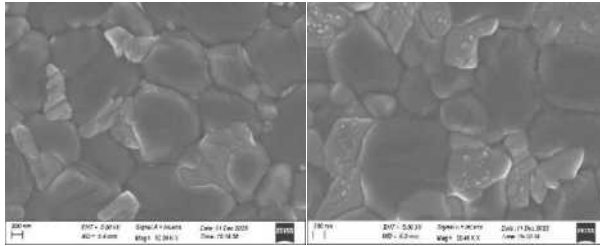


Figure 5 – SEM microscopy of 10s sample (left) and 5s sample (right)

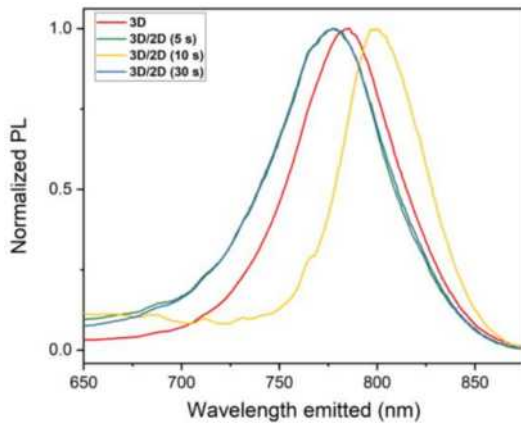


Figure 6 – Steady state photoluminescence showing the optical bandgap of perovskites with different dipping times

To measure the dynamics of charge carriers within the perovskites, steady state and time resolved photoluminescence were performed to measure the light emitted by radiative recombination. This is generally slightly lower energy than the optical bandgap of the material [20]. A clear trend is obtained through the steady state PL on Figure 6, whereby the 2D samples show a larger bandgap than the 3D perovskite, with the 30 second and 5 second sample showing a peak wavelength of 778 nm (1.60 eV) and the 3D reference showing a peak

wavelength of 786 nm (1.58 eV). However, the 10 second sample shows a smaller bandgap than either of the other 2D samples or the reference 3D sample, as well as a narrower full width half maximum (FWHM), which suggests it is closer to an ideal perovskite [21]. This further confirms that the 10 second sample has degraded less than the other 2D samples and it is therefore reasonable to predict that it would follow the trend of the other 2D samples if it had been degraded to the same extent.

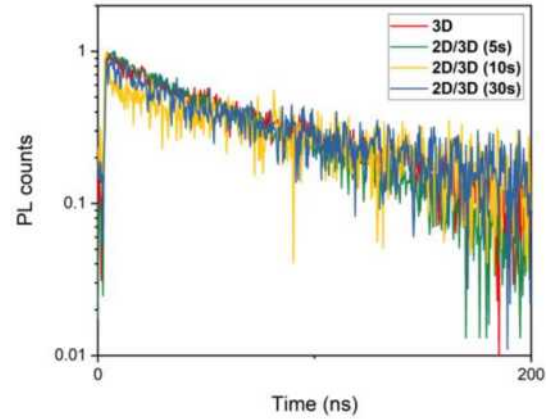


Figure 7 – Time correlated single photon counting (TCSPC) showing the decay of charges in perovskites with different dipping times

As shown in Figure 7, TCSPC was performed on the samples to determine the rate of charge carrier extraction as well as the charge carrier mobility. Literature would suggest that the 2D layers would show lower charge carrier mobility and poorer extraction [6]. All samples showed steep decay rates, indicating effective charge carrier extraction [22]. The 10 second sample had the steepest initial decay rate, however, over time it became less effective and the 5 second sample and the 3D reference became the most effective. Due to the large variation in performance of each sample over the period, no significant trend can be determined.

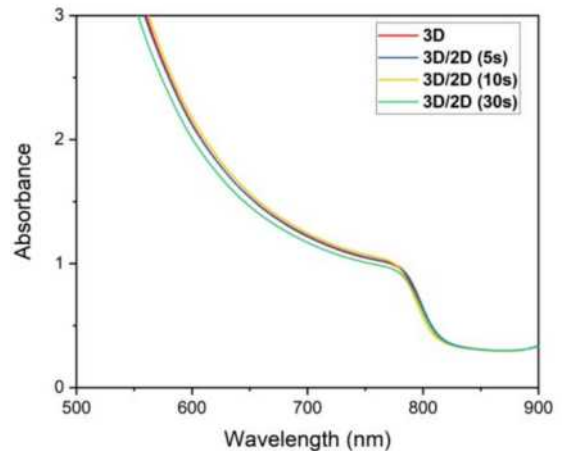


Figure 8 – UV-Vis spectrometry of perovskite films with different dipping times

UV-vis spectroscopy was used to measure the absorbance of the samples to light in the UV and visible range; this is used primarily to characterise the bulk of the perovskite. As shown in Figure 8, the 30 second sample shows slightly lower absorbance across the range displayed and the 10 second shows slightly higher absorbance than the 3D reference and the 5 second sample. As this trend is not consistent with the other trends obtained through characterisation, it is likely that this is a result of random variations in the perovskite that can impact the absorption coefficients, such as film thickness [23]. As such, it can be inferred that there is little to no trend and that the bulk of the perovskite is largely unaffected by the presence of the thin 2D layer.

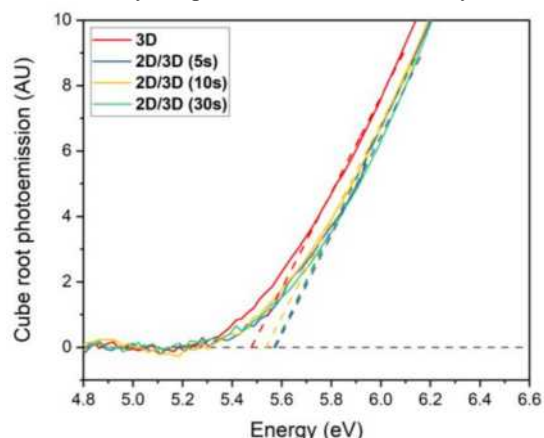


Figure 9 - Ambient pressure photoelectron spectrometry (APS) displaying the valence band edge (E_v) of perovskite films with different dipping times

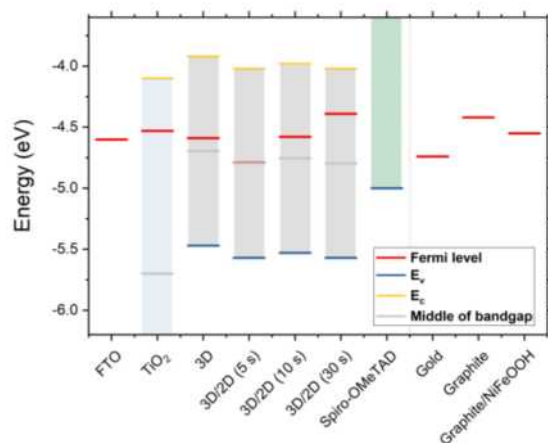


Figure 10 - Energy level diagram of each layer, derived from APS

The final characteristic measured was the energy levels of each layer. This is primarily done through the valence band edge (E_v), which was measured using ambient pressure photoelectron spectrometry (APS), shown on Figure 9. The E_v showed a decrease when 2D treated although 10 second treatment was once again anomalous, showing an E_v in between the 3D reference and the other 2D samples. The decrease in E_v for the 2D layer implies a slight reduction in charge carrier mobility, as it's

energetically favourable for electrons to go down in energy levels, which is consistent with literature and the initial prediction of the 2D layer's properties [5].

As shown in Figure 10, Fermi level (E_F) data was also obtained and showed that as the 2D layer increases in thickness, E_F gets close to the conduction band, which is indicative of a lower charge carrier mobility. This further confirms effects of the 2D layer on the optoelectronic properties of the perovskite.

Owing to the nature of the characterisation techniques requiring the devices to be only built to the 2D layer, the devices can't be tested afterwards. As such, any anomalous behaviour is unable to be compared to linear sweeps on the same samples. As shown in the 10 second sample, there is a lot of variability in between perovskites that should be similar and as such, large datasets would ideally be obtained to offset the presence of random or systematic errors within each set of datapoints. Furthermore, the batch system that was used in this experiment meant that each batch could have systematic errors in different parts of the devices. This means that the devices that were explored in PEC and PV likely would have exhibited a similar quality distribution as the samples that were characterised and may have been imperfect or degraded at the time of testing, resulting in inconsistent datasets that make it challenging to identify trends.

Conclusion

In conclusion, this study revealed a trend based on differing the thickness of a 2D layer formed on a FAMA-based perovskite by varying the dipping time in a FEAI + IPA solution. It was shown that increasing the 2D layer thickness via the dipping time achieves an improvement in the V_{on} and the V_{oc} but with diminishing improvement. For this particular 2D/3D material system and 2D treatment process, it was revealed that the optimum dipping time was approximately 30 seconds, because of the minimal performance gains from increasing the dipping time from 10 to 30 seconds. This confirms the idea that there is a clear trade-off between the passivating effects and the decreased charge mobility with increasing 2D layer thickness.

However, more data needs to be collected on the same and different 2D treatment spans, as well as the reference 3D perovskite to improve the quality of the data. The stability of the devices would be an interesting insight as well, as the purpose of the 2D/3D perovskite is to highlight the increased resistance to water degradation when being used as a photoanode. However, due to time constraints it was difficult to show this expected improvement. Additionally, other types of 2D layers such as 3F-PEA could be explored to verify whether they follow the same trend as the FEAI.

Acknowledgement

We'd like to express our gratitude to the Eslava group for their support throughout this project. In particular, we'd like to thank Matyas Daboczi as well as Salvador Eslava for providing us guidance throughout the experimental process and carrying out characterisation.

References

1. Hurley, S. (2019). Solar energy. [online] Explaining Science. Available at: <https://explainingscience.org/2019/03/09/solar-energy/>.
2. National Renewable Energy Laboratory (2023). Best Research-Cell Efficiency Chart | Photovoltaic Research | NREL. [online] Nrel.gov. Available at: <https://www.nrel.gov/pv/cell-efficiency.html>.
3. Binek, A., Hanusch, F.C., Docampo, P. and Bein, T. (2015). Stabilization of the Trigonal High-Temperature Phase of Formamidinium Lead Iodide. *The Journal of Physical Chemistry Letters*, 6(7), pp.1249–1253. doi:<https://doi.org/10.1021/acs.jpclett.5b00380>.
4. Lin, T., Dai, T. and Li, X. (2023). 2D/3D Perovskite: A Step toward Commercialization of Perovskite Solar Cells. *Solar RRL*, p.2201138. doi:<https://doi.org/10.1002/solr.202201138>.
5. Liu, Y., Akin, S., Pan, L., Uchida, R., Arora, N., Milić, J.V., Hinderhofer, A., Schreiber, F., Uhl, A.R., Zakeeruddin, S.M., Hagfeldt, A., Dar, M.I. and Grätzel, M. (2019). Ultrahydrophobic 3D/2D fluoroarene bilayer-based water-resistant perovskite solar cells with efficiencies exceeding 22%. *Science Advances*, 5(6), p.eaaw2543. doi:<https://doi.org/10.1126/sciadv.aaw2543>.
6. Milot, R.L., Sutton, R.J., Eperon, G.E., Haghighirad, A.A., Martinez Hardigree, J., Miranda, L., Snaith, H.J., Johnston, M.B. and Herz, L.M. (2016). Charge-Carrier Dynamics in 2D Hybrid Metal–Halide Perovskites. *Nano Letters*, 16(11), pp.7001–7007. doi:<https://doi.org/10.1021/acs.nanolett.6b03114>.
7. Sun, S., Salim, T., Mathews, N., Duchamp, M., Boothroyd, C., Xing, G., Sum, T.C. and Lam, Y.M. (2014). The origin of high efficiency in low-temperature solution-processable bilayer organometal halide hybrid solar cells. *Energy Environ. Sci.*, 7(1), pp.399–407. doi:<https://doi.org/10.1039/c3ee43161d>.
8. Stranks, S.D., Eperon, G.E., Grancini, G., Menelaou, C., Alcocer, M.J.P., Leijtens, T., Herz, L.M., Petrozza, A. and Snaith, H.J. (2013). Electron-Hole Diffusion Lengths Exceeding 1 Micrometer in an Organometal Trihalide Perovskite Absorber. *Science*, 342(6156), pp.341–344. doi:<https://doi.org/10.1126/science.1243982>.
9. D’Innocenzo, V., Grancini, G., Alcocer, M.J.P., Kandada, A.R.S., Stranks, S.D., Lee, M.M., Lanzani, G., Snaith, H.J. and Petrozza, A. (2014). Excitons versus free charges in organo-lead tri-halide perovskites. *Nature Communications*, [online] 5(1). doi:<https://doi.org/10.1038/ncomms4586>.
10. Miyata, A., Mitiglu, A., Plochocka, P., Portugall, O., Wang, J.T.-W., Stranks, S.D., Snaith, H.J. and Nicholas, R.J. (2015). Direct measurement of the exciton binding energy and effective masses for charge carriers in organic–inorganic tri-halide perovskites. *Nature Physics*, 11(7), pp.582–587. doi:<https://doi.org/10.1038/nphys3357>.
11. Han, Y., Meyer, S., Dkhissi, Y., Weber, K., Pringle, J.M., Bach, U., Spiccia, L. and Cheng, Y.-B. (2015). Degradation observations of encapsulated planar CH₃NH₃PbI₃ perovskite solar cells at high temperatures and humidity. *Journal of Materials Chemistry A*, [online] 3(15), pp.8139–8147. doi:<https://doi.org/10.1039/C5TA00358J>.
12. Zheng, X., Wu, C., Jha, S.K., Li, Z., Zhu, K. and Priya, S. (2016). Improved Phase Stability of Formamidinium Lead Triiodide Perovskite by Strain Relaxation. *ACS Energy Letters*, 1(5), pp.1014–1020. doi:<https://doi.org/10.1021/acsenergylett.6b00457>.
13. Kumar, A., Gupta, S.K., Dharamaniya, B.P., Pathak, S.K. and Karak, S. (2023). Understanding the origin of defect states, their nature, and effects on metal halide perovskite solar cells. *Materials Today Energy*, [online] 37, p.101400. doi:<https://doi.org/10.1016/j.mtener.2023.101400>.
14. Mao, L., Stoumpos, C.C. and Kanatzidis, M.G. (2018). Two-Dimensional Hybrid Halide Perovskites: Principles and Promises. *Journal of the American Chemical Society*, 141(3), pp.1171–1190. doi:<https://doi.org/10.1021/jacs.8b10851>.

15. Guo Qing Wu, Liang, R., Zhang, Z., Ge, M., Xing, G. and Sun, G. (2021). 2D Hybrid Halide Perovskites: Structure, Properties, and Applications in Solar Cells. *Small*, 17(43), pp.2103514–2103514.
doi:<https://doi.org/10.1002/sml.202103514>
16. Zhang, F., Kim, D.H. and Zhu, K. (2018). 3D/2D multidimensional perovskites: Balance of high performance and stability for perovskite solar cells. *Current Opinion in Electrochemistry*, [online] 11, pp.105–113.
doi:<https://doi.org/10.1016/j.coelec.2018.10.001>
17. Sodhi, N, Automatic Antisolvent Dispense for Pinhole Free OrganicInorganic Halide Perovskite Photoanodes Achieving Days Long Stability and 2D/3D Perovskite Photoanodes
18. Salibara, M., Correa-Baena, J.-P., Wolff, C. M., Stolterfoht, M., Phung, N., Albrecht, S., Neher, D., & Abate, A. (2018). How to make over 20% efficient perovskite solar cells in regular (N–I–p) and inverted (P–I–n) architectures. *Chemistry of Materials*, 30(13), 4193–4201.
<https://doi.org/10.1021/acs.chemmater.8b00136>
19. Daboczi, M., Cui, J., Temerov, F., & Eslava, S. (2023). Scalable All-inorganic halide perovskite photoanodes with >100 h operational stability containing earth-abundant materials. *Advanced Materials*, 35(45).
<https://doi.org/10.1002/adma.202304350>
20. Bergman, L., Chen, X.-B., Morrison, J. L., Huso, J., & Purdy, A. P. (2004). Photoluminescence dynamics in ensembles of wide-band-gap nanocrystallites and powders. *Journal of Applied Physics*, 96(1), 675–682.
<https://doi.org/10.1063/1.1759076>
21. Yoon, Y. J., Shin, Y. S., Park, C. B., Son, J. G., Kim, J. W., Kim, H. S., Lee, W., Heo, J., Kim, G.-H., & Kim, J. Y. (2020). Origin of the luminescence spectra width in perovskite nanocrystals with surface passivation. *Nanoscale*, 12(42), 21695–21702.
<https://doi.org/10.1039/d0nr04757k>
22. Péan, E. V., Dimitrov, S., De Castro, C. S., & Davies, M. L. (2020). Interpreting time-resolved photoluminescence of Perovskite Materials. *Physical Chemistry Chemical Physics*, 22(48), 28345–28358.
<https://doi.org/10.1039/d0cp04950f>
23. Rai, M., Wong, L. H., & Etgar, L. (2020). Effect of perovskite thickness on electroluminescence and solar cell conversion efficiency. *The Journal of Physical Chemistry Letters*, 11(19), 8189–8194.
<https://doi.org/10.1021/acs.jpcclett.0c02363>

CO₂ Capture Using Adsorption: an Outreach Project

Maria Pakradouni and Mikaela Zafet

Department of Chemical Engineering, Imperial College London, U.K.

Abstract

As carbon emissions are rising at an alarming rate, CO₂ capture through solid adsorption can be an important technology to our journey to net-zero. However, despite the increased spotlight that this technology has been receiving, the general public is not familiar with it. The goal of this project is to educate the general public on this topic. Here we create an outreach video on CO₂ capture using adsorption, by using an experimental set-up initially designed by Dr. Petit's research group. In the video, we aim to demonstrate how CO₂ adsorption works, explain the science behind it and highlight its potential in reducing CO₂ emissions. In this paper, the material selection process for the material used in the demonstration is analysed, taking into consideration the constraints of our experimental set-up. To make the video impactful and to demonstrate the power of adsorption to the public, the selection of an appropriate material is crucial. The adsorbent selected had to show high CO₂ uptake and fast adsorption kinetics, while being available in the lab at large enough quantities for the demonstration. This involved extensive screening of materials, modelling the datapoints from their adsorption isotherms, validating the adsorption behaviour of the materials at our disposal and finally testing the selected materials in our experimental set-up and operating conditions. The material selected for the outreach demonstration is Zeolite 13X.

1. Introduction

Anthropogenic emissions are currently responsible for the alarming concentration of greenhouse gases (GHGs) in the atmosphere. These contribute greatly to global warming and are some of the main causes of climate change. Namely, the energy sector is responsible for approximately 75% of GHG emissions, which continue to increase due to the increased world energy demand, leading to rising global temperatures (IEA, 2023). To control global warming and limit the temperature rise to 2°C, as decided in the Paris Agreement in 2015, global emissions need to be reduced by 45% by 2030 and reach net zero by 2050. Achieving this requires governments to commit to emission reduction policies and take active steps to invest in technologies that can remove GHG emissions from the atmosphere. (United Nations, 2023)

Carbon dioxide (CO₂) is one of the most concentrated anthropogenic greenhouse gases in the atmosphere, owing to the extensive emissions by the utilities, manufacturing and energy sectors, mainly through the production of oil, gas, cement and steel. The burning of fossil fuels for power generation and the manufacturing of goods resulted in the emission of 36.8 Gtonnes of carbon dioxide in 2022 (IEA, 2023). Despite efforts to make these processes more environmentally friendly, the reliance on fossil fuels emits CO₂. To counteract these emissions, carbon capture and storage (CCS) or utilisation (CCU) projects are explored, aiming to capture the CO₂ from the emission sources and inject it in the subsurface or reuse it in other processes. There are many technologies available to achieve that, which rely on the separation of CO₂ from the remaining stream gases - nitrogen, methane and hydrogen. Solid adsorption is a promising technology for carbon capture

that researchers have been working on increasingly to apply it on larger scales.

The objective of this report is to demonstrate the importance of CO₂ capture using solid adsorption to raise awareness to the public on its potential to reduce CO₂ emissions. This is done by creating an outreach video and recording a CO₂ adsorption demonstration on an experimental set-up inherited by Dr Petit's group to show the public how adsorbents can be used to capture CO₂. To successfully demonstrate the potential of adsorption, the right adsorbent material had to be selected in order to show a high CO₂ uptake. This involved extensive screening of materials, modelling the datapoints from their adsorption isotherms, validating the adsorption behaviour of the materials at our disposal and finally testing the selected materials in our experimental set-up and operating conditions.

2. Motivation for the Outreach Project

As part of our investigation, we conducted a survey among a group of 25 randomly selected students at Imperial College London to evaluate their understanding of methods to capture CO₂, focusing on adsorption. Their knowledge of anthropogenic emissions was evaluated. As shown in the Supplementary Information (SI), Figure S1, 68% know that the most emitted anthropogenic greenhouse gas is CO₂. However, almost an equal percentage (64%) was not familiar with the concept of carbon capture to reduce its atmospheric concentration and adsorption was not mentioned as a capture technology. This is a first insight into the necessity of educating people on the power of adsorption for CO₂ capture. To examine their adsorption knowledge, we asked the following question: "Imagine 2 vessels: one

empty and one filled with porous beads. If they were both filled with CO₂ at the same pressure, which one would store more CO₂?”. The group was uniformly divided as 50% thought the empty vessel would store more CO₂ than the packed one. This reinstated the importance of this project in increasing awareness on the potential of adsorption as a carbon capture technology.

This survey served as a motivation to educate the general public on how CO₂ can be captured via solid adsorption. We produced a video with a live demonstration that built upon an existing experimental set-up made by Dr Petit’s group which we updated. The demonstration shows how a column filled with a porous adsorbent material stores more CO₂ than an empty one. The set-up and the experiment are described further in Section 4.4. However, for the demonstration to be impactful, the packed column should store a significant quantity of CO₂ and therefore the adsorbent material had to be carefully chosen to have a high CO₂ uptake. To do so, different adsorbent materials were analysed to compare their adsorption capacity and kinetics. Based on this investigation and on availability, an adsorbent was selected for the demonstration.

3. Background on Adsorption

3.1 Adsorption Process

Adsorption through solid sorbents is an effective solution for CO₂ capture. It is a spontaneous process through which solid substances attract to their surface gas molecules for which they have high affinity. The process thus relies on a porous adsorbent material which preferentially binds with the desired gas, in this case CO₂. The gas molecules are binded on the active sites which may be packed much closer together than molecules dispersed in a gaseous state, allowing for more gas to be captured (The Editors of Encyclopaedia Britannica, 2023). Depending on the regeneration scheme, meaning how the adsorbent will desorb the adsorbed molecule and become reusable, adsorption can be broadly classified as pressure swing adsorption, temperature swing adsorption and electrothermal swing adsorption (Sharma, et al., 2021). Solid adsorption is a straightforward process with no liquid waste generation and appears as a promising technique due to its low cost, low energy requirement and applicability over a wide range of temperatures and pressures (Gunawardene, et al., 2022).

3.2 Adsorption Types

Adsorption is highly dependent on surface properties and can occur through physisorption or chemisorption. In physisorption, the CO₂ molecules attach to the pore walls of the adsorbent through weak intermolecular forces such as Van der Waals and pole-pole interactions. It is a naturally reversible process through which gas molecules can be adsorbed and desorbed under pressure or temperature, having a low regeneration cost. Equilibrium can be reached very quickly and increasing temperature leads to reduced surface coverage.

In chemisorption, the surface pores of the adsorbent material undergo chemical grafting or coating by incorporating basic groups, such as amines, in order to interact with the acidic CO₂ molecule. This means that bonding is stronger and has a shorter range, having covalent bonding characteristics. It is exothermic like physisorption. However, it may not be fully reversible and requires high energy for regeneration. Chemisorption usually has an activation energy barrier so equilibrium can be slow and thus increasing the temperature can favour it. Despite its higher energy cost, it is advantageous as it can offer high specificity, through which molecules can be selectively adsorbed (Williams, 2023).

3.3 Solid Sorbent Criteria

There are several criteria to be met to select the most effective CO₂ adsorbent. Depending on the objective of the adsorption process, different characteristics are prioritised. For our experimental demonstration, the most important factors are high CO₂ adsorption capacity, fast adsorption kinetics and safety. This ensures that high amounts of CO₂ gas are captured with an increased speed of adsorption.

Other set-ups might prioritise other criteria. High selectivity towards the gas molecule is often desired to preferentially bind and separate it from the other molecules. Good regeneration ability is also important in many processes as it guarantees a long lifetime of the material, allowing for it to be reused. Cost effectiveness, material sustainability, thermal stability and availability are also important factors (Gunawardene, et al., 2022).

3.4 Common Adsorbent Materials

Promising adsorbent materials that show these properties are activated carbons, metal-organic frameworks (MOFs) and zeolites. Activated carbon materials are advantageous due to their wide commercial availability at a low cost. They are a porous form of carbon which can be manufactured from a variety of carbonaceous raw materials, so depending on the adsorption process, materials of different pore size, structure and specific surface area can be selected. Their hydrophobic nature repels water molecules which would otherwise be in competition with CO₂ for the binding sites. They show high thermal stability but low adsorption capacity and selectivity towards CO₂ at low pressures due to their weak interaction with CO₂.

MOFs are also a large group of materials that have the potential for efficient CO₂ adsorption. These are crystalline porous materials that consist of positively charged metal ions surrounded by organic 'linker' molecules, forming a repeating, cage-like structure. Specifically tuned MOFs have high CO₂ adsorption capacity as pressure increases due to electrostatic CO₂-CO₂ interactions and high surface areas. However, at lower pressures their interaction with CO₂ usually weakens. They are also not as commercially available,

have high synthesis costs and low thermal stability hinting to harder regeneration.

Zeolites are a class of materials that are widely commercially available at a low cost. They are made of aluminosilicates of various structures. The aluminum atoms create a negatively charged framework which is balanced by supplementary non-framework cations like sodium or calcium. They offer moderate CO₂ adsorption capacity at low pressures but high structural stability and very high CO₂ selectivity following appropriate tuning of their properties. However, as they are hydrophilic in nature, competition between H₂O and CO₂ due to the similar size of the molecules can occur. They also require more energy for regeneration due to the very strong interaction of CO₂ with the zeolite framework.

The three adsorbent categories present different advantages and disadvantages, but zeolites are particularly attractive due to their high stability, low cost and the possibility to tune their physicochemical properties to support the desired adsorption process objectives (Boer, et al., 2023).

3.5 Adsorption Separation Types

There are three main types of molecular separations: equilibrium, kinetic and molecular sieving separation, shown in the SI, Figure S2. These happen depending on the adsorbent properties, the kinetic diameter of the molecules involved, their quadrupole moment and their polarisability.

Equilibrium separation relies on the electric field gradient of the adsorbent surface and the electrostatic interaction between the gas molecules and the surface, which is based on the polarisability of the molecules. This type of separation can take place in materials of all pore sizes – micro, meso and macro - as all molecules have access to the pores.

In adsorbents having micropores, kinetic and molecular sieving separation take place. During kinetic separation, the molecule that has a smaller kinetic diameter diffuses faster through the pores and gets more adsorbed. As the pore size becomes smaller, molecular sieving separation is reached, in which the pore size of the adsorbent only allows for the smaller molecules to access its micropores, preventing the larger ones from entering (Boer, et al., 2023).

4. Method

To select the appropriate material for the outreach video demonstration, literature and experimental analysis were combined. The material selection process is shown in Figure 1. The process is limited by some constraints. To begin, the material should be available in the lab. It should have high adsorption capacity and fast adsorption kinetics for a pressure from 0 to 9 bar and a temperature of 298 K which are the experimental conditions of the

demonstration. That way, we can show high CO₂ uptakes with a short time required to reach equilibrium at 9 bar. Finally, it should be safe for the demonstration, so the physical form and volatility of the material should be considered. Taking all these constraints into account, the selection process was the following: material screening, isotherm modelling of literature data, experimental validation of the available materials, and experimental trials in the conditions of the demonstration.

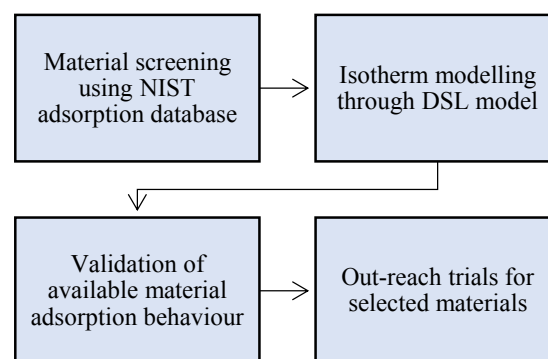


Figure 1: Material selection process for the video demonstration.

4.1 Determination of CO₂ Adsorption Capacity from Literature

As the goal of this project is to show how adsorption can play an important role in CO₂ capture, it was crucial to select a material with high adsorption capacity for the demonstration. Several adsorbent materials such as zeolites, MOFs and activated carbons, were screened using the NIST adsorption data base (NIST, 2023). The purpose was not to find the best material possible, but instead one that performs well in our experimental set-up. Thus, the search was refined and focused on materials that could be readily available in quantities large enough for the demonstration. Given that the demonstration displays the CO₂ uptake of a porous material at 9 bar and 298 K, adsorption isotherms for pure CO₂ were examined at that temperature for a pressure range of 0-9 bar. As our demonstration set-up could not be subjected to vacuum, it was desired that the selected material showed a high CO₂ adsorption uptake at a pressure range of 1-9 bar rather than 0-1 bar, in order to show a larger amount of CO₂ adsorbed in the demonstration and highlight the effect of adsorption to the viewers. From this search, isotherm datapoints from different papers were obtained for each material.

4.2 Equilibrium Isotherm Modelling – Determination of Parameters

A range of variation had to be created to clearly visualise the range of validity of the adsorption capacity of each material and to allow for a more accurate comparison between the materials. This is because the datapoints for the different papers are obtained using different synthesis processes and thus have different CO₂ uptakes for the same material. Using the datapoints obtained from the previous step, the isotherms of the materials that looked

promising for the demonstration were modelled through the Dual Site Langmuir model (DSL). The DSL model is given by the following equations:

$$q_j^* = \frac{q_{sb,j} b_j p}{1 + b_j p} + \frac{q_{sd,j} d_j p}{1 + d_j p} \quad (1)$$

$$b_j = b_{0,j} \exp\left(\frac{-\Delta U_{b,j}}{RT}\right) \quad (2)$$

$$d_j = d_{0,j} \exp\left(\frac{-\Delta U_{d,j}}{RT}\right) \quad (3)$$

For a pure component j , q_j^* is the amount of gas j adsorbed (mmol/g) at pressure p (bar) and temperature T (K), $q_{sb,j}$ and $q_{sd,j}$ are the saturation capacities (mmol/g), b_j and d_j are adsorption coefficients (/bar) described by an expression with two constants, $b_{0,j}$ and $d_{0,j}$ respectively which are the pre-exponential factor (/bar) and $\Delta U_{b,j}$ and $\Delta U_{d,j}$ respectively which is the adsorption energy (J/mol).

The isotherms for each research paper along with their modelling parameters were obtained using a MATLAB code created by Dr Petit's research group (Hwang, et al., 2022). The derived isotherms for each research paper were then averaged per material to obtain an average Langmuir isotherm. Their standard deviation was also calculated to obtain a variation envelope for the average Langmuir isotherm. The envelope was then used as the range of validity for the CO₂ uptake of a material. These equations are shown in the SI, Equations S.1 and S.2.

4.3 Adsorption Isotherms Measurement

Having selected the potential materials from this literature analysis, we needed to ensure that the materials at our disposal matched the envelope of variation. To do so, CO₂ adsorption isotherms for the available materials were measured using the 3Flex Sorption Analyzer. The samples were initially ex-situ degassed overnight and then in-situ degassed for 6 hours at specified temperatures, ranging from 393 K to 623 K, based on their thermal stabilities which are provided in the SI, Table S1. CO₂ gas (research grade, 99.999%, BOC) was used to measure CO₂ adsorption isotherms at 298 K for pressures from 0 to 1 bar. These were compared to the adsorption isotherm envelopes modelled using the literature data in order to ensure the materials selected in the forms available in the lab match our literature results.

4.4 Experimental Trials on the Outreach Demonstration Set-Up

As a final step, the materials were tested experimentally in the set-up for the demonstration to select a final material for the video. Their CO₂ uptakes were measured at 9 bar and compared to each other as well as to their values from the literature search.

The set-up was inherited from Dr Petit's research group. A schematic of it is shown in Figure 2.

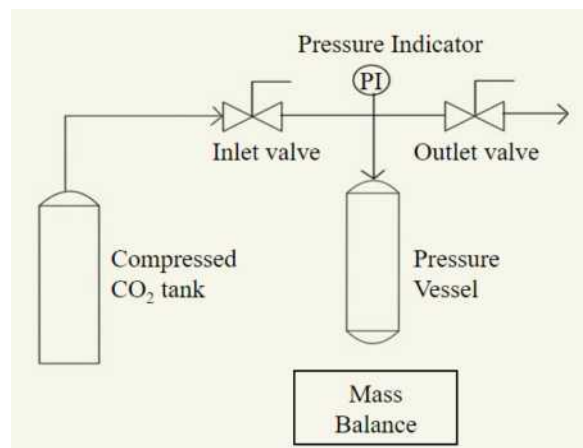


Figure 2: Basic schematic of the set-up for the experimental demonstration, showing the compressed CO₂ tank whose flow was controlled using manual valves and a pressure indicator, the vessel containing the adsorbent, and the mass balance to weigh the amount of gas stored in the vessel.

The main piece of equipment is the pressure vessel, shown in Figure 3. This was filled by the adsorbent material up to a volume of 120 ml for safety reasons. It was sealed with a piece of equipment consisting of metallic tubes from which pure CO₂ gas could be supplied and vented using the inlet and outlet valve respectively, as shown in Figure 2. The pressure could be monitored using the pressure indicator located at the top.



Figure 3: The pressure vessel used in the experimental demonstration.

Prior to conducting the trials, the adsorbent materials are degassed in the vacuum oven at their regeneration temperatures as described in Section 4.5, to ensure their pores are empty and no gases are present. The transparent container is then filled with the degassed adsorbent. The container is initially weighed, the scale is tared and the cable from the compressed CO₂ tank is connected at the inlet tube of the equipment. Compressed CO₂ gas is then introduced in the vessel until it reaches equilibrium at a pressure of 9 bar. The mass of CO₂ adsorbed in the vessel is recorded by weighing it again.

For each material 3 trials were conducted to increase accuracy. This stage allowed for comparison between the possible materials in the conditions that the demonstration would be performed. The final material could then be selected by comparing the CO₂ uptake as well as the kinetics.

4.5 Materials and Gases

The materials used in the experiments described in Section 4.3 and 4.4 are ZIF-8 (40 g, MOF Technologies), Zeolite 13X (50 g, N/A), Zeolite 5A (60 g, N/A), Zeolite 4A (60 g, N/A), Zeolite 3A (60 g, N/A) and carbon dioxide (CO₂) gas (99.995%, BOC). The adsorbent materials and CO₂ gas were used as received, with no further purification. For safety reasons, the materials were in the form of beads and rods rather than powder to reduce their volatility.

At the stage of the experimental trials, before performing the adsorption measurements, the materials had to be reactivated for more than 12 hours under continuous vacuum (4×10^{-4} mbar) at their activation temperatures. The ZIF-8 sample was used in the form of rods and was activated at 408 K. The zeolite 13X, 5A, 4A and 3A samples were used in the form of beads of different particle sizes and were activated at temperatures of 523 K. The activation temperatures for all adsorbent materials were within their range of thermal stability and above the activation temperatures suggested by the manufacturers. This information can be found in the SI, Table S1.

4.6 Video Recording of the CO₂ Adsorption Demonstration

Having selected the material for the video demonstration, the procedure mentioned in Section 4.4 was repeated and recorded. The process was also recorded for an empty container with no adsorbent material inside. This served as a control experiment to allow for comparison between CO₂ capture with and without an adsorbent. In that way, the increased uptake of the adsorbent is highlighted to show the power of adsorption. The video was edited through the DaVinci Resolve 18 Editor.

The recording of the experimental demonstration was part of the outreach video created for the public. The final video contained basic information on CO₂ capture and adsorption, followed by the video from the laboratory. The final video was created using a software called Animaker.

5. Results and Discussion

5.1 Determination of CO₂ Adsorption Capacity

The CO₂ uptakes were identified for different materials at 298 K and pressures of 0, 1 and 9 bar from different research papers available on NIST. These are shown in the SI, Tables S2, S3, S4 and S5. By taking the average of

these for each material, the materials could be compared by examining their CO₂ uptake.

The screening of 20 materials on the NIST adsorption database showed many promising CO₂ absorbents. The list of materials is available in Table S6 in the SI. Materials that showed low CO₂ uptakes were eliminated as, for the purpose of the demonstration, it was desirable to have a material that is highly adsorbent at the operating conditions. Thus, considering the experimental conditions and material quantity and availability, 4 potential materials were selected: ZIF-8, Zeolite 13X, Zeolite 5A and Zeolite 4A. Table 1 shows the average CO₂ adsorption capacities obtained from different papers for these materials at 0, 1 and 9 bar and the difference in CO₂ uptake between them.

Table 1: CO₂ adsorption capacities for the most promising materials at 0, 1 and 10 bar as well as the change in CO₂ uptake between from 0 to 1 bar and 1 to 9 bar.

Material	Average uptake at a pressure of:			Average change for pressures between:	
	0 bar	1 bar	9 bar	0-1 bar	1-9 bar
ZIF-8	0	0.6	4.2	0.6	3.7
Zeolite 13X	0	4.6	6.1	4.6	1.5
Zeolite 5A	0	3.8	4.6	3.8	0.8
Zeolite 4A	0	3.2	4.3	3.2	1.0

The isotherms datapoints derived from the research papers investigated are shown in Figure S3 in the SI. This allowed for a better visualisation of the CO₂ uptake of the chosen materials at 298 K for pressures between 0 and 9 bar.

5.2 Equilibrium Isotherm Modelling

The CO₂ isotherm datapoints obtained from the research papers mentioned in Section 5.1 were modelled using the Dual Site Langmuir model (DSL). The isotherms for each research paper, their modelling parameters and the average Langmuir isotherm along with the envelope of variation were obtained as described in Section 4.2. The fitted isotherms and their fitting parameters are shown in Figure S6 and Tables S7, S8, S9 and S10.

Figure 4 shows the average Langmuir isotherm lines in red and the variation ranges in grey. The envelopes enable a good visualization of the adsorption capacities for each of the chosen materials. As they clearly show a high CO₂ uptake at the experimental conditions, with an evident difference between the uptakes at 1 and 9 bar, it was decided that they were in fact good potential materials for the demonstration.

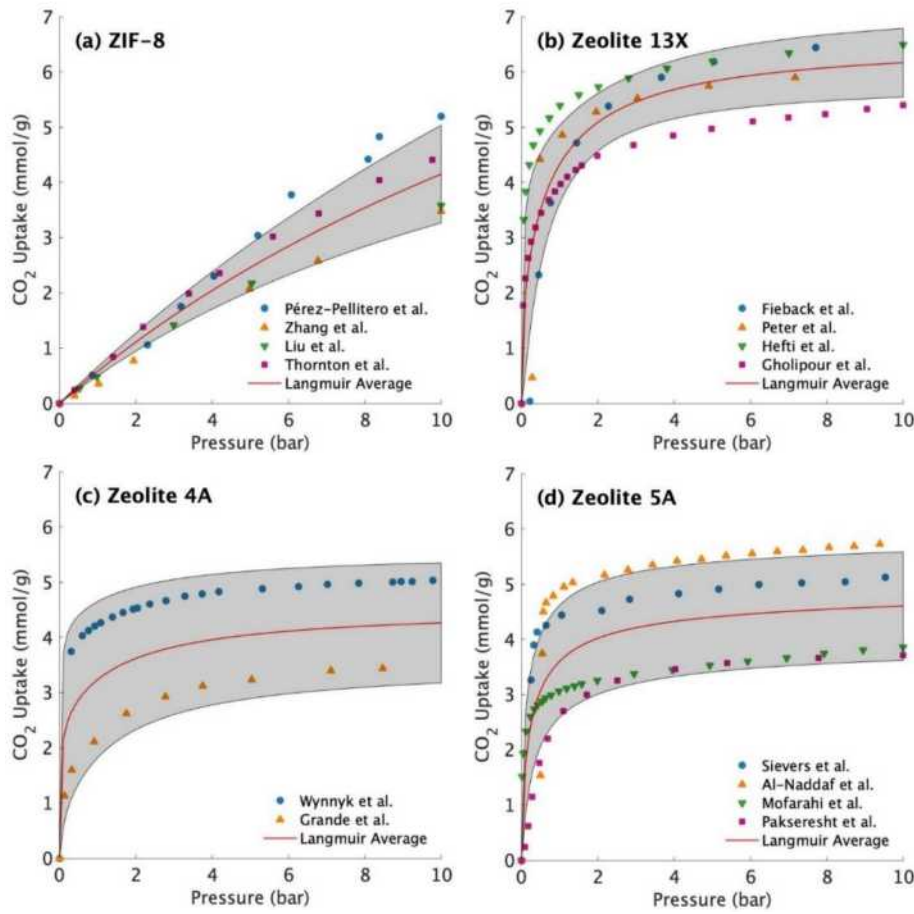


Figure 4: Average isotherm modelled using the DSL model and envelope of variation for (a) ZIF-8, (b) Zeolite 13X, (c) Zeolite 4A and (d) Zeolite 5A, including datapoints from different papers referenced in the legend and in the SI.

From this analysis, the behaviour of the 4 potential materials can be analysed. All isotherms in Figure 4 are Type I. However, ZIF-8 shown in panel (a) shows a different isotherm shape than the zeolites shown in panel (b), (c), and (d), hinting towards the type of adsorption that takes place in each material. The gradient of the isotherm for ZIF-8 is slightly decreasing as pressure increases, meaning that the kinetics are relatively constant, which is a sign of physisorption. On the 3 other isotherms, a sharp increase in the adsorption uptake is detected from 0 to 1 bar, while at higher pressures the uptake has reached a maximum distinguished by the flat plateau, which is a sign of chemisorption. In general, a sharp increase is more favourable, but for the purpose of our demonstration we are interested in the final uptake at 9 bar, which is why all materials show good adsorption behaviour considering they reach an average CO₂ uptake in the range of 4.0-6.0 mmol/g. Because of that,

all 4 materials are still considered as potential candidates for the video demonstration.

5.3 Adsorption Isotherms Measurement

The selected materials are available in the lab in different forms and from different manufacturers than the materials in the research papers. To ensure that available materials behaved in the same way as the literature suggested, their CO₂ adsorption isotherms were measured using the 3Flex Sorption Analyzer. The results obtained for ZIF-8, Zeolite 13X and Zeolite 4A ranged from a pressure of 0 to 1 bar at 298 K and are shown in Figure 5. The measurements were compared to the literature data by superimposing them on the Figure 4 from Section 5.2, containing the isotherm envelope. The x-axis was limited to 1 bar to show the experimental points more clearly and contrast them to literature. Zeolite 5A was not analysed at this stage due to availability limitations of the 3Flex Sorption Analyzer.

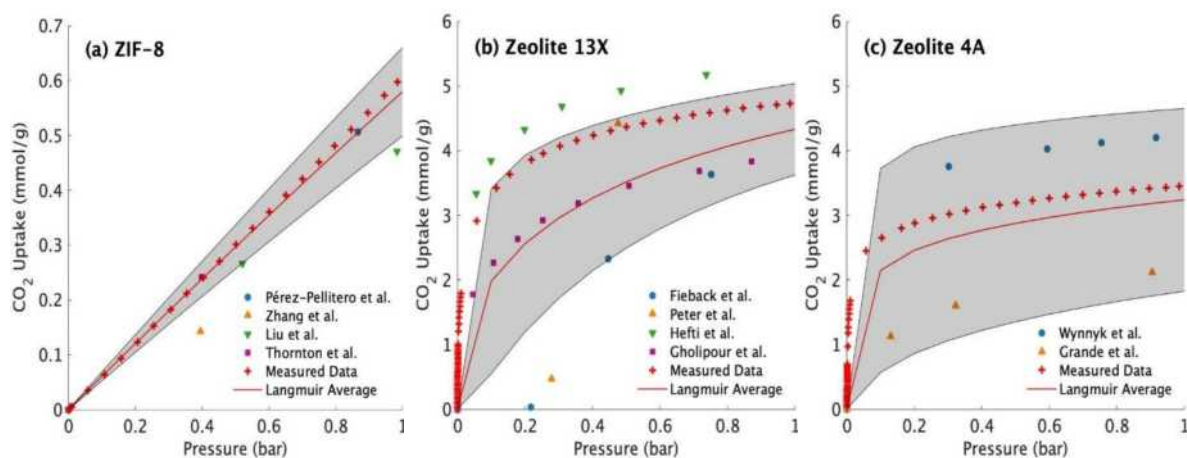


Figure 5: Measured data superimposed on the variation envelopes for (a) ZIF-8, (b) Zeolite 13X and (c) Zeolite 4A.

The isotherms obtained from our measurements are a good match to the literature data. As pressure increases, the measured datapoints are well within the envelope of variation. However, for the zeolites shown in plots (b) and (c), at very low pressures the measured datapoints show higher CO₂ uptake than the literature. This can be explained by the increased amount of datapoints collected by our instrument compared to the literature data. This suggests that the 3Flex instrument has high precision. It can take measurements at small increments of pressure which explains the increased amount of time that was required for the lower pressure measurements as observed during data collection. As the quantities adsorbed at 1 bar fall within the envelopes and the shape and slope of the datapoints approach the literature, the materials available were considered to match with the literature. It was thus assumed that they would also have similar uptakes at higher pressures, if the datapoints were to be extrapolated. Therefore, experimental testing at 9 bar was conducted for each of these materials to select one.

5.4 Experimental Trials on the Outreach Demonstration Set-Up

As ZIF-8, Zeolite 13X and Zeolite 4A performed satisfactorily in the validation analysis, they were examined in the set-up designed for the demonstration to finally determine the ideal one. Zeolite 5A was also investigated in experimental trials as it has similar properties to 4A and showed good adsorption capacities from the literature review. As Zeolite 3A was also available in the lab and has a structure similar to the Zeolites 4A and 5A, it was decided to test it even though no data on its CO₂ capacity were found in the literature. Given that it has a much smaller pore diameter of 3 Å, it was expected that CO₂, whose

kinetic diameter is 3.3 Å, would adsorb much less on it. Zeolites 4A and 5A, however, would adsorb much more CO₂ as they have larger pore diameters of 4 Å and 5 Å respectively.

The experiment described in Section 4.4 was conducted on these 5 adsorbents. As each trial was conducted 3 times per material, Figure 6, panel (a), shows the averaged CO₂ uptakes at 9 bar and 298 K of the materials with error bars. These values can be compared to the values in Figure 6, panel (b), which show the averaged CO₂ uptakes at 1 and 9 bar at 298 K as obtained from research papers as analysed in Section 5.1.

Comparing the two figures, the experimental data match with the literature data as they fall within the error bars of the literature data. It should be noted that the error bars are relatively large due to the different properties and synthesis routes of each material in the different research papers examined, which led to a large standard deviation. It is interesting to note that the trend for the zeolites is similar in trials and literature, but the literature data shows higher values overall. This can be attributed to the form of the materials, given that in literature they were usually powdered samples, while in our experiments we used shaped beads for safety reasons. The powdered samples have a higher surface area and thus a higher adsorption capacity, in the case of pure CO₂ injection, as there are more binding sites available. However, ZIF-8 did not follow a trend compared to the other materials. It had the highest average CO₂ adsorption capacity in trials and the lowest in literature. Nevertheless, its experimental value is close to its literature value.

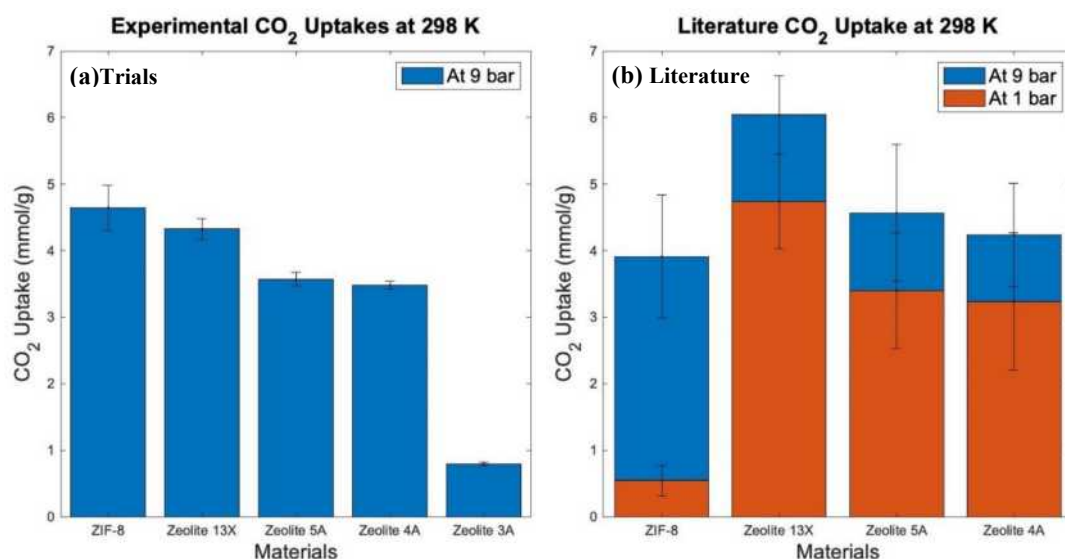


Figure 6: (a) Experimentally determined average CO₂ uptake for the potential materials when reaching equilibrium at 9 bar with error bars and (b) literature determined average CO₂ uptake for the same materials when reaching equilibrium at 1 and 9 bar with error bars.

The experimental results in Figure 6 confirm our theory that Zeolite 3A is a bad CO₂ adsorbent, as it adsorbs less than 1.0 mmol/g. Despite having a smaller pore size than the kinetic diameter of CO₂ it still adsorbs a small amount, owing to the different types of binding sites, and thus pore windows, on its surface. The remaining materials can be considered as high performing CO₂ adsorbents. Comparing the results for the 3 zeolite materials - Zeolite 13X, 4A and 5A - it was determined that 13X shows the highest CO₂ adsorption capacity in both experimental and literature results. Therefore, Zeolites 4A and 5A were eliminated for the purposes of the demonstration.

The final material options remaining were thus Zeolite 13X and ZIF-8. During the trials, we noticed that Zeolite 13X reaches equilibrium faster than ZIF-8, but ZIF-8 shows a bigger CO₂ uptake as shown in Figure 6(a). However, the measurement of the uptake is shown in units of mmol/g. As the outreach experiment displays the CO₂ uptake as measured on the scale, it is more relevant to compare the CO₂ uptakes in mass units while keeping the volume of the material constant. The units are thus consistent with what the audience will directly observe: the mass of CO₂ stored in the vessel. For an initial packed material volume of 120 ml the average CO₂ uptakes in grams are shown in Table 2.

Table 2: Average experimentally determined CO₂ uptakes in mass units at 298 K and 9 bar for the materials tested in the demonstration set-up.

Material	CO ₂ Uptakes (g) at 298 K and 9 bar
ZIF-8	5.9
Zeolite 13X	7.4
Zeolite 5A	7.1
Zeolite 4A	6.9
Zeolite 3A	1.7

From the results relative to mass units, Zeolite 13X will adsorb more mass of CO₂ at the demonstration conditions of 298 K, 9 bar and a volume of 120 ml of adsorbent. Zeolite 13X was therefore chosen for the outreach video.

5.5 Video Recording Outcome

Having selected Zeolite 13X as the adsorbent material for our demonstration, the video was recorded and combined with the general background information regarding CO₂ capture and adsorption. The final video can be found on: [Final Year Research - Maria Pakradouni & Mikaela Zafet.mp4](#).

The difference in the mass of the two containers following the addition of CO₂ illustrated the much higher mass in the vessel containing the porous adsorbent, Zeolite 13X. This highlights the effectiveness of CO₂ adsorption given that it has the capacity to store more CO₂ than the empty container.

6. Conclusion

As a result of this project, an outreach video was created. The goal of the video is to educate the general public on the necessity of reducing carbon emissions and the role adsorption can have in carbon capture. The video includes general information on climate change, background information on adsorption and a live demonstration highlighting the power of adsorbent materials in action. For the demonstration to be impactful, a large amount of CO₂ had to be captured. For that, the adsorbent had to be carefully selected. The chosen adsorbent material was Zeolite 13X.

7. Outlook

With respect to the CO₂ outreach aspects of the project, further engagement of the public can be achieved. More people can be surveyed to raise awareness of CO₂ capture. Given that many people submitted wrong answers, upon completion of the survey the participants can be directed to our video demonstration. Simultaneously, the outreach can become more engaging through live demonstrations, by participating in science fairs, so that people can directly see the experiment take place. Finally, to promote the video to more people, organisations such as climate and sustainability NGO's can be contacted to share it on their websites and show it at their events.

As it was important to select the appropriate material for the outreach demonstration video, extensive adsorption screening was conducted. However, for future work more adsorbents could be screened, especially from other classes of materials. Given the tight timeframe, we focused mostly on zeolites given their aforementioned advantages. However, more focus could be placed on MOFs and activated carbons as these materials can also show high adsorption capacities, are abundant and their physicochemical properties can be tuned.

Acknowledgements

We are thankful for the support and guidance we received throughout this project from Dr. Camille Petit, Ioanna Itskou and Hassan Azzan. We would also like to thank Patricia Carry for her help in the analytical lab.

References

- Al-Naddaf, Q., Rownaghi, A. A. & Rezaei, F., 2020. Multicomponent adsorptive separation of CO₂, CO, CH₄, N₂, and H₂ over core-shell zeolite-5A@MOF-74 composite adsorbents Author links open overlay panel. *Chemical Engineering Journal*, Volume 384.
- Boer, D. G., Langerak, J. & P., P. P., 2023. *Zeolites as Selective Adsorbents for CO₂ Separation*. [Online] Available at: <https://pubs.acs.org/doi/epdf/10.1021/acsaeem.2c03605> [Accessed 5 November 2023].
- Chunshan Song, W. P. S. T. S. J. Z. Y. L. Y.-H. W. B.-Q. X. Q.-M. Z., 2007. *Tri-reforming of Methane over Ni Catalysts for CO₂ Conversion to Syngas With Desired H₂/CO Ratios Using Flue Gas of Power Plants Without CO₂ Separation*. [Online] Available at: <https://www.sciencedirect.com/science/article/pii/S0167299104802702>
- Fieback, J. R. & T., 2013. Multicomponent adsorption measurements on activated carbon, zeolite molecular sieve and metal-organic framework. *Adsorption*, Volume 19, p. 1065–1074.
- Gholipour, F. & Mofarahi, M., 2016. Adsorption equilibrium of methane and carbon dioxide on zeolite 13X: Experimental and thermodynamic modeling. *The Journal of Supercritical Fluids*, Volume 211, pp. 47–54.
- Grande, C. A., Silva, V. M., Gigola, C. & Rodrigues, A. E., 2003. Adsorption of propane and propylene onto carbon molecular sieve. *Carbon*, 41(13), pp. 2533–2545.
- Gunawardene, O. H., Gunathilake, C. A., Vikrant, K. & Amaraweera, S. M., 2022. *Carbon Dioxide Capture through Physical and Chemical Adsorption Using Porous Carbon Materials: A Review*. [Online] Available at: <https://www.mdpi.com/2073-4433/13/3/397>
- Hefti, M., Marx, D., Joss, L. & MAzzotti, M., 2015. Adsorption equilibrium of binary mixtures of carbon dioxide and nitrogen on zeolites ZSM-5 and 13X. *Microporous and Mesoporous Materials*, Volume 215, pp. 215–228.
- Hwang, J., A. H., Pini, R. & Petit, C., 2022. H₂, N₂, CO₂, and CH₄ Unary Adsorption Isotherm Measurements at Low and High Pressures on Zeolitic Imidazolate Framework ZIF-8. *Journal of Chemical & Engineering Data*, 67(7), pp. 1674–1686.

IEA, 2023. *CO₂ Emissions in 2022*. [Online] Available at: <https://www.iea.org/reports/co2-emissions-in-2022> [Accessed 15 November 2023].

Liu, D. et al., 2012. Experimental and molecular simulation studies of CO₂ adsorption on zeolitic imidazolate frameworks: ZIF-8 and amine-modified ZIF-8. *Adsorption*, 1(19), pp. 25-37.

Merck, 2023. *Merck*. [Online] Available at: https://www.sigmaaldrich.com/GB/en/product/sigald/208582?gclid=Cj0KCQiAyeWrBhDDARIsAGP1mWSRc70zO7xLjH6OsMD1UZlvveekUaaDdaM3utW8BQjAUis8_SOKYQgaAipaEALw_wcB

Mofarahi, M. & Gholipour, F., 2014. Gas adsorption separation of CO₂/CH₄ system using zeolite 5A. *Microporous and Mesoporous Materials*, Volume 200, pp. 1-10.

NIST, 2023. *NIST/ARPA-E Database of Novel and Emerging Adsorbent Materials*, s.l.: National Institute of Standards and Technology.

Pakseresht, S., Kazemeini, M. & Akbarnejad, M. M., 2002. Equilibrium isotherms for CO, CO₂, CH₄ and C₂H₄ on the 5A molecular sieve by a simple volumetric apparatus. *Separation and Purification Technology*, 28(1), pp. 53-60.

Pérez-Pellitero, J. et al., 2010. Adsorption of CO₂, CH₄, and N₂ on Zeolitic Imidazolate Frameworks: Experiments and Simulations. *Chemistry - A European Journal*, 16(5), pp. 1560-1571.

Peter, S. A. et al., 2013. Dynamic desorption of CO₂ and CH₄ from amino-MIL-53(Al) adsorbent. *Adsorption*, Volume 19, p. 1235–1244.

Sharma, A. et al., 2021. *Springer Link*. [Online] Available at: <https://link.springer.com/article/10.1007/s10311-020-01153-z> [Accessed 10 November 2023].

Sievers, W. & Mersmann, A., 1994. Single and multicomponent adsorption equilibria of carbon dioxide, nitrogen, carbon monoxide and methane in hydrogen purification processes. *Chemical Engineering & Technology*, 17(5), pp. 325-337.

Stefano E. Zanco, J.-F. P.-C. A. G. B. C. V. B. a. M. M., 2021. *Postcombustion CO₂ Capture: A Comparative Techno-Economic Assessment of Three Technologies Using a Solvent, an Adsorbent, and a Membrane*. [Online] Available at:

<https://pubs.acs.org/doi/10.1021/acsengineeringau.1c00002>

The Editors of Encyclopaedia Britannica, 2023. *Britannica*. [Online] Available at: <https://www.britannica.com/science/adsorption>

Thornton, A. W. et al., 2013. Analytical representation of micropores for predicting gas adsorption in porous materials. *Microporous and Mesoporous Materials*, Volume 167, pp. 188-197.

U.S. Department of Energy, National Energy Technology Lab, 2003. *Adsorption of CO₂, N₂, and O₂ on Natural Zeolites*. [Online] Available at: https://pubs.acs.org/doi/full/10.1021/ef020135l?casa_token=APJ0htoa2aMAAAAA%3AdHreD6Froh2WnLIG4JA6noSalhwyoMNFgvi16Bmk93z8jSvvDE0d7ONUmHaXyqPuRh3dmBjhXzeex4g

United Nations, 2023. *For a livable climate: Net-zero commitments must be backed by credible action*. [Online] Available at: <https://www.un.org/en/climatechange/net-zero-coalition> [Accessed 15 November 2023].

Williams, D., 2023. *Product Characterization: Lecture 5 Porosity and Surface Area Analysis*. [Online] Available at: <https://bb.imperial.ac.uk/ultra/courses/374781/cl/outline>

Wynnyk, K. G., Hojjati, B., Pirzadeh, P. & Marriott, R. A., 2016. High-pressure sour gas adsorption on zeolite 4A. *Adsorption*, Volume 23, p. 149–162.

Zhang, Z. et al., 2013. Enhancement of CO₂ Adsorption and CO₂/N₂ Selectivity. *AIChE*, 59(6), pp. 1830-2266.

Neural networks to simulate and optimise a Pressure-Vacuum Swing Adsorption process

Harashima, Kenji and Tran, Yung

Department of Chemical Engineering, Imperial College London, U.K.

Abstract

Pressure-Vacuum Swing Adsorption (PVSA) shows great potential in post-combustion carbon capture. However, accurately modelling it requires a large amount of computational time using detailed process models, and screening large numbers of adsorbents becomes computationally prohibitive. Using data-driven neural network models have great potential to solve this problem and is the focus of this paper. We build and adapt the previously established machine-assisted adsorption process learning and emulation (MAPLE) framework by using the dual-site Langmuir model, expanding the number of features used, and predicting capture costs, to demonstrate the benefits of using a neural network to make fast and accurate assessments of a PVSA process. A detailed mathematical process model was used to generate training data for our neural networks and a case study was performed to compare the performance of our neural network with that of the detailed model. Our results indicate that our neural networks have performance comparable to that of the detailed model with acceptable levels of uncertainties up to around 14% whilst requiring up to 25,200x less computational time. This vast reduction in computational time shows the great potential of this tool in solving process optimisation problems compared to when using traditional process modelling.

1. Introduction

Carbon capture is a tested method to aid in the decarbonisation of the energy industry that most processes will have to implement in some sort of capacity to achieve net-zero (Allen et al., 2018). Pressure-vacuum swing adsorption (PVSA) is shown to have great potential compared to standard amine-based chemisorption for post-combustion capture of CO₂ (Ruthven et al., 1996). Whilst detailed mathematical modelling is an effective method to describe and predict the behaviours of an adsorption process, it also requires many time-consuming simulations (Haghighpanah et al., 2013; Leperi et al., 2019), and is further exacerbated when looking at multi-objective optimisation that requires thousands of operating points (Agarwal et al., 2010). A standard approach is to couple a detailed process model with an optimisation algorithm, allowing the model to generate operating points for the algorithm to refine to an optimum (Capra et al., 2018).

The limitations of this approach are apparent when evaluating the performance of adsorbents from large databases. There is an ever-increasing number of real and hypothetical adsorbents being synthesised and attempts have been made to conduct large scale screening (Leperi et al., 2019). Simplified models try to circumvent this issue but full process simulations are still required for accurate predictions of certain parameters (Burns et al., 2020). Multiple machine learning (ML) models have also been developed to solve this issue: Subraveti et al. (2019) introduced an optimisation method using a neural network (NN) to simplify the complexity of the problem and enhancing optimisation speed by reducing the complex pressure swing adsorption (PSA) process to a lower

dimensional model. In another development, Leperi et al. (2019) established a NN model that mimics a reduced-order PSA process to reduce computational time.

As such, our goal is to solve this optimisation problem through the use of a NN model that builds on the ML framework detailed by Pai et al. (2020). The machine-assisted adsorption process learning and emulation (MAPLE) framework differs from the aforementioned ML approaches in that the training of the model is not based on the properties of actual adsorbents; instead, it involves parameterising the Langmuir adsorption isotherm itself and using that as an input variable to create a generalised model that does not need training for particular adsorbents. We aim to build on the MAPLE framework by using the more complex dual-site Langmuir (DSL) form rather than a single-site Langmuir (SSL) form, allowing us to better represent the adsorption process and complicated adsorbents (Ritter et al., 2019; Wilkins and Rajendran, 2019). Our model will also be able to make an economic assessment based off the detailed process model from Ward and Pini (2022) and predict the capture cost of CO₂ defined as the number of dollars required to capture one tonne of CO₂ (\$/tonne). This is an aspect that has not been integrated into other previous adsorption-based NNs. As with all ML approaches, an initial investment in computational time for producing the training data is required. Once the NN is trained, it can rapidly make new predictions and solve the multi-objective optimisation problem in significantly less time than numerical methods. We extend the framework established by Ward and Pini (2022) and utilize their detailed process model and optimisation outcomes as a foundation for

constructing and validating our NN. The detailed model will be used to train and validate our NN, followed by a cost optimisation case study of Zeolite 13X, a promising adsorbent that is commonly used for separation of N₂ to produce high purity O₂ (Tong et al., 2018). We then compare our results to Ward's results to verify the accuracy of our NN.

2. Background

2.1. Pressure-Vacuum Swing Adsorption

PVSA is a capture technology that shows great potential especially for post-combustion capture. It offers significant advantages, especially in retrofitting existing plants and providing good capture capacity in comparison to other technologies such as chemical absorption (Riboldi and Bolland, 2017). This study will focus on modelling a typical 4 stage PVSA process; Figure 1 shows a schematic detailing each step of the cycle. The first adsorption step involves passing the CO₂ rich feed gas into the column packed with adsorbent at a high pressure (p_H) in which the CO₂ is preferentially adsorbed. Once the adsorbent is saturated with CO₂, the second forward blowdown step involves closing the feed end and depressurising the column to an intermediate pressure (p_I). A N₂-rich product is then expelled from the product end. During the third reverse evacuation step, the feed end is then opened and the product end closed while depressurising the column even further to a low pressure (p_L). This step desorbs the CO₂ from the adsorbent and allows a CO₂ rich gas to be collected from the feed end. The final feed pressurisation step involves pressurising the feed and repeating the cycle until a cyclic steady state is reached.

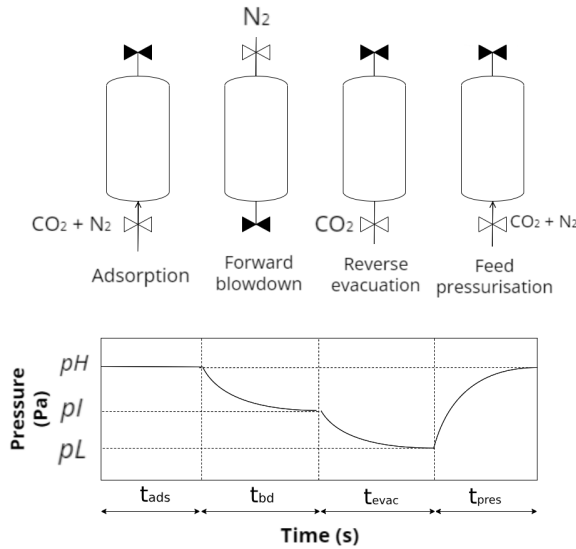


Figure 1: A schematic of the 4 steps of the PVSA process along with the pressure profile during each step

2.2. Detailed Model

We use the numerical model developed and outlined by Ward and Pini (2022) to apply a PVSA process to post-combustion

carbon capture from a coal-fired power plant using dry flue gas and a fixed bed adsorber packed with Zeolite 13X. The flue gas is assumed to be in a 15:85% molar mixture of CO₂ to N₂ and under standard ambient conditions, specifically at a pressure of 1 bar and a temperature of 298.15K. For simulating the adsorption process on the adsorbent bed, the dual-site Langmuir (DSL) isotherm model is employed, where the amount of substance adsorbed by a species i at equilibrium is given by:

$$q_i^* = \frac{q_{b,i}c_i}{1 + \sum_{j=1}^{n_c} b_jc_j} + \frac{q_{d,i}d_i}{1 + \sum_{j=1}^{n_c} d_jc_j} \quad (1)$$

The saturation capacities of species i on site 1 and site 2 of the solid surface are denoted as $q_{b,i}$ and $q_{d,i}$, respectively. The adsorption equilibrium constants for each adsorption site, represented by b_i and d_i , are expressed as functions of temperature through the utilization of the van't Hoff equation,

$$b_i = b_{i,0} \exp\left(\frac{\Delta U_{b,i}}{RT}\right) \quad (2)$$

$$d_i = d_{i,0} \exp\left(\frac{\Delta U_{d,i}}{RT}\right) \quad (3)$$

The change in molar internal energy for species i during adsorption at site 1 and site 2 is denoted as $\Delta U_{b,i}$ and $\Delta U_{d,i}$, respectively. To determine the molar concentration of species i in the gas phase, the ideal gas law is utilized.

$$c_i = \frac{y_i p}{RT} \quad (4)$$

Table 1 contains the parameters for the extended DSL model, which characterizes the adsorption equilibrium of CO₂/N₂ on zeolite 13X (Haghpanah et al., 2013).

Material	Parameter	CO ₂	N ₂
Zeolite 13X	q_b [mol/kg]	3.09	5.84
	q_d [mol/kg]	2.54	-
	b_0 [m ³ /mol]	8.65×10^{-7}	2.50×10^{-6}
	d_0 [m ³ /mol]	2.63×10^{-8}	-
	ΔU_b [J/mol]	-36,600	-15,800
	ΔU_d [J/mol]	-35,700	-

Table 1: DSL parameters of CO₂ and N₂ (Haghpanah et al., 2013)

The achievement of cyclic steady state is determined by tracking 5 key performance indicators (KPIs): purity (Pu_{CO_2}), recovery (Re_{CO_2}), productivity (Pr), total energy usage (E_T) and capture cost ($C_{CO_2}^{cap}$). The simulation concludes once the relative error for each of the 5 KPIs remains below 0.5 percent across 10 successive cycles. The purity represents the percentage of the number of moles of CO₂ compared to the total number of moles of both CO₂ and N₂ present in the product stream. The recovery represents the percentage of moles of CO₂ present in the product stream of the reverse evacuation step compared to the amount in the feed stream. The productivity is defined as the amount of CO₂ (in moles) coming out during the reverse evacuation step, divided by the volume of

Operating parameter	Lower bound	Upper bound
p_H [bar]	1	10
p_I [bar]	0.12	3
p_L [bar]	0.02	0.1
v_F [ms ⁻¹]	0.1	2
t_{ads} [s]	20	100
t_{bd} [s]	30	100
t_{evac} [s]	30	100
L [m]	1	12
$y_{1,F}$	0.03	0.30

Table 2: Operating parameters bounds for neural network data set

the adsorbent and the total time of one cycle, which is the cumulative duration of all four steps in the process. The energy usage is defined as the sum of the energy usage of all 4 steps divided by the mass of CO₂ coming out of the product stream during the reverse evacuation step.

2.3. Detailed model data set

A large data set is needed to create a NN, containing a set of inputs (referred to as features) and expected outputs (referred to as the labels). The labels are the KPIs from the detailed model (Ward and Pini, 2022), while the features for the data set are various parameters specified for the PVSA process. The high pressure (p_H), intermediate pressure (p_I), and low pressure (p_L), time for the first adsorption step (t_{ads}), second blowdown step (t_{bd}), third evacuation step (t_{evac}), and feed velocity (v_F) are used as features describing the PVSA process operating parameters. The adsorbent parameters (ρ , $q_{b,i}$, $q_{d,i}$, $b_{0,i}$, $d_{0,i}$, $\Delta U_{b,i}$, and $\Delta U_{d,i}$) from the DSL isotherm model are added as features describing the adsorbent used. Finally, the column length L , inner radius r_{in} , outer radius r_{out} , and feed composition $y_{1,F}$ are added as parameters related to the adsorption column configuration. In total there are 24 features in the data set and 5 different labels.

A range for most of the operating parameters, except the adsorbent parameters, was created to encapsulate the possible values they could take shown in table 2. The adsorbent parameters were taken from an adsorbent database containing 72 different adsorbents (Ward and Pini 2022). r_{in} and r_{out} are calculated based on the column length L :

$$r_{in} = \frac{L}{6}, \quad (5)$$

$$r_{out} = r_{in} + 0.0175. \quad (6)$$

3. Neural Networks

This section gives a background on the mathematics behind NNs, training, validation and testing of the NN, and NN hyperparameters.

3.1. Basic neural network structure

NNs can create surrogate models with a large data set, using the features to generate predicted labels and then using the true labels from the data set to improve the accuracy of its

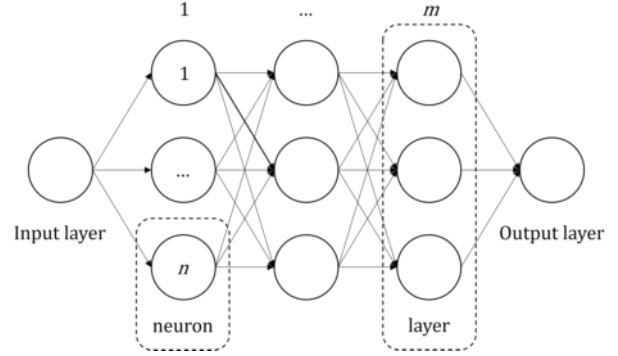


Figure 2: Simple NN diagram

predictions. Figure 2 is used to discuss the structure and components within a NN. The NN comprises of an input layer, m hidden layers containing n neurons, and an output layer. Every layer in the NN is made up of neurons which function differently depending on which layer the neurons are in.

The neurons in the input layer represent each input variable or feature in the input data set. For an input data set containing l features, the input layer will contain l neurons and produce an input vector \mathbf{x} . It is typical to normalise all input data using the mean, μ , and standard deviation, σ , of each feature before passing it through the NN as it helps train the NN faster and scales inputs equally preventing features with large values dominating over other features. The normalised feature \mathbf{x}_{norm} is given by,

$$\mathbf{x}_{norm} = \frac{\mathbf{x} - \mu}{\sigma}. \quad (7)$$

The neurons in the hidden layers have an associated weight vector, \mathbf{w} , and a bias term, b that are used to compute an output vector \mathbf{a} . The weight vector represents the strength between connections of neurons between layers, while the bias term helps shift the output to better model the outputs. First, each neuron in the hidden layer calculates a value z . For the neurons in a layer m ,

$$z = \mathbf{x} \cdot \mathbf{w}_n^{[m]} + b_n^{[m]}, \quad (8)$$

where $\mathbf{w}_n^{[1]}$ is the weight vector for the n^{th} neuron in the m^{th} hidden layer, and $b_n^{[1]}$ is the bias term for the n^{th} neuron in the m^{th} hidden layer. $\mathbf{x} \cdot \mathbf{w}_n^{[m]}$ is the dot product between the input vector \mathbf{x} and weight vector $\mathbf{w}_n^{[m]}$. For the first neuron in the first hidden layer, the dot product is,

$$\mathbf{x} \cdot \mathbf{w}_1^{[1]} = x_1 w_{1,1}^{[1]} + x_2 w_{1,2}^{[1]} + \dots + x_l w_{1,l}^{[1]}. \quad (9)$$

Note that the weight vector for the first neuron has the same amount of terms as the input vector \mathbf{x} so the dot product can be computed. Therefore, all the weight vectors in the first layer have l weight terms and the dot product for the n^{th} neuron in the first hidden layer is,

$$\mathbf{x} \cdot \mathbf{w}_n^{[1]} = x_1 w_{n,1}^{[1]} + x_2 w_{n,2}^{[1]} + \dots + x_l w_{n,l}^{[1]}. \quad (10)$$

The number of weight terms in the weight vector \mathbf{w} for a neuron depends on the number of neurons in the previous layer. The weight vectors for each neuron in the rest of the hidden layers all have n number of weight terms in their weight vector \mathbf{w} .

An activation function $g(z)$ is used to add non-linearity to the NN model and help it capture complex relations between the features and labels. For our NN, we use the Rectified Linear Unit (ReLU) function as the activation function for all neurons in the hidden layers as it only outputs non-negative values leading to faster training of the NN,

$$g(z) = \max(0, z). \quad (11)$$

Using the ReLU function, the neurons in hidden layer m computes the output vector terms $a_n^{[m]}$,

$$\mathbf{a}^{[m]} = \begin{bmatrix} a_1^{[m]} \\ a_2^{[m]} \\ \vdots \\ a_n^{[m]} \end{bmatrix}, a_n^{[m]} = g(z) = \max(0, \mathbf{x} \cdot \mathbf{w}_n^{[m]} + b_n^{[m]}). \quad (12)$$

The first hidden layer produces the output vector $\mathbf{a}^{[1]}$ from the input vector \mathbf{x} , the second hidden layer produces $\mathbf{a}^{[2]}$ from $\mathbf{a}^{[1]}$, and so on until the m^{th} hidden layer produces $\mathbf{a}^{[m]}$ from $\mathbf{a}^{[m-1]}$.

The final output layer consists of only one neuron and takes $\mathbf{a}^{[m]}$ from the m^{th} hidden layer to calculate the final output value being the KPIs for our NNs. The neuron in the output layer also contains an activation function that can differ depending on its use case. We use the ReLU function in the output layer for the productivity, energy usage, and capture cost NNs as the predicted values should only be positive, but for purity and recovery we instead use the sigmoid function as the output activation function,

$$g(z) = \frac{1}{1 + e^{-z}} = \frac{1}{1 + e^{-(\mathbf{a}^{[m]} \cdot \mathbf{w}^{[m+1]} + b^{[m+1]})}}, \quad (13)$$

Where $\mathbf{w}^{[m+1]}$ is the weight vector for the output layer neuron, and $b^{[m+1]}$ is the bias term for the output layer neuron. The sigmoid function calculates a value between 0 and 1 which is useful for the purity and recovery NNs as these values also range only from 0 to 1 making it much more suitable for these NNs.

3.2. Neural network training

The NN uses a cost function C measuring the error between the predicted labels from the NN and the true labels from the output data set to train the NN. A commonly used cost function is the mean squared error (MSE) function:

$$C = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2, \quad (14)$$

where y is the true label from the output data set, \hat{y} is the predicted label from the NN, and N is the total number of points in the training data set. \hat{y}_i is a function of all the weight terms w in the weight vectors \mathbf{w} and all the bias terms b for each

neuron in all the layers. An optimisation algorithm is chosen to minimise the cost function C and train the NN, finding the optimal values for all w and b . The Adam (short for adaptive moment estimation) optimiser algorithm is known for its adaptive learning rates that change while training the NN allowing it to quickly converge to find optimal parameters (Kingma and Ba, 2017). The learning rate is a hyperparameter that controls how much w and b change as the algorithm tries to minimise the cost function. Further hyperparameters are discussed in Section 3.4

3.3. Validating and testing the neural network

The total data set is usually split into parts; a large part of the data set goes to the training set which is used to train the NN, while the rest is split into a validation set and test set. The usual split for these sets is 80% goes to the training set, while 10% set goes to the validation and test set each.

The validation set prevents the NN from learning the training data set too well. If the loss function continually decreases while training the NN, it could lead to overfitting where the NN would only be accurate in predicting points from the training data set, but poor in predicting new points that it has not encountered before. Every time the NN is trained, the NN is used on the validation set to check the accuracy of its predictions and evaluate the NNs ability to generalise new data it has not seen before.

There are various metrics used to measure the accuracy of the NN. One such metric that is usually used for the validation set is the Root Mean Squared Error (RMSE) function,

$$\text{RMSE} = \sqrt{\sum_{i=1}^N \frac{(y_i - \hat{y}_i)^2}{N}}. \quad (15)$$

Note that N , y_i , and \hat{y}_i in this equation and all the other metrics discussed in this section refer to the variables in the validation/test data set. The RMSE equation is similar to the MSE equation in penalising larger errors more than small errors, but may be hard to interpret as it depends on the scale of the output. Other metrics that are used to determine the accuracy of each NN are the mean absolute percentage error (MAPE) function, and coefficient of determination R^2 , a widely used metric that measures how well a model predicts an output:

$$\text{MAPE} = \frac{1}{N} \sum_{i=1}^N \left| \frac{y_i - \hat{y}_i}{y} \right|. \quad (16)$$

$$R^2 = 1 - \frac{\sum_{i=1}^N (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^N (y_i - \bar{y})^2}, \quad (17)$$

Where \bar{y} is the mean of the true labels in the validation/test set. The MAPE gives the spread of the errors in a relative percentage making it easier to understand the total error in predictions for each NN and compare the errors easily between NNs.

The NN is then used on the final test data set to see its final predictive abilities on a completely new, unseen data set. Note that the test set is only used once the NN has been fully optimised to give an unbiased measure of the NN's performance.

3.4. Optimising the neural network

The validation set is also used to tune and optimise the hyperparameters of the NN, optimising the performance of the NN further. Hyperparameters refer to variables that do not change while the NN is being trained, and are usually specified before training a NN. There are several hyperparameters that can be changed that help to improve the predictive abilities of the NN:

- *Number of neurons*: The number of neurons in the hidden layers helps with capturing complex relationships between the input and output data.
- *Number of layers*: The number of layers refers to the number of hidden layers in the NN and also helps capture complex relationships like the number of neurons
- *Learning rate*: Learning rate controls the rate at which the weight w and bias b parameters in the neurons are updated during training.
- *Epochs*: An epoch refers to the number of times the input data is passed through the NN to predict the output, and minimise the cost function.
- *Batch size*: Batch size refers to the number of samples from the training set that is used while training the NN.
- *Regularisation*: Regularisation is a term that is added to the cost function that helps to prevent the NN from over-fitting.
- *Output layer activation function*: The output layer activation function discussed in Section 3.1 can also be treated as a hyperparameter that is configured before the NN is trained.

3.5. TensorFlow implementation

For our project, we chose to code the NNs in Python as it offers a package that extensively supports NN creation, hyperparameter tuning, and hyperparameter visualisation. TensorFlow is an open source machine learning package that supports a wide array of machine learning applications including making NNs.

4. Methods

4.1. Data set creation

To create the data set for the NN to be trained, validated, and tested on, the following steps were used:

1. Values for each feature were pseudo-randomly picked from the ranges in Table 2 and the adsorbent database mentioned in Section 2.3 using Latin hypercube sampling to generate a uniform data set. p_L was always less than p_H as these ranges overlap and would not help the NN to learn how to accurately predict the KPIs.
2. The data set of operating points was evaluated by Ward's detailed adsorption model to produce the value of each KPI. The use of Imperial College London's High Computing Performance (HPC) was used which significantly reduced the evaluation time for each set of operating points. The data set was then filtered to remove any sets of operating points that exceeded the satisfactory ranges for each KPI shown in Table 3. After filtering, a data

KPI	Acceptable range
Productivity, Pr [mol/m ³ /s]	0 - 9
Energy usage, E_T [kWh/tonne]	0 - 5000
Purity, Pu_{CO_2} [%]	0 - 100
Recovery, Re_{CO_2} [%]	0 - 100
Capture cost, $C_{CO_2}^{cap}$ [\$/tonne]	0 - 600

Table 3: Acceptable range for each KPI, the ranges used to remove points outside the acceptable range as these values were outliers produced from the detailed model

Hyperparameter	Values
# of neurons	10, 20, 30, 40
# of layers	2, 4, 6, 8
# of epochs	500, 1000, 1500, 2000
Learning rate	0.1, 0.01, 0.001, 0.0001
Batch size	250, 500, 750, 1000
Regularisation	0.1, 0.01, 0.001, 0.0001

Table 4: Discrete values randomly picked for hyperparameter tuning

set of 11,454 operating points with their associated KPIs was created.

3. Finally, the features were normalised using Equation 7. 80% (9,163) of the total data set was used for training the NNs, 10% (1,145) for validating the NNs, and the last 10% (1,146) for testing the NNs.

4.2. Neural network training, validating, and testing

Using the data set, the NNs for each KPI were created and the general structure of these NNs can be seen in Figure 3.

To tune the hyperparameters discussed in Section 3.4, a technique known as random search was used. The hyperparameters were randomly selected from a range of discrete values in Table 4 500 times for each KPI, totalling 1500 NN models. All the NNs were then trained on the training set, optimising all the weight w and bias b terms in each neuron of every layer in the NN, and the performance was evaluated on the validation set using the RMSE (Equation 15). The set of hyperparameters with the lowest RMSE was selected as the final NN model hyperparameters for each KPI, narrowing down the 1500 NNs to just 5. Lastly, the MAPE (Equation 16) and R^2 (Equation 17) were then used to test the final performance of the NNs using the test set.

4.3. Optimisation Case Study

Three optimisation case studies were performed to validate the performance of our NNs against the performance of Ward's detailed model. The industry standard NSGA-II algorithm implemented in Python was used to perform these optimisation studies. The NSGA-II algorithm was initialised with 72 initial samples, running for 50 generations with 72 off-

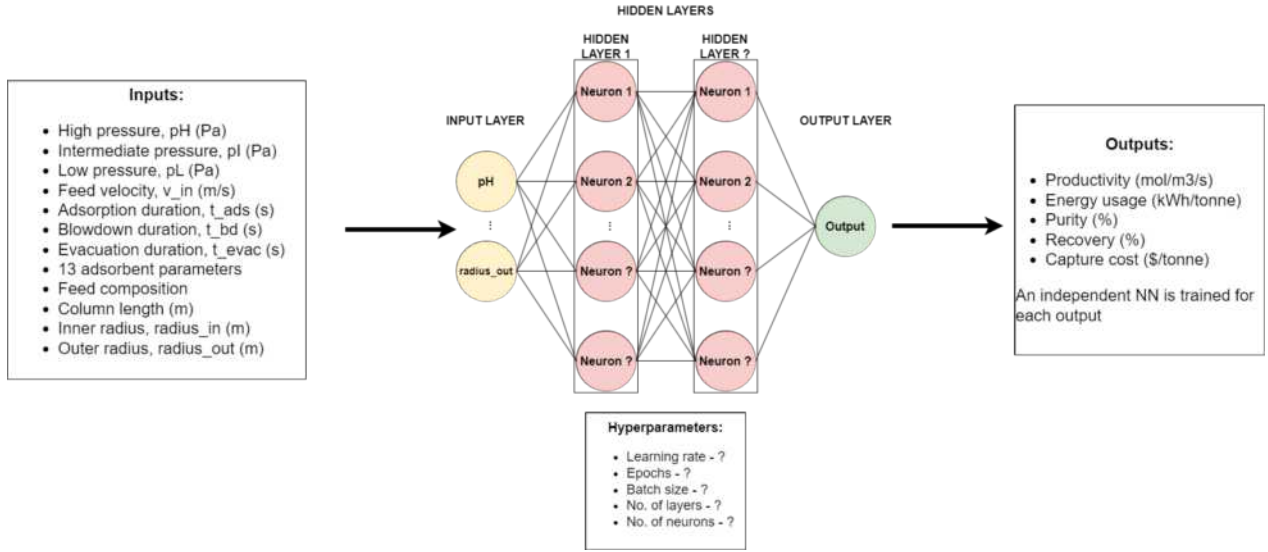


Figure 3: General NN structure for predicting the KPIs from the detailed model

springs per generation. The results were then compared with Ward's own cost optimisation results that used the Thompson Sampling Efficient Multiobjective Optimisation (TSEMO) algorithm.

4.3.1. Unconstrained purity/recovery optimisation

The first optimisation problem involved maximising both the purity and recovery of the PVSA process to check the feasibility of the chosen adsorbent. Post-combustion carbon capture is subject to regulatory requirements of at least 90% recovery and 95% purity for the captured CO₂, so performing this optimisation will reveal whether there exists operating points that satisfies both of these requirements. If not, the chosen adsorbent is considered infeasible, deeming the next two steps redundant and a new adsorbent would have to be chosen. The problem is formally defined as follows:

$$\min_{\theta} (F_1 = -\text{Pu}_{\text{CO}_2}, F_2 = -\text{Re}_{\text{CO}_2}) \quad (18)$$

$$\text{s.t. } \theta_L \leq \theta \leq \theta_U \quad (19)$$

Where Pu_{CO_2} and Re_{CO_2} are the purity and recovery of CO₂ respectively. F_1 and F_2 are the two objectives function that we wish to minimise. Note that minimising the negative of a function yields the same result as maximising the positive of the function. θ_L and θ_U are the lower and upper bounds of each feature respectively. The list of features that were varied along with their accompanying bounds are shown in Table 2. The adsorbent parameters were fixed for zeolite 13X shown in Table 1, while P_L and $y_{1,F}$ were fixed to 0.05 bar and 0.15 respectively to simulate the same optimisation problem that Ward's detailed model solved. Upon solving this multi-objective problem, a Pareto front is generated that represents the set of solutions with the best trade-off between the two

objective functions. If the Pareto front passes through the feasible region of $\text{Pu}_{\text{CO}_2} \geq 95\%$ and $\text{Re}_{\text{CO}_2} \geq 90\%$, then it can be concluded that the chosen adsorbent is feasible and the next optimisation case can be performed.

4.3.2. Constrained productivity/energy usage optimisation

A constrained productivity/energy usage optimisation is performed next. It is desired to maximise the productivity to achieve the highest CO₂ capture whilst using the least amount of adsorbent, reducing the length of each cycle. We also minimise the total energy consumed for each tonne of CO₂ captured to increase profitability and commercial attractiveness of the process. A set of constraints are also introduced to ensure that the set of solutions found by the algorithm will abide by the aforementioned purity and recovery requirements, and solving the multi-objective problem will once again yield a set of optimal Pareto solutions. The full optimisation problem is defined as follows:

$$\min_{\theta} (F_1 = -\text{Pr}, F_2 = E_T) \quad (20)$$

$$\text{s.t. } \theta_L \leq \theta \leq \theta_U \quad (21)$$

$$\text{Pu}_{\text{CO}_2} \geq 95\% \quad (22)$$

$$\text{Re}_{\text{CO}_2} \geq 90\% \quad (23)$$

A penalty function is implemented to impose the constraints and sums a large value into the objective function when a constraint is violated, making sure that the algorithm avoids such operating points and focuses more on points that do not violate the constraints. Thus, the problem is formulated as follows:

$$F_1 = -\text{Pr} + \phi_1 \quad (24)$$

$$F_2 = E_T + \phi_2 \quad (25)$$

$$\phi = \begin{cases} \max(0, 0.95 - \frac{Pu_{CO_2}}{100} + \max(0, 0.9 - \frac{Re_{CO_2}}{100})^2 \\ 0.25 \times \max(0, 95 - Pu_{CO_2}) + \max(0, 90 - Re_{CO_2})^2 \end{cases} \quad (26)$$

4.3.3. Capture cost optimisation

The final step of the case study involves solving a single-objective optimisation problem minimising capture cost of the process to screen out economically unviable adsorbents and is an important factor when choosing an adsorbent for the process. This step also signifies an advancement in current literature as there is little work using surrogate NN models to evaluate industrial-scale economic performance of a PVSA process. The cost optimisation problem also involves a penalty function which is applied when one or both of the constraints are violated. The purity and recovery constraints are identical as per previous problems:

$$\min_{\theta} C_{CO_2}^{cap} + \phi \quad (27)$$

$$\text{s.t. } \theta_L \leq \theta \leq \theta_U \quad (28)$$

$$\phi = 10 \times [\max(0, 95 - Pu_{CO_2}) + \max(0, 90 - Re_{CO_2})]^2 \quad (29)$$

5. Results

5.1. Optimised neural networks

Table 5 shows the optimised hyperparameters for the NNs of each KPI and the RMSE on the validation set. It may seem that the energy usage and capture cost NNs performance may be poor as the validation RMSE values are much larger than the NNs for productivity, purity, and recovery, but it should be noted again that the RMSE is scale dependent. Considering the range of acceptable values for each KPI from Table 3, the RMSE values are much smaller than the overall KPI ranges that the NNs predict over.

Table 6 shows the RMSE, MAPE, and coefficient of determination R^2 of each NN on the test set. Again, the ability to interpret the performance of the NNs using the RMSE is difficult without knowing the scales of the KPIs, so the MAPE and R^2 offer better understanding of each NN's predictive abilities. The MAPE of each NN is below ranges from 6 - 14% while the R^2 of the NNs are greater than 0.9 suggesting that the NN's can accurately predict each KPI.

5.2. Optimisation case study

5.2.1. Unconstrained purity/recovery optimisation

The results of the optimisation study are compared with Ward and Pini's study in Figure 4. The blue triangles show the Pareto front obtained from our NN coupled with the NSGA-II algorithm whereas the orange triangles show the optimisation results obtained by Ward and Pini using the TSEMO algorithm coupled with their detailed model. The black dotted lines represent the boundaries of the purity and recovery constraints described in Section 4.3.2, and the red shaded region represents the area of uncertainty of our model based on the MAPE from Table 6. The Pareto front generated by our NN model is in good agreement with Ward and Pini's Pareto front following the same trends. It can also be concluded that

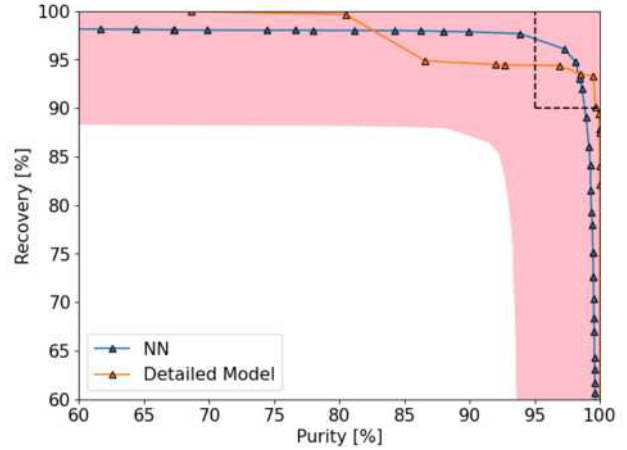


Figure 4: Unconstrained purity/recovery Pareto fronts for Zeolite 13X at a p_L of 0.05 bar.

Zeolite 13X is a feasible adsorbent for this process as there are several points that satisfy the process requirements. More importantly, using the NN model proved to be monumentally quicker in terms of computational time as it took roughly 50s for the algorithm to finish running whereas the detailed model took around 23-24hrs. This represents a 1600x reduction in computing time and demonstrating the great potential of this approach.

5.2.2. Constrained productivity/energy usage optimisation

The results of the constrained productivity/energy usage optimisation step are shown in Figure 5. Again, there is good agreement from the two Pareto fronts and both exhibit similar shapes, with our NN exhibiting vastly shorter computational times: around 50s compared to the 350hrs taken for the detailed model (Ward and Pini, 2022).

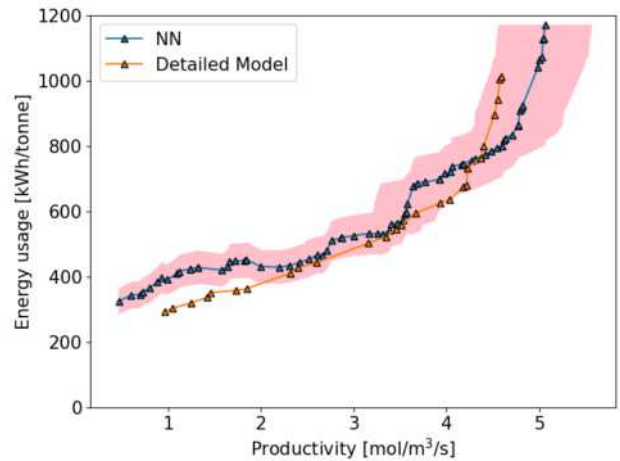


Figure 5: Unconstrained purity/recovery Pareto fronts for Zeolite 13X at a p_L of 0.05 bar

KPI	Hyperparameter							Validation RMSE
	Neurons	Layers	Epochs	Learning rate	Batch size	Regularisation	Output activation function	
Productivity, Pr [mol/m ³ /s]	40	8	2000	0.001	250	0.001	ReLU	0.09
Energy usage, E_T [kWh/tonne]	40	4	500	0.1	750	0.01	ReLU	258
Purity, Pu _{CO₂} [%]	40	6	2000	0.001	500	0.0001	Sigmoid	4
Recovery, Re _{CO₂} [%]	40	8	1500	0.0001	250	0.0001	Sigmoid	5
Capture cost, $C_{CO_2}^{cap}$ [\$/tonne]	40	8	1500	0.01	750	0.1	ReLU	25.9

Table 5: Optimised hyperparameters for each neural network with validation set RMSE of each KPI

KPI	Test RMSE	Test MAPE [%]	Test R ²
Productivity, Pr [mol/m ³ /s]	0.10	10	0.987
Energy usage, E_T [kWh/tonne]	258	12	0.943
Purity, Pu _{CO₂} [%]	4	6	0.981
Recovery, Re _{CO₂} [%]	5	10	0.972
Capture cost, $C_{CO_2}^{cap}$ [\$/tonne]	27.6	14	0.906

Table 6: Test set RMSE, MAPE, and R² for each KPI's neural network

Output	Neural Network w/ NSGA2	Detailed Model w/ TSEMO
Capture cost [\$/tonne]	41 ± 5	45.80
Productivity [mol/m ³ /s]	1.18 ± 0.1	1.27
Energy usage [kWh/tonne]	753 ± 75	693
Purity [%]	94.9 ± 6	96.5
Recovery [%]	90.1 ± 9	89.3

Table 7: Constrained capture cost optimisation results

5.2.3. Capture cost optimisation

The results from the single-objective problem of minimising capture cost are shown and compared to Ward and Pini's results in Table 7. Again the KPIs are similar to those of Ward and Pini's showing that the NN can be used effectively to screen adsorbent based on their techno-economic performance. It is also noted that running this optimisation only took around 57s exhibiting exponentially faster computational times with an acceptable level of error.

6. Discussion

6.1. Comparison to MAPLE framework neural networks

To compare our NNs with the MAPLE NNs from [Pai et al. \(2020\)](#), we use the adjusted coefficient of determination R_{adj}^2 :

$$R_{adj}^2 = 1 - \frac{(1 - R^2)(N - 1)}{(N - k - 1)} \quad (30)$$

where N is the number of samples in the test set, and k is the number of features. The R_{adj}^2 can be considered to be a better predictor of accuracy for our NN models as it also takes into consideration the number of features that go into the model, while penalising features that may not contribute as much. While the productivity, purity, and recovery all have R_{adj}^2 greater than 0.94, the energy usage and capture cost NNs have R_{adj}^2 values lower than 0.90. It should be noted that the training set used on the NNs was only 9163 points, so a larger data set would increase the accuracy of the capture cost and energy usage NNs, while also improving the productivity, recovery, and purity NNs. The R_{adj}^2 also questions whether some features in the data set may not be contributing to the performance of the NNs. Two such features are the column radii, r_{in} and r_{out} as these are dependent on the column length L based on Equations 5 and 6. These three variables have the same value when normalised so the NNs do not get any further useful information from r_{in} and r_{out} , introducing extra complexity to the NNs without improving the performance of them. Therefore, it is important to feed the NNs independent features that provide useful information related to the labels, and to verify how the features themselves are related to each other

KPI	R^2_{adj}
Productivity	0.976
Energy usage	0.887
Purity	0.962
Recovery	0.945
Capture cost	0.817

Table 8: Adjusted coefficient of determination for each neural network

so as to not introduce redundant features.

Finally, Table 9 shows a summarised list of comparisons with the MAPLE NNs. First, the number of inputs that can be specified for our NN is more than 2x the number of inputs the MAPLE NNs use allowing for greater flexibility in defining the PVSA process for a wide range of adsorbents.

Second, we employ the DSL isotherm model instead of the SSL model allowing us to better model the equilibrium between CO_2 and N_2 better while also modelling more complex adsorbents. Although this may introduce more complexity into the NNs, this allows our NNs to better represent a wider range of adsorbents giving it much greater applicability to screen various adsorbents for various operating conditions.

Third, we build on the simple sensitivity analysis that was performed for the MAPLE NNs that only looked at optimising the number of samples used in the training set and number of neurons in each layer to improve R^2_{adj} . We instead optimise multiple hyperparameters essential to the configuration of each NN. However, the optimal number of neurons shown in Table 5 is also the upper bound of discrete values it could take, suggesting that exploring a higher number of neurons could further increase the accuracy of the NNs by modelling any complex relationships between the features and true labels better.

Last, we compare the R^2_{adj} of our NN with the MAPLE framework NNs which shows that our NNs have a lower accuracy than the MAPLE NNs. However, the complexity of our NNs is much greater and they were trained on a smaller training set than the MAPLE NNs, so a much larger data set would be required to reach the accuracy of the MAPLE NNs. Nevertheless, the NNs developed over the course of the project perform very well considering the complexity of the NNs and the small training set size.

6.2. Reduction in computational time

It is clear through the results of this study and through others such as [Pai et al. \(2020\)](#) that utilising NNs for optimisation of an adsorption process require significantly less computational time than incumbent methods that use detailed process modelling. Figure 6 shows that our NNs have the capability to solve optimisation problems tens of thousands of times faster than a detailed model can with similar outcomes and accuracy. As such this tool has great potential to vastly shorten computational efforts in adsorption process optimisation problems, feasibility studies and large-scale adsorbent screening. The

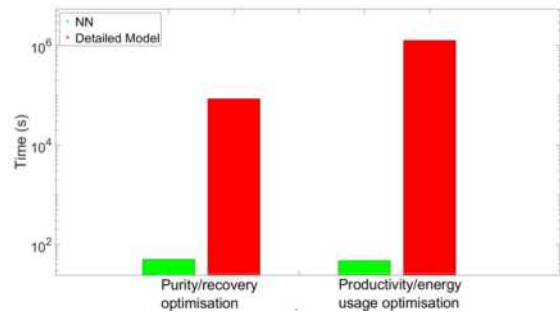


Figure 6: Single core CPU times of our NN vs the detailed model

uncertainties detailed in 6 are reasonable and comparable to that described in [Cleeton et al. \(2022\)](#), suggesting that experimental uncertainties that arise from adsorption isotherm parameters being determined empirically will influence the uncertainties of the Pareto front. Hence, we conclude that the uncertainties in our NN predictions is a potential point of improvement - through the use of more training data or by further hyperparameter tuning. However, there will always exist some intrinsic uncertainties that are outside of the scope of this study, so it is important to not be overly analytical about the nominal values that we obtain but rather focus on the validity and great potential of the approach.

7. Conclusion

In conclusion, this research has demonstrated the significant advantages of utilising data-driven NN models in optimising PVSA processes for carbon capture. Our study, building on the MAPLE framework, shows that NNs can offer rapid assessments of PVSA processes with accuracy comparable to that of a traditional process model. The model is a significant leap forwards in terms of its ability to predict the economic aspect of carbon capture, specifically the capture cost of CO_2 . Similarly it represents an improvement from the MAPLE framework through its ability to capture more complex adsorption behaviours through the use of the dual-site Langmuir model and the inclusion of more operating parameters.

Our models show acceptable uncertainty levels of around 14% at the highest, comparable to other similar works. These uncertainties have potential for improvement through the use of more training data or further optimisation of the NNs hyperparameters.

In light of these findings, it is evident that ML, especially NN, can play a pivotal role in advancing carbon capture technologies. By significantly reducing the time and computational resources required for process optimisation, this tool opens up potential for rapid screening of large adsorbent databases and will surely be useful for the deployment of effective carbon capture solutions.

References

Agarwal, A., Biegler, L.T., Zitney, S.E., 2010. Superstructure-based optimal synthesis of pressure swing adsorption cy-

NNs developed from our project	MAPLE framework NNs
Takes 24 input features and outputs 5 KPIs (Pr, E_T , Pu_{CO_2} , Re_{CO_2} , $C_{\text{CO}_2}^{\text{cap}}$)	Takes 10 inputs and outputs 4 KPIs ((Pr, E_T , Pu_{CO_2} , Re_{CO_2})
Use of DSL isotherm to model CO_2/N_2 equilibrium	Use of SSL isotherm to model CO_2/N_2 equilibrium
Several hyperparameters were optimised for each NN using random grid search	Sensitivity analysis completed on number of training samples and number of neurons in MAPLE NNs
NNs have a range of 0.8 - 0.95 for R_{adj}^2 for a training set of around 9200 data points	MAPLE NNs have a $R_{\text{adj}}^2 \geq 0.995$ with 10000 data points in training set

Table 9: Comparison of our NNs with MAPLE framework NNs (Pai et al., 2020)

- cles for precombustion co2 capture. Industrial & Engineering Chemistry Research 49, 5066–5079.
- Allen, M., Dube, O., Solecki, W., Aragón-Durand, F., Cramer, W., Humphreys, S., Kainuma, M., et al., 2018. Special report: Global warming of 1.5 c. Intergovernmental Panel on Climate Change (IPCC) .
- Burns, T.D., Pai, K.N., Subraveti, S.G., Collins, S.P., Krykunov, M., Rajendran, A., Woo, T.K., 2020. Prediction of mof performance in vacuum swing adsorption systems for postcombustion co2 capture based on integrated molecular simulations, process optimizations, and machine learning models. Environmental science & technology 54, 4536–4544.
- Capra, F., Gazzani, M., Joss, L., Mazzotti, M., Martelli, E., 2018. Mo-mcs, a derivative-free algorithm for the multiobjective optimization of adsorption processes. Industrial & Engineering Chemistry Research 57, 9977–9993.
- Cleeton, C., Farmahini, A.H., Sarkisov, L., 2022. Performance-based ranking of porous materials for psa carbon capture under the uncertainty of experimental data. Chemical Engineering Journal 437, 135395.
- Haghpahan, R., Majumder, A., Nilam, R., Rajendran, A., Farooq, S., Karimi, I.A., Amanullah, M., 2013. Multiobjective optimization of a four-step adsorption process for postcombustion co2 capture via finite volume simulation. Industrial & Engineering Chemistry Research 52, 4249–4265.
- Kingma, D.P., Ba, J., 2017. Adam: A method for stochastic optimization. [arXiv:1412.6980](https://arxiv.org/abs/1412.6980)
- Leperi, K.T., Chung, Y.G., You, F., Snurr, R.Q., 2019. Development of a general evaluation metric for rapid screening of adsorbent materials for postcombustion co2 capture. ACS Sustainable Chemistry & Engineering 7, 11529–11539.
- Pai, K.N., Prasad, V., Rajendran, A., 2020. Generalized, adsorbent-agnostic, artificial neural network framework for rapid simulation, optimization, and adsorbent screening of adsorption processes. Industrial & Engineering Chemistry Research 59, 16730–16740. URL: <https://doi.org/10.1021/acs.iecr.0c02339>, doi:10.1021/acs.iecr.0c02339, arXiv:<https://doi.org/10.1021/acs.iecr.0c02339>
- Riboldi, L., Bolland, O., 2017. Overview on pressure swing adsorption (psa) as co2 capture technology: state-of-the-art, limits and potentials. Energy Procedia 114, 2390–2400.
- Ritter, J.A., Bumiller, K.C., Tynan, K.J., Ebner, A.D., 2019. On the use of the dual process langmuir model for binary gas mixture components that exhibit single process or linear isotherms. Adsorption 25, 1511–1523.
- Ruthven, D.M., Farooq, S., Knaebel, K.S., 1996. Pressure swing adsorption. John Wiley & Sons.
- Subraveti, S.G., Li, Z., Prasad, V., Rajendran, A., 2019. Machine learning-based multiobjective optimization of pressure swing adsorption. Industrial & Engineering Chemistry Research 58, 20412–20422.
- Tong, L., Zhang, P., Yin, S., Zhang, P., Liu, C., Li, N., Wang, L., 2018. Waste heat recovery method for the air pre-purification system of an air separation unit. Applied Thermal Engineering 143, 123–129. URL: <https://www.sciencedirect.com/science/article/pii/S1359431117337316>, doi:<https://doi.org/10.1016/j.applthermaleng.2018.07.072>
- Ward, A., Pini, R., 2022. Efficient bayesian optimization of industrial-scale pressure-vacuum swing adsorption processes for co2 capture. Industrial & Engineering Chemistry Research 61, 13650–13668. URL: <https://doi.org/10.1021/acs.iecr.2c02313>, doi:10.1021/acs.iecr.2c02313, arXiv:<https://doi.org/10.1021/acs.iecr.2c02313>
- Wilkins, N.S., Rajendran, A., 2019. Measurement of competitive co2 and n2 adsorption on zeolite 13x for postcombustion co2 capture. Adsorption 25, 115–133.

Cold Chain Integration of Liquefied Natural Gas Supply Chains

Nicole Godfrey and Qing Li Wei

Department of Chemical Engineering, Imperial College London, U.K.

Abstract: Liquefied natural gas (LNG) has grown to account for a much larger part of global natural gas (NG) supply chains in recent years with Europe's demand skyrocketing in 2022. With the onset of the Russia-Ukraine war and increasing geopolitical tensions globally, sourcing energy from further supply regions to ensure reliable energy supply has become more common. LNG is a highly efficient and economical way of transporting energy across vast distances especially when there is lack of access to traditional gas pipelines. Current processes fail to utilise the cold energy of LNG with roughly 13% energy equivalent of the gas consumed throughout the supply chain. Studies have mainly focused on cold energy extraction during regasification of LNG. This paper introduces an integrated LNG supply chain that reduces energy requirements by 29.7% and carbon emissions by 30.8%. The economics of this proposed design is analysed and assessed against current industry leading processes. The results show a 29.1% reduction in capital expenditure and 20.1% reduction in operating expenditure. Additional income is generated through the integration of the LNG supply chain with a valuable cryogenic process.

Keywords: Natural Gas, Liquefied Natural Gas, Cold Energy, Supply Chain Integration, Air Separation

Introduction:

Natural gas is considered the lowest carbon content fossil fuel with approximately 52.91 kg CO₂/MMBtu compared to coal and oil which has an average CO₂ content of 95.92 and 70.07 kg CO₂/MMBtu respectively^[1]. Natural gas composition can vary depending on region but typically consists of 85-95%^[2] methane with small proportions of ethane, propane, butane and nitrogen. Since natural gas is primarily methane, the calorific value of natural gas is therefore greater than other fossil fuels at roughly 50-55 MJ/kg^[3]. This makes natural gas an ideal fuel for electricity power generation compared to conventional coal fired power plants. Ultimately natural gas is a fossil fuel, its impact on climate change and overall carbon emissions is considerably greater than that of renewable energy alternatives. However, the benefits of increasing natural gas as a proportion of a country's energy composition especially in the mid-term is undeniable. Natural gas has the potential to play a pivotal role in enabling a smoother transition to a greener economy as it is a reliable and abundant source of energy. In 2021 the global proven natural gas reserve stood at 188.1 trillion cubic metres^[4] with Russia, Iran and Qatar accounting for half of global reserves.

Regions	Production	Consumption
North America	1203.9	1099.4
S. & Cent America	162.0	161.7
Europe	220.4	498.8
CIS	805.9	551.2
Middle East	721.3	560.6
Africa	249.0	162.5
Asia	681.3	907.1
World	4043.8	3941.3

Table 1: Table showing production and consumption of natural gas in billion cubic metres by region in 2022^[5]. Notably, North America, CIS and Middle East are the main supply regions. Europe and Asia are the main demand regions. This imbalance in supply and demand of energy product is what drives international trade of natural gas.

The UK is a large consumer of natural gas with approximately 72 bcm consumed in 2022^[5]. Prior to the invasion of Ukraine in 2022, most natural gas imports to

continental Europe and the UK were primarily through traditional gas pipelines specifically from Russia. This reliance caused natural gas prices to spike to €300/MWh in August 2022 compared to 2020 levels of \$30/MWh^[6]. In response to wean itself from Russian gas, the UK sought to find alternatives to traditional pipeline natural gas. One alternative was to source energy supplies from further afield in the form of liquefied natural gas (LNG). LNG is natural gas that has been cooled to approximately -161°C at atmospheric pressures which turns natural gas into the liquid state. LNG has significant advantages compared to traditional natural gas as its volume is approximately 1/600th of natural gas. It is an efficient method of transporting energy across vast distances where pipeline options are not available or are not economical. However, current technologies employed in industry have significant energy requirements. Typically, natural gas is consumed to provide energy. Roughly 13% of the energy of LNG is consumed during the liquefaction, shipping and regasification process^[7].

The UK increased LNG imports in 2022 by 74% from the previous year to 25.6 bcm, which accounted for 45% of total natural gas imports in the year^[8]. The United States, Qatar and Peru are the primary import sources of LNG to the UK importing 50%, 30% and 8% respectively in 2022. The UK has the second largest LNG regasification infrastructure in Europe after Spain with sites located in Milford Haven and Isle of Grain. One prominent trade route and the basis of this study is the Qatargas 2 Ras Laffan LNG terminal delivering LNG to South Hook LNG terminal in Milford Haven.

The global demand for LNG has surged in recent years with demand maintaining an average growth rate of 5.3% between 2012-2022^[5]. Most of this demand has been from Asian countries such as Japan and China who are either lacking in domestic natural resources or are trying to move away from traditional fossil fuels which are more polluting. Even when considering improvements in existing energy technologies, global energy demand is projected to grow by 15% by 2050^[9]. Since the commencement of the Paris agreement in 2016 many countries have had greenhouse

gas emission targets at the forefront of their minds. One method of meeting these targets is to increase the usage of LNG which will ultimately drive demand for LNG and funnel investments into technological developments. The primary drivers for LNG growth are liquefaction and regasification capacity, rising energy consumption fuelled by a growing population and fast approaching emissions target milestones. To confront these challenges face on, optimisation of the individual parts of the LNG supply chain as well as utilisation of cold energy of LNG have been analysed. Developing an integrated supply chain has the ability to further enhance energy efficiency. The potential to integrate other value generating processes with the LNG supply chain to produce marketable products is also analysed.

Background:

Current LNG supply chains consist of a liquefaction site at the region of production where LNG is then loaded onto a carrier bound for the importing country. Upon arrival the LNG is piped off the carrier and into a regasification site. The LNG is evaporated back into natural gas and configured to specific country grid specifications before delivery to end customers. Significant infrastructure is required to facilitate this supply chain and so ensuring minimal energy loss during these processes is of great importance.

Different liquefaction processes are used in industry today with APCI accounting for the majority of the market. There are typically three main liquefaction processes: pure refrigerant cascade, mixed refrigerant and nitrogen expander cycles. The most widely used process is APCI - C3MR by Air Products which utilises a propane precooling loop followed by a mixed refrigerant to provide cooling. This process accounts for 41% of the world's installed capacity in 2018^[10]. The Conoco Philips optimised cascade process is the second most utilised process accounting for 22% of installed capacity globally. This process utilises a pure refrigerant three stage cascade process. The technology of focus for this paper and the current leader in terms of LNG production capacity is the AP-X process which utilises a propane precooling loop followed by a mixed refrigerant loop to provided cooling and a nitrogen expander loop to enable subcooling of the natural gas. This technology has a maximum capacity of 10Mtpa per train and the first deployment of this technology was at the Qatargas 2 Ras Laffan site, where two trains produce a combined capacity of 15.6Mtpa. This site is part of an integrated energy supply chain connecting Qatar's north field gas fields to the UK. It is serviced by a fleet of 14 Q-Max and Q-Flex LNG carriers which can hold 266,000m³ and 216,200m³ respectively^[11]. At the regasification site, LNG can be evaporated back into natural gas through two common methods. Open rack vaporisers can utilise sea water to raise the temperature of LNG. This method however requires a slightly elevated sea temperature which the UK does not have. The South Hook LNG terminal uses a submerged combustion

vaporiser which uses fuel gas to heat up water in order to raise the temperature of LNG. The entire process from liquefaction to regasification consumes roughly 13%^[7] of the total available energy with liquefaction consuming 10% of the energy. Shipping and regasification both consume roughly 1.5% of the equivalent gas energy.

Process	Technology
Liquefaction	AP-X
Shipping	Q-Max/Q-Flex
Regasification	Submerged Combustion Vaporiser

Table 2: Table showing specific technologies employed at each stage of the Qatar to UK LNG supply chain.

One method of reducing this energy consumption is by utilising the LNG cold energy for other valuable processes that require a high degree of cooling or for cold storage facilities. Significant research has been conducted on the utilisation of cold energy of LNG at the point of delivery. The first ideas stemmed from Nakaiwa et al. where a system to integrate LNG delivery with an air separation unit combined with a conventional LNG combined cycle was studied. The research found that thermal efficiency of the power generator could be increased from 40% to 46.8%^[12] through the introduction of oxy-fuel combustion and that the CO₂ emissions would be reduced by 13%. Research conducted by Xiong and Hua analysed a cryogenic air separation unit utilising LNG cold energy to produce liquid nitrogen and oxygen gas. Their study found that LNG cold energy could be fully utilised when recycled nitrogen was compressed to above 6.5MPa^[13]. This paper showed that utilisation of LNG cold energy was not only technically and economically feasible but also decreases energy consumption compared to conventional cryogenic air separation processes by more than half. Other studies such as the one done by Shingan et al used exergy analysis and parametric optimisation to design an optimal air separation process integrated with LNG regasification^[14] but did not take into account the natural gas delivery requirements for importing countries which would result in more energy needed to vaporise the LNG.

Prior studies primarily focused on extracting cold energy utility from the regasification side but missed the opportunity to reduce energy consumption throughout the whole LNG supply chain. By integrating an air separation unit with LNG regasification to produce nitrogen and oxygen, the nitrogen product stream can be used as a medium to transport cold energy back to the liquefaction site. Liquid nitrogen can then be used to support the cooling of natural gas during liquefaction. Nitrogen is easier to transport in its liquid state as it occupies roughly 1/700th of the volume as gaseous nitrogen. The benefit of doing this is that it will reduce energy requirements on the liquefaction side which typically contributes the most to total energy consumed in LNG production (10%^[7]) and also ensure full utilisation of shipping vessels on the return journey. Once the cold energy has been utilised, the ambient nitrogen can be sold as feedstock into various industries such as fertiliser, mining, metals and electronics.

With Qatar being located in the middle east, many channels for trade and shipping are available.

The oxygen on the other hand can be sold to natural gas-powered power plants utilising a combined cycle gas turbine (CCGT). Oxygen can be used for oxy-fuel combustion which enables the fuel to combust at greater temperatures which results in greater energy efficiency. This also eliminates the need for amine scrubbing which is a notoriously expensive process. The fuel burns cleaner than in air with the flue gas containing primarily CO₂ and H₂O which can be separated relatively easily. The CO₂ can be capture and stored using existing carbon capture and storage (CCS) technologies such as sequestration in saline aquifers. A potential site for such an integrated supply chain is Teesside in the northeast of England where clusters of industries and businesses utilising CCS technology are already located.

Oxygen can also be sold to highly polluting industries such as the cement manufacturing industry which accounts for approximately 7% of total global emissions^[15]. This enables further value extraction from the LNG supply chain and creates an integrated supply chain that is capable of producing marketable products and provides an additional source of income. Figure 1 below shows the entire value chain and proposed integrated supply chain. Previous works have focused on extracting cold energy utility of LNG from the regasification side to increase energy efficiency. However, there has been little analysis around the costs associated with implementing these proposed designs. There has also been little consideration into the additional income generated through introduction of an air separation unit. This is vital if organisations and businesses were to actually consider pursuing this cold energy supply chain integration. This is the primary objective of this study. An analysis of costs associated with implementation as well as energy and emissions intensities will be quantified through assessment against leading industry processes.

Methodology:

Following the main objective of this project, current industry processes were selected and recreated in Aspen Plus V11 to analyse the performance of the existing supply chain. Aspen is generally good at modelling hydrocarbons as well as heat exchange, compression and expansion processes. The Peng-Robinson thermodynamic equation of state was used to calculate thermodynamic properties since it was designed to calculate fluid properties for all natural gas processes. The LNG supply chain is characterised by three main stages:

- Liquefaction: AP-X process consisting of a pre-cooling propane loop, mixed refrigerant refrigeration cycle and a nitrogen expander loop.
- Transportation: QatarEnergy Q-Max LNG vessel.
- Regasification: Submerged combustion vaporiser.

In addition, a conventional cryogenic separation process was modelled and analysed in order to provide a basis for comparison with the integrated supply chain. Energy requirements for each process and the costs associated with them were modelled using Aspen's built in economic analyses functions in tandem with calculations performed utilising reported numbers from market insight and industry. Both the standalone and integrated process were designed to produce and deliver a capacity of 15.6Mtpa of LNG to the UK at the UK's natural gas delivery requirements. Table 3 shows the assumptions for various parts of the supply chain as well as different components modelled with their respective operating parameters. Equation 1 was used to carrying out energy balances around control volumes within the processes to calculate unknown parameters. Where \dot{Q}_{CV} denotes heat supplied, \dot{W}_{CV} denotes work done, \dot{m} refers to the mass flowrate and h refers to the enthalpy of component entering and leaving the control volume.

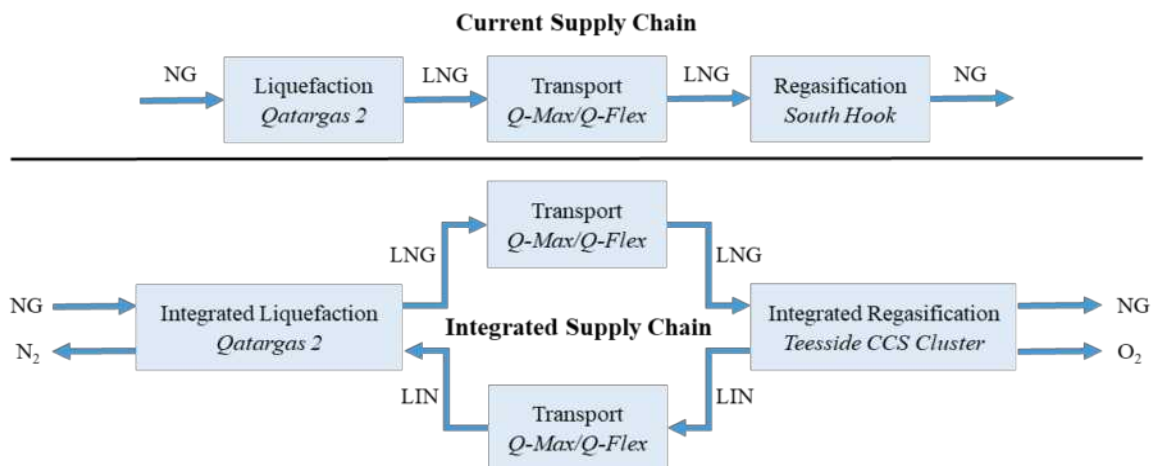


Figure 1: Diagram showing integrated LNG supply chain versus current supply chain. The integrated supply chain includes an additional air separation unit at the regasification site which produces oxygen and nitrogen products. Oxygen can be sold domestically whilst nitrogen is transported back to the liquefaction site.

$$0 = \dot{Q}_{CV} + \dot{W}_{CV} + \sum \dot{m}_{in} h_{in} - \sum \dot{m}_{out} h_{out} \quad (1)$$

Assumptions

Liquefaction	Values
NG feed temperature	43°C
NG feed pressure	65 bar
LNG outlet temperature	-166°C
LNG outlet pressure	1.05 bar
HX min approach temperature	10°C
Isentropic efficiency of compressor (C3 cycle)	78%
Isentropic efficiency of compressor (MR cycle)	75%
Isentropic efficiency of compressor (N2 cycle)	85%
Isentropic efficiency of expander (N2 cycle)	85%
Mechanical efficiency	90%
Transportation	
Boil rate per day	0.15 %/vol
Average transit time	18 days
Regasification	
HX min approach temperature	10°C
Pump outlet pressure	88.8 bar
Delivery temperature of NG	10°C
Air separation	
Air inlet temperature	30°C
N2 rich stream outlet temperature	-196°C
O2 rich stream outlet temperature	-191°C
LPC pressure	5.8 bar
HPC pressure	1.3 bar

Table 3: Table showing process design assumptions used in the modelling for each part of the supply chain within Aspen Plus V11. These assumptions are assumed across both standalone and integrated designs for fair comparison.

Liquefaction

Natural gas enters a liquefaction plant after being pre-treated through the removal of hydrogen sulphide, slug and mercury. Carbon dioxide, water and heavy hydrocarbons are also removed before the liquefaction process as these impurities can cause freezing issues. The AP-X process is known for being the liquefaction technology with the greatest production capacity. It is comprised of three cycles: a propane precooling cycle; a mixed refrigerant liquefaction cycle and a nitrogen expansion cycle for subcooling. Ambient natural gas is pre-cooled to -30°C before entering the mixed refrigerant loop. The phase change of propane at different pressures provides the cold energy to cool down the natural gas and mixed refrigerant. Once pre-cooled the natural gas is heat exchanged with the cold mixed refrigerant stream via two heat exchangers until -110°C. A Brayton gas expansion cycle utilising nitrogen is then used to enable subcooling of the natural gas to -166°C before going to a flash separator to remove any boil off gases. The use of nitrogen improves efficiency and greatly increases capacity compared to its predecessor, the C3MR process. In addition, the use of nitrogen also reduces the required flowrates of the other refrigerants.

The Qatargas 2 site has a production capacity of 15.6Mtpa of LNG. Using a cost assumption of \$482/tonne per annum of LNG^[16], the capital expenditure to construct the site can be calculated. The operating cost for the liquefaction site can be calculated through summation of

utility duty which includes cooling, heating and electricity. The industrial electricity costs in Qatar were reported to be 0.07 QAR/kWh^[17]. Water and steam utility pricing was assumed to be 2.12×10^{-7} \$/kJ and 1.9×10^{-6} \$/kJ respectively based on Aspen's default energy pricing. Other significant costs considered included operation and maintenance costs which was assumed to have a yearly cost equal to 3% of total capital expenditure. Equation 2 shows how the energy intensity of the process can be calculated through dividing the total energy consumption TE by the mass flowrate of liquified natural gas \dot{m}_{LNG} . Total energy consumption for the process is simulated by Aspen.

$$Energy\ Intensity\ \left[\frac{kWh}{t_{LNG}} \right] = \frac{TE}{\dot{m}_{LNG}} \quad (2)$$

Using equation 3 the carbon intensity associated with the duty was calculated through using net equivalent mass of carbon dioxide emissions and the mass of LNG) involved. E_i is the CO₂ emission intensity of utility i , \dot{m}_{bp} is the net mass flowrate of by-product stream and E_{bp} is the CO₂ emission intensity of by-product stream.

$$Carbon\ Intensity\ \left[\frac{tCO_2e}{t_{LNG}} \right] = \frac{\sum E_i Q_i + E_{bp} \dot{m}_{bp}}{\dot{m}_{LNG}} \quad (3)$$

The emission intensity of electricity generation in Qatar is reported to be 489.87g of CO₂e/kWh^[18] for gas generated electricity. Water and steam carbon emissions can be calculated using the carbon tracking function in Aspen. This function uses the recommended US-EPA-Rule which assumes natural gas as the fuel source and 8000 operating hours per year^[19].

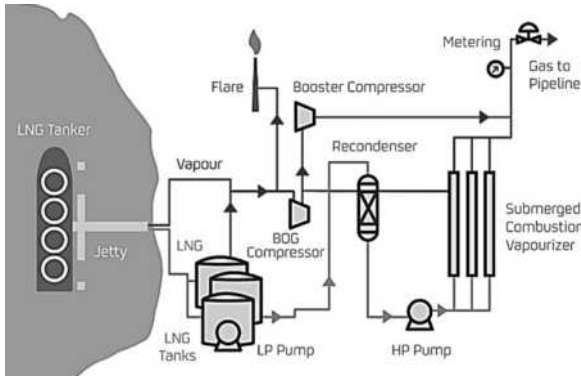
Transportation

LNG is predominantly transported via oceanic shipping routes via large LNG carriers capable of withstanding the cryogenic temperatures of the product. The tanks on these carriers must be heavily insulated and capable of maintaining the LNG at -162°C to prevent unnecessary boil off gas. Most do this via onboard reliquefaction systems which condenses the boil off gas back into LNG. The boil off gas can also be used as fuel to provide ship propulsion. The Q-Max and Q-Flex vessels used by QatarEnergy are membrane type LNG carriers capable of transporting up to 266,000m³ and do not use boil off gas as fuel to limit energy loss from transportation. These vessels can be modelled as large storage containers with a constant rate of boil off, which has been shown to be approximately 0.1-0.15% volume per day of the tank's capacity^[20]. The typical cost of one of these vessels is \$1,500 per m³ to construct^[21]. The average daily freight charge is assumed to be \$112,000, based on industry averages, and a charge of \$500,000 for passage through the Suez Canal. Other costs associated with operations were not considered for simplification. The average transit time between Qatar and the UK is 18 days^[22]. The average

carbon intensity of these vessels arriving in the UK is reported to be 79 kgCO₂/boe^[23].

Regasification

Upon arrival, LNG is pumped into a regasification plant where LNG is vapourised back into natural gas via heat exchange typically with sea water in an open rack vaporiser. In south hook, the regasification site of focus, sea water temperatures are too low for open rack vaporisers to be deployed. LNG throughput would need to be throttled significantly for effective heat transfer.



Instead, submerged combustion vaporiser typically powered by natural gas is used to heat up a water bath that evaporates the LNG.

Figure 2: Schematic showing a standard regasification facility utilising submerged combustion vaporiser to convert LNG back into natural gas^[24].

This results in roughly 1.5%^[7] of the fuel being consumed during regasification. This stage of the process is also the point in which the natural gas is configured to the delivery specifications of the receiving country. In the case for the UK, natural gas outlet temperatures should be 1-38°C and must not exceed maximum operating pressures at the delivery point^[25]. The capital expenditure for the site in south hook was reported to cost \$1bn to construct^[26]. Yearly operational and maintenance expenses can be estimated based on similar assumptions made for the liquefaction process. The cost of electricity utility is calculated to be \$0.18/kWh assuming power to be generated on site using a gas turbine generator with a 49% thermal efficiency^[27]. The average UK gas price in 2023

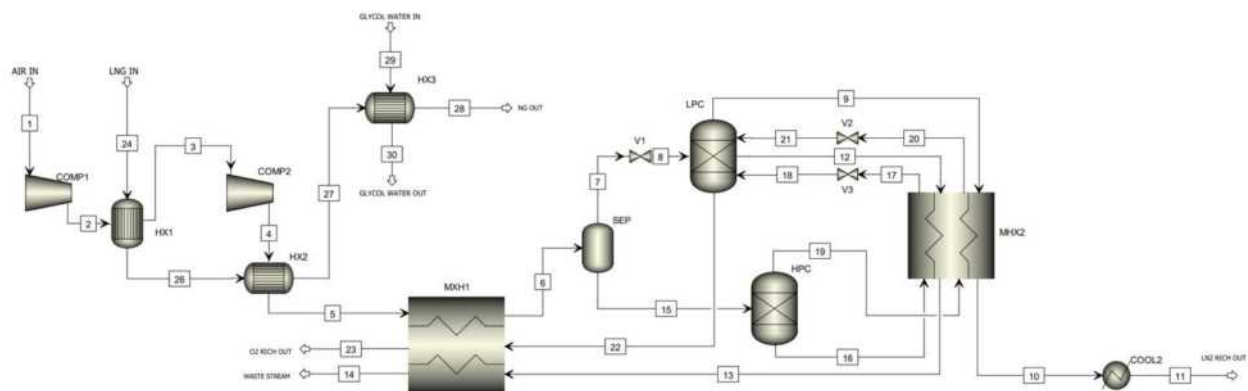
is £0.07/kWh^[28]. Carbon intensity associated with electricity consumption can be calculated using the UK's grid emission factor in the manufacturing sector reported to be 0.21kg CO₂e/kWh^[29].

Cryogenic air separation

In a conventional cryogenic air separation unit, air is first compressed to the desired pressure before being purified of any dust, water and oxides that may cause freezing problems. The air stream is then cooled and separated to form an oxygen enriched stream and a nitrogen enriched stream. The oxygen rich stream then gets passed through an expander and fed into a low-pressure distillation column. The nitrogen rich stream is distilled in a high-pressure distillation column and then flows into the low-pressure distillation column to form a high purity nitrogen product stream. This process requires a large amount of energy to provide cooling to the gases which is approximately 0.55kWh/Nm³^[30]. Having a source of cold energy to assist in this process would reduce energy intensity greatly but careful heat exchanger design is required. A balance between the minimum temperature to achieve a liquid nitrogen stream and sufficient subcooling is desired. This would enable efficient transfer of cold energy back to the liquefaction site without excessive cooling. A gaseous oxygen stream capable of supplying nearby industries is desired. The oxygen can be utilised by industry clusters via pipelines such as the ones found at Teesside, UK.

Integrated regasification

Using Aspen, the LNG liquefaction process is integrated with a conventional cryogenic air separation process to create an integrated regasification process seen in figure 3. By specifying the inlet and outlet temperature of the natural gas, the total cold energy available for air separation can be calculated. Using the desired outlet conditions of the nitrogen and oxygen streams, the maximum production capacity of the integrated regasification process is found. In this process, pre-purified and filtered ambient air at 30°C is cooled using HX1 and HX2 before being separated into two streams enriched in oxygen and nitrogen. Two columns at high and low pressures are used to perform the separation, splitting



the air into 3 streams: a nitrogen rich stream at -196°C; an oxygen rich stream at 15°C and a waste stream. In this configuration LNG is heated to 10°C using ambient air and some additional glycol water which has enhanced heat transfer properties. The liquid nitrogen is then transferred back onto the LNG carrier to be returned back to Qatar where the nitrogen can be harnessed to provide cooling to the natural gas.

The estimated capital expenditure for equipment costs was calculated using the cost of process equipment reported on cost estimation websites and through quotes obtained from industrial suppliers. The equipment purchasing cost *EPC* then needs to be scaled using equation 4 to account for other major costs associated with capital expenditure, such as material and engineering costs. Bejan, Tsatsaronis and Moran.^[31] considered direct costs *DC* and indirect costs *IC* as a function of equipment purchasing costs which yielded a Lang factor of 4.4. The utility costs and carbon intensity can be found in the same way as the standalone regasification and air separation process for comparison. Due to the cryogenic conditions required for the air separation, the cooling utility required to achieve extremely low temperatures was high and thus would result in high energy costs which Aspen V11 built in function calculated.

$$CAPEX(\$) = DC + IC \approx 4.4EPC \quad (4)$$

Integrated liquefaction

The AP-X process uses a mixed refrigerant loop to provide cooling of natural gas and a nitrogen expansion compression loop to provide subcooling. This requires immense amounts of energy to get the refrigerants to low enough temperatures to enable heat transfer. With the integrated liquefaction process liquid nitrogen can be delivered at -196°C directly into the Ras Laffan site reducing the huge energy requirements. The replacement of the Brayton gas expansion cycle with a simple heat exchange process drastically decreases the energy requirements to cool the natural gas. The mixed refrigerant loop can also be replaced with a liquid nitrogen stream that

natural gas. Once utilised the nitrogen leaves the system at ambient conditions which can then be sold on to other industries. The estimated capital expenditure is calculated from the net elimination of process equipment from the standalone AP-X process. The sizing parameters of new equipment introduced was used to calculate the cost in the same manner stated in the integrated regasification design. The utility required for the system was simulated in Aspen and the corresponding costs was calculated using Qatari electricity tariffs of \$0.019/kWh. The carbon intensity was also calculated based on the electricity used in the process using the report carbon intensity of gas generated electricity in Qatar^[18].

Results and Discussion:

Once the standalone and integrated supply chains were modelled and simulated on Aspen, the results were compared against one another to evaluate the benefits of the integrated system. Performance metrics were calculated for both systems, specifically focusing on energy consumption, cost of production and carbon emissions.

Energy requirements of the integrated system were 29.7% lower than the standalone processes which included the cryogenic air separation unit. Without accounting for the air separation unit in the baseline standalone process, the reduction in energy intensity from the integrate process would be 11.9%. The main source of decrease in energy consumption came from the replacement of the nitrogen subcooling loop in the traditional AP-X process, which contributed to 51.1% of total liquefaction energy consumption prior to integration. Analysis was conducted to evaluate the effect on energy intensity from the replacement of just the nitrogen subcooling loop, the subcooling loop plus the mixed refrigerant cooling loop as well as replacement of the whole AP-X process with liquid nitrogen as the cooling medium. The maximum cold energy available from the integrated regasification site could only replace fully the refrigeration and subcooling loops with the whole process needing 334.8 MW of cold energy.

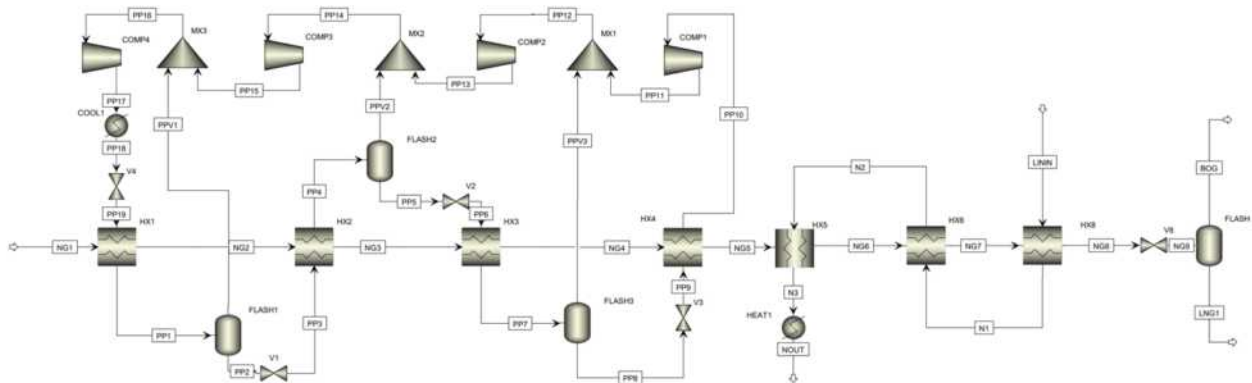


Figure 4. Process flow diagram showing the integrated liquefaction of LNG utilising the liquid nitrogen produced in the integrated regasification process in the UK. The liquid nitrogen stream replaces the Brayton gas expansion cycle which enables the subcooling of natural gas. The liquid nitrogen stream also replaces the mixed refrigerant refrigeration cycle.

Stream	T °C	P bar	\dot{m} kg/s	x
LNG inlet	-166	2.2	436.3	0.000
LNG post HX with air	-21.5	2.2	436.3	1.000
NG outlet	10	2.2	436.3	1.000
Air inlet	29.9	1.0	1450	1.000
Air post compression	27	5.9	1450	1.000
Air post HX with LNG	-173.7	5.8	1450	1.000
O2 rich into LPC	-190.8	1.3	707.3	0.102
O2 rich out of LPC	-191.7	1.3	275.8	0.000
O2 outlet	15	1.2	275.8	1.000
N2 rich into HPC	-173.7	5.8	486.5	1.000
N2 rich out of HPC	-193.4	1.3	486.5	0.999
N2 rich out of LPC	-191.7	1.3	800.2	1.000
N2 outlet	-196.0	1.0	800.2	0.000

Table 4: Table showing key integrated regasification stream data. The table provides data on temperature T , pressure P , mass flowrate \dot{m} and vapour fraction x .

The cold energy available was also constrained by the number of shipping vessel available for the return journey back to Qatar. The new system should not exceed the current number of vessels arriving from Qatar into the UK as this would incur additional shipping costs and result in more complex transportation requirements.

The biggest improvements in costs can be seen at the liquefaction site (figure 5) where capital expenditure decreased by 33.5% primarily due to the removal of the nitrogen subcooling loop which consisted of capital-intensive compressors and expanders corresponding to an equipment purchasing cost of \$562 million. Annual operating costs for the liquefaction process was calculated to be 36.1% lower compared to the conventional AP-X process primarily due to the decrease in utility costs which is composed of steam, water and electricity. The total cost of electricity decreased by 36.8% due to needing less electricity to drive the compressors and expanders for the cooling of nitrogen. The decrease in electricity

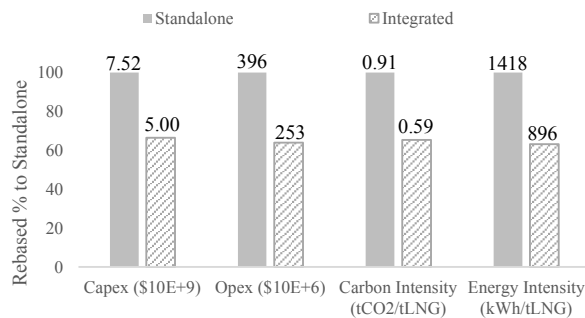


Figure 5: Graph showing changes in capital and expenditure, carbon intensity and energy intensity between the integrated and standalone process for the liquefaction process.

Stream	T °C	P bar	\dot{m} kg/s	x
NG inlet	43.0	65.0	448.5	1.000
NG post precooling	-33.2	65.0	448.5	0.996
NG post refrigeration	-111	65.0	448.5	0.000
NG post subcooling	-166.0	4.5	448.5	0.000
LNG outlet	-165.9	1.1	448.5	0.000
N2 inlet	-195.8	1.0	772.0	0.000
N2 post subcooling	-195.7	1.0	772.0	0.547
N2 post refrigeration	-33.6	1.0	772.0	1.000
N2 outlet	15	1.0	772.0	1.000
Propane into precooling	17.6	7.9	4000	0.110
Propane into HX2	6.8	5.8	400	0.078
Propane into HX3	-17.2	2.7	160	0.149
Propane post precooling	-26.2	1.3	72	1.000

Table 5: Table showing key integrated liquefaction stream data. The table provides data on temperature T , pressure P , mass flowrate \dot{m} and vapour fraction x .

consumption also resulted in a decrease in carbon emissions which was calculated to be 25.6% lower than the AP-X process. Tables 4 and 5 below summarise the operating conditions for the integrated design on both the regasification and liquefaction side. The key performance metrics is summarised in table 6 and shown side by side for the standalone and integrated process. The standalone process considers the addition of an air separation unit at the regasification site as the baseline for comparison.

Key metrics	Standalone	Integrated
Capex (\$bn)	10.93	7.66
Opex (\$bn)	1.42	1.13
Carbon intensity (tCO2/tLNG)	1.127	0.781
Energy intensity (kWh/tLNG)	2205	1550

Table 6: Table comparing the key performance metrics between the standalone current LNG supply chain and the integrated supply chain.

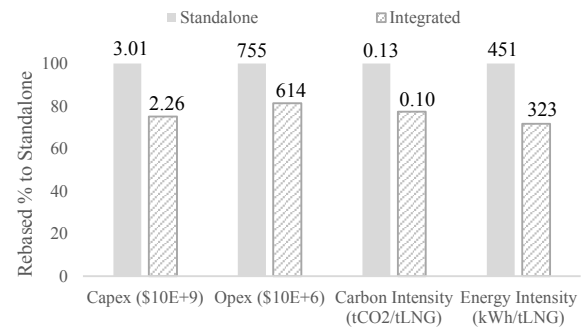


Figure 6: Graph showing changes in capital and expenditure, carbon intensity and energy intensity between the integrated and standalone process for the regasification process.

Opportunities in O₂ and N₂

Once the cold energy has been extracted during the liquefaction process, ambient nitrogen is then available to be sold as a feedstock at >99.95% purity into a wide variety of industries. The fertiliser industry is a huge buyer of nitrogen gas for the Haber-Bosch process which produces ammonia. Ammonia is then used to make ammonium nitrate; the most common fertiliser used in industry and contains roughly 33.5% nitrogen^[32]. Upon analysing the prices of nitrogen feedstock into various industries, it was determined the most economical option was to sell into the fertiliser industry. Currently, prices for ammonium nitrate have averaged around £360 per ton^[33] and demand for the product is expected to grow by 3.6% over the next 10 years^[34]. The average prices of nitrogen purchased by each industry was calculated and shown in table 6. With the capacity of the integrated regasification process in mind, the expected additional annual income is found to be \$230m. Due to the location of Qatar, flexible delivery into the market with the most opportunities is easily adaptable.

Industry	Price
Fertiliser	\$9.4/tonne
Mining	\$6.5/tonne ^[35]
Electronics manufacturing	\$5.2/tonne ^[36]

Table 7: Table showing different purchasing prices for various industries for >99.95% purity of nitrogen per tonne.

In addition, the production of high purity oxygen is also an exciting opportunity. Oxygen can enable oxy-fuel combustion which eliminates the need for expensive amine scrubbing units used to remove CO₂ from flue gas. For perspective, a 1.4GW air-fuelled coal fired power plant can expect to pay upwards of \$50/tonne CO₂ captured using amine scrubbing^[37]. Oxy-fuel combustion produces flue gas, consisting of CO₂ and water, reducing SO_x and NO_x emissions due to removal of pollutants and nitrogen in the gas feed. With oxy-fuel combustion, CO₂ produced during combustion can be easily separated from water and captured for CCS processes, decreasing capital expenditure for post combustion capture. Current oxy-fuel CCGT demonstrates power plant efficiencies up to 55% at a turbine inlet temperature of 1625K^[38]. Combustion can also occur at higher temperatures netting a higher thermal efficiency. Based on the research conducted by Nakaiwa et al., 1996 it has been shown that the efficiency improvements of oxy-fuel combustion in a natural gas CCGT increase by 1 percent per 36.6K increase in temperature. A decrease in CO₂ exhaust emission of 0.054% per Kelvin increase^[12] can also be expected. To reach similar power plant efficiencies of current air fuelled CCGT (up to 62.2%^[39]), turbine inlet temperature will need to be raised to 1888K which would net a 14.2% decrease in CO₂ exhaust emissions.

Another industry of interest is the cement industry, which accounts for 7% of global emissions. By utilizing oxy-fuel combustion they can reduce CO₂ emissions, supporting efforts to decarbonise the heavily polluting

industry. Current high purity oxygen market prices range from £25-£40 per tonne^[40], expecting \$353m in additional income generated per year through the sale of oxygen gas.

Sensitivity analysis

In the base case only the nitrogen subcooling loop in the liquefaction process was replaced. This would use only 75.5% of the produced nitrogen from the integrated regasification site. Operating expenditure and carbon intensity was sensitised by the amount of LIN flowing into the liquefaction plant and replacing the mixed refrigerant loop as shown in figure 7. The operating expenditure is directly related to the cost of utility and therefore the amount of energy needed to cool down the natural gas. The graph shows the sensitivity analysis for varying LIN flowrates. With full utilisation of LIN at 797kg/s which is the maximum carrying capacity of the LNG vessels over a year, assuming 131 ships. Below 772kg/s there is a cross over in temperature within the heat exchangers therefore this range was not considered.

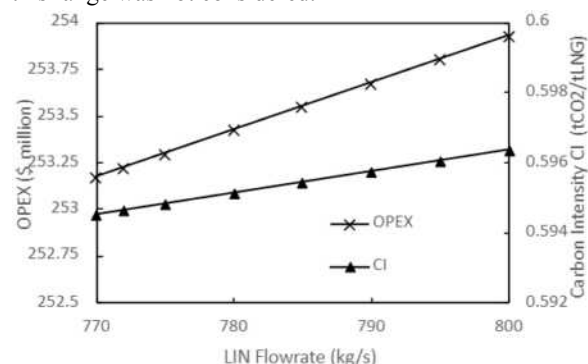


Figure 7: Graph showing changes in operating expenditure (left) and carbon intensity (right) at various liquid nitrogen flowrates.

Carbon dioxide reduction

Net carbon intensity for the integrated supply chain was 30% lower compared to the current supply chain. Even after considering the addition of an air separation unit. The transfer of cold energy back to the liquefaction site resulted in a decreased of 35% in carbon emissions from the liquefaction process, the most emissions intensive part of the supply chain. This reaffirms the hypothesis that the energy and emissions improvement through integration can not only be realised at the regasification site like most studies focus on, but also at the liquefaction site.

Conclusion

The integrated system shows a 29.1% decrease in capital expenditure, 20.1% in operating expenditure, 29.7% energy consumption as well as 30.8% in carbon emissions which were the main criteria when assessing the feasibility of the new design. Through implementation of this integrated design, additional value can be extracted from the cold energy of LNG whilst reducing the effects on the environment. The integration is easily adaptable from current industry processes due to the simplicity of the design and the implementation of an air separation unit can

be justified when the regasification site is located nearby heavily polluting industries that could take advantage of the oxygen produced. The integrated design will have a total capital expenditure of \$7.66bn. Employee, utility and material cost is estimated to be \$1.31bn per annum. Additional income generated from the sale of nitrogen and oxygen stands at \$583m per annum. Using a conservative estimate for the price of LNG at \$3.5/MMBtu the cost of investment can be recuperated after 4 years of operation from implementing this new design. It has been shown that through integration of the supply chain, greater energy efficiencies can be achieved leading to notable cost savings.

Outlook

Additional considerations need to be investigated further regarding the use of current LNG vessel to contain LIN instead. New vessels may need to be developed and the cost associated with these new builds need to be considered. A possible implementation of this new supply chain is in the Teesside net zero cluster as it is nearby industries that could benefit from the products produced at the integrated regasification site. However, Q-type LNG vessels are notoriously large and sufficient berthing depth at Teesside may not be available to enable delivery of LNG and other alternatives such as offshore floating regasification terminals could be explored.

In addition, implementation of oxy-fuel combustion in CCGT is constrained a 55% thermal efficiency due to materials used in construction. Until better materials are developed that can operate at the upper temperature limits of oxy-fuel combustion, the efficiency gains offered may not be realised.

Whilst the integrated process offers carbon emissions reduction of 30% the emissions from the return journey and additional journeys to deliver products as well as to construct infrastructure may lower this saving significantly. Further analysis is required for specific use cases to justify the implementation of the new design.

As prices for deployment of renewable energy surges due to supply constraints, inefficient government policies and geopolitical tension, demand for natural gas in the medium term will increase. Since the invasion of Ukraine, OECD nations are looking toward the US and Middle East for natural gas alternatives which cannot be delivered by pipeline therefore the need to an efficient LNG supply chain will become even more important.

Acknowledgments

The authors would like to acknowledge the contributions and support given by Imperial College London Chemical Engineering Department in particular Dr Mahdi Sharifzadeh as well as Dr Depak Lal.

References

[1] EIA. Available at: https://www.eia.gov/environment/emissions/co2_vol_mass.php (Accessed: 13 December 2023).

[2] Pospíšil, J. et al. (2019) 'Energy demand of liquefaction and regasification of natural gas and the potential of LNG for Operative Thermal Energy Storage', *Renewable and Sustainable Energy Reviews*, 99, pp. 1–15. doi:10.1016/j.rser.2018.09.027.

[3] World Nuclear Association. Available at: <https://world-nuclear.org/information-library/facts-and-figures/heat-values-of-various-fuels.aspx> (Accessed: 13 December 2023).

[4] bp Statistical Review of World Energy 2021. Available at: <https://www.bp.com/content/dam/bp/business-sites/en/global/corporate/pdfs/energy-economics/statistical-review/bp-stats-review-2021-full-report.pdf> (Accessed: 13 December 2023).

[5] EI Statistical Review of World Energy 2023. Available at: <https://www.energyinst.org/statistical-review> (Accessed: 13 December 2023).

[6] EU Natural Gas - Price - Chart - Historical Data - News. Available at: <https://tradingeconomics.com/commodity/eu-natural-gas> (Accessed: 13 December 2023).

[7] Balcombe, P. et al. White Paper 1: Methane and CO2 emissions from the Natural Gas Supply Chain, Imperial College London. Available at: <https://www.imperial.ac.uk/sustainable-gas-institute/our-research/white-paper-series/white-paper-1-methane-and-co2-emissions-from-the-natural-gas-supply-chain/> (Accessed: 13 December 2023).

[8] Department for Energy Security and Net Zero (2023) UK energy in brief 2023, GOV.UK. Available at: <https://www.gov.uk/government/statistics/uk-energy-in-brief-2023> (Accessed: 13 December 2023).

[9] ExxonMobil Global Outlook: Our View to 2050. Available at: <https://corporate.exxonmobil.com/what-we-do/energy-supply/global-outlook> (Accessed: 13 December 2023).

[10] John M. Campbell & Co. An introduction into the air products and Chemicals, Inc., mixed refrigerant LNG liquefaction process - what is it, and how does it work? Available at: <https://www.jmcampbell.com/tip-of-the-month/2020/07/an-introduction-into-the-air-products-and-chemicals-inc-mixed-refrigerant-lng-liquefaction-process-what-is-it-and-how-does-it-work/> (Accessed: 13 December 2023).

[11] QatarEnergy LNG N(2) Trains. Available at: <https://www.qatarenergylng.qa/english/operations/lng-trains> (Accessed: 13 December 2023).

[12] Nakaiwa, M. et al. (1996) 'Evaluation of an energy supply system with Air Separation', *Energy Conversion and Management*, 37(3), pp. 295–301. doi:10.1016/0196-8904(95)00787-3.

[13] Xiong, Y.Q., Hua, B., 2014. Simulation and Analysis of Cryogenic Air Separation Process with LNG Cold Energy Utilization. *AMR* 881–883, 653–658. <https://doi.org/10.4028/www.scientific.net/amr.881-883.653>

[14] Shingan, B., Vijay, P. and Pandiyan, K. (2023) 'Cryogenic Air Separation Process Integrated with cold utilization of Liquefied Natural Gas: Design, Simulation and Performance Analysis', *Arabian Journal for Science and Engineering*, 48(12), pp. 16921–16940. doi:10.1007/s13369-023-08218-5.

- [15] Hanifa, M. et al. (2023) 'A review on CO₂ Capture and sequestration in the construction industry: Emerging approaches and commercialised technologies', *Journal of CO₂ Utilization*, 67, p. 102292. doi:10.1016/j.jcou.2022.102292.
- [16] Steuer, Claudio. (2019) "Cost of Liquefaction Plant Projects." *Outlook for Competitive LNG Supply*, Oxford Institute for Energy Studies, pp. 8–15. JSTOR, <http://www.jstor.org/stable/resrep31040.11>. Accessed 13 Dec. 2023.
- [17] Qatar General Electricity & Water Corporation - المؤسسة العامة للكهرباء والماء. Available at: <https://www.km.qa/ExportedSites/Custom/Pages/Tariff.aspx> (Accessed: 13 December 2023).
- [18] Salma Saleh. (2023) Qatar: Emissions intensity from Electricity Generation, Statista. Available at: <https://www.statista.com/statistics/1302666/qatar-emissions-intensity-from-electricity-generation/#:~:text=In%202020%2C%20the%20emissions%20intensity,compared%20to%20the%20previous%20year> (Accessed: 13 December 2023).
- [19] Couper, J. et al. (2005) 'Introduction' (2005) *Chemical Process Equipment*, pp. 1–16. doi:10.1016/b978-075067510-9/50033-4.
- [20] Song, Q. et al. (2022) 'A comparative study on energy efficiency of the maritime supply chains for liquefied hydrogen, ammonia, methanol and natural gas', *Carbon Capture Science & Technology*, 4, p. 100056. doi:10.1016/j.ccs.2022.100056.
- [21] Staff, L.P. (2023) QatarEnergy eyes orders for Giant LNG carriers, LNG Prime. Available at: <https://lngprime.com/asia/qatarenergy-eyes-orders-for-giant-lng-carriers/92268/> (Accessed: 13 December 2023).
- [22] The supply chain South Hook LNG. Available at: <https://www.southhooklng.com/operations/the-supply-chain/> (Accessed: 13 December 2023).
- [23] Natural gas carbon footprint analysis - the North Sea Transition Authority (2023) North Sea Transition Authority. Available at: <https://www.nstauthority.co.uk/the-move-to-net-zero/net-zero-benchmarking-and-analysis/natural-gas-carbon-footprint-analysis/#:~:text=In%20particular%20the%20analysis%20shows,intensity%20of%2079%20kgCO2%2Fboe> (Accessed: 13 December 2023).
- [24] LNG process Saint John LNG. Available at: <https://www.saintjohnlng.com/lng-process> (Accessed: 13 December 2023).
- [25] Agency, E. (2020) Material comparators to assist in an end-of-waste assessment. Material comparators for Fuels: natural gas, GOV.UK. Available at: <https://www.gov.uk/government/publications/defining-product-comparators-to-use-when-applying-waste-derived-materials-to-land> (Accessed: 13 December 2023).
- [26] Global Energy Monitor (2023) South Hook LNG Terminal. Available at: https://www.gem.wiki/South_Hook_LNG_Terminal#cite_note-9 (Accessed: 13 December 2023).
- [27] Statista Research Department. (2023) UK: Thermal efficiency of gas turbine stations 2022, Statista. Available at: <https://www.statista.com/statistics/548943/thermal-efficiency-gas-turbine-stations-uk/#:~:text=In%202022%2C%20combined%20cycle%20gas,gas%20was%20converted%20to%20electricity>. (Accessed: 13 December 2023).
- [28] Warren, J. (2023) Average cost of gas per kwh UK 2023, Energy Guide. Available at: [https://energyguide.org.uk/average-cost-gas-kwh/#:~:text=In%20terms%20of%20the%20average,electricity%20\(for%20average%20use\)](https://energyguide.org.uk/average-cost-gas-kwh/#:~:text=In%20terms%20of%20the%20average,electricity%20(for%20average%20use)). (Accessed: 13 December 2023).
- [29] Barclay, J. (2023) New UK grid emissions factors 2023, ITP Energised. Available at: <https://www.itpenergised.com/new-uk-grid-emissions-factors-2023/> (Accessed: 13 December 2023).
- [30] Tesch, S., Morosuk, T. and Tsatsaronis, G. (2020) 'Comparative evaluation of cryogenic air separation units from the Exergetic and economic points of View', *Low-temperature Technologies* [Preprint]. doi:10.5772/intechopen.85765.
- [31] Bejan, A., Tsatsaronis, G. and Moran, M.J. (1996) *Thermal design and optimization*. New York: John Wiley & Sons.
- [32] 'Fertilisers and Manures' (2023) Lockhart and Wiseman's Crop Husbandry Including Grassland, pp. 81–114. doi:10.1016/b978-0-323-85702-4.00014-5.
- [33] GB fertiliser prices AHDB. Available at: <https://ahdb.org.uk/GB-fertiliser-prices> (Accessed: 13 December 2023).
- [34] Ammonium nitrate market analysis Growth & Forecast, 2032. Available at: <https://www.chemanalyst.com/industry-report/ammonium-nitrate-market-525/#:~:text=H1%202022%3A%20In%202022%2C%20the,roughly%2010%20million%20thousand%20tonnes> (Accessed: 13 December 2023).
- [35] Mine Safety and Health Administration - Mine inerting information. Available at: <https://arlweb.msha.gov/Seals/SealsInertingGases.pdf> (Accessed: 13 December 2023).
- [36] Styles, C. (2023) Revisiting the costs of Nitrogen Gas, Purity Gas. Available at: <https://puritygas.ca/revisiting-the-costs-of-nitrogen-gas/> (Accessed: 13 December 2023).
- [37] Panja, P., McPherson, B. and Deo, M. (2022) 'Techno-economic analysis of amine-based CO₂ Capture Technology: Hunter plant case study', *Carbon Capture Science & Technology*, 3, p. 100041. doi:10.1016/j.ccs.2022.100041.
- [38] Mancuso, L. et al. (2015) Oxy-Combustion Turbine Power Plants. Available at: https://ieaghg.org/docs/General_Docs/Reports/2015-05.pdf (Accessed: 13 December 2023).
- [39] Rooij, D.D. (2023) Natural gas combined cycle, Managerisks and maximize ROI for your PV and energy storage projects. Available at: <https://sinovoltaics.com/learning-center/technologies/natural-gas-combined-cycle/> (Accessed: 13 December 2023).
- [40] Serrano, J.R. et al. (2022) 'Thermo-Economic Analysis of an oxygen production plant powered by an Innovative Energy Recovery System', *Energy*, 255, p. 124419. doi:10.1016/j.energy.2022.124419.

A Machine Learning Platform for the Optimisation and Innovation of Ionizable lipids for Efficient RNA Delivery

John Huang and Oliver Tomes

Department of Chemical Engineering, Imperial College London, U.K.

Abstract mRNA vaccines and therapies such as COVID-19 vaccines demonstrate significant potential in treating and preventing diseases. Lipid Nanoparticles (LNP) are crucial for encapsulating, protecting, and facilitating the endosomal escape and release of mRNA. Within the LNP, the most important ingredient is the ionizable lipid which has a major impact on delivery. This study utilised machine learning to optimize LNP design, focusing on ionizable lipid features. Delivery was quantified using the transfection efficiency metric. Having reviewed over 35 research papers, 439 data points were collected, encompassing 291 LNP formulations and 208 unique ionizable structures. These were categorized into *in vivo*, *in vitro*, and specific cell and organ type datasets. A Random Forest model, developed using these categorized datasets, identified crucial ionizable lipid features affecting transfection efficiency, notably 'Tail Length', 'Unsaturated', and 'Number of Tails', aligning with existing literature. This coherence indicates the model's potential in capturing essential characteristics of ionizable lipids. However, there existed discrepancies, primarily due to data limitations. This suggested the need for a more robust dataset to enhance predictive accuracy. The inclusion of additional LNP features like pK_a and relative abundance further demonstrates the potential for a more comprehensive analysis. The study's trajectory points towards the creation of a more extensive dataset and refinement of the predictive model, highlighting the potential of machine learning in advancing ionizable lipid-based delivery systems.

Keyword – Drug Delivery, Lipid Nanoparticles, Ionizable Lipid, Machine Learning, Random Forest, Modelling

1 Introduction and Background

The success of the Pfizer-BioNTech and Moderna COVID-19 vaccines showed the great potential and crucial role of RNA-based therapeutics in the prevention and treatment of future diseases. However, the cytosolic delivery of the active ingredients remains a serious challenge. Genetic material is fragile and can be easily broken down before reaching the target cell.

Nanoparticle-based drug delivery systems have emerged as a method of encapsulating RNA therapeutics and enabling efficient delivery to target cells. In 2018, ONPATTRO (Patisiran) became the first small interfering RNA-based therapeutic to be approved by the U.S. Food and Drug Administration (FDA).^[1] To date, three more medications have followed: Givosiran, Lumasiran, and Inclisiran.^[2] - all of which have been designated by the FDA as First-in-Class Medication.

However, with new global challenges emerging, acceleration in the development of this technology is crucial for the success of upcoming pharmaceuticals.

Ionizable Lipid Nanoparticles (LNPs) are a pH responsive nanoparticle technology. LNPs structure consists of four main ingredients: Phospholipid, Cholesterol, PEGylated lipid, and arguably most important: Ionizable lipid. Whilst the other components offer many structural and mechanical advantages^[1], the ionizable lipid plays a key role in ensuring successful transfection of the genetic payload. After endocytosis, the ionizable

lipid facilitates the release of the cargo from the endosome into the cytosol - a process commonly known as endosomal escape.^{[3],[4],[5]}

In recent years, many ionizable lipids have been experimented with, resulting in a vast amount of literature detailing ionizable lipid structures and their transfection success rates. Currently, laborious synthesis techniques are used to generate ionizable lipids. Whilst combinatorial chemistry may offer a way of rapidly producing ionizable lipid libraries^[6], the synthesis process for many ionizable lipids remains time consuming and costly. With so much literature in this area, there is potential for computational methods to aid in the development of successful ionizable lipids.

There exist previous works that explore the possibility of applying machine learning algorithms such as lightGBM and artificial neural networks to build a prediction model on the Lipid Nanoparticle (LNP) performance.^{[7],[8]} Moreover, some looked into the ionizable lipids' feature importance score and identified that the "outside carbons" which is number of carbons in tails, as the important feature when it comes to designing ionizable lipid.^[8]

This study investigated the most critical attributes of ionizable lipid design based on relevant literature. Distinct datasets were formulated containing the ionizable lipids and their corresponding efficacy. In addition, other datasets containing LNP properties were formulated, and the development of a predictive model was attempted. **Figure 1** shows the workflow of the process.

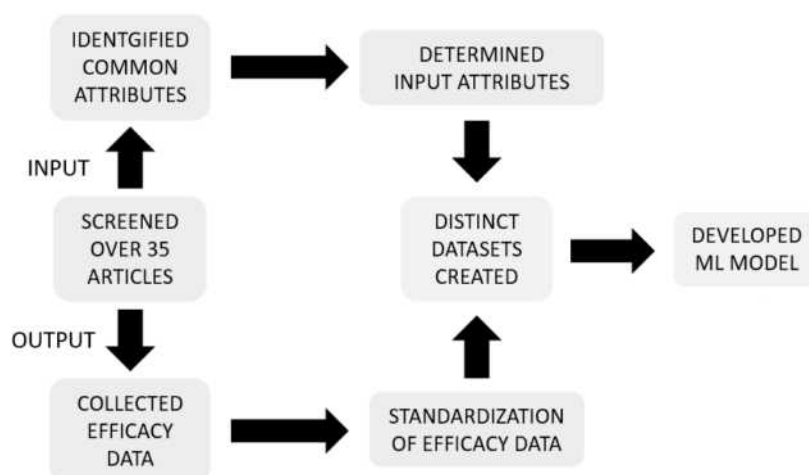


Figure 1. The workflow of the predictive model building process

2 Materials and Methods

2.1 Software

Python 3.9 (64-bit) was used throughout this project for all data processing and modelling purposes. The scikit-learn package was used to develop and validate machine learning models. The graphical analyser WebPlotDigitizer-4.6 was used to interpolate data reported in graphics.

2.2 Data selection from available literature

Data was collected from articles using PubMed, Google Scholar data base. To be considered for this study they satisfy the following criteria:

1. mRNA/siRNA payload delivered by LNPs containing the ionizable lipid.
2. All LNPs contain ionizable lipid, PEGylated lipid, phospholipid, and cholesterol.
3. The structure of the ionizable lipid is reported in graphics.
4. The paper compared the ionizable lipid to an appropriate benchmark (commercially available ionizable lipid).
5. Study reported results on LNP efficacy in either text or graphics.

The above aspects were either reported in the main text, citations or supplementary materials of the article.

2.3 Data preprocessing

2.3.1 Identification of input variables

Through literature review, data on several different ionizable lipid and LNP features were gathered. For the ionizable lipid this consisted of areas such as structural classes, size, and functional groups. Features were created based on already established ideas in literature such as biodegradable or branched tails. For LNPs, features such as the composition of

different ingredients and pK_a were collected, as well as other parameters such as the loading degree (N/P).

For the sake of this preliminary investigation, a feature would only be implemented into the model if a significant number of sources containing the feature were found. The criterion of at least ten sources was decided using engineering judgement.

The following ionizable features are defined the following way:

1. “Tail Length” number of carbons in the tail. If the tails are identical use one of them otherwise use the longer tail.
2. “Biodegradable” means a lipid that contains ester bond.

After the final features were established, any incomplete datapoints with missing features were removed from the dataset.

2.3.2 Transfection efficiency standardization

Transfection efficiency is defined as a measure of the successful delivery and expression of genetic material into cells. It is a very important metric for quantification of the LNPs performance. Due to ease of visualization, most papers report delivering mRNA for Fluorescent Proteins and observe the level of luminescence. Hence this is selected as the indication of transfection efficiency in this research. However, it is acknowledged that different groups’ work under different laboratories can bring significant variance even for the same type of experiment. So, to prepare the data for comparison standardisation was a compulsory step.

The following equations were used to calculate the Relative Performance (RP) depending on the payload investigated:

$$RP(mRNA) = \frac{\text{Transfection efficiency of the LNP}}{\text{Benchmark transfection efficiency}} \quad (1)$$

$$RP(siRNA) = \frac{\text{Benchmark Transfection Efficiency}}{\text{Transfection efficiency of the LNP}} \quad (2)$$

The ‘benchmarks’ in Equation (1) and (2) above are usually the commercially available ionizable lipids reported such as MC3, SM-102, ALC-0315, etc. The performance of LNP were then classified to “GOOD”, “MEDIUM”, “BAD” within each paper according to their RP score. [9],[10],[11],[12]

2.4 Principle Component Analysis (PCA) of the dataset

PCA was conducted on both *In vitro* and *In vivo* datasets using the Scikit-Learn package in Python.

2.5 Machine Learning Modelling

2.5.1 Random Forest

Random Forest models were built from Scikit-Learn package in python. `sk.learn_model_selection` was used to split the data into a training and validation set of ratio 80:20.

2.5.2 Performance Metrics

The following metrics were used to assess the model:

Confusion matrix:

$\forall \text{Classes "1", "2", "3"}:$

1 predicted as 1	1 predicted as 2	1 predicted as 3
2 predicted as 1	2 predicted as 2	2 predicted as 3
3 predicted as 1	3 predicted as 2	3 predicted as 3

This was used to calculate the Accuracy, shown in Equation (2), and the Mean Square Error, shown in Equation (3).

Accuracy:

$$\text{Accuracy} = \frac{\text{No. of accurate predictions}}{\text{No. of data points in testing set}} \quad (2)$$

MSE:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (3)$$

Where:

n is the number of data point;

Y_i is the observed value;

\hat{Y}_i is the predicted value.

Cross-validation (CV) was performed, and the average CV score was calculated using Equation (4). 5 validations were used for all models. Furthermore, the feature importance was calculated using Equation (5). This was done to mitigate the impact of anomalies on the model's performance. Then an Average of CV score was calculated to be used as the final accuracy for the model, shown in Equation (4):

Average CV score:

$$\text{Average CV score} = \frac{\sum \text{CV scores}}{\text{Number of validations}} \quad (4)$$

Equation (5) was used to calculate the feature importance in the Random Forest model:

Feature Importance^[13]:

$$f_{i_i} = \frac{\sum_{j: \text{node } j \text{ splits on feature } i} n_{i_j}}{\sum_{k \in \text{all nodes}} n_{i_k}} \quad (5)$$

3 Results

3.1 Features

The following Figure 2 and Figure 3 showed all of the ionizable features and LNP features considered and implemented.

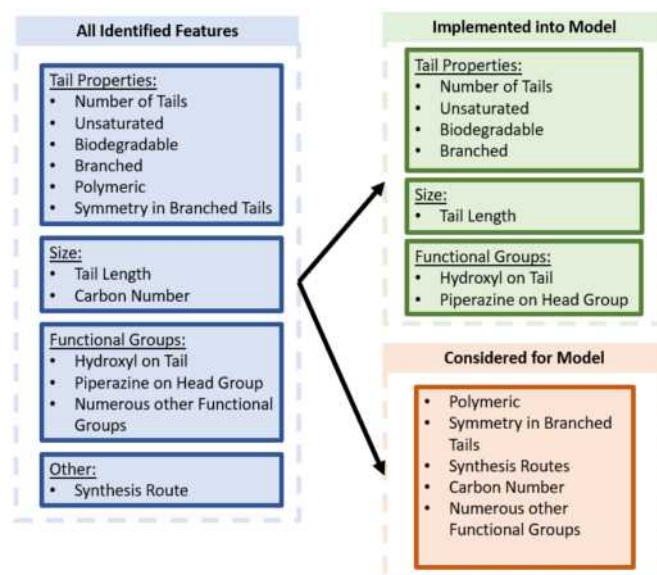


Figure 2. Considered and implemented ionizable features

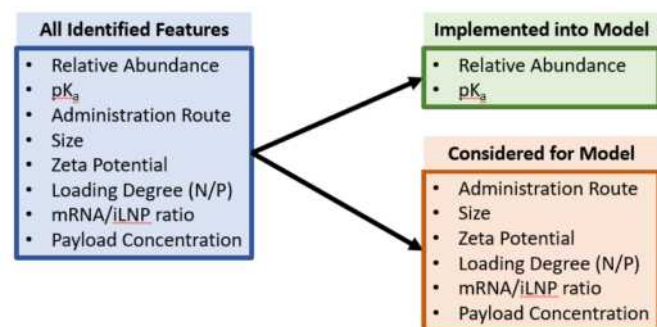


Figure 3. Considered and implemented LNP features

The following tree diagram Figure 4 shows the features established after literature review:

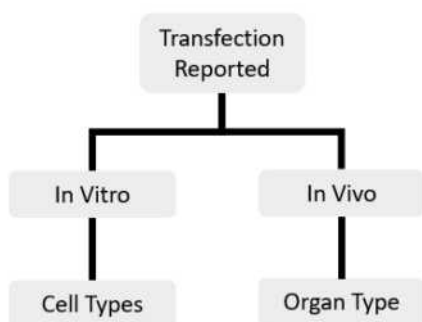


Figure 4. Tree Diagram of output categories

3.2 Overview of dataset

There were in total 439 data points collected from literature which included 291 different LNP formulations. Among these LNPs, 208 distinct ionizable lipid structures were covered. And the distribution of these data is shown in **Figure 5**.

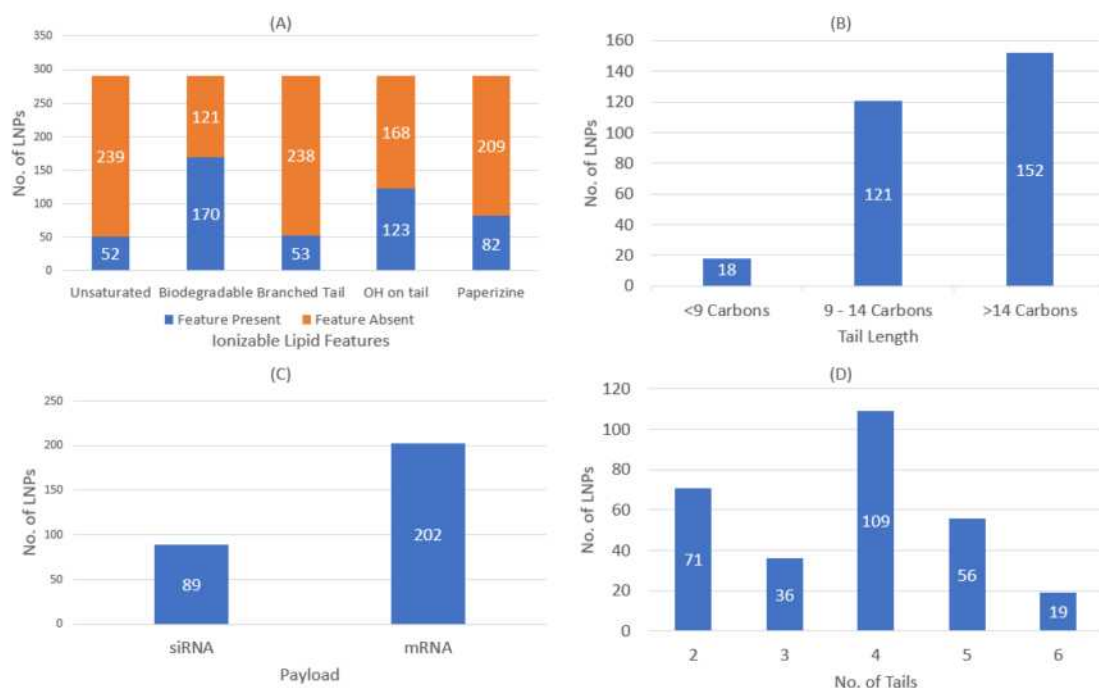


Figure 5. Overview of dataset statistics for (A) Ionizable lipid feature present; (B) Tail Length; (C) Payload; (D) Number of tails

3.3 PCA results

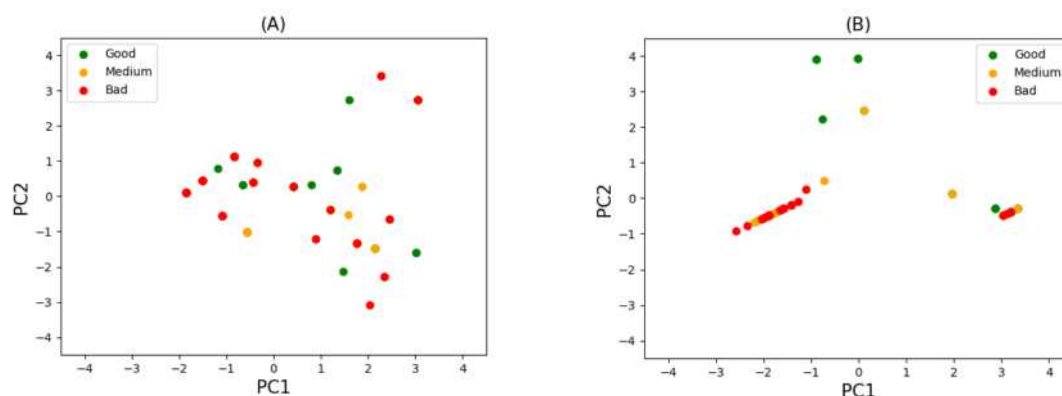


Figure 6. Principal Component Analysis of (A) *In vitro* Features. Principal Component 1 (PC1) and Principal Component 2 (PC2) accounted for 44% and 21% of total variance, respectively; (B) *In vivo* Features. Principal Component 1 (PC1) and Principal Component 2 (PC2) accounted for 71% and 18% of total variance, respectively.

3.4 Modelling

3.4.1 Ionizable Lipid Features

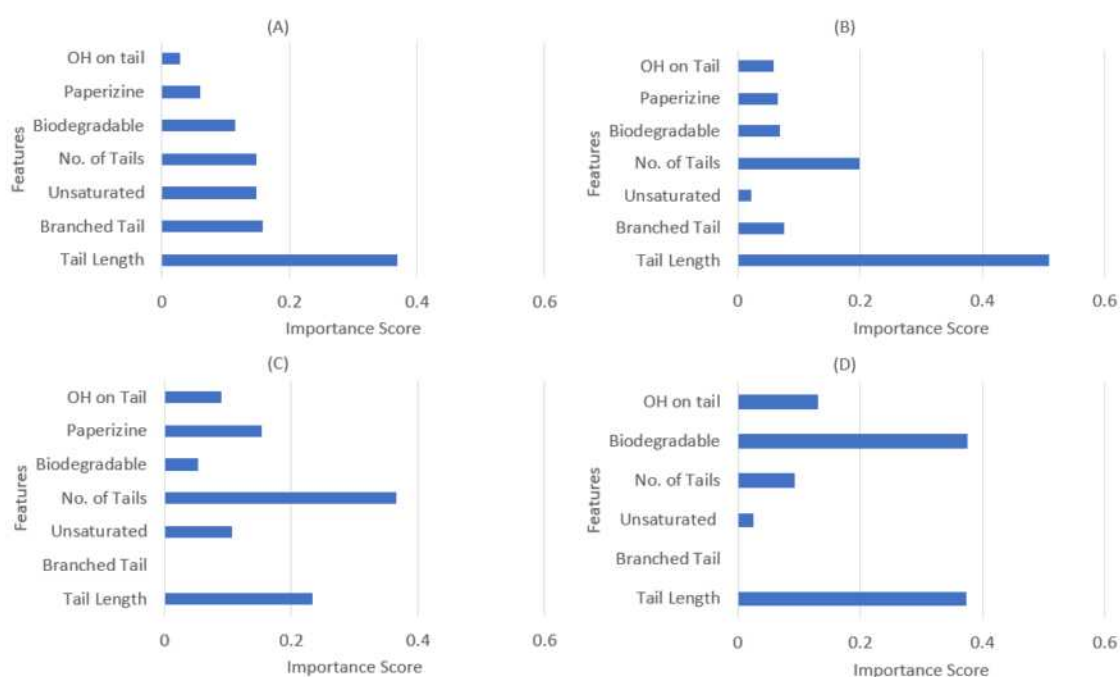


Figure 7. Feature importance scores for: (A) *in vivo* overall; (B) *in vivo* spleen; (C) *in vivo* liver; (D) *in vitro* HepG2

Table 1. Accuracies and ranges of validity of Random Forest model trained with ionizable lipid features data

Models	Data point	Accuracy	Range of Validity	
			Tail Length	Multi-tails
<i>In vivo</i> (Overall)	70	0.5429	9 - 31	2 - 5
<i>In vivo</i> (Liver)	101	0.5648	6-25	2 - 5
<i>In vivo</i> (Spleen)	76	0.5000	6-31	2 - 5
<i>In vitro</i> (Overall)	192	0.2960	6 - 18	2 - 6
<i>In vitro</i> (HepG2)	68	0.6011	12-17	4 - 6

3.4.2 LNP Feature

Table 2. Accuracies and ranges of validity of Random Forest model trained with LNP feature data

Models	Data point	Accuracy	Range of Validity				
			pK _a	PEG %	Phospholipid %	Ionizable Lipid %	Cholesterol %
<i>In vitro</i> (Overall)	60	0.4333	4.60 - 9.18	1 - 3	8 - 56.5	16 - 61	6 - 60
<i>In vivo</i> (Overall)	54	0.5182	5.83 - 9.18	1.1 - 2.5	8 - 46.5	35 - 55	10.9 - 38.5
<i>In vivo</i> (Liver)	62	0.5269	5.4 - 9.18	1.1 - 2.5	8 - 46.5	35 - 55	10.9 - 38.5
<i>In vivo</i> (Spleen)	55	0.5636	4.4 - 7.25	1.5 - 2.5	10 - 32.8	40 - 54.6	11 - 42.67

4 Discussion

4.1 Input features

Many of the features chosen resemble those associated with clinically approved lipids Dlin-MC3-DMA, ALC-0315, and SM-102. This included unsaturated, branched, and biodegradable tails.^[14] Furthermore, the multi-tail feature associated with commercial lipids such as C12-200 was captured through the Number of Tails feature. Amongst the ionizable lipids collected, the number of tails varied from 2-6. To capture this variety, the feature was treated as continuous. Due to lack of available data, polymeric tails were not included in the model.

Numerous studies have investigated optimisation of the tail length for delivery.^{[15],[16],[17]} Many trends have been hypothesized in the literature and this was investigated by accounting for the tail length as a feature.

Piperazine was chosen due to its ability of stabilising the LNP and as well as improving the transfection efficiency mainly by enhancing cellular uptake and endosomal escape.^[18] This means a more efficient release of drug encapsulated within the LNP.

Hydroxyl groups on the tail chosen due to its association with increased transfection. Previous studies have shown the promotion of hydrogen bonding can enhance ionizable lipid performance.^[19] Many other functional groups were considered such as hydroxyl in the head group, however lack of availability of lipids containing these groups led to them not being included in the modelling section.

Regarding the LNP properties, relative abundance of its ingredients is included as an attempt to optimise composition. pK_a is commonly cited as one of the determining factors of transfection.^[20] To account for variability due to the wide range of ionizable lipids in the dataset, it was included in the LNP model.

Many other features were considered during literature review. Despite loading degree (N/P) and mRNA/LNP ratio being an important parameter^[21], the lack of availability of this data meant it was not sensible to be included for modelling.

4.2 PCA

The principal component analysis was the first step in the modelling process. It provided an opportunity to get a good visualization of the different features reduced to two dimensions. **Figure 6** depicts the PCA and shows clear sparsity in the dataset. Clusters of the same colour are not distinguishable based on this visualisation.

4.3 Model selection

When approaching the modelling, the sparse result from the PCA suggests a linear regression tool such as partial least squares regression would not be the most suitable.^[22]

Due to the use of both discrete and continuous inputs and outputs, a classification model was implemented. Given the sparsity of the dataset, a Decision Tree would likely to overfit. In such dispersed dataset Decision Tree might capture the noise and outliers and hence lead to poor model accuracy. Hence, a random forest was the most suitable model to use.

4.4 Random Forest

As **Table 1** and **Table 2** shows, the accuracy of the *in vivo* Random Forest models ranges from 50-60%. This shows moderate reliability when used as a predictive tool. This is likely due to the dispersity of the training dataset observed earlier in the PCA. Data are collected for several types of cells such as HEK, HeLa and HepG2, as well as both payload: mRNA and siRNA. This is depicted in **Figure 5 (C)**.

Whilst the *in vitro* model gave a very low accuracy, when looking at the specific cell type HepG2, the model accuracy increased more than two-fold. This shows that cell type has big impact on the output of the model and hence *in vitro* model should be based on the same type of cell. This leads a clear pathway of further improvement of the model accuracy via data collection.

4.4.1 Ionizable Lipid Models

From the feature importance graphs in **Figure 7** it can be seen that “Tail Length” is the critical feature across the three *in vivo* model. According to literature the rule of thumb when designing an ionizable lipid is with a small head group and a big tail group.^[7] According to another study using similar approach it is found that the “Tail Length” has the highest importance score.^[8]

It is also reported that drugs containing piperazine derived structure are mainly metabolised in liver,^[24] which can be substantiated from **Figure 7(c)** that according to our model the importance score of piperazine is higher than that in the spleen.

If taking a closer look at **Figure 7 (A)** it is found that apart from “Tail Length”, the other most important features in order are branched and unsaturated tails which are in line with the common features of commercialised LNPs and even FDA approved ionizable lipids such as MC3 (unsaturated), ALC-3015 (branched) and SM-102 (branched).

Furthermore, the *in vivo* models show that biodegradable ionizable lipids has relative low importance score across the three models given that

there is a good split of 121 to 170 data it can be concluded from the model that biodegradability (i.e., ester bond) doesn't have an as great impact on the transfection efficiency. This prediction is also in line with literature reporting an ester bond is more readily hydrolyzed *in vivo* and hence would be more likely to break down which affects the delivery efficiency.^[25] The primary role of the biodegradable lipid is to improve clearance and reduce cytotoxicity of the LNP.^[14]

For in spleen **Figure 7 (B)**. The model showed that "Tail Length" is dominantly the most important feature followed by "No. of tails" all the other data shows less than 0.1 importance score. However, there are quite limited amount of data on the short tail ionizable lipid (number of carbons < 9) so the reliability of the importance score can potentially be improved by collecting data selectively on more short tailed ionizable lipids.

When viewing **Figure 7(C)** and **Figure 7(D)**, it is shown that there isn't comparability between the *in vivo* and *in vitro* model by comparison between liver and HepG2 (cell line that is often used to replicate the *in vivo* liver activities). Noticeably the data used for HepG2 are LNPs loaded with siRNA while for *in vivo* liver data the LNPs are loaded with mRNA. This might explain the discrepancies between the two sets of data.

At this stage, due to the limitations and biases present in the dataset, it would not be possible to comment on the comparability of the HepG2 cells and *in vivo* liver experiments. Notice that there are very little data on branched tail for *in vivo* liver and HepG2 cells, which is the leading cause of 0 importance score for those two graphs **Figure 6 (C)(D)**. The accuracy and reliability of these two models can potentially be further improved by collecting data on branched tail ionizable lipids and other features.

4.4.2 LNP Models

Table 1 shows the results of evaluate LNP properties, despite differences in ionizable lipids. It shows a certain level of correlation between pKa, relative abundances and the transfection performance of the LNP as a whole. This was driven by the recognition that LNP performance is a result of a complex interplay of various factors. The observed correlation between features highlights the role of LNP features such as relative abundances and pKa play alongside ionizable lipids. This approach enriches our understanding of LNPs, demonstrating that while ionizable lipids are crucial, there is also potential for optimising LNP composition. This underscores the importance of considering all aspects of LNP design for optimal performance.

4.4.3 Limitations and Recommendations

The primary limitation identified in this study is the constrained scope of our dataset. To enhance the robustness and performance of our model, particularly for *in vivo* liver and *in vitro* HepG2 analysis, it is recommended to target data collection to fill existing gaps. Specifically, there's a notable deficiency in data regarding branched tail ionizable lipids. Proactively acquiring this data could significantly enrich our dataset, leading to more comprehensive insights and potentially improved model accuracy.

Moreover, the method used to standardise the transfection data as discussed in 2.3.2 can be further improved by comparing the transfection efficiency between the benchmark lipids. The *in vivo* transfection data of all the common benchmarks including MC3, SM-102, ALC-0315 and cKK-E12 were reported in a paper^[23] and hence can be used as a base of comparison to re-calculate the relative transfection efficiency of all the lipids across papers.

One of the other limitations of this study is the sole reliance on the Random Forest model, without exploring alternative machine learning algorithms. When sufficient data is gathered, it is beneficial to compare the results of multiple models, such as lightGBM and artificial neural networks. Also, the scikit-learn package was treated as a "black box" which limits our insight into how the decision-making process was carried out by the model.

It should be noted that the model was built based on the assumption that the features of LNPs were not significant to the output. From **Table 2**, the LNP features did affect the output, and in future study it will be necessary to compare the results obtained from the ionizable lipid model against the same LNP features (e.g. relative abundance) across the experiments for a more accurate model.

The gaps in the dataset limited the number of features used in the models. Of the 20 features considered only 9 were implemented into the modelling stage due to the lack of available data. The subsequent dataset represents some of the most relevant ionizable lipid designs in the field however less mainstream and more niche features are not accounted for. Further features to include may be more to do with size – accounting for both the head and tail. This could be incorporated as a head to tail ratio, and an overall lipid size parameter such as molecular weight or carbon number.^[6]

The enhancement of our dataset with more comprehensive and representative data could significantly advance our modelling efforts. A refined model, encompassing a broader array of ionizable lipid and LNP features, has the potential to uncover deeper correlations among these variables. Such an enriched model is expected to not only provide a more nuanced understanding of the underlying mechanisms but also markedly improve

our predictive capabilities in the realm of LNP-based delivery systems.

5 Conclusions and Outlook

This study, through the establishment of a Random Forest model, reinforced the significance of 'Tail Length' as the paramount feature, which align with existing research. Branched tails and unsaturated tails are identified as secondary yet crucial features. The current model, while demonstrating moderate accuracy in predicting LNP transfection performance, aligns well with established literature in terms of feature importance.

This coherence suggests that the model has successfully captured key aspects of ionizable lipid characteristics. The discrepancies observed, primarily attributed to data limitations, highlight the potential for enhanced predictive accuracy with a more robust dataset. Thus, this study underscores the viability of machine learning approaches in this domain and indicates that refining the model and enriching the dataset could significantly advance our predictive capabilities in ionizable lipid-based delivery systems.

LNP models, incorporating additional characteristics, such as pK_a and relative abundance, demonstrated the potential for a more holistic analysis. The future trajectory of this research involves creating a more robust dataset and refining the predictive model. This approach holds promise as a preliminary screening tool for designing new ionizable lipids, as detailed in section 4.4.3. Implementing these recommendations could significantly improve our understanding and application of machine learning in LNP-based delivery systems.

6 References

- [1] Xu, L. et al. (2021) 'Lipid nanoparticles for Drug Delivery', *Advanced NanoBiomed Research*, 2(2). doi:10.1002/anbr.202100109.
- [2] Kim, Y.-K. (2022) 'RNA therapy: Rich history, various applications and Unlimited Future prospects', *Experimental & Molecular Medicine*, 54(4), pp. 455–465. doi:10.1038/s12276-022-00757-5.
- [3] Samaridou, E., Heyes, J. and Lutwyche, P. (2020) 'Lipid nanoparticles for nucleic acid delivery: Current perspectives', *Advanced Drug Delivery Reviews*, 154–155, pp. 37–63. doi:10.1016/j.addr.2020.06.002.
- [4] Schlich, M. et al. (2021) 'Cytosolic delivery of nucleic acids: The case of ionizable lipid nanoparticles', *Bioengineering & Translational Medicine*, 6(2). doi:10.1002/btm2.10213.
- [5] Zhang, Z. et al. (2022) 'Application of lipid-based nanoparticles in cancer immunotherapy', *Frontiers in Immunology*, 13. doi:10.3389/fimmu.2022.967505.
- [6] Öztürk, A.A., Gündüz, A.B. and Ozisik, O. (2019) 'Supervised machine learning algorithms for evaluation of solid lipid nanoparticles and particle size', *Combinatorial Chemistry & High Throughput Screening*, 21(9), pp. 693–699. doi:10.2174/1386207322666181218160704.
- [7] Wang, W. et al. (2022) 'Prediction of lipid nanoparticles for mRNA vaccines by the machine learning algorithm', *Acta Pharmaceutica Sinica B*, 12(6), pp. 2950–2962. doi:10.1016/j.apsb.2021.11.021.
- [8] Lewis, M.M., Beck, T.J. and Ghosh, D. (2023) Applying machine learning to identify ionizable lipids for nanoparticle-mediated delivery of mRNA [Preprint]. doi:10.1101/2023.11.09.565872.
- [9] Billingsley, M.M. et al. (2020) 'Ionizable lipid nanoparticle-mediated mRNA delivery for human CAR T cell engineering', *Nano Letters*, 20(3), pp. 1578–1589. doi:10.1021/acs.nanolett.9b04246.
- [10] Hu, B. et al. (2022) 'Thermostable ionizable lipid-like nanoparticle (Iland) for RNAi treatment of Hyperlipidemia', *Science Advances*, 8(7). doi:10.1126/sciadv.abm1418.
- [11] Hashiba, K. et al. (2022) 'Branching ionizable lipids can enhance the stability, fusogenicity, and functional delivery of mRNA', *Small Science*, 3(1). doi:10.1002/ssmsc.202200071.
- [12] Kim, M. et al. (2021) 'Engineered ionizable lipid nanoparticles for targeted delivery of RNA therapeutics into different types of cells in the liver', *Science Advances*, 7(9). doi:10.1126/sciadv.abf4398.
- [13] Płoński, P. (2020) Random Forest feature importance computed in 3 ways with python, MLJAR. Available at: <https://mljar.com/blog/feature-importance-in-random-forest/> (Accessed: 14 December 2023).

- [14] Han, X. et al. (2021) 'An ionizable lipid toolbox for RNA delivery', *Nature Communications*, 12(1). doi:10.1038/s41467-021-27493-0.
- [15] Tilstra, G. et al. (2023) 'Iterative design of ionizable lipids for intramuscular mRNA delivery', *Journal of the American Chemical Society*, 145(4), pp. 2294–2304. doi:10.1021/jacs.2c10670.
- [16] Paunovska, K. et al. (2022) 'The extent to which lipid nanoparticles require apolipoprotein E and low-density lipoprotein receptor for delivery changes with ionizable lipid structure', *Nano Letters*, 22(24), pp. 10025–10033. doi:10.1021/acs.nanolett.2c03741.
- [17] Da Silva Sanchez, A.J. et al. (2023) 'Substituting racemic ionizable lipids with stereopure ionizable lipids can increase mRNA delivery', *Journal of Controlled Release*, 353, pp. 270–277. doi:10.1016/j.jconrel.2022.11.037.
- [18] Ni, H. et al. (2022) 'Piperazine-derived lipid nanoparticles deliver mRNA to immune cells in vivo', *Nature Communications*, 13(1). doi:10.1038/s41467-022-32281-5.
- [19] Cornebise, M. et al. (2021) 'Discovery of a novel amino lipid that improves lipid nanoparticle performance through specific interactions with mRNA', *Advanced Functional Materials*, 32(8). doi:10.1002/adfm.202106727.
- [20] Shobaki, N., Sato, Y. and Harashima, H. (2018) 'Mixing lipids to manipulate the ionization status of lipid nanoparticles for specific tissue targeting', *International Journal of Nanomedicine*, Volume 13, pp. 8395–8410. doi:10.2147/ijn.s188016.
- [21] Hanafy, M.S. et al. (2023) 'Effect of the amount of cationic lipid used to complex siRNA on the cytotoxicity and proinflammatory activity of siRNA-lipid nanoparticles', *International Journal of Pharmaceutics*, X, 6, p. 100197. doi:10.1016/j.ijph.2023.100197.
- [22] Nie, B. et al. (2022) A novel regression method: Partial least distance square regression methodology [Preprint]. doi:10.2139/ssrn.4296907.
- [23] Lam, K. et al. (2023) 'Unsaturated, trialkyl ionizable lipids are versatile lipid-nanoparticle components for therapeutic and vaccine applications', *Advanced Materials*, p. 2209624. doi:10.1002/adma.202209624.
- [24] Preedy, V.R. et al. (2016) 'Chapter 5: Drugs of Abuse and the Internet', in *Neuropathology of Drug Addictions and Substance Misuse*. Amsterdam: Elsevier Academic Press, pp. 50–60.
- [25] Semple, S.C. et al. (2010) 'Rational design of cationic lipids for siRNA delivery', *Nature Biotechnology*, 28(2), pp. 172–176. doi:10.1038/nbt.1602.

Turning a new page on PAGE: Investigating the effect of oligonucleotide structure on gel mobility

Andrew Fung and Solen Marqueste

Department of Chemical Engineering, Imperial College London, U.K.

Abstract

Polyacrylamide gel electrophoresis (PAGE) is a common laboratory technique, favoured for its straightforward operation, used to analyse and separate oligonucleotides, with the variation in gel mobility that distinguishes samples from one another depending on the size – or steric bulk – of molecules. However, it is often the case where oligonucleotides hybridise in unexpected manners, leading to a wide range of possible 3D structures. As a result, there remains a degree of ambiguity when interpreting PAGE results. In order to quantify structural effects on gel mobility, oligonucleotides were purposefully manipulated with the hybridisation between single-stranded DNA (ssDNA) to form double-stranded DNA (dsDNA) studied. The two main structures explored were branched and looped oligonucleotides, with decreases in gel mobility observed as the branched overhang length and loop circumference increased. Next, this study investigated the impact of multiple poly(A) strands binding to a single complementary poly(dT) by increasing the proportion of poly(A) to poly(dT), as well as increasing adenine content within the poly(A): both changes resulted in an exponential reduction in gel mobility as multi-strand formation is favoured. Ultimately, this study's data demonstrates that steric bulk arising from non-linearities in structures has the greatest effect on gel migration compared to molecular weight, which was negligible. This study thus lays the groundwork for future studies exploring the impact of complex oligonucleotide structures on gel mobility.

Keywords: DNA hybridisation, Gel Mobility, PAGE, Secondary Structures, Steric Bulk

1. Introduction

1.1 Background

The SARS-Cov2 pandemic triggered a record-breaking speed in the history of RNA and DNA drug development, prompting the advancement of analytical technologies such as electrochemical biosensors to quickly determine features of interest such as sample size and structure [1]. However, these technologies are relatively nascent, presenting issues such as extensive sample processing steps and limited appropriate target analytes [1]. As such, more traditional techniques such as polyacrylamide gel electrophoresis (PAGE), which can be conducted in the majority of biochemistry laboratories, are still relied upon.

PAGE is a widely-used laboratory technique to separate biological macromolecules, including proteins and oligonucleotides, in an electric field. Oligonucleotides - short single- or double-stranded DNA or RNA molecules – have a negatively charged phosphate backbone, thus they will be attracted to and migrate to the positively charged anode placed at the bottom of the gel [2]:

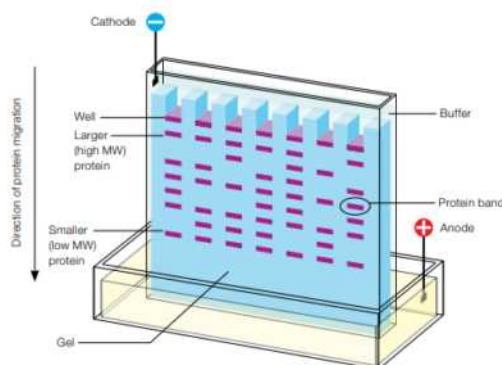


Figure 1 from [3]: Schematic of a PAGE protein separation

It is understood gels act as a size-selective sieve, that the rate of gel migration is a combination of characteristics of the electrophoresis system and the sample analyte, including the size and conformation of the nucleotide sample, concentration of the gel (i.e., size and density of the gel) and strength of the electric field [3]. As a result, the finished gel will present bands at varying heights. Of all the contributing factors above, the ‘size’ of the sample is the most ambiguous term and is hardest to quantify, as this is not merely the length of the oligonucleotide, which may present different spatial orientations.

The exact mechanism by which DNA fragments migrate through the gel is still up to debate, with the “biased reptation” model developed by *Noolandi et al.* currently prevailing. In free solutions, the mobility of DNA μ_0 is size-independent, however, electrophoretic gels reduce DNA mobility by decreasing the volume available

during migration [4]. To understand the model, two characteristic dimensions for DNA are needed: the contour length L , which for DNA is equal to the number of the bases in the chain times the distance per base, and the persistence length p describing the bending stiffness of DNA [5]. Linear DNA fragments of negligible width and length smaller than the average gel pore size \bar{a} act as a rod that can enter all pores and mobility as if unaffected by the gel.

For $L > p$, the fragment folds on itself and assumes a globular shape with radius of gyration R_g , with migration only occurring if enough pores with radius $a \geq R_g$ exist. For globular fragments, separation occurs via a sieving-like process and is described by the Ogston model [6]. For $R_g \gg \bar{a}$, the Ogston model predicts dramatically reduced gel mobility. However, this is not the case as large DNA fragments can migrate through the gel: this is explained by reptation theory, wherein molecules follow wormlike displacements within paths laid out within the cross-linked polymeric gel, moving headfirst through gel pores instead of migrating as a globule from one pore to another [7]. At this transition between Ogston and reptation theory, the mobility is given by [4]:

$$\frac{\mu(N)}{\mu_0} \propto \frac{1}{N} \propto \frac{1}{L} \quad \text{For } R_g > \bar{a}$$

, where N is the necessary number of pores to house the fragment. This inverse relationship between DNA length and its electrophoretic mobility in gels was first documented by de Gennes [7].

For significantly large molecules, the DNA globule deforms, and a non-linear mobility-size relationship appears, wherein mobility of large fragments in finite electric fields is affected more by field strength than fragment size: *Lumpkin, Dejardin and Zimm (LDZ)* proposed that the head of the DNA fragment aligns itself to the field and mobility should decrease monotonically yet non-linearly, eventually plateauing [8]. However, the biased reptation model from *Noolandi et al.* found by simulation that the end-to-end length in direction (x) of the electric field h_x changes during migration, with intermediate-sized molecules occasionally compacting with small end-to-end vectors rather than the expected elongated, linear form [9]. Compact configurations move much slower than their elongated counterparts and thus bands travelling faster do not always correspond to lower molecular weight DNA: this is known as band inversion and is less commonly seen for very long chains as they are less likely to compact [8].

Therefore, biased reptation offers a model for all regimes of DNA gel migration (Figure 2). However, the model is predominantly founded on computer simulation and as such, “systematic experimental studies of the full mobility versus molecular size” spectrum would greatly further understanding of DNA gel electrophoresis. Furthermore, the band inversion behaviour was only verified for linear ds-DNA fragments in agarose gels, thus mitigations suggested to

eliminate band inversion may be limited to experiments using similar structures, opening up many avenues of exploration. For example, de Gennes has demonstrated that melting DNA strands to cause unravelling and thus branched structures drastically reduces mobility, with an exponential decline in reptational transport as the length of the branch attached to the principal chain increases [4].

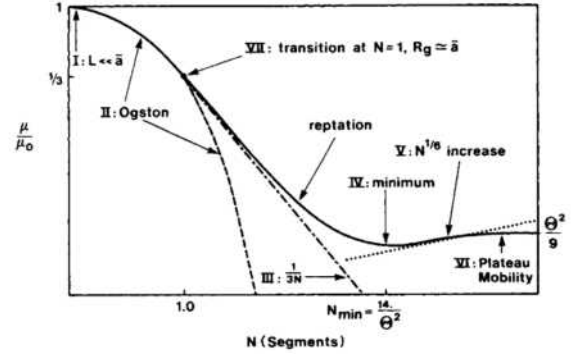


Figure 2 from [4]: Biased Reptation Model of DNA Gel Electrophoresis: log-log graph of relative mobility versus number of DNA segments

One should also note the biased reptation model utilises agarose as the gel medium. Compared to agarose, which has a larger pore size and are thus suitable for separating nucleic acids and larger proteins, polyacrylamide has a smaller pore size and is ideal for separating smaller proteins and nucleic acids. It has been observed that curved DNA molecules preferentially interact with the curved polyacrylamide gel matrix, thus presenting abnormally slow mobilities [10]. The added complexity of polyacrylamide gels motivates further research into this PAGE specifically, as well as the effect of other operating conditions such as temperature and salt concentration on DNA mobility.

1.2 Aims and Objectives

Hence, the aim of this study is to identify additional factors that influence oligonucleotide hybridisation and secondary structure formation as well as the effect that different resultant structures have on gel migration, aiming to reduce ambiguity in qualitative and quantitative DNA PAGE analysis. Different spatial conformations leading to steric hindrance have even been demonstrated to affect oligonucleotide function [11], further motivating deeper understanding of this behaviour to pave the way for the engineering of better characterised DNA products.

2. Materials and Methods

2.1 Designing oligonucleotides

As outlined above, previous experiments investigating band inversion only utilised linear dsDNA. As the combinations of different DNA/RNA products has evolved rapidly over recent years, it is more pertinent now to extend these experiments to a wider range of oligonucleotides.

During the course of this investigation into PAGE, various sequences of oligonucleotides were designed: Table S1 contains the list of all the oligonucleotides used in this study that will be referred to throughout this report. Of particular note are the poly(A) oligonucleotides consisting of repeated adenine (A) nucleotides and poly(dT)s consisting of repeated thymine (T) nucleotides, with A and T bases bonding with two hydrogen bonds; DNA nucleotides were used instead of RNA for increased stability and ease of storage [12]. Oligonucleotides were ordered from Integrated DNA Technologies (IDT) [13].

Linear oligonucleotides

Various linear oligonucleotides had to be designed. For this purpose, random DNA sequences were generated using the random DNA tool on Bioinformatics [14]; sequences were inputted into VectorBuilder to ensure that the linear structure would be the most likely to form, as molecules will stabilise themselves to form the lowest free energy configuration [15].

“Looped” oligonucleotides

Oligonucleotides that would form a heteroduplex with one loop were designed; a randomised opening sequence followed by a string of thymine bases and concluding with a reverse complementary sequence to the opening sequence would lead to a single thymine loop pinched at the top by a complementary backbone. As the number of consecutive thymine bases in the nucleotide sequence increases, so does the diameter of the loop.

These sequences were inputted into VectorBuilder to ensure the Gibbs free energy of formation was sufficiently negative so that these looped structures would be likely to form.



Figure 3: Designed looped secondary structure: loop(20). All oligonucleotides were drawn on BioRender. Blue: thymine; green and orange: reverse complementary sequences

2.2 Gel Preparation

Native-Polyacrylamide Gel Electrophoresis (PAGE) was used to separate oligonucleotide structures, more specifically, 15% PAGE gel was prepared (containing 10%vol of 10x TBE, 50%vol of 30% Acry/Bis, 39%vol of DI Water, 1%vol of 10% APS, 0.04%vol of TEMED) and for experiments which required a denaturing gel, a 15% PAGE gel with 8M urea was used (same composition as above with urea added). Note that a 15% gel was chosen as this concentration provides optimal resolution for the size range of oligonucleotides explored [16]. The gels were then transferred to a casting stand and left to polymerise for 30-45 minutes,

at which point full solidification had occurred and the gels had distinguished wells.

2.3 Sample preparation

Oligonucleotides were dissolved in phosphate-buffered saline (PBS) and all sample concentrations were measured using a UV-Vis spectrometer. Samples were then prepared to contain a fixed amount of material: 500ng of oligonucleotide and/or 550ng of poly(dT) unless specified otherwise, as well as a 1x concentration of Purple RNA Loading Dye.

Denaturation

For denaturing PAGE gel, samples were incubated in a thermocycler at 72°C for five minutes, then kept at 20°C for 5 min before being put on ice for 5 min to facilitate denaturation and ensure that all oligonucleotides would be linearised so that no secondary structures may form [17].

Hybridisation

For samples requiring hybridisation (e.g. poly(A)s with poly(dT)s), samples were incubated in a thermocycler at 40°C for 5 min and then kept at 20°C for 5 min.

2.4 PAGE Procedure and Analysis

Once fully solidified, gels were transferred to a BioRad Vertical Electrophoresis Cell [18], which was then filled with TBE 1x, where the samples were loaded after having washed out the wells. The electrophoresis cell was then run at a fixed voltage of 100V until the purple loading dye neared the bottom of the gel, indicating completion.

Staining

Gels were then removed from the gel box and stained by placing them in a solution of 50mL TBE 1x with 5µL of SYBR Safe, and agitating them for 30 min.

Visualisation and ImageJ analysis

Gels were imaged using a NuGenius+ UV transilluminator [19]: the fluorescent intensity of each gel band corresponds directly to the density of that sample within the gel. Image analysis was conducted using ImageJ with the mean gray scale or intensity of each band corresponding to the concentration of each complex formed. This was verified by analysing the brightness of linear oligonucleotide comp2 at different sample masses. As seen in Figure 4, the mean gray scale or brightness is directly proportional to an increasing sample mass, thus this is a valid tool of measurement for subsequent quantitative analysis.

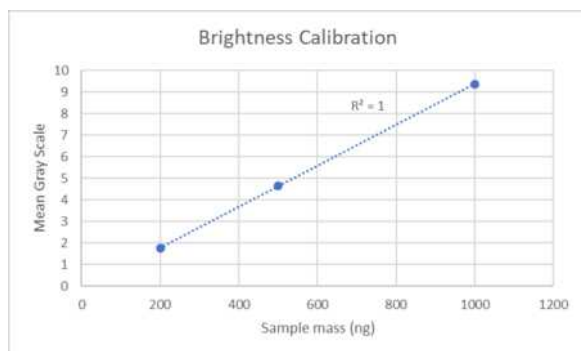


Figure 4: Brightness Calibration Curve of comp2 at three different sample masses

Note that each experiment was conducted twice to mitigate the effect of any random errors. Brightness measurements taken for each band were an average of both equivalent gels and background readings were subtracted to get the raw brightness of bands. Furthermore, final images presented in the body of this report were colour inverted from their original scans, thus the term ‘brightness’ is used throughout despite gel figures consisting of black bands.

3. Results and Discussion

Firstly, the matter of increasing size was explored. In the *Noolandi* model, DNA size was equivalent to contour length and varied by continuously adding segments of fixed length [4]. However, structure conformation can occur via multiple pathways and as such band inversion may occur to different extents depending on the way DNA is synthesised. For example, in the formulation of double-stranded DNA/RNA products via the hybridisation of single complementary oligonucleotides, there is the possibility of mismatched base pairing, leading to unexpected sequence conformations [20]: the distortions of these “heteroduplexes” cause them to have reduced mobility compared to “homoduplexes”, which have perfectly paired bases from end to end [21]; the reduction in polyacrylamide gel mobility is proportional to the degree of divergence between the two constituent single-stranded sequences [22].

3.1 Experimenting with overhangs

To initially determine the effect of different oligonucleotide hybridised structures on gel mobility, an experiment was devised wherein gel electrophoresis would be used to visualise hybridised structures between dT(60) and oligonucleotides with increasing poly(A) content (see supplementary Table S1). The expectation was that a band would show at 110 bases if the 50 base-long poly(A) and 60 base-long poly(dT) had hybridised, alongside one band at 50 bases representing leftover non-hybridised poly(A) and one band at 60 bases representing leftover non-hybridised dT(60). This expectation is based on the common assumption that PAGE separates samples based on their molecular

weight.

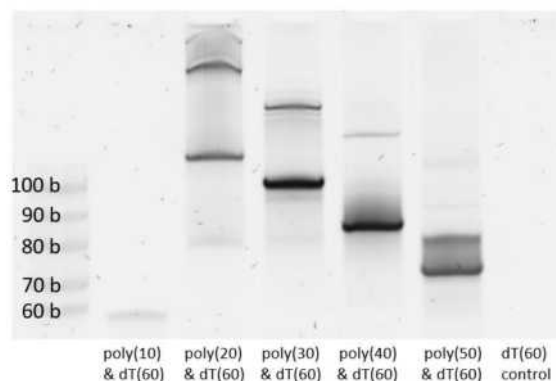


Figure 5: Hybridisation of poly(A)s with dT(60) in PAGE

After hybridising the poly(A)s with dT(60), it was concluded that poly(10) contains insufficient adenine bases to hybridise properly with the poly(dT), with at least twenty hydrogen bonds being necessary to maintain the hybridised structure. An extremely bright stairwell can be seen (Figure 5) for poly(20) to poly(50): this will become a recurring theme in many experiments. Despite all poly(A)s hybridising to form an oligonucleotide complex of equal 110 bases weight, only the poly(20) well displayed a bright band around 110 bases, with bands for subsequent samples falling in this stairwell. It can also be observed that the wells present with a second bright band above the “hybridised band” and that those higher bands not only also form a stairwell but are decreasing in brightness between poly(20) and poly(50). One can also observe that the dT(60) control has seemingly disappeared. This will be further discussed in Section 3.4.

Thus, two hypotheses were posed. Firstly, that the main contributor to mobility within the gel was the amount of resistance that the sample would present, which would be related to its surface area or ‘steric bulk’. Secondly, that a fraction of poly(dT)s had more than one poly(A) attached to it and this was the source of the second higher bands with drastically reduced mobility.

The first hypothesis would explain the presence of the stairwell. Indeed, when mixed, the poly(A)s and poly(dT)s may not fully hybridise but rather only partially do so, resulting in structures with a short double-stranded segment and single-stranded overhangs. This was corroborated by IDT’s OligoAnalyzer tool [23], with these branched heteroduplex structures having extremely low Gibbs free energies of formation and thus formation is very likely. Additionally, the brightness of each band was compared to the free energy of its suspected corresponding structure to demonstrate that the probability of each structure forming does in fact match the observed abundance following brightness analysis. Note that the poly(50) result was treated as anomalous as the presented brightness was abnormally low, as the band appears more as a smear in Figure 5. As seen in Figure 6, brightness decays rapidly as structures

becoming less thermodynamically favourable, with an x-axis intercept of $\Delta G = -21.22 \text{ kcal/mol}$. This is notably more negative than the expected free energy for a hybridisation between poly(10) and dT(60) where $\Delta G = -17.5 \text{ kcal/mol}$, suggesting that this combination is below a threshold necessary for the hybridised structure to form and thus does not present in the gel. However, note that the extrapolated line of best fit in Figure 6 is only based on three data points and further experiments involving poly(A)s with a wider range of adenine content should be conducted to confirm this theory.

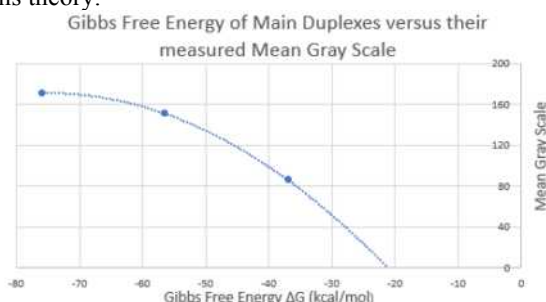


Figure 6: Comparing free energy to brightness of main duplex bands

Bulkier complex structures with more nucleotide overhangs are hypothesised to have higher resistance, migrate more slowly through the gel and hence present bands near the top. Thus, structures with the same molecular weight can present different results following PAGE analysis. The length of the overhang is expected to decrease from poly(20) to poly(50), as modelled in Figure 7, which would thus explain why they present a descending stairwell as the mobility increases accordingly. This corroborates experimental findings from de Gennes, whereby DNA mobility decreases as the length of the overhang branching off the main strand increases [24].

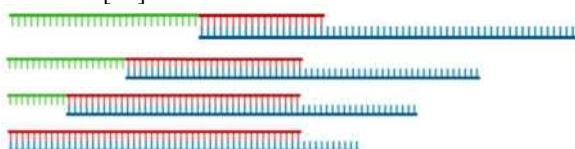


Figure 7: From top to bottom: hybridised poly(20) with dT(60); hybridised poly(30) with dT(60); hybridised poly(40) with dT(60); hybridised poly(50) with dT(60). Blue: thymine; red: adenine; green: random non-complementary sequence

The second hypothesis would explain the presence of the higher stairwell as perhaps multiple poly(A)s attach to a single poly(dT) to form higher molecular weight structures (Figure 8). These would still present decreasing overhangs between poly(20) to poly(50) and thus the same stairwell behaviour.

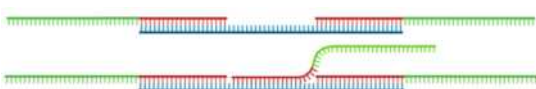


Figure 8: Top: Two poly(20)s attached to a single dT(60); Bottom: Three poly(20)s attached to a single dT(60)

Furthermore, it is hypothesised that poly(A)s with greater adenine content are less likely to exhibit these multi-strand complexes as more thymine bases on the dT(60) strand would be occupied by a single poly(A), discouraging an additional one from attaching. This in turn would explain the decreasing brightness of the secondary stairwell bands between poly(20) to poly(50). Following ImageJ analysis comparing the formation of the proposed multi-strand complex with the expected duplex (Figure 9), this hypothesis was supported, with an exponential decrease in the brightness and thus abundance of multi-strand duplexes as the number of adenine bases increased.

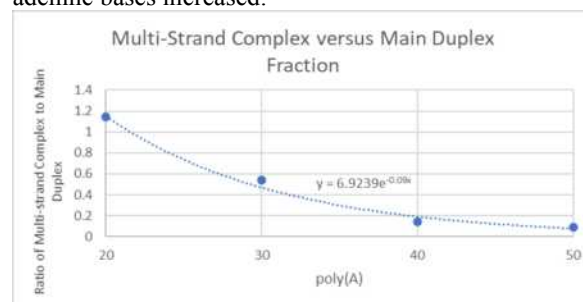


Figure 9: Fraction of multi-strand to main heteroduplex formation between poly(A)s and dT(60)

Unfortunately, OligoAnalyzer was unable to predict the formation of these multi-strand complexes and thus the analysis from Figure 6 could not be applied here. This is an area of improvement for future oligonucleotide modelling tools.

3.2 Hypothesis 1: Steric bulk is the main driver of gel mobility

To thoroughly investigate the first hypothesis and determine whether gel electrophoresis distinguishes samples based on molecular weight or steric bulk, two experiments were designed. The first involved running single-stranded and double-stranded DNA fragments of the same length in the gel and the second entailed purposefully exploring secondary structures and heteroduplexes to quantify the impact of steric bulk on gel mobility.

Single-strand versus double-strand

The first experiment was conducted as follows: a 50 base long single-stranded oligonucleotide sequence (comp1) was randomly generated alongside its reverse complementary counterpart (comp2). Both strands were designed to be linear with no other secondary structures forming. When both complementary strands are mixed, this should result in a fully hybridised 50bp homoduplex. Single-stranded oligonucleotide comp2, as well as the hybridised double-strand, were run in the same gel (Figure 10).

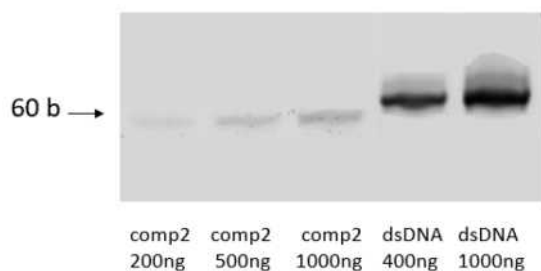


Figure 10: Comparing single stranded and double stranded oligonucleotides of equal length

From Figure 10, indicates that the single-stranded and double-stranded samples both show a single band at around 50 bases, despite the dsDNA being twice as heavy as the single-stranded comp2. Additionally, the linear dsDNA 50bp duplex has greater gel mobility than even the poly(50) & dT(60) branched duplex from Figure 5, with duplexes exhibiting greater branching decreasing in mobility further still. This supports the theory that weight had little to no impact on gel mobility. Thus, the focus of this study will continue to be the effect of oligonucleotide structure on mobility.

Looped Oligonucleotides

The second experiment devised to confirm that steric bulk is the main driver behind gel mobility, was to test the effect of “looped” oligonucleotide structures on gel migration.

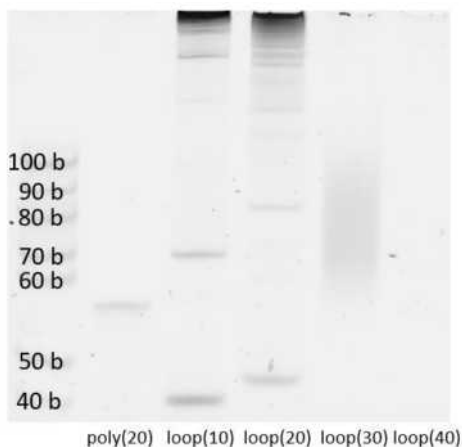


Figure 11: Formation of looped oligonucleotide secondary structures

In this experiment, all loop sequences are 50 bases long but form a thermodynamically favourable secondary heteroduplex structure, wherein the circumference of the looped portion increases from 10 to 40 bases.

The result of this experiment is that loop(10) and loop(20) present two bands which form stairwells of decreasing mobility, as well as extremely bright bands at the top of the gel, loop(30) shows a smear and loop(40) does not appear at all. This phenomenon is addressed in Section 3.4.

The existence of multiple stairwells likely arises from the fact that the loops are able to attach to each other in more linear configurations instead of forming the desired secondary structure. From IDT’s OligoAnalyzer tool [23], up to 19 alternate homo-dimers – duplexes resulting from hybridisation of identical strands - structures are possible, although it is important to note that many of these structures have small free energies of formation and are not as likely to form. These kinked alternative structures, such as in Figure 12, may also present overhangs, adding further resistance similarly to the structures discussed in Section 3.1.

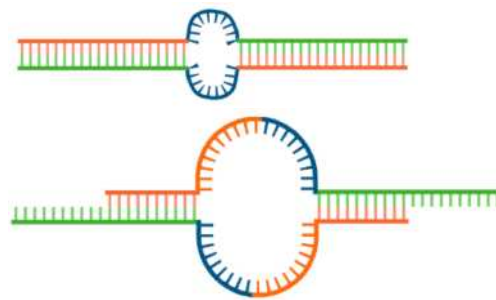


Figure 12: Two possible kinked homo-dimer formed from loop(10)s. The top structure is the most stable and thus likely to form

As an experimental control to compare the looped oligonucleotides to their linear counterparts, this experiment was repeated using urea as a denaturing agent. Urea is used in PAGE to denature secondary and hybridised structures by reducing them to single-stranded fragments [17], thus separating samples purely on molecular weight. It was thus expected that all samples would present bands at a single height. However, this was not the case. As seen in Figure 13, a similar lower stairwell may be seen, whereas all other structures have been eliminated.

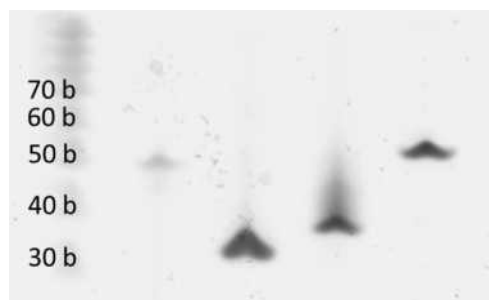


Figure 13: Looped Oligonucleotides in denatured gel

Mickel et al. have demonstrated previously that circular closed DNAs do not follow the same wormlike mechanism proposed by the reptation model, due to the lack of free ends on the DNA chain. Below a critical molecular weight, the reduced radius of gyration R_g compared to its linear counterpart leads to greater mobility for circular DNAs, suggesting that the greater mobility bands seen in Figure 13 could be the expected looped secondary structure. However, this phenomenon

is dependent on the size of the loop, as well as the gel concentration, with the exact relationship dictating this balance not fully understood [25].

Compared to the aforementioned kinked alternative structures, the main looped secondary structure of interest is expected to demonstrate the least steric bulk and thus migrate furthest down the gel. To further support this, the Gibbs free energy of formation is lower for the expected loop structure ($\Delta G = -39.2 \text{ kcal/mol}$) compared to even the most stable kinked structure ($\Delta G = -36.91 \text{ kcal/mol}$) [23]. As result, this loop structure would be the least susceptible to urea denaturation. However, these values are not dissimilar, and this should by no means be seen as a definitive proof, with further quantitative structure analysis recommended to confirm this hypothesis, as discussed in Section 4.2.

Going by this explanation, the secondary stairwells of loop(10) and loop(20) at 70b and 80b respectively likely correspond to the top diagram of Figure 12, whereby instead of folding to form the expected looped secondary structure in Figure 3, two identical oligonucleotides hybridise together to form a homodimer, with the reverse complementary sequences interrupted by the thymine loop causing a kink that occupies additional space in an otherwise linear structure: this ‘kinked’ structure has reduced gel mobility compared to the linear poly(20) structure. From Section 3.2, it has been demonstrated that molecular weight has a negligible effect on mobility compared to structure. As a result, it is most likely that the space occupied by the kink is the main contributor to the decreased mobility. This kink expands as the size of the loop increases, offering further gel resistance and explaining the decrease in the secondary stairwell mobility from loop(10) to loop(20). The same trend is observed for the structures within the denatured gel (Figure 13).

Finally, the bright cluster at the top of loop(10) and loop(20) wells arising from the stacking of multiple bands are likely due an agglomeration of the remainder of the extremely bulky kinked structures predicted by OligoAnalyzer. The smear from loop(30) may have resulted from these structures eventually migrating into the well rather than being stuck at the top, although the exact reason for this is unknown and again necessitates further quantitative structure analysis. Despite these extraneous features, the main feature of interest, namely the decreasing mobility of larger loops, further supports Hypothesis One.

3.3 Hypothesis 2: Oligonucleotides form multi-stranded structures

Addressing Hypothesis Two, it is clear that it is incorrect to assume that a single poly(A) strand always binds with a single poly(dT) strand. In this study, a simplified

approach was taken to infer the probability of these different structures forming.

Redesigned poly(dT)s

Instead of using dT(60), new poly(dT) oligonucleotides were designed so that there is the same number of thymine bases as adenine for each poly(A) strand such a perfectly complementary A-T sequence forms, with the remaining bases made up of repeating randomly generated 10bp strands (Table S1) to form a non-complementary overhang:

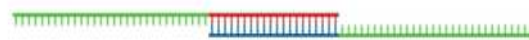


Figure 14: Poly(20) with new designed dT(20)

As illustrated in Figure 14, this should guarantee the binding of a single poly(A) to a single poly(dT), whilst still presenting the stairwell resulting from overhang resistance: experiments verified this.

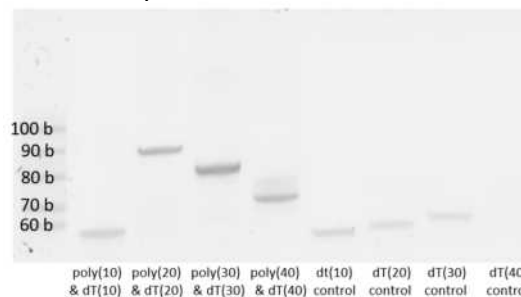


Figure 15: Hybridisation of poly(A)s with accompanying poly(dT)s

From Figure 15, one can see that the secondary stairwell has been eliminated whilst still maintaining a primary stairwell, supporting the hypothesis that multiple poly(A)s can bind to a single poly(dT) and that this was the cause for the secondary ladder. One can also observe that the poly(10) is still not hybridising, confirming that a 10 base bond is not stable enough. Finally, one can see that dT(40) did not stain. This is explored in further details in Section 3.4.

One should also note that the primary ladder now starts slightly lower around 90b; a decrease was expected as hybridised should now present less resistance due to minimised overhang (now 10 bases shorter than before as the tailored poly(dT)s are 50 bases long where dT(60) was 60 bases long), thus travelling faster down the gel.

Varying ratios

Another factor hypothesised to contribute to this multi-strand complex formation is the ratio of poly(dT) to poly(A) mixed in each aliquot: having excess poly(A) is likely to lead to more than one poly(A) binding to a single poly(dT).

To test out this hypothesis, an experiment was designed wherein poly(20) and dT(60) were used as this would leave the greatest number of thymine bases available for bonding from multiple poly(A)s whilst also ensuring the

resulting complex would be stable. From Figure 16, several phenomena may be observed. Firstly, the expected main hybridised structure at 110bp that was also observed in Figure 5 presented more intense bands when dT(60) was in excess, with this structure disappearing completely in an excess of poly(A) with ratios 1:5 and 1:10. Conversely, the band corresponding to the secondary stairwell multi-strand complex increased in brightness as poly(A) was added in excess. With an excess of poly(A), significant smearing at the very top of the well can be observed, showing an abundance of heavy structures which get stuck at the top of the well.

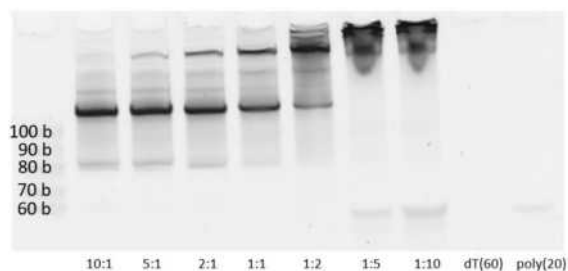


Figure 16: Hybridisation between dT(60) and poly(20) with changing sample ratios. From left to right: 10:1, 5:1, 2:1, 1:1, 1:2, 1:5, 1:10

3.4 Nucleic Acid Staining Dye

As alluded to throughout this report, a recurring phenomenon that may be seen in Figures 5, 15 and 16 is the disappearance of poly(dT)s, as well as loop(40) in Figure 11. The common denominator between these cases is that involve species with a greater number of repeating bases, prompting the question of whether this affects the nucleic staining method, rendering them invisible in the gel results.

Nucleic acid staining dyes operate via either intercalation or minor groove binding, with the former involving insertion of a small molecule between adjacent base pairs, causing structural changes in the nucleic acid [26]. Minor Groove Binders (MGBs), such as SYBR Safe [27], hydrogen bond to base pair edges - especially A-T - alongside non-covalent interactions with the minor groove wall of the DNA [28]. Various factors, including the dye's physicochemical properties and geometry, affect the mode of nucleic acid binding [29] and future studies could well benefit from testing of a variety of staining dyes on the structures discussed in this report.

4. Conclusions and Outlook

4.1 Conclusion

With its main strengths laying not only in its low cost, but its extreme simplicity and accessibility, lending to its prevalence as a 'litmus test' to determine basic nucleic acid properties, PAGE is an invaluable technique that is still not fully understood.

In this study, the effect of oligonucleotide structures on polyacrylamide gel mobility was explored, with the aim of raising awareness towards potential challenges that could cause ambiguity when interpreting results. It was demonstrated, by purposely introducing complexity in hybridised and secondary structures, that the steric bulk is the main contributor to gel retardation. This was demonstrated with both branched and looped oligonucleotides, with both instances showing that any spatial deviations from a purely linear structure led to changes in gel mobility, despite having the same molecular weight. As the region of interest introducing steric bulk, namely the branch length and loop circumference, was increased, the gel mobility was reduced.

From the experiments conducted, the gel mobility for structures of similar molecular weight is lowest for the branched structures, followed by the linear dsDNA equivalent. Looped structures displayed the greatest gel mobility although it is important to note that, due to time constraints, only a limited number of looped structures were explored. Future studies could entail investigating oligonucleotides with loop circumference increasing in finer increments, as well as increasing the number of bases of the oligonucleotide, allowing for larger loops.

A second hypothesis proposed the existence of alternate hybridisation structures, such as the case where multiple poly(A)s hybridise onto a single poly(dT). Not only was this hypothesis supported by the existence of multiple gel stairwells corresponding to bulkier structures, but also multi-strand complex formation was encouraged by increasing the ratio of poly(A)s to a single poly(dT). Furthermore, the probabilities of these different complexes forming was inferred by using online tools such as VectorBuilder and OligoAnalyzer to predict their Gibbs free energy of formation, with further definitive evidence via quantitative ImageJ analysis provided, wherein the brightness of measured bands corresponds directly to the abundance of sample present. This conclusion highlights the importance of proper oligonucleotide design and consideration of sample preparation, as the formation of multiple, unexpected bands can lead to ambiguous gel results and incorrect structure identification.

Overall, this study opens up many avenues for exploration, especially as the surge in DNA/RNA product interest, such as mRNA vaccines or CRISPR-Cas9 gene editing and its applications, will require clear analytical understanding of these structures. As the ease of use and low cost of PAGE lend to its attractiveness as a laboratory-scale technique, it is important now more than ever to ensure that PAGE results can be relied upon.

4.2 Limitations and Future Work

As mentioned briefly when analysing the looped oligonucleotide structures, this study would further benefit from an alternative technique to accurately and quantitatively determine the structure of resultant hybridised fragments to predict the effect on its mobility

within electrophoretic gels; single-molecule localisation microscopy (SMLM) can fluorescently label oligonucleotides with a single dye and measure intensity of dye [30], however this method is extremely expensive and relatively inaccessible. On the other hand, PAGE and other electrophoretic gel techniques are relatively fast and simple to conduct.

Although ImageJ analysis was treated as a quantitative method to determine band brightness, band selection was still determined manually and thus a degree of human error was introduced. This again encourages the use of more objective techniques such as SMLM to support this study's findings.

Finally, as a direct continuation of this study, further experiments should be conducted with an increased range of oligonucleotide structures, not only branched and looped. It would be interesting to quantitatively measure the mobility of different structures, both within the same family and universally, to draw comparisons to the theoretically biased reptation model.

Acknowledgements

We are extremely grateful to Dr Ying Tu, Dr Karen Polizzi and all members of the Polizzi Lab for their continued guidance and support.

References

- [1] Santhanam, M., Algov, I. & Alfonta, L., "DNA/RNA Electrochemical Biosensing Devices a Future Replacement of PCR Methods for a Fast Epidemic Containment," *Sensors (Basel)*, 2020. <https://doi.org/10.3390/s20164648>
- [2] Lee, P. Y., Costumbrado, J., Hsu, C.-Y. & Kim, Y. H., "Agarose Gel Electrophoresis for the Separation of DNA Fragments," *Journal of Visualized Experiments*, 2012. <https://doi.org/10.3791/3923>
- [3] Bio-Rad, "A Guide to Polyacrylamide Gel Electrophoresis and Detection," [Online]. Available: https://www.bio-rad.com/webroot/web/pdf/lsr/literature/Bulletin_6040.pdf. [Accessed November 2023].
- [4] Slater, G. W. & Noolandi, J., "The biased reptation model of DNA gel electrophoresis: mobility vs molecular size and gel concentration," *Biopolymers*, vol. 28, no. 10, pp. 1781 - 1791, 1989. <https://doi.org/10.1002/bip.360281011>
- [5] Seol, Y., Li, J., Nelson, P. C., Perkins, T. T. & Betterton, M. D., "Elasticity of Short DNA Molecules: Theory and Experiment for Contour Lengths of 0.6–7 μm ," *Biophysical Journal*, vol. 93, no. 12, pp. 4360 - 4373, 2007. <https://doi.org/10.1529/biophysj.107.112995>
- [6] Ogston, A. G., "The spaces in a uniform random suspension of fibres," *Transactions of the Faraday Society*, vol. 54, pp. 1754 - 1757, 1958. <https://doi.org/10.1039/TF9585401754>
- [7] de Gennes, P. G., "Reptation of a Polymer Chain in the Presence of Fixed Obstacles," *The Journal of Chemical Physics*, vol. 55, pp. 572 - 579, 1971. <https://doi.org/10.1063/1.1675789>
- [8] Slater, G. W., Rousseau, J. & Noolandi, J., "On the stretching of DNA in the reptation theories of gel electrophoresis," *Biopolymers*, vol. 26, pp. 863-872, 1987. <https://doi.org/10.1002/bip.360260607>
- [9] Doi, M., Kobayashi, T., Makino, Y., Ogawa, M., Slater, G. W. & Noolandi, J., "Band Inversion in Gel Electrophoresis of DNA," *Physical Review Letters*, vol. 61, pp. 1893 - 1896, 1988. <https://doi.org/10.1103/PhysRevLett.61.1893>
- [10] Stellwagen, N.C., "Electrophoresis of DNA in agarose gels, polyacrylamide gels and in free solution," *Electrophoresis*, vol. 30, pp. S188 - S195, 2009. <https://doi.org/10.1002/elps.200900052>
- [11] Troisi, R., Napolitano, V., Rossitto, E., Osman, W., Nagano, M., Wakui, K., Popowicz, G. M., Yoshimoto, K. & Sica, F., "Steric hindrance and structural flexibility shape the functional properties of a guanine-rich oligonucleotide," *Nucleic Acids Research*, vol. 51, no. 16, pp. 8880 - 8890, 2023. <https://doi.org/10.1093/nar/gkad634>
- [12] Technology Networks - Genomics Research, "DNA vs. RNA – 5 Key Differences and Comparison," [Online]. Available: <https://www.technologynetworks.com/genomics/articles/what-are-the-key-differences-between-dna-and-rna-296719#:~:text=DNA%20is%20more%20stable%20due,from%20DNA%20during%20protein%20synthesis.> [Accessed November 2023].
- [13] Integrated DNA Technologies, "DNA Oligonucleotides," [Online]. Available: <https://eu.idtdna.com/pages/products/custom-dna-rna/dna-oligos/custom-dna-oligos>. [Accessed October 2023].
- [14] BioInformatics, "Sequence Manipulation Suite - Random DNA Sequence," [Online]. Available:

- https://www.bioinformatics.org/sms2/random_dna.html. [Accessed October 2023].
- [15] VectorBuilder, “DNA Secondary Structure,” [Online]. Available: <https://en.vectorbuilder.com/tool/dna-secondary-structure.html>. [Accessed October 2023].
- [16] Integrated DNA Technologies, “Running agarose and polyacrylamide gels,” [Online]. Available: <https://eu.idtdna.com/pages/education/decode/article/running-agarose-and-polyacrylamide-gels>. [Accessed November 2023].
- [17] Summer, H., Grämer, R. & Dröge, P., “Denaturing Urea Polyacrylamide Gel Electrophoresis (Urea PAGE),” *Journal of Visualized Experiments*, 2009. <https://doi.org/10.3791/1485>
- [18] Bio Rad, “Electrophoresis Chambers,” [Online]. Available: <https://www.bio-rad.com/en-uk/category/electrophoresis-chambers?ID=N3F2N0E8Z>. [Accessed November 2023].
- [19] Syngene, “NuGenius/NuGenius+ - Gel imaging at a touch,” [Online]. Available: <https://www.syngene.com/product/nugenius-gel-imaging/>. [Accessed November 2023]
- [20] Bhattacharyya, A. & Lilley, D. M., “The contrasting structures of mismatched DNA sequences containing looped-out bases (bulges) and multiple mismatches (bubbles).,” *Nucleic Acids Research*, vol. 17, 1989. <https://doi.org/10.1093/nar/17.17.6821>
- [21] Upchurch, D. A., Shankarappa, R. & Mullins, J. I., “Position and degree of mismatches and the mobility of DNA heteroduplexes,” *Nucleic Acids Research*, vol. 28, 2000. <https://doi.org/10.1093/nar/28.12.e69>
- [22] Delwart, E. L., Pan, H., Sheppard, H. W., Wolpert, D., Neumann, A. U., Korber, B. & Mullins, J. I., “Slower Evolution of Human Immunodeficiency Virus Type 1,” *Journal of Virology*, 1997. <https://doi.org/10.1128/jvi.71.10.7498-7508.1997>
- [23] Integrated DNA Technologies, “OligoAnalyzer Tool,” [Online]. Available: <https://eu.idtdna.com/pages/tools/oligoanalyzer?returnurl=%2Fcalc%2Fanalyze>. [Accessed December 2023].
- [24] Lerman, L. S. & Frisch, H. L., “Why does the electrophoretic mobility of DNA in gels vary with the length of the molecule?,” *Biopolymers*, vol. 21, no. 5, 1982. <https://doi.org/10.1002/bip.360210511>
- [25] Mickel, S., Arena Jr., V & Bauer, W., “Physical properties and gel electrophoresis behavior of R 1 2-derived plasmid DNAs,” *Nucleic Acids Research*, vol. 4, no. 5, 1977. <https://doi.org/10.1093/nar/4.5.1465>
- [26] Mukherjee, A. & Sasikala, W. D., “Chapter One - Drug–DNA Intercalation: From Discovery to the Molecular Mechanism,” *Advances in Protein Chemistry Structural Biology*, vol. 92, 2013. <https://doi.org/10.1016/B978-0-12-411636-8.00001-8>
- [27] Haines, A. M., Tobe, S. S., Kobus, H. J. & Linacre, A., “Properties of nucleic acid staining dyes used in gel electrophoresis,” *Electrophoresis*, vol. 36, no. 6, 2016. <https://doi.org/10.1002/elps.201400496>
- [28] Lauria, A., Montalbano, A., Barraja, P., Dattolo, G. & Almerico, A. M., “DNA minor groove binders: an overview on molecular modeling and QSAR approaches,” *Current Medicinal Chemistry*, vol. 14, no. 20, 2007. <https://doi.org/10.2174/092986707781389673>
- [29] Del Castillo, P., Horobin, R. W., Blázquez-Castro, A. & Stockert, J. C., “Binding of cationic dyes to DNA: distinguishing intercalation and groove binding mechanisms using simple experimental and numerical models,” *Biotechnic & Histochemistry*, vol. 85, no. 10, 2010. <https://doi.org/10.3109/10520290903149620>
- [30] Lelek, M., Gyparakı, M. T., Beliu, G., Schueder, F., Griffié, J., Manley, S., Jungmann, R., Sauer, M., Lakadamyali, M & Zimmer, C., “Single-molecule localization microscopy,” *Nat Rev Methods Primers*, vol. 1, 2021. <https://doi.org/10.1038/s43586-021-00038-x>

Modelling the Solubility of Cholesterol in Primary Alcohols using the SAFT- γ Mie Equation of State

Sathya Thakrar and Anand Vadukul

Department of Chemical Engineering, Imperial College London, U.K.

Abstract Predicting the solubility of cholesterol is important due to its role in the pathogenesis of gallstones and atherosclerosis. This is in particular due to the solid–solid phase transition that occurs at around body temperature. In this work, we develop a predictive model for the solubility behaviour of cholesterol in primary alcohols including 1-butanol, 1-pentanol and 1-hexanol by employing the SAFT- γ Mie group-contribution approach. For longer chain alcohols, solvate influences on the crystalline structure become significant which our solid phase description does not account for. We validate the use of combining rules by modelling the vapour pressure and isobaric vapour-liquid phase equilibria of pure 2,2-dimethylcyclohexanol and cyclohexanol + isophorone respectively, as these contain the uncharacterised moieties of cholesterol. We demonstrate the transferability of these parameters by obtaining an excellent prediction for the liquid heat capacity of cholesterol prior to modelling cholesterol solubility. The robustness of the model is quantified by comparing theoretical predictions with experimental data. Excellent agreement between the calculations and experimental data are observed for cholesterol in 1-butanol, along with good agreements for cholesterol in 1-pentanol and 1-hexanol. For these systems, the SAFT- γ Mie approach captures the correct order of magnitude of solubility and trajectory of the solubility curves. We thus demonstrate the versatility of the SAFT- γ Mie approach in modelling the solubility of large molecules.

1. Introduction

Cholesterol plays an essential role in biological membranes and is a precursor for the synthesis of steroid hormones and bile acids [1]. It is an organic molecule consisting of a steroid nucleus, a rigid tetracyclic hydrocarbon structure, in between an aliphatic hydrocarbon tail on one end, and a hydroxyl group on the other. This hydroxyl group contributes to the amphipathic nature of the molecule [2].

The presence of cholesterol crystals in gallstones and atherosclerotic plaques [3], makes the solubility behaviour of cholesterol particularly relevant. The existence of a solid-solid phase transition around body temperature is a central feature of this solubility behaviour, with both solid phases having been detected in gallstones [4], and has been the focus of many experimental studies [4]–[6]. This polymorphic nature of cholesterol arises due to the rotational isomerism exhibited by the aliphatic tail [7].

The crystalline properties of cholesterol have been shown to be influenced by the solvent itself [6], [8]. It is, therefore, interesting to study the solubility of cholesterol in a homologous series of solvents as they exhibit regular patterns in chemical and thermodynamic properties. This provides a basis by which the solvent effects on the crystalline structure can be assessed. Primary alcohols are a pertinent homologous series due to their polarity which decreases with carbon chain length. Consequently, the amphipathic nature of cholesterol may lead to an interesting relationship between solubility and alcohol chain length. This allows for systematic solvent selection in physiological applications, including the dissolution of gallstones.

The development of theoretical tools to accurately model the thermodynamic behaviour of substances is valuable insofar as it can alleviate the need for experimental efforts. Equations of state (EoS) are particularly useful for this. The statistical associating fluid theory (SAFT) is one such example and can be used to model complex fluids including associating

fluids, such as those exhibiting hydrogen bonding. Several iterations of SAFT have been developed since its inception by Chapman *et al.* in 1989 [9]. These include SAFT-VR, which treats molecules as homonuclear chains of bonded segments interacting through a square-well potential [10], and PC-SAFT which uses a hard-chain reference fluid [11]. In this study, we utilise the SAFT- γ Mie EoS, comprising of a group-contribution approach, whereby molecules are broken down into their constituent functional groups which interact via a Mie potential (a generalised Lennard-Jones potential) [12]. By employing this group-contribution approach, we eliminate the need for experimental data specific to the molecules of interest. This is particularly useful in our case, due to the scarcity of pure cholesterol data. This highlights a key strength of the EoS. Additionally, this EoS has been shown to accurately model the thermodynamic properties and fluid-phase behaviours of complex mixtures [13].

In Sections 2 and 3, we present the SAFT- γ Mie theory and model as well as the solid-liquid phase equilibria relations utilised. In Section 4, we validate the use of combining rules as sufficient initial estimates for the underlying group parameters within the SAFT- γ Mie EoS. We also present the prediction of the liquid heat capacity of pure cholesterol using the SAFT- γ Mie approach, in addition to the predictions for solubility of cholesterol in primary alcohols. Concluding remarks are provided in Section 5.

2. Theory

2.1 SAFT- γ Mie Theory and Model

At the foundation of the SAFT- γ Mie Equation of State, molecules are depicted as associating heteronuclear chains of fused-spherical segments interacting through variable-range Mie potentials [12]. These segments are accompanied by associative sites representing short-range interactions (i.e. hydrogen bonding) modelled by the square-well potential. Based

on a perturbation approach, the total Helmholtz free energy A , of a fluid mixture consisting of non-ionic, associating chains of fused-spherical segments, is expressed as a sum of four contributions [14]:

$$A = A^{\text{ideal}} + A^{\text{monomer}} + A^{\text{chain}} + A^{\text{association}} \quad (1)$$

where A^{ideal} is the free energy of an ideal gas mixture, A^{monomer} accounts for the repulsive and attractive interactions of monomeric segments through the Mie potential, A^{chain} is the change in free energy when chains (molecules) are formed from fusing Mie segments and $A^{\text{association}}$ is the contribution to the free energy upon the association. Detailed expressions for each of these contributions are delineated in previous works [12], [15], [16].

Utilising the group-contribution approach, identical segments can be used to model distinct functional groups within a molecule [12]. An example of this is demonstrated in Figure 1 for the breakdown of cholesterol (our main compound of study) into its resultant functional groups.

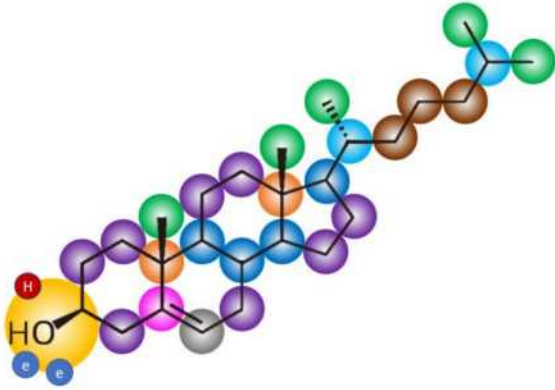


Figure 1: SAFT- γ Mie molecular model of cholesterol. This is modelled as one cCHOH (in yellow), eight cCH₂ (in purple), four cCH (in dark blue), five CH₃ (in green), three CH₂ (in brown), two CH (in light blue), one CH= (in grey), one C= (in pink) and two C (in orange) groups. Smaller red and blue circles indicate association sites, with the former labelled e, representing electronegative (acceptor) sites and the latter labelled H, representing hydrogen (donor) sites. Note, spheres are not to scale and are drawn for illustrative purposes.

Groups are considered to act in isolated and individual interactions with other groups, unaffected by all other interactions within the system. Thus, individual group parameters are transferable across systems (molecules or mixtures) containing these groups. This simplifies molecule characterisation given a robust database is established. By accumulating the appropriate groups of a specified molecule, it is assumed its gross thermodynamic properties can be evaluated and attained.

In view of the transferability, we also consider the systems of 2,2-dimethylcyclohexanol and cyclohexanol + isophorone as they contain relevant moieties of cholesterol, making them appropriate for validating combining rules utilised.

The dispersion force between two segments of groups k and l is a function of the distance between both

segments and is the overall combination of attractive and repulsive effects. This is delineated through the Mie (pair) potential Φ_{kl}^{Mie}

$$\Phi_{kl}^{\text{Mie}}(r_{kl}) = C_{kl} \epsilon_{kl} \left[\left(\frac{\sigma_{kl}}{r_{kl}} \right)^{\lambda_{kl}^r} - \left(\frac{\sigma_{kl}}{r_{kl}} \right)^{\lambda_{kl}^a} \right] \quad (2)$$

where r_{kl} is the distance between the centers of two segments, ϵ_{kl} is the dispersion energy and depth of the potential well, σ_{kl} is the size parameter (specifically the segment diameter in this case), and λ_{kl}^r and λ_{kl}^a are the repulsive and attractive exponents of the intersegment interaction. The prefactor C_{kl} is a function of the exponents and ensures $-\epsilon_{kl}$ is the minimum of the interaction despite the respective attractive and repulsive effects:

$$C_{kl} = \frac{\lambda_{kl}^r}{\lambda_{kl}^r - \lambda_{kl}^a} \left(\frac{\lambda_{kl}^r}{\lambda_{kl}^a} \right)^{\frac{\lambda_{kl}^a}{\lambda_{kl}^r - \lambda_{kl}^a}} \quad (3)$$

The strong associating interactions, typically hydrogen bonding, are depicted as extra, eccentric associating sites affixed on segments that express such interactions. The amount of different site types $N_{\text{ST},k}$ within a group k , and the number of sites of each type, $n_{k,a}$, $n_{k,b}$, ..., $N_{\text{ST},k}$, are further included to fully characterise associating groups. Short-ranged square-well potentials model the interactions between site type a on segment k and site type b on segment l :

$$\phi_{kl,ab}^{\text{HB}}(r_{kl,ab}) = \begin{cases} -\epsilon_{kl,ab}^{\text{HB}} & \text{if } r_{kl,ab} \leq r_{kl,ab}^c \\ 0 & \text{if } r_{kl,ab} > r_{kl,ab}^c \end{cases} \quad (4)$$

where $-\epsilon_{kl,ab}^{\text{HB}}$ is the association energy and depth of the square-well potential well, $r_{kl,ab}$ is the center-center separation of the two sites a and b , and $r_{kl,ab}^c$ is the cutoff range of the interaction, which can also be depicted in terms of bonding volume, $K_{kl,ab}$ [17]. The distance of site a from the center of segment k is given by $r_{kk,aa}^d$ and a similar term is given for site b on segment l . Following Wertheim's thermodynamic perturbation theory (TPT1) [18], the relative locations of sites on a segment are not considered because discrete sites are independent of one another in their bonding effect.

2.2 Solid-Liquid Equilibria

The total Helmholtz free energy expression presented in Equation (1) is in fact a function of the temperature T , volume V and number of molecules N , represented as a vector. Calculating standard thermodynamic relations with respect to each of these variables leads to equations in common state variables. For example, pressure is the first order derivative of A with respect to the volume V as shown by

$$P = - \left(\frac{\partial A}{\partial V} \right)_{T,N} \quad (5)$$

and chemical potential of compound i is the first order derivative of A with respect to the composition

$$\mu_i = - \left(\frac{\partial A}{\partial N_i} \right)_{T,V,N_{j \neq i}} \quad (6)$$

Developing these relations further gives rise to the derivation of phase equilibrium properties. For instance, of initial relevance to this study is the liquid heat capacity c_p^L and vapour pressure P^{sat} .

The main equilibrium quantity of interest is the solubility, which is denoted as the molar fraction \mathbf{x}^{sat} of the compound of interest in a given solvent at a particular temperature T and pressure P . This arises from the chemical equilibrium condition whereby the pure solid phase (S) chemical potential μ_i^{S} of the compound is equated with the liquid phase (L) chemical potential μ_i^{L} . The liquid phase chemical potential is given by

$$\mu_i^{\text{L}}(T, P, \mathbf{x}^{\text{sat}}) = \mu_i^*(T, P) + RT \ln a_i^{\text{sat}}(T, P, \mathbf{x}^{\text{sat}}) \quad (7)$$

where $\mu_i^*(T, P)$ is the pure liquid chemical potential as a reference state, R is the molar gas constant and $a_i^{\text{sat}}(T, P, \mathbf{x}^{\text{sat}})$ is the activity of compound i in solution. The activity models the non-ideal behaviour of real fluids and is given as the product of the solute mole fraction $x_i^{\text{sat}}(T, P, \mathbf{x}^{\text{sat}})$ and activity coefficient $\gamma_i(T, P, \mathbf{x}^{\text{sat}})$, $a_i^{\text{sat}} = x_i^{\text{sat}} \gamma_i$. It is therefore logical to express activity in this way as it exhibits the solubility explicitly as the molar fraction. Equating Equation (7) to the pure solid chemical potential (since its activity is 1, the natural log term would disappear) as discussed previously, we obtain

$$\ln x_i^{\text{sat}}(T, P, \mathbf{x}^{\text{sat}}) + \ln \gamma_i(T, P, \mathbf{x}^{\text{sat}}) = \frac{\mu_i^{\text{S}}(T, P) - \mu_i^{\text{L}}(T, P)}{RT} \quad (8)$$

where the difference in the pure chemical potentials of the liquid and solid phase can be reformulated as the partial molar Gibbs free energy of fusion $-\Delta g_i^{\text{fus}}(T, P)$ of pure compound i as presented below

$$\ln x_i^{\text{sat}}(T, P, \mathbf{x}^{\text{sat}}) + \ln \gamma_i(T, P, \mathbf{x}^{\text{sat}}) = -\frac{\Delta g_i^{\text{fus}}(T, P)}{RT} \quad (9)$$

and this can further be expressed utilising the fundamental relation $\Delta g_i^{\text{fus}}(T, P) = \Delta h_i^{\text{fus}}(T, P) - T \Delta s_i^{\text{fus}}(T, P)$ to attain

$$\ln x_i^{\text{sat}}(T, P, \mathbf{x}^{\text{sat}}) + \ln \gamma_i(T, P, \mathbf{x}^{\text{sat}}) = -\frac{\Delta h_i^{\text{fus}}(T, P)}{RT} + \frac{\Delta s_i^{\text{fus}}(T, P)}{R} \quad (10)$$

where $\Delta h_i^{\text{fus}}(T, P)$ is the enthalpy of fusion and $\Delta s_i^{\text{fus}}(T, P)$ is the entropy of fusion.

Since the system is residing at a temperature T below the fusion temperature T_i^{fus} , it is impracticable to explicitly measure the fusion enthalpy $\Delta h_i^{\text{fus}}(T, P)$ and entropy $\Delta s_i^{\text{fus}}(T, P)$ experimentally. However, as the system is a supercooled liquid at these conditions, it suggests a thermodynamic cycle can be followed instead. This involves taking an alternative path to reach the same desired state, by employing known expressions and values at ambient conditions. Ignoring pressure effects by assuming ambient pressure and solution incompressibility, expressions are extracted invoking values at the fusion temperature T_i^{fus} :

$$\Delta h_i^{\text{fus}}(T, P) = \Delta h_i^{\text{fus}}(T_i^{\text{fus}}, P) - \int_T^{T_i^{\text{fus}}} \Delta c_{p,i}(T', P) dT' \quad (11)$$

and

$$\Delta s_i^{\text{fus}}(T, P) = \Delta s_i^{\text{fus}}(T_i^{\text{fus}}, P) - \int_T^{T_i^{\text{fus}}} \frac{\Delta c_{p,i}(T', P)}{T'} dT' \quad (12)$$

where $\Delta c_{p,i}(T, P)$ is the difference between the heat capacity of the pure liquid phase $c_{p,i}^{\text{L}}(T, P)$ and the pure solid phase $c_{p,i}^{\text{S}}(T, P)$. At the fusion point, $\Delta g_i^{\text{fus}}(T, P) = 0$ and therefore

$$\Delta h_i^{\text{fus}}(T, P) = T \Delta s_i^{\text{fus}}(T, P) \quad (13)$$

Substituting Equations (11), (12) and (13) into Equation (10), we obtain a form of the solubility equation for solid-liquid equilibria [19] as

$$\begin{aligned} \ln x_i^{\text{sat}}(T, P, \mathbf{x}^{\text{sat}}) + \ln \gamma_i(T, P, \mathbf{x}^{\text{sat}}) = & -\frac{\Delta h_i^{\text{fus}}(T_i^{\text{fus}}, P)}{R} \left(\frac{1}{T} - \frac{1}{T_i^{\text{fus}}} \right) \\ & + \frac{1}{RT} \int_T^{T_i^{\text{fus}}} \Delta c_{p,i}(T', P) dT' \\ & - \frac{1}{R} \int_T^{T_i^{\text{fus}}} \frac{\Delta c_{p,i}(T', P)}{T'} dT' \end{aligned} \quad (14)$$

In this study, the value of the heat capacity difference is known at T_i^{fus} , hence the approximation $\Delta c_{p,i}(T, P) \approx \Delta c_{p,i}(T_i^{\text{fus}}, P)$ is used, conveying the assumed T insensitive nature of $\Delta c_{p,i}$. This allows for convenient integration of Equation (14) to the desired form of the solubility equation for solid-liquid equilibrium, given as

$$\begin{aligned} x_i^{\text{sat}}(T, P, \mathbf{x}^{\text{sat}}) \cdot \gamma_i(T, P, \mathbf{x}^{\text{sat}}) = & \exp \left[\frac{\Delta h_i^{\text{fus}}(T_i^{\text{fus}}, P)}{R} \left(\frac{1}{T_i^{\text{fus}}} - \frac{1}{T} \right) \right. \\ & \left. + \frac{\Delta c_{p,i}(T_i^{\text{fus}}, P)}{R} \left(\frac{T_i^{\text{fus}}}{T} - 1 - \ln \left(\frac{T_i^{\text{fus}}}{T} \right) \right) \right] \end{aligned} \quad (15)$$

Equation (15) has been further rearranged to resemble the form present within gPROMs [20] our main SAFT predictions tool. For the remainder of this study, we shall refer to Equation (15) as the standard SLE equation.

3. Methods

3.1 SAFT- γ Mie Group Parameters

In Table 1, we present the relevant SAFT- γ Mie group interactions for the systems of interest in a triangular matrix. This includes the groups of 2,2-dimethylcyclohexanol, cyclohexanol, isophorone as well as the groups pertaining to the main systems of study, cholesterol with the primary alcohols. (Note that the molecular depiction of cholesterol, illustrating its corresponding groups is presented in Figure 1. Cells coloured blue indicate previously published work [12], [13], [15], whilst cells coloured grey containing a ‘‘CR’’ indicate the use of combining rules. This is seen in the cCHOH-C and cCHOH-C= unlike interactions respectively. Due to Table 1 existing as a triangular matrix, the leading diagonal, therefore, contains the like interactions, whilst all other cells are unlike interactions between distinct groups.

Table 1: Group interactions required to model cholesterol in primary alcohol solvents along with 2,2-dimethylcyclohexanol, cyclohexanol, isophorone using the SAFT- γ Mie approach. Blue cells represent interactions previously published work [12], [13], [15], grey cells containing "CR" represent interactions calculated using combining rules and white cells represent interactions that are not needed in this work.

cCHOH	CH ₂	CH	CH ₃	cCH	cCH ₂	CH=	C=	C	CH ₂ OH	cCO
cCHOH	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue
CH ₂	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue
CH	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue
CH ₃	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue
cCH	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue
cCH ₂	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue
CH=	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue
C=	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue
C	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue
CH ₂ OH	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue
cCO	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue

Published Work

Combining Rules

Not needed

In Section 2.1 prior, we introduce the idea that segments interact through pair potentials, specifically the Mie potential for dispersion interactions and the square-well potential for strong, associative interactions. In order to obtain the parameters that underpin these potentials, combining rules are employed. These provide estimates of unlike interactions, utilising expressions of their like analogues. Starting with the Mie parameters, the unlike segment diameter σ_{kl} is obtained using the Lorentz rule. This is an arithmetic mean of the parallel like diameters presented as

$$\sigma_{kl} = \frac{\sigma_{kk} + \sigma_{ll}}{2} \quad (16)$$

In contrast, geometric-mean criterion are applied for the remaining parameters. The unlike dispersion energy ϵ_{kl} is obtained through an altered Berthelot-like geometric mean rule which rationalises the irregularity in segment sizes [12]. This is given as

$$\epsilon_{kl} = \frac{\sqrt{\sigma_{kk}^3 \sigma_{ll}^3}}{\sigma_{kl}^3} \sqrt{\epsilon_{kk} \epsilon_{ll}} \quad (17)$$

Again, a geometric-mean rule is applied for the calculation of the unlike repulsive and attractive exponents, λ_{kl}^r and λ_{kl}^a respectively. This is applied on the van der Waals energy as follows [21]

$$\lambda_{kl}^j = 3 + \sqrt{(\lambda_{kk}^j - 3)(\lambda_{ll}^j - 3)}, \quad \text{where } j = (a, r) \quad (18)$$

The associative square-well parameters can be calculated in very similar fashion, employing geometric and arithmetic mean expressions. The unlike association energy $\epsilon_{kl,ab}^{\text{HB}}$ is conveyed through a geometric-mean as [15]

$$\epsilon_{kl,ab}^{\text{HB}} = \sqrt{\epsilon_{kk,aa}^{\text{HB}} \epsilon_{ll,bb}^{\text{HB}}} \quad (19)$$

whilst the unlike bonding volume $K_{kl,ab}$ is conveyed through an arithmetic-mean as [15]

$$K_{kl,ab} = \left(\frac{\sqrt[3]{K_{kk,aa}} + \sqrt[3]{K_{ll,bb}}}{2} \right)^3 \quad (20)$$

In practice, combining rules are typically exploited as an initial approximation of the unlike parameters [12]. More sophisticated optimisation methods are favoured to refine these estimates with greater accuracy using experimental data i.e. parameter estimation. However, in the discussion that follows, it can be argued that parameter estimation would perform similarly to the combining rules in our case.

Shape factors indicate the contribution of each group to the overall thermodynamic property of the compound they reside in. A value of 1 indicates a very strong contribution, whereas a value close to 0 indicates negligible contribution. For the unlike interaction between distinct groups, the relevant indicator of this effect is the product of the shape factors. In Table 2, we present the individual and appropriate cross interactions for the groups currently utilising combining rules. As can be observed, the shape factor products $S_{\text{cCHOH}C=}$ and $S_{\text{cCHOH}C}$ are small and very small respectively. We can, therefore, denote that the contributions to the free energy of these interactions are also small. Therefore, the unlike dispersion energies, $\epsilon_{\text{cCHOH}C=}$ and $\epsilon_{\text{cCHOH}C}$, can in theory have any value because it is predicted that the results will not change (significantly) for properties considered. Thus, it is deemed that the combining rule values are good enough and parameter estimation is not required. Additionally, parameter estimation will not be significant for the same reasons; it is improbable to optimise a parameter if it is almost independent of the result obtained. Furthermore, utilising combining rules demands less computational effort which simplifies the characterisation process.

Table 2: Shape factors for the cCHOH, C= and C groups along with the shape factor products for the cCHOH-C= and cCHOH-C unlike interactions.

Group k	Shape factor	$S_{k \neq \text{cCHOH}} \times S_{\text{cCHOH}}$
cCHOH	0.68123	-
C=	0.15330	0.10443
C	0.040720	0.027739

3.2 Solid-Solid Phase Transition

Due to the enantiotropic nature of cholesterol, the standard SLE equation (Equation (15)), in isolation, is not sufficient to describe the entire solubility curve. In order to depict the additional solid form cholesterol possesses, another SLE equation is required. Thus, Domanska *et al.* [22] empirically modified the standard SLE equation as follows [22]

$$x_i^{\text{sat}}(T, P, \mathbf{x}^{\text{sat}}) \cdot \gamma_i(T, P, \mathbf{x}^{\text{sat}}) = \exp \left[\frac{\Delta h_i^{\text{fus}}(T_i^{\text{fus}}, P)}{R} \left(\frac{1}{T_i^{\text{fus}}} - \frac{1}{T} \right) + \frac{\Delta h_i^{\text{tr}}(T_i^{\text{tr}}, P)}{R} \left(\frac{1}{T_{i,s}^{\text{tr}}} - \frac{1}{T} \right) + \frac{\Delta c_{P,i}(T_i^{\text{fus}}, P)}{R} \left(\frac{T_i^{\text{fus}}}{T} - 1 - \ln \left(\frac{T_i^{\text{fus}}}{T} \right) \right) \right] \quad (21)$$

This incorporates two additional variables; $\Delta h_i^{\text{tr}}(T_i^{\text{tr}}, P)$, which is the enthalpy of the solid–solid phase transition and T_i^{tr} , which is the solid–solid transition temperature. Whilst T_i^{tr} is system dependent (its value changes depending on the solvent in system with a solute), in the case of cholesterol, $\Delta h_i^{\text{tr}}(T_i^{\text{tr}}, P)$ is currently reported as the pure transition enthalpy $\Delta h_{\text{cholesterol}}^{\text{tr}}(T_{\text{cholesterol}}^{\text{tr}}, P)$. This is because the mixture effects upon the property have not yet been investigated. For the remainder of this study, we shall refer to Equation (21) as the Domanska’s SLE equation and subscripts denoting cholesterol as “chol”.

Both the standard SLE equation (Equation (15)) and Domanska’s SLE equation (Equation (21)) [22] are used in tandem to produce the correct solubility curve for systems experiencing solid–solid phase transitions before fusion. For temperatures below $T_{\text{chol}}^{\text{tr}}$, polymorph β described by Domanska’s SLE equation (Equation (21)) [22] is the most stable, since this achieves the minimum Gibbs free energy of the system. In contrast, for temperatures above $T_{\text{chol}}^{\text{tr}}$, polymorph α , described by the standard SLE equation (Equation (15)), is now the most stable. Visually, this appears as a kink like curve in a temperature–composition phase diagram. The collective use of Equations (15) and (21) shall form the basis of our SLE model.

Our SLE model is highly sensitive to $\Delta h_{\text{chol}}^{\text{fus}}(T_{\text{chol}}^{\text{fus}}, P)$, due to its high order of magnitude. Therefore, we emphasise the importance of using the true value of $\Delta h_{\text{chol}}^{\text{fus}}(T_{\text{chol}}^{\text{fus}}, P)$. As eight distinct measurements for $\Delta h_{\text{chol}}^{\text{fus}}(T_{\text{chol}}^{\text{fus}}, P)$ were reported by eight different authors without uncertainties [23]–[30], we find it necessary to model a confidence interval for which the true value of $\Delta h_{\text{chol}}^{\text{fus}}(T_{\text{chol}}^{\text{fus}}, P)$ resides in. Due to the limited sample size, a student-t distribution is the most appropriate probability distribution to employ. Our confidence interval for a t-distribution is defined such that the probability the true value lies between the upper and lower bounds is 95%. This is expressed as

$$P \left[\bar{X} - t_{\alpha/2, n-1} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\alpha/2, n-1} \frac{S}{\sqrt{n}} \right] = 1 - \alpha \quad (22)$$

where $1 - \alpha$ represents the chosen confidence level of 95%, n is the sample size represented by the 8 measurements, \bar{X} is the mean of the sample computed as 25.8 kJ mol^{-1} , S is the sample standard deviation calculated as 3.49 kJ mol^{-1} , μ is the true value of $\Delta h_{\text{chol}}^{\text{fus}}(T_{\text{chol}}^{\text{fus}}, P)$ and $t_{\alpha/2, n-1}$ is the t – score for a two-tailed test with $n - 1$ degrees of freedom. This is obtained from statistical tables as a value of 2.365, for the aforementioned parameters. Evaluating Equation (22) further, we obtain a confidence interval for the true value of $\Delta h_{\text{chol}}^{\text{fus}}(T_{\text{chol}}^{\text{fus}}, P)$ as $[22.9, 28.7] \text{ kJ mol}^{-1}$.

The sample mean along with the upper and lower bounds of our confidence interval are inputted into our

SLE model to produce three solubility curves for each cholesterol + primary alcohol system. The latter form the upper and lower bounds of the confidence region, within which, the true solubility curve based on the true $\Delta h_{\text{chol}}^{\text{fus}}(T_{\text{chol}}^{\text{fus}}, P)$ value lies, and the curve produced from the mean value is that which is used in direct comparison with experimental data. This allows for a more justified evaluation of our predictions.

3.3 Absolute Average Deviations

To assess the accuracy of the theoretical approach provided by the SAFT- γ Mie Equation of State, average absolute deviations are calculated for all systems of interest. The percentage absolute deviation (%AAD) of a property p for a system s is given by

$$\% \text{AAD}_s[p] = \frac{1}{N_{s,p}^D} \sum_{i=1}^{N_{s,p}^D} \left| \frac{X_{s,p,i}^{\text{exp}} - X_{s,p,i}^{\text{calc}}}{X_{s,p,i}^{\text{exp}}} \right| \times 100 \quad (23)$$

where $N_{s,p}^D$ is the total number of experimental points for the property, p , of interest, $X_{s,p,i}^{\text{exp}}$ is the i th measured value in the property p data vector and $X_{s,p,i}^{\text{calc}}$ is the analogous value, calculated with the SAFT- γ Mie approach. %AAD, however, normalises the deviation with respect to the experimental value. In the case of our solubility calculations, this results in considerably large deviations due to the low orders of magnitude observed. Therefore, we favour assessing the crude average absolute deviation (AAD) given by

$$\text{AAD}_s[p] = \frac{1}{N_{s,p}^D} \sum_{i=1}^{N_{s,p}^D} |X_{s,p,i}^{\text{exp}} - X_{s,p,i}^{\text{calc}}| \quad (24)$$

4. Results & Discussion

4.1 Combining Rule Evaluation of Small Systems

In Figure 2, we present the results for the small systems, namely, pure 2,2-dimethylcyclohexanol and a binary mixture of cyclohexanol + isophorone. This is to evaluate the use of combining rules for the cCHOH-C unlike interaction in the case of the former, and both the cCHOH-C and cCHOH-C= unlike interactions in the case of the latter. In Figure 2(a), we present the prediction of vapour pressures for pure 2,2-dimethylcyclohexanol. Whilst there are only two experimental values to compare this to, the order of magnitude and general trend is correctly captured by the SAFT- γ Mie approach. With a moderately large %AAD of 12.78%, the importance of using larger data sets is noted to distinguish between an accurate trend and anomalous results. This was, however, the only experimental data set found for this system. As such, particular emphasis is placed on Figure 2(b) which presents a larger data set in the form of the isobaric vapour-liquid phase equilibria for cyclohexanol + isophorone. Furthermore, this system contains both the unlike interactions of interest, and is therefore, more relevant to our work. Despite a slight deviation from the experimental data for the pure saturation temperature of

isophorone, we obtain an excellent description of the binary mixture using the SAFT- γ Mie approach. This slight deviation remains systematic throughout the entire phase envelope. With %AADs of 1.05% for the bubble temperatures T^{bub} , and 1.29% for the dew temperatures T^{dew} , these results reaffirm the negligible contribution of the cCHOH-C and cCHOH-C= unlike interactions to the total free energy, deduced by the small shape factor products presented in Table 2 in Section 3.1. The %AADs are summarised in Table 3. In light of these results, the use of combining rules for these two unlike interactions is considered appropriate and is therefore used for predicting the properties of systems involving cholesterol.

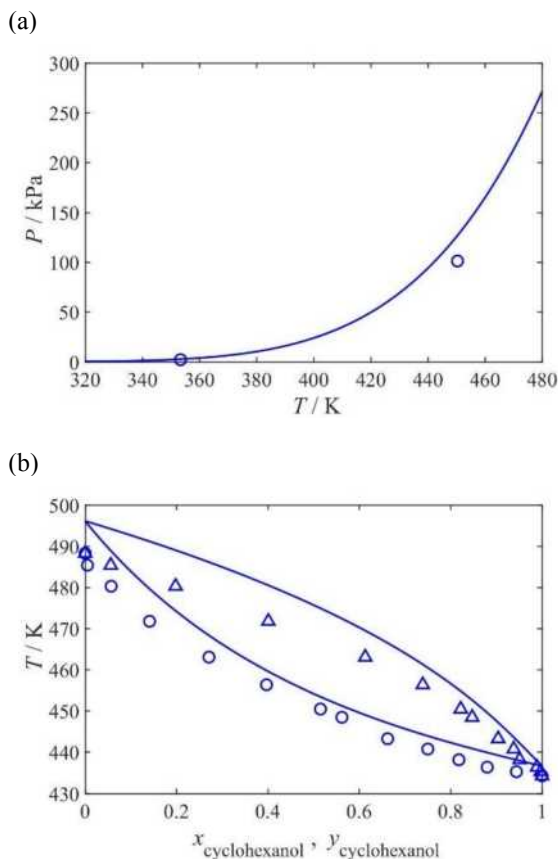


Figure 2: (a) Vapour pressures for 2,2-dimethylcyclohexanol. Circles represent experimental data [31]. (b) Isobaric vapour-liquid equilibria of cyclohexanol + isophorone at $P = 101.33$ kPa. Circles and triangles represent experimental bubble and dew temperatures respectively [32]. The continuous curves represent the SAFT- γ Mie calculations.

Table 3: Percentage absolute average deviations from experimental values of calculated vapour pressures for 2,2-dimethylcyclohexanol, and bubble and dew temperatures for cyclohexanol + isophorone.

Compound or system	Property	%AAD
2,2-dimethylcyclohexanol	P^{vap}	12.78
Cyclohexanol + isophorone	T^{bub}	1.05
Cyclohexanol + isophorone	T^{dew}	1.29

4.2 Heat Capacity

Pure cholesterol data is extremely scarce, with only one reliable source reporting the molar heat capacity $c_{P, \text{chol}}$ found in the literature [29]. However, this was primarily for heat capacity in the solid phase $c_{P, \text{chol}}^{\text{S}}$, with only one measurement reported for the heat capacity in the liquid phase $c_{P, \text{chol}}^{\text{L}}$ (at $T = 424.44$ K, just above the fusion temperature, $T_{\text{chol}}^{\text{fus}} = 421.15$ K). We present this data in Figure 3 alongside the SAFT- γ Mie prediction for $c_{P, \text{chol}}^{\text{L}}$ (note that the SAFT- γ Mie approach cannot be applied to solids). Although, there is only one data point to compare this to, we observe an excellent prediction at that corresponding temperature with a %AAD of 1.49%. This promising result for a pure property of cholesterol gives us confidence in the use of the SAFT- γ Mie approach in predicting the properties of cholesterol in mixtures, e.g. solubility.

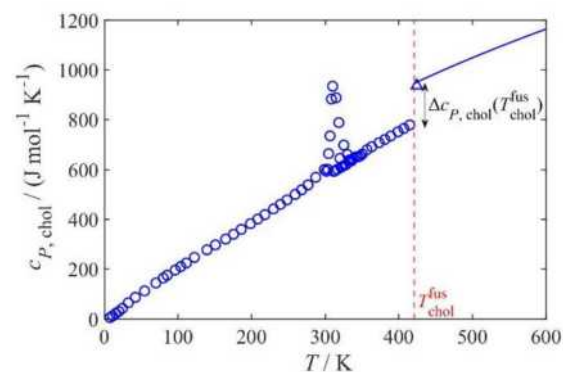


Figure 3: Molar heat capacity of cholesterol. Circles represent solid phase data and the triangle represents liquid phase experimental data [29]. Continuous curve represents SAFT- γ Mie calculations. Dashed red line denotes $T_{\text{chol}}^{\text{fus}}$.

In Figure 3, we observe a peak in $c_{P, \text{chol}}$ at 310.2 K (note the proximity to body temperature). Miltenburg *et al.* [29] attributes this to the solid-solid phase transition which has also been observed by Domanska *et al.* [22]. Indeed, the temperature at which this peak occurs roughly corresponds to the transition temperatures reported by Domanska *et al.* [22] at which this solid-solid phase transition occurs.

In Figure 3, we also denote the difference between the heat capacity of the liquid phase and the heat capacity of the solid phase at the fusion temperature $\Delta c_{P, \text{chol}}(T_{\text{chol}}^{\text{fus}})$. However, the heat capacity data was not reported at $T_{\text{chol}}^{\text{fus}}$ exactly. The closest temperature to $T_{\text{chol}}^{\text{fus}}$ for which $c_{P, \text{chol}}^{\text{S}}$ was reported was 418.92 K, with a value of 795.19 $\text{J mol}^{-1} \text{K}^{-1}$ and as mentioned previously, the only value of c_P^{L} was reported at 424.44 K, with a value of 936.75 $\text{J mol}^{-1} \text{K}^{-1}$ [29]. Since these temperatures are only within a few Kelvin of $T_{\text{chol}}^{\text{fus}}$, their corresponding heat capacity values can be approximated as the values at $T_{\text{chol}}^{\text{fus}}$. We, therefore, calculate that $\Delta c_{P, \text{chol}}(T_{\text{chol}}^{\text{fus}})$ is approximately equal to 141.56 $\text{J mol}^{-1} \text{K}^{-1}$.

Knowing the value of $\Delta c_{P, \text{chol}}(T_{\text{chol}}^{\text{fus}})$ is important as it is needed in the prediction of solubility. A term

involving $\Delta c_{P, \text{chol}}(T_{\text{chol}}^{\text{fus}})$ appears in both the standard SLE equation (Equation (15)) and Domanska's SLE equation (Equation (21)) [22]. However, often this $\Delta c_{P, \text{chol}}(T_{\text{chol}}^{\text{fus}})$ term is neglected, particularly when experimental values for $\Delta c_{P, \text{chol}}(T_{\text{chol}}^{\text{fus}})$ are unknown [33]. In Figure 4, we present two solubility predictions for cholesterol in 1-butanol: the red curve denotes the solubility prediction which neglects the $\Delta c_{P, \text{chol}}(T_{\text{chol}}^{\text{fus}})$ term, and the blue curve denotes the solubility prediction which accounts for it. Not only do we observe a large deviation between the two curves, but the blue curve is also in much better agreement with the experimental data, denoted by the circles and squares [8], [22].

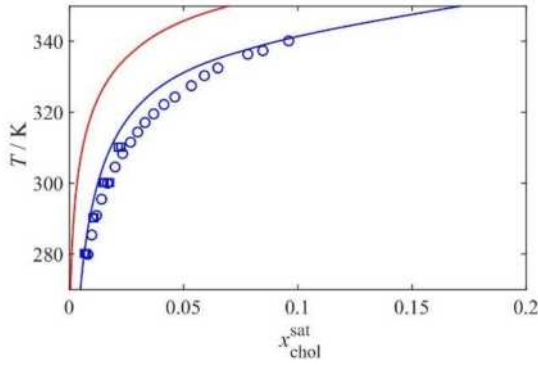


Figure 4: Solid-liquid equilibria for cholesterol in 1-butanol at $P = 101.325$ kPa. Continuous curves represent the SAFT- γ Mie calculations including (blue) and neglecting (red) the heat capacity term. Circles [22] and squares [8] represent experimental data.

The deviation between the two curves can be quantified using the absolute relative deviation (%ARD) given by

$$\% \text{ARD} = \left| \frac{x_{\text{chol}}^{\text{sat, NO}} - x_{\text{chol}}^{\text{sat, YES}}}{x_{\text{chol}}^{\text{sat, YES}}} \right| \times 100, \quad (25)$$

where $x_{\text{chol}}^{\text{sat, NO}}$ is the solubility prediction of cholesterol at a given T , neglecting $\Delta c_{P, \text{chol}}(T_{\text{chol}}^{\text{fus}}, P)$ and $x_{\text{chol}}^{\text{sat, YES}}$ is the solubility prediction of cholesterol at a given T which accounts for it. We calculate a %ARD of 73.6% at 298.15 K. This follows the trend reported by Febra [33], whereby, the contribution of the $\Delta c_{P, \text{chol}}(T_{\text{chol}}^{\text{fus}}, P)$ term is large for compounds with a large T_i^{fus} and $\Delta c_{P, i}(T_i^{\text{fus}}, P)$. Although, $T_{\text{chol}}^{\text{fus}}$ is not particularly large, the $\Delta c_{P, \text{chol}}(T_{\text{chol}}^{\text{fus}}, P)$ certainly is. We, therefore, observe a large contribution of the $\Delta c_{P, \text{chol}}(T_{\text{chol}}^{\text{fus}}, P)$ term when predicting the solubilities of cholesterol. Consequently, it is important to account for this term when carrying out such predictions.

4.3 Solubility Predictions

In Figure 5, we present the solubility of cholesterol in different primary alcohols (1-butanol to 1-decanol) at body temperature ($T = 310.15$ K). We present the predictions for this using the SAFT- γ Mie approach (black circles), alongside experimental data from

Domanska *et al.* (blue squares) [22] and Flynn *et al.* (red triangles) [8].

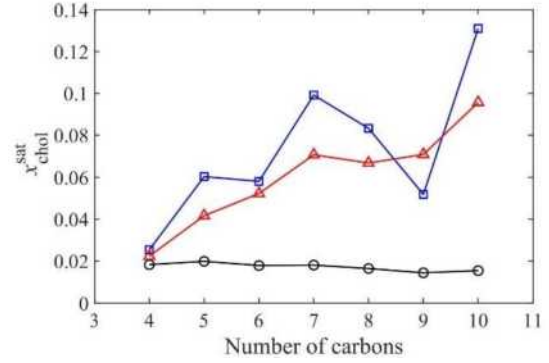


Figure 5: Solubility profile of cholesterol in primary alcohols at $T = 310.15$ K. Black circles represent SAFT- γ Mie calculations. Blue squares [22] and red triangles [8] represent experimental data. Straight lines have been drawn to guide the eye.

Both Domanska *et al.* [22] and Flynn *et al.* [8] observed a general increase in solubility for an increase in alcohol chain length, from 1-butanol to 1-heptanol. However, we observe a noticeable scattering in the experimental data with the exception of the 1-butanol and 1-hexanol systems. We, therefore, present the predicted SLE curves as a function of temperature for these systems (Figures 6(a) and 6(c) respectively). We also choose to present the predicted SLE curve for the 1-pentanol system (Figure 6(b)) to highlight this degree of scattering and the challenge in conducting experimental measurements at the low concentrations observed. This may explain the limited experimental data available for these systems. Furthermore, the 1-pentanol system falls in between the 1-butanol and 1-hexanol systems with regards to chain length. We, therefore, present these three systems for continuity.

Beyond 1-heptanol, both Domanska *et al.* [22] and Flynn *et al.* [8] observed erratic trends in solubility for increasing alcohol chain length. Flynn *et al.* [8] attributes this to the significant effects of solvate formation. The formation of these solvates result in crystalline changes, thereby altering the structure of the solid phase. These changes become significant for larger solvents [8]. This may explain the significant deviation between the experimental data and the SAFT- γ Mie calculations for chain lengths beyond 1-heptanol and why the solubility predictions remain almost constant for increasing chain length. Without knowing the fusion properties of these solvates from experimental data, the solid phase cannot be modelled accurately, since the SAFT- γ Mie approach is only applied in the modelling of fluids (note that our current description of the solid phase relies on parameters obtained experimentally e.g. $\Delta h_{\text{chol}}^{\text{fus}}$, $T_{\text{chol}}^{\text{fus}}$, $\Delta h_{\text{chol}}^{\text{tr}}$, $T_{\text{chol}}^{\text{tr}}$ and $\Delta c_{P, \text{chol}}(T_{\text{chol}}^{\text{fus}})$). Consequently, we do not consider systems beyond 1-heptanol.

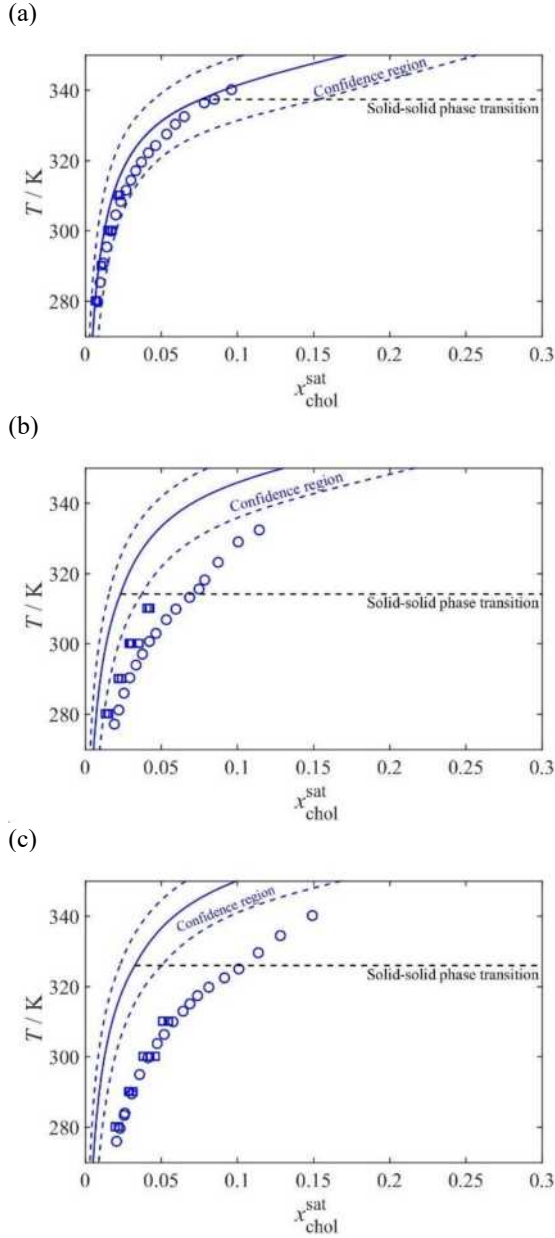


Figure 6: Solubilities of cholesterol in (a) 1-butanol, (b) 1-pentanol, and (c) 1-hexanol. Solid curves represent the SAFT- γ Mie calculations using the sample mean $\Delta h_{\text{chol}}^{\text{fus}}$ value. The dashed curves represent SAFT- γ Mie calculation using the limits of the 95% confidence interval for the true value of $\Delta h_{\text{chol}}^{\text{fus}}$. Circles [22] and squares [8] represent experimental data.

For the cholesterol + 1-butanol system, Domanska *et al.* [22] reports a solid-solid phase transition occurring at $T_{\text{chol}}^{\text{tr}} = 337.4$ K. In Figure 6(a) we present an excellent prediction of the solubility of cholesterol in 1-butanol, with a %AAD of 19.28% and an AAD $x_{\text{chol}}^{\text{sat}}$ of 0.005324 (all AAD $x_{\text{chol}}^{\text{sat}}$ values reported are with respect to the calculated SLE curve using the mean $\Delta h_{\text{chol}}^{\text{fus}}$ value i.e. the solid blue curve). Although this %AAD may seem large, it should be noted that %AAD is not an appropriate measure of deviation in this context due to the extremely low orders of magnitude of solubilities. A slight deviation from the experimental data in absolute terms, can result in a significant %AAD. When considering the AAD $x_{\text{chol}}^{\text{sat}}$ for this system, an excellent prediction can

still be concluded since the order of magnitude of the AAD $x_{\text{chol}}^{\text{sat}}$ is lower than that of the measured/predicted solubilities as highlighted by observing the x -axis of Figure 6(a).

In Figures 6(b) and 6(c), we present poorer, albeit slightly, solubility predictions for cholesterol in 1-pentanol and 1-hexanol respectively. These systems exhibit a solid-solid phase transition at $T_{\text{chol}}^{\text{tr}} = 314.2$ K and $T_{\text{chol}}^{\text{tr}} = 326.0$ K respectively [22]. In both cases, the SAFT- γ Mie approach underpredicts the solubility of cholesterol. The latter (Figure 6(c)) appears to be particularly poor when considering its larger deviation from the wider confidence region. Nevertheless, with an AAD $x_{\text{chol}}^{\text{sat}}$ of 0.02654 and 0.03707 for the 1-pentanol and 1-hexanol systems respectively, the prediction captures the correct order of magnitude as well as the correct trajectory of the SLE curves. Additionally, with such a wide confidence region due to the variation in experimental values of $\Delta h_{\text{chol}}^{\text{fus}}$, the true value of $\Delta h_{\text{chol}}^{\text{fus}}$ could result in an SLE curve closer to the upper bound of the confidence region i.e. the dashed curve furthest to the right. This would result in lower AAD $x_{\text{chol}}^{\text{sat}}$ values. We, therefore, conclude a good agreement observed for these systems. The %AADs and AAD $x_{\text{chol}}^{\text{sat}}$ values are summarised in Table 4.

Table 4: Absolute average deviations from experimental values of calculated solubilities of cholesterol in 1-butanol, 1-pentanol and 1-hexanol.

System	%AAD $x_{\text{chol}}^{\text{sat}}$	AAD $x_{\text{chol}}^{\text{sat}}$
Cholesterol + 1-butanol	19.28	0.005324
Cholesterol + 1-pentanol	59.71	0.02645
Cholesterol + 1-hexanol	68.32	0.03707

In Figure 5, increasingly large deviations between the SAFT- γ Mie prediction and the experimental data are observed for increasing alcohol chain length. These deviations become significant and erratic from 1-heptanol onwards, which is attributed to significant solvate formation effects. The case for this argument is now strong when we consider the excellent agreement between the solubility prediction and experimental data for the 1-butanol system and reasonable agreements for the 1-pentanol and 1-hexanol systems (refer to Figure 6). This would suggest that our description of the liquid phase (provided by the SAFT- γ Mie approach) for these systems is accurate. Hence, we can assume the underlying parameters e.g. ϵ_{kl} , σ_{kl} and λ_{kl}^{r} , are also accurate. This is supported by the accurate modelling of small systems in Figure 2 and the excellent prediction of pure liquid heat capacity of cholesterol in Figure 3. Following these results, we would expect our liquid phase description to be accurate when predicting the solubility of cholesterol in longer chain alcohols i.e. beyond 1-heptanol. With an accurate liquid phase description for these systems, we identify the poor description of the solid phase as the main drawback of our SLE model. This reinforces the idea that the formation of solvates in longer chain systems result in

significant crystalline changes impacting the properties of the solid phase. Since these properties are unknown, we cannot capture their effects in our current description of the solid phase.

Indeed, there is a high degree of uncertainty in our description of the solid phase, with significant variation found amongst experimental values of $\Delta h_{\text{chol}}^{\text{fus}}$ and $\Delta h_{\text{chol}}^{\text{tr}}$. Variation was also found amongst experimental values of $T_{\text{chol}}^{\text{tr}}$, however, only Domanska *et al.* [22] reported solvent specific values of $T_{\text{chol}}^{\text{tr}}$ relevant to our work but made no mention of the uncertainty in their measurements. We, therefore, have a lack of confidence in the true values for these parameters and note this as another drawback of our SLE model.

5. Conclusions

Understanding the thermodynamic properties of cholesterol is important due to its role in the behaviour of biological membranes and the pathogenesis of gallstones [8]. The existence of a solid-solid phase transition at approximately body temperature ($T = 310.15 \text{ K}$), makes the solubility behaviour of cholesterol particularly relevant.

The SAFT- γ Mie calculations reported in this work expand our understanding of the thermodynamic properties of systems involving cholesterol. We validate the use of combining rules for the cCHOH-C and cCHOH-C= unlike interactions by presenting an excellent prediction of the bubble and dew temperatures of a binary cyclohexanol + isophorone system, with %AADs of 1.05% and 1.29% respectively.

We obtain an excellent prediction for the liquid heat capacity of cholesterol at $T = 424.44 \text{ K}$, with a %AAD of 1.49%. We also observe a significant contribution of the difference in the heat capacity of the liquid and solid phases (at fusion temperature) to the solubility prediction of cholesterol in primary alcohols, highlighting its necessity in our SLE model.

We present the solubility predictions of cholesterol in primary alcohols, 1-butanol to 1-decanol, at body temperature ($T = 310.15 \text{ K}$), as well as the predicted SLE phase diagrams for cholesterol in 1-butanol, 1-pentanol and 1-hexanol. With AAD $x_{\text{chol}}^{\text{sat}}$ values of 0.005324, 0.02645 and 0.03707, we present an excellent prediction for the 1-butanol system and good predictions for the 1-pentanol and 1-hexanol systems respectively. The SAFT- γ Mie approach, therefore, captures the order of magnitude of cholesterol solubility and the trajectory of these SLE curves correctly. However, for longer chain alcohols, the formation of solvates becomes significant, influencing the crystalline structure, which our solid phase description does not capture. We highlight this as the main drawback of our SLE model, but not of the SAFT- γ Mie approach. Understanding the properties of these solvates would enable us to provide a better description of the solid phase leading to more accurate solubility predictions.

The results from this work demonstrate the versatility of the SAFT- γ Mie approach, insofar as it enables us to model molecules for which experimental

data is scarce, thus solidifying its status as a state-of-the-art equation of state.

Acknowledgements

We would like to thank the MSE Group including Dr Thomas Bernet, Shubhani Paliwal, Ahmed Alyazidi for their continuous support and guidance throughout this work.

References

- [1] A. V. Prabhu, W. Luu, D. Li, L. J. Sharpe, and A. J. Brown, 'DHCR7: A vital enzyme switch between cholesterol and vitamin D production', *Prog Lipid Res*, vol. 64, pp. 138–151, Oct. 2016, doi: 10.1016/j.plipres.2016.09.003.
- [2] L. Andrade, 'Understanding the role of cholesterol in cellular biomechanics and regulation of vesicular trafficking: The power of imaging', *Biomedical Spectroscopy and Imaging*, vol. 5, pp. S101–S117, Dec. 2016, doi: 10.3233/BSI-160157.
- [3] A. Grebe and E. Latz, 'Cholesterol Crystals and Inflammation', *Curr Rheumatol Rep*, vol. 15, no. 3, p. 313, Mar. 2013, doi: 10.1007/s11926-012-0313-z.
- [4] L. Y. Hsu and C. E. Nordman, 'Phase transition and crystal structure of the 37 degrees C form of cholesterol', *Science*, vol. 220, no. 4597, pp. 604–606, May 1983, doi: 10.1126/science.6836303.
- [5] K. van Putte, W. Skoda, and M. Petroni, 'Phase transition and CH₃-rotation in solid cholesterol', *Chemistry and Physics of Lipids*, vol. 2, no. 4, pp. 361–371, Nov. 1968, doi: 10.1016/0009-3084(68)90011-X.
- [6] N. Garti, L. Karpuij, and S. Sarig, 'Phase transitions in cholesterol crystallized from various solvents', *Thermochimica Acta*, vol. 35, no. 3, pp. 343–348, Feb. 1980, doi: 10.1016/0040-6031(80)87134-6.
- [7] S. Bridgwater, 'Accurate free energy methods for model organic solids', phd, University of Warwick, 2014. Accessed: Dec. 10, 2023. [Online]. Available: <http://webcat.warwick.ac.uk/record=b2752105~S1>
- [8] G. L. Flynn, Y. Shah, S. Prakongpan, K. H. Kwan, W. I. Higuchi, and A. F. Hofmann, 'Cholesterol solubility in organic solvents', *Journal of Pharmaceutical Sciences*, vol. 68, no. 9, pp. 1090–1097, 1979, doi: 10.1002/jps.2600680908.
- [9] W. G. Chapman, K. E. Gubbins, G. Jackson, and M. Radosz, 'SAFT: Equation-of-state solution model for associating fluids', *Fluid Phase Equilibria*, vol. 52, pp. 31–38, Dec. 1989, doi: 10.1016/0378-3812(89)80308-5.
- [10] A. Gil-Villegas, A. Galindo, P. J. Whitehead, S. J. Mills, G. Jackson, and A. N. Burgess, 'Statistical associating fluid theory for chain molecules with attractive potentials of variable range', *The Journal of Chemical Physics*, vol. 106, no. 10, pp. 4168–4186, Mar. 1997, doi: 10.1063/1.473101.

- [11] J. Gross and G. Sadowski, 'Perturbed-Chain SAFT: An Equation of State Based on a Perturbation Theory for Chain Molecules', *Ind. Eng. Chem. Res.*, vol. 40, no. 4, pp. 1244–1260, Feb. 2001, doi: 10.1021/ie0003887.
- [12] V. Papaioannou *et al.*, 'Group contribution methodology based on the statistical associating fluid theory for heteronuclear molecules formed from Mie segments', *The Journal of Chemical Physics*, vol. 140, no. 5, p. 054107, Feb. 2014, doi: 10.1063/1.4851455.
- [13] A. J. Haslam *et al.*, 'Expanding the Applications of the SAFT- γ Mie Group-Contribution Equation of State: Prediction of Thermodynamic Properties and Phase Behavior of Mixtures', *J. Chem. Eng. Data*, vol. 65, no. 12, pp. 5862–5890, Dec. 2020, doi: 10.1021/acs.jced.0c00746.
- [14] W. G. Chapman, K. E. Gubbins, G. Jackson, and M. Radosz, 'New reference equation of state for associating liquids', *Ind. Eng. Chem. Res.*, vol. 29, no. 8, pp. 1709–1721, Aug. 1990, doi: 10.1021/ie00104a021.
- [15] S. Dufal *et al.*, 'Prediction of Thermodynamic Properties and Phase Behavior of Fluids and Mixtures with the SAFT- γ Mie Group-Contribution Equation of State', *J. Chem. Eng. Data*, vol. 59, no. 10, pp. 3272–3288, Oct. 2014, doi: 10.1021/je500248h.
- [16] S. Dufal *et al.*, 'The A in SAFT: developing the contribution of association to the Helmholtz free energy within a Wertheim TPT1 treatment of generic Mie fluids', *Molecular Physics*, vol. 113, no. 9–10, pp. 948–984, May 2015, doi: 10.1080/00268976.2015.1029027.
- [17] G. Jackson, W. G. Chapman, and K. E. Gubbins, 'Phase equilibria of associating fluids', *Molecular Physics*, vol. 65, no. 1, pp. 1–31, Sep. 1988, doi: 10.1080/00268978800100821.
- [18] M. S. Wertheim, 'Thermodynamic perturbation theory of polymerization', *The Journal of Chemical Physics*, vol. 87, no. 12, pp. 7323–7331, Dec. 1987, doi: 10.1063/1.453326.
- [19] J. M. Prausnitz, *Molecular thermodynamics of fluid-phase equilibria*, 3rd ed. in Prentice-Hall international series in the physical and chemical engineering sciences. Upper Saddle River, NJ: Prentice Hall, 1999.
- [20] 'gPROMS'. Process Systems Enterprise, 1997 2023.
- [21] T. Lafitte *et al.*, 'Accurate statistical associating fluid theory for chain molecules formed from Mie segments', *The Journal of Chemical Physics*, vol. 139, no. 15, p. 154504, Oct. 2013, doi: 10.1063/1.4819786.
- [22] U. Domańska, A. Pobudkowska, and P. Gierycz, 'Experimental solid–liquid phase equilibria of {cholesterol+binary solvent mixture: 1-Alcohol (C4–C10)+cyclohexane}', *Fluid Phase Equilibria*, vol. 289, no. 1, pp. 20–31, Feb. 2010, doi: 10.1016/j.fluid.2009.10.009.
- [23] W. Chen, B. Su, H. Xing, Y. Yang, and Q. Ren, 'Solubilities of cholesterol and desmosterol in binary solvent mixtures of n-hexane+ethanol', *Fluid Phase Equilibria*, vol. 287, no. 1, pp. 1–6, Dec. 2009, doi: 10.1016/j.fluid.2009.08.016.
- [24] M. Iwahashi *et al.*, 'Thermodynamic Properties of Steroids: 5-cholesten-3 β -ol, 5 α -cholestan-3 β -ol, and 5 β -cholestan-3 α -ol', *Journal of Oleo Science*, vol. 50, no. 9, pp. 693–699, 2001, doi: 10.5650/jos.50.693.
- [25] J. Kaloustian, A.-M. Pauli, P. Lechene de la Porte, H. Lafont, and H. Portugal, 'Thermal analysis of anhydrous and hydrated Cholesterol', *Journal of Thermal Analysis and Calorimetry*, vol. 71, no. 2, pp. 341–351, Feb. 2003, doi: 10.1023/A:1022818902212.
- [26] E. Kosal, C. H. Lee, and G. D. Holder, 'Solubility of progesterone, testosterone, and cholesterol in supercritical fluids', *The Journal of Supercritical Fluids*, vol. 5, no. 3, pp. 169–179, Sep. 1992, doi: 10.1016/0896-8446(92)90004-4.
- [27] K. Vezzù, A. Bertucco, and F. P. Lucien, 'Solid–liquid equilibria of multicomponent lipid mixtures under CO₂ pressure: Measurement and thermodynamic modeling', *AIChE Journal*, vol. 54, no. 9, pp. 2487–2494, 2008, doi: 10.1002/aic.11543.
- [28] L. Peng, X. Jiangjun, M. Fangquan, L. Xi, and Z. Chaocan, 'Study on the thermodynamic properties of cholesterol', *J Therm Anal Calorim*, vol. 93, no. 2, pp. 485–488, Aug. 2008, doi: 10.1007/s10973-007-8675-6.
- [29] J. C. van Miltenburg, A. C. G. van Genderen, and G. J. K. van den Berg, 'Design improvements in adiabatic calorimetry: The heat capacity of cholesterol between 10 and 425 K', *Thermochimica Acta*, vol. 319, no. 1, pp. 151–162, Oct. 1998, doi: 10.1016/S0040-6031(98)00402-X.
- [30] J. Gmehling, 'Pure compound data from DDB'. 1983.
- [31] K. v. Auwers and E. Lange, 'Zur Kenntnis hydroaromatischer Verbindungen: Über hydroaromatische Alkohole und Ketone mit einer gem.-Dimethylgruppe', *Justus Liebigs Annalen der Chemie*, vol. 401, no. 3, pp. 303–326, 1913, doi: 10.1002/jlac.19134010305.
- [32] M. Hu, H. Tian, J. Sun, R. Zhang, Q. Zhao, and W. Xu, 'Vapor–Liquid Equilibrium Measurements of Cyclohexene–Isophorone and Cyclohexanol–Isophorone Binary Systems and Predictions for Cyclohexene–Cyclohexanol–Isophorone Ternary System', *J. Chem. Eng. Data*, vol. 65, no. 6, pp. 3103–3108, Jun. 2020, doi: 10.1021/acs.jced.0c00103.
- [33] S. Febra, 'Ring formation in a statistical associating fluid theory framework', Chemical Engineering, Imperial College London, 2019. Accessed: Dec. 10, 2023. [Online]. Available: <http://hdl.handle.net/10044/1/68078>

Optimisation of Ionic Liquids in a Closed-Loop Dye Recycling Process

Humaira Mansuri and Nursyazana Mahzan

Department of Chemical Engineering, Imperial College London, U.K.

Abstract: Polyethylene terephthalate (PET) is the largest single fibre type used in the global textile industry. To dye PET's hydro-phobic fibres, complementary disperse dyes and numerous auxiliary agents are utilised. Current garment production and dyeing processes consume vast quantities of water, energy, and dyestuff materials with little effort to recycle the dye, discarded PET and abate their water and chemical usage. Leaching of wastewater containing harmful chemicals into the environment ensues. Since PET and disperse dyes do not readily biodegrade, it has spurred innovation to recycle the disperse dyes extracted from PET textile fibres by employing ionic liquids (ILs) and eliminating auxiliary chemical usage. This study explored a step towards a closed-loop dye recycling process by developing a washing method to maximise IL recovery and minimise water consumption. Recoveries of [DMBA][MeSO₃] and [TEA][MeSO₃], which were employed in the extraction of C.I. Disperse Blue 56 from PET and the resulting dye bath's reuse in dyeing virgin PET was compared, while varying the increment volume of wash water. The minimum IL loss achieved was 3.35%, utilising the smallest increment of wash water volume investigated (5 ml/g_{fabric}) and implementing [DMBA][MeSO₃]. Similar [DMBA][MeSO₃] and [TEA][MeSO₃] retrieval rates in the wash waters indicate the washing process is not mass transfer limited. UV-vis spectroscopy concluded that incremental wash water volume sizes and type of IL does not affect the dyed fabric's colour strength. Partition coefficient evaluation suggests the stagewise washing procedure is physically limited.

1. Introduction

To suit economic objectives, negligent practices by the fashion industry have encouraged user overconsumption, making it the second most polluting industry^[1]. Annually, it expends a staggering 79 trillion litres of water, accounting for approximately 20% of total global wastewater production^[1]. From fibre to textile production, over 15,000 chemicals are involved during clothing manufacture: consisting of sizing agents, dyestuffs, pigments, basic chemicals and auxiliary chemicals for pretreatment, dyeing and finishing^[2]. As many textile mills are in developing countries, most do not possess advanced wastewater treatment units, so the environmentally hazardous wastewater is discharged directly into rivers and seas, seriously harming aquatic life, exacerbating eutrophication and threatening human health^[3]. Aside from compounds that do not easily biodegrade, others participate in destructive reactions, such as disperse azo dyes, which can be reduced to carcinogenic amines and non-aromatic dyes, that contain toxic heavy metals^[4]. Alongside production emissions, the industry manages its unused pre-consumer waste, like unsold or returned garments and fabric off-cuts by incineration and landfill deposition, recycling only 14.7% in 2018 according to the EPA^[5].

Polyester dominates the textile market, comprising 54% of global fibre production [6] and around 80% of synthetics, due to its relative low cost, performance and versatility. To match polyester's hydrophobic properties, complementary disperse dyes which are sparingly soluble in water, are used to dye its hydrophobic fibres [7]. The two main classes of disperse dyes, divided by their functional groups, are azo-dyes and anthraquinone dyes. Their synthesis requires hydrocarbons derived from non-renewable petroleum sources. Although, stricter environmental regulations are encouraging textile manufacturers to adopt effluent treatment and water recycling systems to mitigate the industry's environmental impact, these are costly solutions and will propagate to increase the purchase price for consumers.

Inherent change in the textile industry will be achieved when responsibility of a garment is maintained

all the way to the end of its life and extended past the manufacturing stage. Synthetic dye recycling and their extraction “as a pretreatment for material recycling methods and their subsequent reuse to dye new fibres provides a new circularity dimension in the textile industry” [7]. This dye recycle process invented by Abouelela et al. [7], aims to curb the environmental dangers posed by leaching harmful disperse dyes found in PET manufacturing wastewater, by using low-cost protic ionic liquids (ILs) to selectively extract dyes from synthetic textile fibres and recycle them. Once optimised, the technology intends to provide a way to recycle decoloured PET which meets recycling criteria and achieve a sustainable circular dyeing route by utilising textile waste and reducing virgin synthetic dye consumption [7].

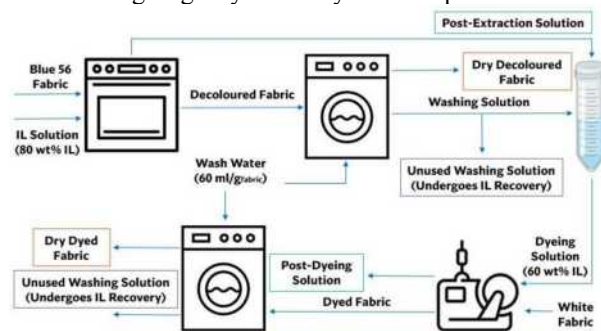


Figure 1. Schematic diagram of dye recycling process

The research contained in this report seeks to help streamline the IL, water and energy usage throughout the closed loop process, recycling resources where possible by using a Design of Experiment approach. The principal focus of the experiments was to gauge the feasibility of achieving near complete recovery of IL, minimising losses, by varying the incremental wash water volume. A second IL was also investigated. The experiments consisted of selective extraction of dye from polyester fabric. This uncontaminated decoloured PET can be hypothetically sent to recycling facilities and the extracted dye-rich solution was used as a colour specific dye bath for dyeing virgin polyester. The whole dye recycling process is shown in Figure 1. The effect of the washing

procedures on dyeing performance was also investigated using UV-vis spectrometry.

2. Background

The dyeing process is the dispersion of dye into liquid solvent, followed by diffusion of the dye molecules from the solvent to the substrate, which is the free space volume of the fabric fibres^[8]. In batch dyeing, textile is loaded into the dyeing vessel and allowed to reach equilibrium with the solution containing the dye and other auxiliary chemicals. When the desired shade of textile is achieved, the textile is washed to remove unfixed dye molecules and chemicals. Solid loading is the ratio of the mass of the substrate fabric to the volume of the solution containing the dye. Notably, the solid loading correlates with dyeing vessel size, water and energy consumption of the process and therefore influences process economics considerably for exhaustion methods. Continuous and semi-continuous dyeing processes are used for meeting high production volumes, in which the dyestuff is applied to the textile as it is passed through rollers rotating at speeds of 50-250 m/min^[9]. Heat is applied to fix the dyes to the fabric by drying or steam, consuming a large amount of energy. Instead of the solid loading parameter, the wet pickup percentage is used to measure the mass of dye-containing solution picked up by 100 grams of fabric^[9]. Typical polyester dyeing uses pressurised jet-dyeing machines in batch operation. The high temperature (120-150 °C) employed in this process enables sufficient swelling of the fibres to allow dye penetration, by exceeding the glass transition temperature of polyester (60-80 °C)^[10]. After heating, the polyester cools and returns to its crystalline structure with the dye molecules trapped between the fibres. This method eradicates the need for various carriers and phenol-based swelling-agents that were used prior to its invention. However, removal of the unfixed dye sometimes requires alkaline chemicals in reduction cleaning for darker coloured textiles^[10] and the process still has a significant carbon footprint due to its energy requirements. Other common polyester dyeing methods are the Carrier, High Pressure High Temperature (HTHP) and Pad thermosol^[11]. These conventional methods are resource intensive and rely on the additional polluting auxiliary chemicals, discussed earlier, to increase dyeability and improve dispersion of the low solubility-disperse dyes in solvents like water.

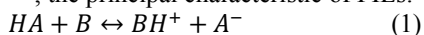
Consumer demand, regulatory pressure, competitive advantage, and innovation in the fashion industry are all playing a part to dismantle the fashion industry's traditional dyeing processes for textiles and are accelerating the development of environmentally conscious dyeing processes that consume less synthetic auxiliary chemicals and have lower carbon and water footprints. One such dyeing process uses supercritical CO₂ (sCO₂) as a solvent for disperse dyes, that are utilised for hydrophobic polymers like PET and other synthetic fibres^[12]. The dissolved dye in the circulating CO₂ can penetrate the PET fibres as they swell, aiding dye diffusion and its even distribution. Expansion at low pressure facilitates easy removal of sCO₂ and recovery of excess dye^[12]. The end product is comparable in colour strength, and fastness to conventional wet dyeing methods. As it is a dry process, it consumes zero water and therefore also eliminates the

energy-consuming drying step^[13]. Another advantage is that it employs a simpler dye concoction, with the absence of auxiliary chemical, has quicker dyeing times, making this process extremely efficient and eco-friendly. However, the high capital needed for equipment costs and process construction, alongside the operating costs associated with the elevated pressures (100-300 bars) the process requires, impede commercialisation of sCO₂ dyeing in the industry's low-cost dyeing landscape^[12]. Migration of polymers from inside the fibres to the PET fabric's surface as well as its transfer to dyeing equipment also presents a technical issue^[14]. Other, dry techniques like digital micro-printing and microencapsulation suffer from similar disadvantages due to their complexity^[11]. Another dyeing method that minimises water usage is solvent-assisted dyeing. An important advantage of this technique over sCO₂ dyeing, is that in principle adaptation of aqueous dyeing machinery can be achieved relatively easily with minor modifications and lower costs. Solvent selection for fibre dyeing, benefits from increased dye bath reusability and increased dye uptake as they explicitly influence colour fastness, dyeing cost and effluent control. Initially chlorinated hydrocarbons were proposed as suitable solvents for polyester but understandably did not popularise, due to their toxicity and flammability^[15]. Elimination of auxiliary chemicals like wetting and levelling agents and improved diffusion of disperse dyes in water-alcohol solvent mixtures, is causing alcohol-assisted dyeing to gain recognition more recently^[16]. Alternative solvents to volatile organic alcohols, investigated by Ferrero and Periolatto, found that glycerol as an additive, is as effective as ethanol in replacing auxiliary chemicals^[16].

Although extensive research has been conducted into dye removal methods from textile wastewaters due to their devastating environmental impact, little investigative efforts have gone into dye removal technologies from solid substrates, like textile fibres. Robinson^[17] briefly mentioned the possible use of ILs to preserve the colouring of cotton fibres or remove their original colouring, to produce neutral fibres^[17]. However, the paper largely focused on using ILs to separate and recycle polymer fibres from polymer-cellulose textile blends. Xiuzhu et al. investigated dye removal from polyester fibres employing much harsher sodium formaldehyde sulfoxylate and acetone chemicals^[18]. Despite this method's effectiveness at decolourising a range of polyester fabric shades, it does not preserve the dye chemistry whatsoever, and the investigation's sole aim was to enable polyester textile recycling.

Ionic liquids are a group of salts that have bulky anions and cations and don't possess much symmetry in their structure, sterically limiting their ability to form ordered structures^[19]. Meaning at temperatures less than 100 °C, they exist in the liquid state and have low melting points. Another reason for this, is the delocalisation of the cationic and anionic charges spreading over more than one atom (apart from halides), which induces a reduction in lattice energy. Varying alkyl chains can be substituted with the hydrogen atoms to introduce rational degrees of freedom at low temperatures. These alkyl chains are less symmetric, thus decreasing the melting point of the ionic liquid^[20]. The cation in ILs is usually comprised of a per

alkylated organic ion like amines, imidazoles or pyrimidines, and the anion is usually polyatomic and can be either inorganic or organic such as $[\text{MeSO}_3^-]$ [21]. A key advantage of ILs is that they can be specifically tailored for a particular application due to the unique properties experienced from pairing different anions and cations. Applications include their use as solvents, catalysts, in electrochemistry and lignocellulose pretreatment [7][21]. Their combined ionic-organic nature enables their engagement in a variety of inter and intra-molecular interactions from weak isotropic forces to strong, specific and anisotropic forces [22]. As a result, they are often referred to as “designer solvents”. Unlike organic solvents, ILs are not flammable due to their negligible vapour pressures, they do not require vapour disposal measures. However, the immediate toxic risk they pose to aquatic life when released into the environment, in some instances is higher than molecular solvents, deeming quantitative recovery necessary. Thus, efficient separation of products, such as dyed fibres, from the IL, is essential to optimise recovery and curtail water consumption during the washing stage. Their largely involatile disposition also means the amount of IL able to be recovered and recycled is increased, which is advantageous, because of their associated high cost. Despite this fact, there is a subclass of ILs, called protic ILs (PILs) which cost substantially less than conventional aprotic ILs and whose price is comparable with common inexpensive organic solvents [21]. PILs are conceived by proton transfer from the acid to the base via an acid-base neutralisation reaction [22] (Equation 1), causing availability of the proton which allows formation of inter and intra hydrogen bonding between the anions and cations as well as dissolved solutes [23]; the principal characteristic of PILs.



$[\text{DMBA}][\text{MeSO}_3]$ and $[\text{TEA}][\text{MeSO}_3]$ are two PILs that share the aforementioned IL properties and have high solubility for disperse dyes [24], facilitating efficient extraction of the dyes from the fabric and enhancing the dyeing process on polyester fabric. Their relatively low viscosities aid the mass transport rate of the reaction and dyeing process [24]. Furthermore, they possess good thermal stability [25], enabling dyeing processes at higher temperatures required for dye fixation on polyester, contributing to better dye adsorption and colour fastness. Moreover, their chemical compatibility with polyester fabric and disperse dyes, to minimise adverse reactions that could affect the quality of dyeing or the fabric itself, makes them suitable for this projects’ purpose. Figure 2 shows their chemical structures.

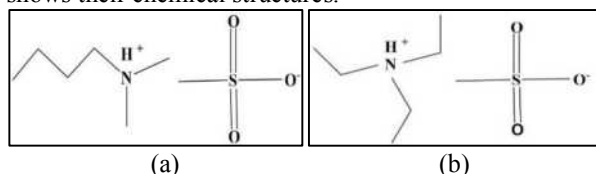


Figure 2. Chemical structure of IL. (a) $[\text{DMBA}][\text{MeSO}_3]$. (b) $[\text{TEA}][\text{MeSO}_3]$

Dyeing using ionic liquids has multiple advantages over sCO_2 dyeing and alcohol-based dyeing, as the dyeing process can be operated at atmospheric pressure and existing aqueous dyeing machinery can be easily modified

for IL-assisted dyeing, costing much less than sCO_2 dyeing implementation. Compared to IL applications in various other fields, research for their use in dyeing textiles is still relatively scarce and they have only been studied modestly. Opwis et al. conducted a study exploring the efficacy of commonly utilised aprotic ionic liquids in dyeing polyester. The ionic liquids tested were comprised of a 1-ethyl-3-methyl-imidazolium cation combined with various anions including acetate, chloride, sulfate, methyl sulfate, and phosphate. Among these, methyl sulfate-based ionic liquids exhibited superior performance, showcasing enhanced dyeing outcomes and achieving deeper shades particularly at elevated temperatures ranging from 140 to 160 °C [26]. Moreover, their research highlighted that dyeing in the ionic liquid medium yielded superior colour shades and comparable fastness results in comparison to conventional aqueous-based dyeing techniques. Separate studies by Yuan et al. demonstrated the utility of $[\text{C}_4\text{C}_{1\text{im}}]\text{Cl}$ for enhancing dyeing of wool [27]. Meanwhile, Bianchini et al. conducted a more extensive examination, screening eight aprotic ionic liquids for dyeing cotton, wool, and polyester. Among the tested compounds, the 1-(2-hydroxyethyl)-3-methylimidazolium chloride IL demonstrated the most promising dyeing results across various fabrics under atmospheric pressure at 100 °C in initial screening tests [28].

3. Methodology

3.1 Materials

N-Butyldimethylamine (DMBA) and triethylamine (TEA) were obtained from TCI Chemical (Tokyo, Japan) while methanesulfonic acid (MeSO_3H) was obtained from Sigma Aldrich (Missouri, USA). Disperse blue 56 (spun polyester woven) and white (100% polyester) fabrics were purchased from Testfabrics, Inc (USA). Deionised water was provided in the laboratory. DMSO NMR solvent was purchased from Sigma-Aldrich (USA).

3.2 Making Ionic Liquid Solution

For $[\text{DMBA}][\text{MeSO}_3]$ synthesis, dropwise equimolar MeSO_3H was added to DMBA in a round bottom flask and cooled in an ice bath. It was left to stir continuously overnight. Deionised (DI) water was added to the solution until it had a composition of 80 wt% ionic liquid and 20 wt% water. It was proved that high ionic liquid concentration can achieve high efficiency of dye extraction [7]. Volumetric Karl Fischer titration (V20 Mettler Toledo, USA) was used to measure the water content while 400 MHz ^1H -NMR spectrometer (Tokyo, Japan) was used to ensure the acid-base ratio of the solution is 1. The same procedure was used to synthesise $[\text{TEA}][\text{MeSO}_3]$.

3.3 Conductivity Calibration Curve

Pre-measured deionised (DI) water was added to a 50 ml beaker. The conductivity of DI water using a conductivity meter was measured. A quantity of the ionic liquid (IL) solution was taken using a 1 ml syringe. Then, known mass (1 to 2 drops) of IL solution from the syringe was added into the water. The conductivity of the solution was measured, ensuring the solution was well-mixed before the reading was taken. These steps were repeated, gradually increasing the amount of IL solution added to the water until the conductivity reading was out of the

instrument's range (0 – 2 mS). Equation 2 was used to determine the concentration of IL in the solution. Enabling, a linear graph of conductivity against IL concentration to be plotted.

$$C_{IL} = \frac{0.8m_{IL,sol}}{(m_{IL,sol} + m_{DI})} \times 100 \quad (2)$$

where C_{IL} is the IL concentration in wt%, $m_{IL,sol}$ is the mass of IL solution added while m_{DI} is the mass of DI water. The 80 wt% of IL in the initial IL solution was considered in the calculation. The conductivity calibration curve was carried out for both [DMBA][MeSO₃] and [TEA][MeSO₃], respectively.

3.4 Dye Extraction Process

Three samples with 10 g IL solution in each pressure tube, were prepared. (From section 3.4 to 3.6, IL refers to [DMBA][MeSO₃] unless stated otherwise.). The blue polyester fabric was cut into small pieces (approximately 1 × 1 cm) to increase the surface area for dye extraction and mixed in the IL solutions with a fabric loading of 15% (1.5 g of fabric in each sample) as commercially, the fabric loading ranges from 10% to 30%^[29]. Equation 3 was used to obtain the mass of fabric. The oven was preheated to 150 °C and the pressure tubes were put into the oven for 1 hour. This temperature proved to have high efficiency of dye extraction without decomposing the dye^[7]. The pressure tubes were taken out of the oven using gloves and left to cool before starting the washing process.

$$m_{bf} = \frac{FL}{100} \times m_{IL,sol}^f \quad (3)$$

where m_{bf} is the mass of blue fabric used, FL is the fabric loading in % while $m_{IL,sol}^f$ is the feed mass of IL solution.

3.5 Washing Procedure

After the pressure tubes had cooled down, the post-extraction solutions were decanted into empty falcon tubes and the mass of these solutions was measured. The fabric was squeezed using a spatula to ensure as much of the solutions were decanted. Then, 20 ml/g_{fabric} (30 ml) of DI water was used in each sample to wash the fabric. They were mixed well using a spatula to ensure the fabric was fully immersed in the solution. Following this, the pressure tubes were put into the sonicator (Fisherbrand) for 2 minutes to further ensure the solution was well-mixed. After that, they were placed back in the oven (temperature fixed at 150 °C) for 10 minutes. They were then taken out and allowed to cool. Following this, the first washing solutions were collected into empty falcon tubes and the mass of these solutions was measured. These steps were repeated with the second and third washes until the total volume of water used to wash the fabric reached 60 ml/g_{fabric} (90 ml) for each sample. The mass of the fabric was recorded after the fabric completely dried. The procedures described in sections 3.4 and 3.5 were repeated for 10 ml/g_{fabric} (15 ml) and 5 ml/g_{fabric} (7.5 ml) volumes of water used in each wash by keeping the total wash water volume constant at 60 ml/g_{fabric}. Thus, the number of washes was 6 for 10 ml/g_{fabric} and 12 for 5 ml/g_{fabric}.

3.6 Dye Recycling Process

The post-extraction solutions (concentrated with blue dyes) were recycled to dye virgin fabric. These solutions were diluted using the wash water collected from the 1st wash, so its composition became 60 wt% IL and 40 wt% water. Equation 4 was used to determine the mass of the 1st washing solution required for dilution.

$$m_{w1} = m_{Ex,sol} \left(\frac{C_o}{C_D} - 1 \right) \quad (4)$$

where m_{w1} is the mass of first washing solution required to dilute the post-extraction solution, $m_{Ex,sol}$ is the mass of post-extraction solution, C_o is the initial IL concentration which is 80 wt% while C_D is the desired IL concentration which is 60 wt%.

After the dilution, these solutions were referred to as dyeing solutions. The total mass of dyeing solutions was determined using Equation 5 below. Dyes are insoluble in water. Therefore, higher dyeing efficiency can be achieved by increasing the water content to reduce the solubility of dye, shifting the equilibrium towards dyeing^[7]. The dyeing solutions were then transferred to the dyeing pots. Then, a piece of white fabric was added into each dyeing pot with a fabric loading of 15%. Equation 6 was used to obtain the mass of white fabric required for dyeing.

$$m_{ds} = m_{Ex,sol} + m_{w1} \quad (5)$$

$$m_{wf} = \frac{FL}{100} \times m_{ds} \quad (6)$$

where m_{wf} is the mass of white fabric used, FL is the fabric loading in % while m_{ds} is the mass of dyeing solution.

The dye pots were then loaded into the dyeing machine (Roches Pyrotec⁴, Yorkshire, UK) with a rotation speed of 40 RPM and heating speed of 3.1 °C/min to the set temperature (130 °C). After that, the dyeing machine was cooled down to 30 °C at cooling speed of 3 °C/min. After the process finished, the dye pots were taken out from the machine and the post-dyeing solutions were removed into empty falcon tubes, respectively. The mass of these solutions was measured. The dyed fabrics were transferred into pressure tubes to start the washing process, which had the same procedure as in section 3.5. Similarly, the volume of water in each wash was also varied (20 ml/g_{fabric}, 10 ml/g_{fabric} and 5 ml/g_{fabric}). This entire process (section 3.4 to 3.6) was repeated using [TEA][MeSO₃] as the IL.

3.7 Measuring the Conductivity

The conductivity of IL can be measured directly using conductivity meter, which has range of 0 to 2 mS. However, some of the solutions were too concentrated with IL such that the conductivity meter reading was out of range. Therefore, the solutions had to be diluted first with known mass of DI water. Calibration curve obtained from section 3.3 was used to determine the IL concentration, based on the conductivity reading. Equation 7 was also used for the solutions that required dilution.

$$C_{O,IL} = \frac{(m_{sol} + m_{DI})}{m_{sol}} \times C_{D,IL} \quad (7)$$

where $C_{O,IL}$ is the actual IL concentration in the original solution before dilution while $C_{D,IL}$ is the diluted concentration of IL, both in units of wt%. m_{sol} is the mass of solution used for dilution while m_{DI} is the mass of DI water used to dilute the solutions.

3.8 Colour Strength of Dyed Fabric

UV-Vis spectroscopy was used to measure the reflectance of the dyed fabrics. A background spectrum was first acquired from an empty sample holder. The Kubelka-Munk function was used to obtain the colour strength (K/S)^[29] (Equation 8).

$$K/S = \frac{(1-R_{min})^2}{2R_{min}} \quad (8)$$

where K is the absorption coefficient of light, S is the scattering coefficient of light while R_{min} is the minimum reflectance in the spectrum (highest absorbance) at wavelength range of 620 to 650 nm. The larger the K/S value, the higher the intensity of the dye in the fabric.

4. Results and Discussion

From this point forward in the report [DMBA][MeSO₃] will be referred to as DMBA and [TEA][MeSO₃] will be referred to as TEA. The term ‘IL’ refers to either [DMBA][MeSO₃] or [TEA][MeSO₃] and similarly ‘ILs’ refers to both [DMBA][MeSO₃] and [TEA][MeSO₃].

4.1 Calibration Curve

Ionic compounds conduct electricity when they are in solution form as ions can move freely and carry the charge^[30]. Thus, the correlation between conductivity and concentration of ionic liquid solutions of DMBA and TEA respectively can be determined, as shown in Figure 4.1. Both ionic liquids showed linear relationship, but DMBA has higher conductivity compared to TEA. Types of ions in solutions will affect the conductivity as they have

different ability to transmit the charge due to the sizes and interaction with water molecules^[31]. DMBA is larger since the arrangement of its carbon chain is longer compared to TEA, which has more compact carbon chain, seen in Figure 2. Therefore, the bond strength between DMBA and water molecules is weaker, enables the ions to move more freely in solution. The equations shown in Figure 4.1 were used to obtain the concentration of IL.

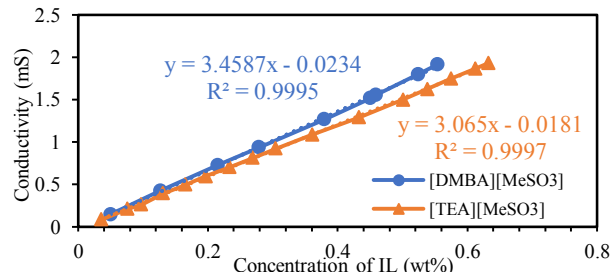


Figure 4.1. Conductivity calibration curve for [DMBA][MeSO₃] and [TEA][MeSO₃]

4.2 Concentration of IL in Solutions

As explained in section 3.5, the total amount of water used for washing the IL off the post-extraction decoloured fabric and post-dyeing dyed fabric was set at 60 ml/g_{fabric}.

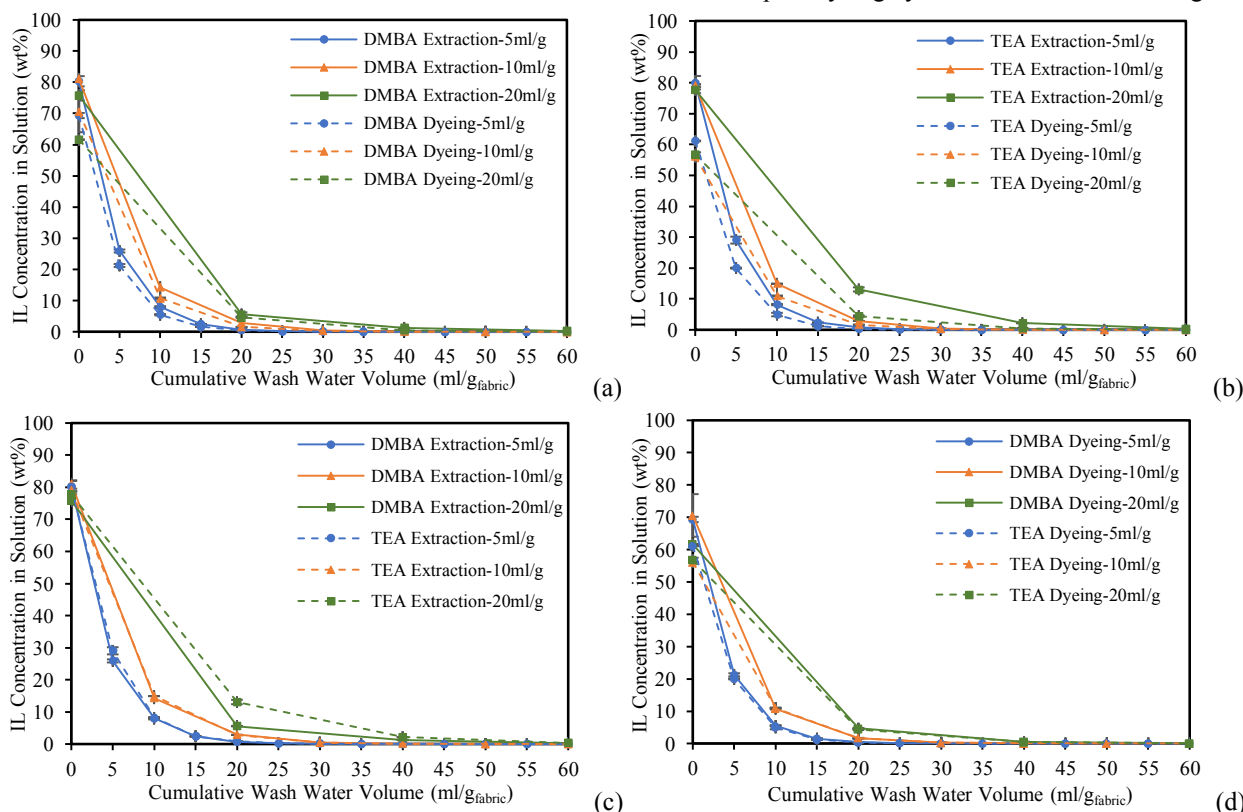


Figure 4.2. IL concentration in solutions post-extraction and post-dyeing. (a) Comparison of DMBA extraction and dyeing. (b) Comparison of TEA extraction and dyeing. (c) Comparison of DMBA and TEA extractions. (d) Comparison of DMBA and TEA dyeings

Figure 4.2 shows the concentration of IL removed from the fabrics, into the wash waters post-extraction and post-dyeing, for both ILs tested. A first glance at Figure 4.2 solidifies the intuitive prediction of there being a progressive decrease in IL concentration in the wash waters after each wash. At constant volumes of wash water in Figure 4.2a, there is a reduction in the DMBA concentration as the stagewise increment wash water

decreases from 20 ml/g_{fabric} to 10 ml/g_{fabric} to 5 ml/g_{fabric}. This is seen zooming in at 20 ml/g_{fabric} cumulative wash water volume, examining the extraction data in Figure 4.2a. The DMBA concentration is 6.28% for 20 ml/g_{fabric}, 2.81% for 10 ml/g_{fabric} and 0.70% for 5 ml/g_{fabric}. This occurs because when reaching 20 ml/g_{fabric} cumulative wash water volume, there are four times as many washes in the stagewise operation employing 5 ml/g_{fabric} water-

fabric ratio and twice as many for the 10 ml/g_{fabric} compared to the 20 ml/g_{fabric} increment. Consequently, each time new wash water is added, at each wash stage, the concentration difference of IL between the fabric and the surrounding water is reset, re-establishing the driving force for the IL to migrate into the water and out of the fabric. This concentration difference lessens over subsequent wash stages, as the fabric retains less IL from the previous washes, translating to less IL being removed in the subsequent washes individually. Although, this is true for all three water-fabric ratios investigated, it is deduced that washing in 5 ml/g_{fabric} increments of water was the most effective at removing IL into the wash water. This is due to its higher frequency of washes, which resulted in 0.00% IL being washed out in its last 4 washes (Figures 4.2a and 4.2b). This is supported by comparing the DMBA concentrations in the last wash waters at the end of the washing process, as being 0.24%, 0.01% and 0.00% for the 20 ml/g_{fabric}, 10 ml/g_{fabric} and 5 ml/g_{fabric} stagewise operations, respectively. Another way to think of this, is the cumulative recovery of IL, explained later in section 4.3.2. As expected, this is also seen in the DMBA dyeing profiles in Figure 4.2a as well as, the TEA extraction and dyeing profiles shown in Figure 4.2b, since all the conditions for washing were kept constant. The data points plotted at 0 ml/g_{fabric} wash water represent the IL concentration of the dye baths directly after the extraction and dyeing processes, before washing. At 0 ml/g_{fabric}, the extraction profiles, for the three different water-fabric ratios, should all start at 80% IL concentration, confirmed using the Karl-Fischer machine. However, the data expressed in these graphs at 0 ml/g_{fabric} was deduced by recording the conductivity of the solutions using a conductivity meter which, although has an $R^2 > 0.999$ (Figure 4.1), did introduce uncertainty, which permeated through the calculations to the IL concentrations ($79.09 \pm 2.84\%$). The conductivity meter's uncertainty was suspected during lab experiments and confirmed by measuring the conductivity of tap water, which produced a reading of 0 mS, contrary to its actual conductivity. Nonetheless, the uncertainty's manifestation in the IL concentrations of the solutions is consistent across all the processed data as the conductivities were recorded by the same conductivity meter. Therefore, this does not affect the conclusions drawn from these figures.

Figures 4.2c and 4.2d demonstrate the near identical removal ability of DMBA compared to TEA from the fabric, employing the same washing procedure across the extraction and dyeing processes, evident in the profiles closely overlapping one another. In Figure 4.2c this is apparent for the 10 ml/g_{fabric} and 5 ml/g_{fabric} increments. There is a noticeable difference in the post-extraction wash 1 TEA concentration compared to DMBA for the 20 ml/g_{fabric} water-fabric ratio. This is because during the extraction process employing DMBA, a smaller pressure tube was used initially. After extraction, it was realised that the smaller size could not accommodate the 30 ml of wash water needed for the 1.5 g_{fabric} corresponding to the 20 ml/g_{fabric} water-fabric ratio. As a result, after squeezing the post-extraction dye bath (DMBA-dye mixture) in the smaller pressure tube, the fabric was transferred to a larger pressure tube for the 1st wash. As the TEA experiments were conducted after the DMBA experiments, this

foresight led to the use of the larger pressure tube for the TEA-extraction process (20 ml/g_{fabric} water-fabric ratio) so it could accommodate enough water for the washing process after extraction. This was decided in accordance with the standard operating procedure (in section 3.4), of performing the extraction step and post-extraction washes in the same pressure tube, for each sample being investigated. However, in hindsight, the method used for the 20 ml/g_{fabric} TEA extraction and washing should have been carried out in the same way as the 20 ml/g_{fabric} DMBA extraction and washing, done prior, for their concentrations after the 1st wash to be comparable. Manual recovery of the TEA post-extraction dye bath (TEA-dye mixture) for the 20ml/g_{fabric} sample proved to be more difficult with the use of the larger pressure tube. This is due to the inconsistency of the manual wringing-action employed between the washes for the 20 ml/g_{fabric} DMBA and TEA samples. In the larger pressure tube (TEA), the spatula used, was not long enough to allow the remaining fabric to be squeezed in the same manner as in the smaller pressure tube (DMBA) which resulted in less force being applied, by hand, and transmitted via the spatula to the fabric, during coercion of the extracted dye bath out of the fabric, for collection. Thus, leaving more of the extracted dye bath, containing TEA within the fabric and on its surface. As a result, when water was added to the pressure tube for the 1st wash, it appears in Figure 4.2c as more than double the amount of TEA (13.09%) being removed than DMBA (5.57%). Thus, the deviation between those two points was because more TEA was left in the fabric before the 1st wash and not due to TEA being easier to remove. TEA's greater viscosity, which entailed more difficulty in its detachment from the fabric as well as being responsible for leaving residual solution on the inner walls of the pressure tubes, reinforces this reasoning. Unexpectedly, the similar gradients of the slopes between the wash stages in the profiles in Figure 4.2c indicates that this property did not affect TEA's washing rate, showing as very similar to DMBA's washing rate. Hence, suggesting that the wash stages in the washing procedure removed the same amount of IL for both ILs and implies that the washing process is not mass transfer limited. This conclusion can only be drawn for the 5ml/g and 10ml/g_{fabric} profiles as the 20 ml/g_{fabric} profiles are incomparable due to the procedural inconsistency, associated with its 1st wash, explained earlier. Figure 4.2d confirms this behaviour for the post-dyeing washing process, also displaying similar washing rates between both ILs, after the 1st wash, in which the 20 ml/g_{fabric} profiles are valid for comparison, as the pressure tube size during dyeing was consistent for both TEA and DMBA.

It is important to remember that, as described in section 3.6, the 60% IL concentration of the dye bath employed in the dyeing process, was made by diluting the post-extraction dye bath with calculated amounts of wash water (Equation 4) from the 1st wash, for each of the three increment volumes. As a result, the systematic error introduced by the conductivity meter, explained earlier, has impacted the data points at 0 ml/g_{fabric}, in Figure 4.2d, in which all the dyeing profiles should start at 60%, but range from 56.74% to 70.57%, because of this. Particular to the dyeing profiles, it was assumed that the wash water from the 1st washes, used for dye bath dilution, was pure

DI water, however it discernibly contained IL and some extracted dye that was washed off during the 1st post-extraction wash, thereby altering the actual concentration of the dye bath prepared for the dyeing process. Equation 4 (section 3.6), which was used to obtain the corresponding mass of wash water required for the dilution of the post-extraction dye bath samples, did not account for this. Another source of error was fabric mass loss due to loose threads separating from the fabric pieces during cutting, transfer and squeezing actions described in the methodology, section 3, which was also unaccounted for in the mass balance. The error bars on Figure 4.2, also represent minor deviations induced from water evaporating during wash water retrieval from the pressure tubes. This is because although, the sealed pressure tubes were allowed to cool after coming out of the oven, they were still warm while the pressure tubes were opened to retrieve the wash waters into the falcon tubes. Thus, slightly changing the IL concentration.

Despite the 5 ml/g_{fabric} stagewise operation suggesting complete removal of all the IL from the fabric from its last four washes (washes 9, 10, 11 & 12) containing 0% IL for both DMBA and TEA across the extraction and dyeing processes across Figure 4.2. Through performing mass balances around the processes, it was deduced that although marginal, there were still small quantities of IL remaining in the fabric. This is understandable because of the difficulty in removing IL trapped between the fibres of the fabric's structure. This was to be expected but is not seen in Figure 4.2 because of the functional limits associated with the conductivity meter. It is only able to produce a reading if a solution's conductivity is between 0 and 2 mS. Further, the length and diameter of the outer casing surrounding the actual probe, limited the depth the conductivity probe could reach inside the falcon tube to encounter the solution inside. Coupled with the smaller stagewise wash water volume of 7.5 ml (1.5 g_{fabric}) corresponding to the 5 ml/g_{fabric} water-fabric ratio, the amount of sample was relatively little. During experiments, it was deduced that at least 10 ml of sample solution had to be in the falcon tube for the probe to be able to reach the sample. Ostensibly, this could not be done unless dilutions were performed to increase the volume of sample. Small quantities of DI water were added incrementally to the sample until the diluted solutions were enough quantitatively and within range of the conductivity meter, in a trial-and-error fashion. As the IL concentration decreased in the latter washes, dilution was not required to obtain a reading from the conductivity meter. However, without increasing the volume of the sample by diluting with DI water, the probe was not able to contact the sample. So, dilution was continued until the diluted samples' IL concentrations produced a reading of 0 mS. With a reading of 0 mS, back calculation of the diluted samples' IL concentration to find the actual solutions' IL concentration was not possible, thereby giving a reading of 0% for the 5 ml/g_{fabric} samples' last 3-4 washes. Moreover, the conductivity meter's functional limits and concentrated samples also necessitated dilutions for the 10 ml/g_{fabric} and 20 ml/g_{fabric} increments too, as explained in section 3.7. Although, the experiments were repeated 3 times for each sample, experimental error during sample collection via volumetric pipettes and

weighing using the mass balance while performing dilutions of the samples add to the uncertainty of the actual IL concentrations of the solutions.

4.3 Recovery of IL

Equations 9 and 10 were used to determine the recovery of IL in the process. To obtain the crude IL recovery, the terms for IL in the washing solutions were ignored.

$$R_{IL,Ex} = \frac{(C_{IL,Ex} \times m_{Ex,sol}) + \sum_{i=1}^n C_{IL,i} m_i}{(80 \times m_{IL,sol}^f)} \quad (9)$$

$$R_{IL,Dye} = \frac{(C_{IL,Dye} \times m_{Dye,sol}) + \sum_{j=1}^n C_{IL,j} m_j}{(60 \times m_{ds})} \quad (10)$$

where $R_{IL,Ex}$ and $R_{IL,Dye}$ are the recoveries of IL after extraction and dyeing, respectively in wt%, $C_{IL,Ex}$ and $C_{IL,Dye}$ are IL concentration in the post-extraction and post-dyeing solutions, respectively in wt%, $m_{Ex,sol}$ is the mass of post-extraction solution, $m_{IL,sol}^f$ is the feed mass of IL solution that was 10 g, $m_{Dye,sol}$ is the mass of post-dyeing solution, m_{ds} is the mass of dyeing solution, $C_{IL,i}$ and $C_{IL,j}$ are the IL concentrations in the washing solutions in wt%, m_i and m_j are mass of washing solutions wherein, i and j are the wash stage numbers after extraction and dyeing (1 to n), respectively.

4.3.1 Crude Recovery of IL

The crude recovery of IL in the post-extraction solutions should be similar regardless of the three interval wash water volumes used (20 ml/g_{fabric}, 10 ml/g_{fabric} and 5 ml/g_{fabric}) since the initial conditions were the same.

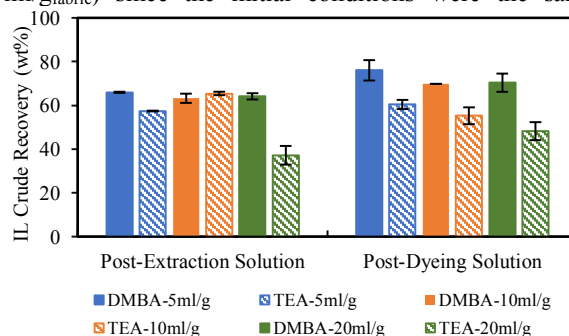


Figure 4.3.1. Crude recovery of IL

Observing Figure 4.3.1, this is true for the crude recovery of DMBA in post-extraction solutions as they are almost the same at $64.44 \pm 1.39\%$. On the other hand, there are significant differences between the crude TEA recoveries for the post-extraction solutions specifically, for the sample that was washed with 20 ml/g_{fabric} incremental water. As mentioned in section 4.2, the use of a larger pressure tube during the extraction process of the TEA-20ml/g_{fabric} sample posed increased difficulty while recovering the crude IL-dye mixture. This rationalises the observed low crude recovery of TEA (37.19%) in the post-extraction solution for the 20 ml/g_{fabric} sample. Variations in the crude IL recoveries for the post-dyeing solutions have primarily arisen due to irregularities in extracting the liquid from the fabric via manual squeezing. Transfer of the solutions from the pressure tubes to the falcon tubes also contributed to this. It also depends on the mass of the dyeing solution (Equation 10), which differs if the mass of the post-extraction dye bath varies (Equation 5). Overall, a greater crude recovery of DMBA was attained compared to TEA. Branching of alkyl side chains in TEA (Figure 2)

leads to its reduced rotational freedom, resulting in TEA possessing higher viscosity than DMBA^[32]. This made the process of removing the solutions more challenging for the TEA experiments, which led to a notable amount of TEA remaining on the inner side walls of the pressure tube and residing on the fabric's surface. This loss of solution mass, coupled with some droplets of the solution not being caught in the falcon tubes, which is relevant to both ILs' samples, during collection, accounts for the errors displayed in Figure 4.3.1.

4.3.2 Total Recovery of IL

Despite not achieving high crude recovery of IL, the remaining IL was recovered in the washing solutions. From Figure 4.3.2, the recovery of IL in the system increased with increasing wash stages, demonstrating a notable higher rate of recovery in the initial washes before reaching a plateau.

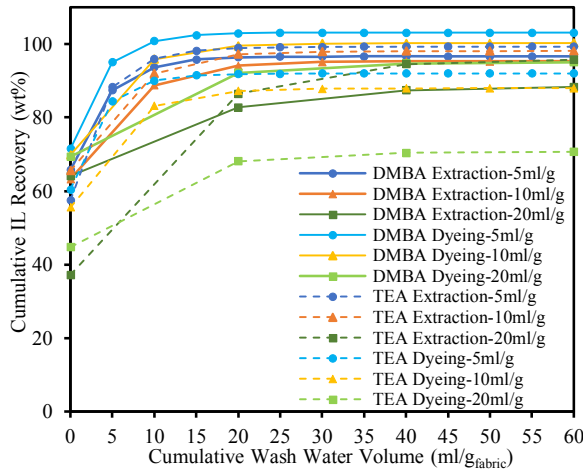


Figure 4.3.2. Total recovery of IL in the system

A greater IL recovery was also obtained using smaller wash water volume increments. This is evident as 100% of DMBA was recovered in the post-dyeing process after 12 washes using 5 ml/g_{fabric} water while only 95.02% of DMBA was recovered when 20 ml/g_{fabric} water was used (3 washes). Allowing the deduction to be made that the 5 ml/g_{fabric} water-fabric ratio was the most effective at removing IL from the fabric as almost all the IL was recovered, in total, at the end of the washing procedure. The recovery of DMBA was also higher in the post-dyeing process when compared to the post-extraction process. In contrast, the recovery of TEA was lower at the end of the post-dyeing washes. This is seen as the recovery of TEA post-dyeing and after the washing process is 70.67% whereas it is 95.72% post-extraction and after washing, using the same wash water volume (20 ml/g_{fabric}).

4.4 IL Lost in System

The total IL lost in the whole system was obtained using Equation 11 below.

$$L = \frac{(m_{IL,decfab} + m_{IL,dyedfab})}{0.8m_{IL,sol}} \times 100 \quad (11)$$

where L is the total IL lost in the whole process from initial IL solution and $m_{IL,decfab}$ and $m_{IL,dyedfab}$ are the mass of IL in decoloured and dyed fabrics, respectively.

For simplicity, it was assumed that the IL was retained in the decoloured and dyed fabrics after both the post-extraction and post-dyeing processes. An increase in the

IL recovery in overall washing solutions reduces the IL lost in the system. As seen in Figure 4.4, using 5 ml/g_{fabric} of water in both post-extraction and post-dyeing processes indeed results in the lowest amount of DMBA lost in the system (3.35%), whereas 20 ml/g_{fabric} shows the highest TEA loss in the process (15.26%). From further inspection, DMBA exhibits lower loss compared to TEA. Although, 3.35% is already the minimum loss of IL that can be achieved in this experiment, this value still represents a considerable quantity of IL loss and implies that the system must continuously supply an extra 3.35% of fresh DMBA to compensate for the IL loss.

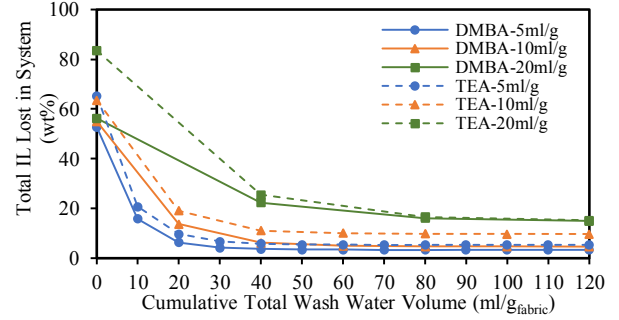


Figure 4.4. Total IL lost in the process

4.5 Colour Strength

The dyeing performance of the extracted dye bath that was recycled and used to colour virgin polyester was analysed using its colour strength (K/S) as described in section 3.8.

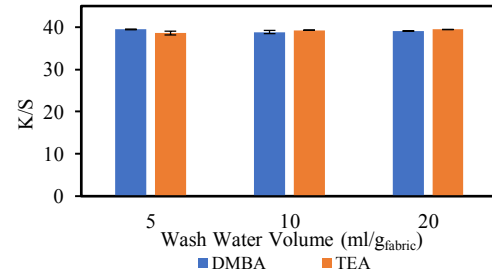


Figure 4.5. K/S of dyed fabric

The near constant K/S values (Figure 4.5) ranging from 39.66 ± 0.08 for the 5 ml/g_{fabric} DMBA sample to 38.29 ± 0.45 for the 5 ml/g_{fabric} TEA sample calculated for the three different increment wash water volumes and the two different ILs tested, demonstrates that the number of washes and stagewise wash water volumes did not affect the colour strength of the dyed polyester fabric. In addition, there is no trend in the K/S data to suggest that DMBA is better than TEA or vice versa, in their dyeing performance. One explanation for the minor differences in the K/S values across Figure 4.5 and the error bars shown, is how the fabric samples were placed inside the UV-vis spectrometer. This is because it relies on a laser hitting the fabric sample at an angle to record the samples' reflectance. The fabrics were subjected to an intensive washing procedure where between each wash, the fabric was being prodded by a spatula in the pressure tube to wring out as much of the IL-wash water solution, and reheated in the oven for 10 minutes, resulting in the final dried fabrics being very creased. The randomness of the creases and folds produced between the fabric samples may have impacted the angle the laser hit each fabric

sample causing slight variations in the reflectance (R), used in the calculation (Equation 8).

4.6 Partition Coefficient

The distribution of IL from the fabric into the wash waters during the washing process was investigated by calculating the partition coefficients (Equation 12) used in the stagewise washing operations for DMBA and TEA.

$$K = \frac{C_{IL,sol}}{C_{IL,fab}} \quad (12)$$

where K is the partition coefficient, $C_{IL,sol}$ is the IL concentration in the washing solutions while $C_{IL,fab}$ is the IL concentration in either decoloured or dyed fabric.

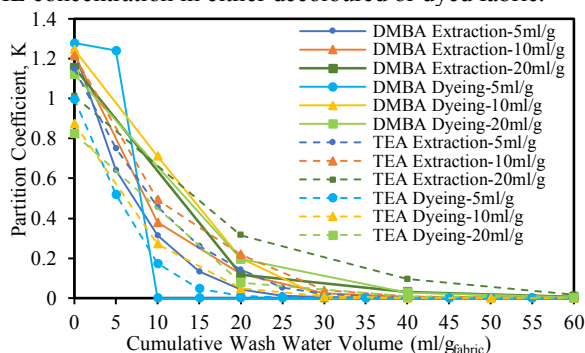


Figure 4.6. K values in the system

Figure 4.6 shows the partition coefficient decreasing with progressive wash stages meaning the washing process is physically limited. This can be explained because of the reducing concentration gradient associated with subsequent wash stages coupled with, the difficulty the IL experiences to dislodge itself from between the fabric's fibres and allow its migration into the wash water because of their bulky structures (Figure 2), making it easier for the IL molecules to get caught in the fibres. For DMBA, the dyeing data points all have a higher partition coefficient than their extraction counterparts. This can be understood because after the dyeing process, the fabric is saturated with dye molecules, so there is less volume within the fabric's structure that is occupied with IL and more of it is either on the surface or in the surrounding solution. This trend not seen in the TEA profiles, possibly due to its higher viscosity. Contrastingly, in post-extraction, the dye molecules are drawn out of the fabric so there is more IL occupying and trapped in the fibre's free space volume before the washing starts, therefore more IL is drawn out in the washes after dyeing, than after extraction. Contrary to expectation, the TEA extraction and dyeing profiles appear to have higher partition coefficients than DMBA when comparing each sample, which conflicts with TEA's more bulky chemical structure and lower IL recovery. But, upon deeper inspection of the data points at 0 ml/g_{fabric}, this transpires because the K values for DMBA are higher in the extraction solutions. Meaning, that most of the DMBA was removed in the extracted dye baths before the washing commenced. So, during the washing, since less IL was present in both the wash water solutions and in the fabric for DMBA, the K values are lower in comparison to TEA.

5. Conclusions and Outlook

In conclusion, DMBA performed superior to TEA by inducing minimum loss of IL during its employment.

Therefore, it is the preferred ionic liquid for the dye recycling process, from the two ILs investigated overall. Furthermore, higher IL recoveries were achieved using smaller increment volumes of wash water, with the lowest increment of 5 ml/g_{fabric} being the most effective at removing IL from the fabrics. It was also proved that the colour strength of dyed fabric remains unaffected by the stagewise wash water volumes and type of IL. Sought to prevent requirement of fresh IL in the system which would incur additional costs, the objective of attaining negligible (less than 1%) IL loss was not reached in this study.

Refinement of the methods utilised while conducting experiments, to reduce errors may improve IL recovery. Replacement of the manual squeezing action used to extract the liquid mixtures from the fabrics, with less aggressive vacuum filtration, press machine/hydraulic press or simply using pipettes to absorb the remaining liquid residing on the fabric's surface and inner apparatus walls would help to reduce the human error and procedural error reflected in IL concentrations. They would also prevent the fabric fibres from experiencing unnecessary mechanical stress which is an important factor to consider for commercialisation purposes. Measures to minimise accidental sample loss, residual sample loss on equipment and fabric mass loss that were unaccounted for in the mass balance, are also necessary to increase IL recovery. Systematic error introduced by the measuring instruments, mainly from the conductivity meter, which was employed to ascertain the IL concentrations of solutions, propagated into the results. Also, its functional limitations called for dilutions of most samples, which increased the uncertainty of the results (explained in section 4.2). Volumetric Karl Fischer titration would offer more precise determination of IL concentrations of solutions, albeit requiring frequent calibration before its use (once a week). This is achieved by measuring the water content in the IL solutions and with a simple calculation, the IL concentration can be obtained. After the washing stage, it was assumed that the remaining IL was lost within the fabric. Scanning Electron Microscopy (SEM), along with Fourier Transform Infrared Spectroscopy (ATR-FTIR) can be used to quantitatively verify the presence of IL in the fabric. SEM can capture the surface morphology of the fabric and ATR-FTIR would be useful for identifying the chemical structure (functional group) and composition of the IL, provided it is indeed present in the fabric^[33]. The obtained results can be compared with the calculated data in section 4.4 to further validate this assumption. Counter current washing is a method that could improve the recovery of IL. The least contaminated water from the final wash is recycled for the second-to last wash and this continues, in sequence, until it reaches the initial wash stage, at which point it is discharged^[34]. This method is cost-effective and relatively straightforward to implement in multi-stage washing processes. However, it is not applicable for this experiment because the partition coefficient varies in each washing stage, as explained in section 4.6. Thus, further studies need to be performed before its implementation. It is also useful to investigate the impact of auxiliary chemicals circulating in the process, that are leached into the extracted dye bath and solutions from the textile waste starting material, and how they affect IL recovery as the wash waters containing IL will be evaporated.

6. Acknowledgements

We express our sincere gratitude to Antonio Ovejero, Aida R. S. Aboulela and Jiaying Chen for their invaluable support and guidance for the duration of this project.

7. References

- [1] Bailey, K., Basu, A. & Sharma, S. (2022) The Environmental Impacts of Fast Fashion on Water Quality: A Systematic Review. *Water*, 14 (7), 1073. doi:10.3390/w14071073.
- [2] Niinimäki, K., Peters, G., Dahlbo, H., Perry, P., Rissanen, T. and Gwilt, A. (2020). The Environmental Price of Fast Fashion. *Nature Reviews Earth & Environment*, [online] 1(4), pp.189–200. doi:https://doi.org/10.1038/s43017-020-0039-9.
- [3] Al-Tohamy, R., Ali, S.S., Li, F., Okasha, K.M., Mahmoud, Y.A.-G., Elsamahy, T., Jiao, H., Fu, Y. and Sun, J. (2022). A critical review on the treatment of dye-containing wastewater: Ecotoxicological and health concerns of textile dyes and possible remediation approaches for environmental safety. *Ecotoxicology and Environmental Safety*, 231, p.113160. doi:https://doi.org/10.1016/j.ecoenv.2021.113160.
- [4] Chiappe, Cinzia & Bianchini, Roberto & Cevasco, Giorgio & Pomelli, Christian & Rodriguez Douton, Maria Jesus. (2015). Ionic Liquids Can Significantly Improve Textile Dyeing: An Innovative Application Assuring Economic and Environmental Benefits. *ACS Sustainable Chemistry & Engineering*, 3, 150729120304009. 10.1021/acssuschemeng.5b00578.
- [5] EPA (2018). Textiles: Material-Specific Data. [online] US EPA. Available at: <https://www.epa.gov/facts-and-figures-about-materials-waste-and-recycling/textiles-material-specific-data>.
- [6] Textile Exchange. (n.d.). Preferred Fiber and Materials. [online] Available at: <https://textileexchange.org/knowledge-center/reports/preferred-fiber-and-materials/>.
- [7] Rafat Said Aboulela, A. (2020). Development of Sustainable Chemical Technologies Using Low-cost Ionic liquids for Waste Decontamination and Valorization. Thesis.
- [8] Teli, M.D. (2008). Textile coloration industry in India. *Coloration Technology*, 124(1), pp.1–13. doi:https://doi.org/10.1111/j.1478-4408.2007.00114.x.
- [9] Shamey, Renzo & Zhao, Xiaoming. (2014). Modelling, Simulation and Control of the Dyeing Process. Elsevier, pp. 1–30
- [10] P. Senthil Kumar and S. Suganya (2017). Sustainable Fibres and Textiles | ScienceDirect. [online] Available at: <https://www.sciencedirect.com/book/9780081020418/sustainable-fibres-and-textiles>.
- [11] Mahapatra, N.N. (2016). Textile Dyes (1st ed.). WPI Publishing. <https://doi.org/10.1201/b21336>
- [12] Gao, D., Yang, D.F., Cui, H.S., Huang, T.T. and Lin, J.X., 2015. Supercritical carbon dioxide dyeing for PET and cotton fabric with synthesized dyes by a modified apparatus. *ACS Sustainable Chemistry & Engineering*, 3(4), pp.668-674.
- [13] Banchero, M., 2013. Supercritical fluid dyeing of synthetic and natural textiles—a review. *Coloration Technology*, 129(1), pp.2-17.
- [14] Montero, G., Hinks, D. and Hooker, J., 2003. Reducing problems of cyclic trimer deposits in supercritical carbon dioxide
- [15] Chemical Safety Facts. (n.d.). Chlorinated Solvents. [online] Available at: <https://www.chemicalsafetyfacts.org/chemicals/chlorinated-solvents/>.
- [16] Ferrero, F., Periolatto, M., Rovero, G. and Giansetti, M., 2011. Alcohol-assisted dyeing processes: a chemical substitution study. *Journal of Cleaner Production*, 19(12), pp.1377-1384.
- [17] Robinson, E.G., 2020. Textile recycling via ionic liquids.
- [18] Fei, X., Freeman, H.S. and Hinks, D., 2020. Toward closed loop recycling of polyester fabric: Step 1. decolorization using sodium formaldehyde sulfoxylate. *Journal of Cleaner Production*, 254, p.120027.
- [19] Brandt, A., Ray, M.J., To, T.Q., Leak, D.J., Murphy, R.J. and Welton, T., 2011. Ionic liquid pretreatment of lignocellulosic biomass with ionic liquid–water mixtures. *Green Chemistry*, 13(9), pp.2489-2499.
- [20] Hunt, P.A., Kirchner, B. and Welton, T., 2006. Characterising the electronic structure of ionic liquids: an examination of the 1-butyl-3-methylimidazolium chloride ion pair. *Chemistry—A European Journal*, 12(26), pp.6762-6775.
- [21] Plechkova, N.V. and Seddon, K.R., 2008. Applications of ionic liquids in the chemical industry. *Chemical Society Reviews*, 37(1), pp.123-150.
- [22] Hayes, R., Warr, G.G. and Atkin, R., 2015. Structure and nanostructure in ionic liquids. *Chemical reviews*, 115(13), pp.6357-6426.
- [23] Greaves, T.L. and Drummond, C.J., 2008. Protic ionic liquids: properties and applications. *Chemical reviews*, 108(1), pp.206-237.
- [24] Al-Etaibi, A.M. and El-Asasery, M.A., 2023. Can Novel Synthetic Disperse Dyes for Polyester Fabric Dyeing Provide Added Value?. *Polymers*, 15(8), p.1845.
- [25] Al-Etaibi, A.M. and El-Asasery, M.A., 2022. Microwave-assisted synthesis of azo disperse dyes for dyeing polyester fabrics: Our contributions over the past decade. *Polymers*, 14(9), p.1703.
- [26] Opwis, K., Benken, R., Knittel, D. and Gutmann, J.S., 2017. Dyeing of PET fibers in ionic liquids. *International Journal of New Technology and Research*, 3(11), p.263192.
- [27] Yuan, J., Wang, Q. and Fan, X., 2010. Dyeing behaviors of ionic liquid treated wool. *Journal of Applied Polymer Science*, 117(4), pp.2278-2283.
- [28] Bianchini, R., Cevasco, G., Chiappe, C., Pomelli, C.S. and Rodriguez Douton, M.J., 2015. Ionic liquids can significantly improve textile dyeing: an innovative application assuring economic and environmental benefits. *ACS Sustainable Chemistry & Engineering*, 3(9), pp.2303-2308.
- [29] Chen, J. (2023). *Ionic Liquid Textile Dyeing and Sustainable Dye Recycling*. Thesis.
- [30] Study Mind. (n.d.). *Ionic Compound Properties (GCSE Chemistry)*. [online] Available at: <https://studymind.co.uk/notes/ionic-compound-properties/>.
- [31] APERA INSTRUMENTS. (2018). *Conductivity, TDS, Salinity, Resistivity*. [online] Available at: <https://aperainst.com/blog/cat/Conductivity%2C+TDS%2C+Salinity%2C+Resistivity/#:~:text=There%20are%20three%20main%20factors>.
- [32] Virginia Gschwend, F.J. (2017). Towards an Economical Ionic Liquid Based Biorefinery. Thesis.
- [33] Villalta, E., Riba Moliner, M. and Lis Arias, M.J., 2022. Recovery of cellulose from polyester/cotton fabrics making use of ionic liquids. *Polymer Science: Peer Review Journal*, 4(3, article 000590).
- [34] www.fibre2fashion.com (n.d.). Water efficiency in textile processes. [online] www.fibre2fashion.com. Available at: <https://www.fibre2fashion.com/industry-article/3406/water-efficiency-in-textile-processes>.

Engineering Magnetically Steerable Biohybrid Cells

Osei Kendell and Kadiy'a Roberts

Department of Chemical Engineering, Imperial College London, U.K.

Abstract

Bacterial cells have proven to be useful for several therapeutic and diagnostic applications. These cells have the potential to have a significant positive impact on areas of treatment within the medical field such as the treatment of tumours, autoimmune diseases, and targeted drug delivery. Creation of bacterial biohybrids, done by the attachment of synthetic material to the bacteria, expands the scope of applications in which these cells can be used. For instance, bacterial biohybrids can be made to respond to external stimuli and thus be used in therapeutics for the controlled delivery of cargo to target sites within the body. However, current biohybrids underperform in this area, hence, it is within this that the work presented in this paper lies. In this paper, we report on engineering magnetically steerable biohybrid cells by interfacing *E. Coli* bacteria with magnetoliposomes and by interfacing lipid encapsulated bacteria with magnetic nanoparticle (cationic exchange). When placed in a microfluidic cell and a magnetic field applied, biohybrids prepared using both methods responded favourably, moving towards the magnet. Additionally, when comparing the optical density of the bacteria before interfacing with nanoparticle to that after magnetic separation it was clear that some bacteria became magnetised and responded positively to the applied magnetic field. This work illustrates favourable methods for the preparation of magnetically steerable bacterial biohybrid cells.

Key Words: Bacterial biohybrids; Magnetic nanoparticle (MNP); Magnetotaxis; Liposome; Encapsulation; Optical density (OD).

1 Introduction

Bacterial cells have been widely studied for centuries, and in more recent times, there has been extensive work on the use of bacteria in therapeutics. One approach to this is the use of bacterial biohybrids which are bacterial cells interfaced with synthetic material in an attempt to increase their functionality. Extensive work has been done on the use of bacterial biohybrids as carriers in the form of microrobots. In this application, the intrinsic propulsion, targeting and sensing abilities of the bacteria aid in their successful application as micro-swimmers [1].

In nature, there exists bacterial magnetic nanoparticles (BNPs), which are another class of nanoparticles widely utilized in therapeutic applications. These are produced within naturally occurring magnetotactic bacteria which contain magnetosomes which are organelles that consist of a magnetic nanoparticle core surrounded by a lipid bilayer. These naturally occurring particles have been used for applications in targeted drug delivery, magnetic resonance imaging (MRI) and magnetic hyperthermia [2,3].

Overtime, bacterial cells have evolved to swim efficiently using food sources within their environment at the micrometre scale as an energy source [4]. This self-propulsion capability makes it advantageous to use bacteria in therapeutics. When coupled with magnetic control, this property can be significantly improved, making the movement of bacteria more efficient while allowing for increased control of the bacteria's special positioning.

The research presented in this paper is aimed towards coupling the self-propelling abilities of bacterial cells with

magnetic control through attachment with magnetic nanoparticles (MNPs) thus developing magnetically steerable biohybrid cells. This would allow for bacteria to be used for applications such as targeted drug delivery and the treatment of tumours. Being able to successfully guide these bacterial biohybrids can also reduce some deleterious effects of present biohybrids including premature cargo release and non-specific treatment due to an inability to guide bacteria to a target site. Two approaches were investigated in this research. The first approach involves interfacing bacterial cell with magnetoliposomes (encapsulated nanoparticles) based on electrostatics. The second involves interfacing lipid encapsulated bacteria with MNPs through carbodiimide facilitated coupling.

2 Background

The use of magnetotactic bacteria in therapeutics is an area of research with growing interest. Several applications of this have been studied. Bacterial therapeutics have proven useful in areas such as treating autoimmune and inflammatory disease, cancer treatment, combatting viral and bacterial infections and treatment of metabolic disorders [5]. One notable application involved using these bacterial cells to enhance image contrast for magnetic resonance imaging (MRI) [6]. The research presented in the mentioned paper, by M. R. Benoit *et al*, gave light to several positive aspects with the targeting ability of the bacteria being at the forefront. This targeting ability is one of the many features that makes bacterial therapeutics advantageous. In this research, it was found that 2-6 days after being administered intravenously to

mice, the bacteria had accumulated in tumours thus increasing MRI signal.

Due to the self-propelling nature of bacteria, they have proven to enhance the delivery of therapeutics to hypoxic tumour regions as they can counteract tumour interstitial fluid pressure (TIFP). This pressure is responsible for preventing the efficient delivery of therapeutics by liposomes [7]. Research has shown that the propensity of bacteria to seek out, via aerotaxis, the hypoxic regions in tumours, makes it suitable for applications in delivering cargo to these regions that otherwise go untreated [8]. In the same paper, the magneto-aerotactic behaviour of *Magnetococcus marinus* strain MC1 bacteria was used to successfully guide drug-loaded nanoliposomes to the target tumour site. When compared to passive agents the MC1 cells had superior penetration due to magnetotaxis and aerotaxis. Further to this, other bacteria genera such as *Escherichia Coli* (*E. Coli*) and *Salmonella* have been found to preferentially accumulate in tissues due to chemotaxis and aerotaxis [8]. This gives insight into the applicability of coupling magnetic guidance with the natural motility of bacterial cells for targeted drug delivery to deep tumours.

In addition to the benefits in targeted drug delivery, the use of bacteria can aid in increasing the efficacy of drug delivery. Bacteria can be used to deliver drugs that are easily degraded in the stomach, bloodstream or in transit through the upper gastrointestinal tract [9]. Additionally, this method of drug administration reduces systemic drug exposure and has potential to decrease the number of side effects experienced by patients when compared to other administration methods [9].

The work of Akolpoglu et. al, 2022 speaks to the development magnetically steerable bacterial biohybrids as microrobots. They stated that when interfaced with therapeutics, contrast agents and targeting components bacterial biohybrids become ideal candidates for medical micro robotics. In this work, they presented one method of developing magnetic bacterial biohybrids which involved using MNPs with covalently bound streptavidin. To increase functionality by including nanoliposomes, a biotin-streptavidin-biotin complex was used thus allowing for the integration of artificial units onto the bacterial cells. This paper reported a 92.2 % success with conjugating the MNP cells and 86.3 % success with conjugating both MNPs and nanoliposomes indicating that engineering magnetotactic bacterial biohybrids in this way is viable.

One area of concern surrounding the use of bacteria in therapeutics is their survival when facing extreme conditions within the body such as strong acids and alkalis, antibiotics, and ethanol [10]. The work done in this paper highlights the benefits of coating bacteria with lipids to combat these negative effects thus increasing their resistance to these conditions while maintaining their viability.

While there has been substantial research on the use of bacteria in therapeutics, there is limited research available on the specific topic of magnetotactic bacterial biohybrids for similar applications. The aim of our

research is to engineer magnetically steerable biohybrid cells using MNPs. While there has been some work in this area, our aim is to encapsulate either the bacterial or MNP component of the biohybrid to determine the viability of these methods of biohybrid engineering.

3 Materials and Methodology

3.1 Materials

Chemicell fluidMAG-Amine 25 mg/mL magnetic nanoparticles, Chemicell fluidMAG-PAS 25 mg/mL magnetic nanoparticles, 25 mg/mL Avanti Polar Lipids 18:1 ($\Delta 9$ -Cis) PC (DOPC), 25 mg/mL Avanti Polar Lipids 18:1 TAP (DOTAP), 25 mg/mL Avanti Polar Lipids 18:1 ($\Delta 9$ -Cis) PE (DOPE), Rhodamine B, Phosphate Buffered Saline (PBS), Ampicillin-resistant *E. Coli* (MG1655), GFP-expressing *E. Coli*, 2% Uranyl Acetate (UA), 0.1% Phosphotungstic Acid (PTA).

3.2 Equipment

Zetasizer Ultra – Malven Panalytical, UK, Avanti Mini-Extruder – Avanti Polar Lipids, UK, JEOL STEM 2100Plus Electron Microscope, Nikon Eclipse Ti2 Light Microscope, VWR Digital Vortex Mixer, VWR Ultrasonic Cleaner, Eppendorf Centrifuge 5415 D, WPA CO8000 Cell Density Meter, NanoDrop One Microvolume UV-Vis Spectrophotometer.

3.3 Nanoparticle characterization

3.3.1 Dynamic Light Scattering (DLS)

Prior to dilution, the MNP stock solution was vortexed for 2 minutes at 2500 rpm using the digital vortex mixer to ensure homogeneity. Once completed, 1 mL samples of varying concentrations were produced by diluting the stock solution in predetermined volumes of water. The concentrations produced were: 10 mg/mL, 5 mg/mL, 2.5 mg/mL, 1 mg/mL, 0.5 mg/mL, and 0.25 mg/mL. These samples were vortexed for 2 minutes and sonicated for 10 minutes using the ultrasonic cleaner to ensure thorough mixing. To acquire the desired information regarding particle size, concentration, particle count rate and zeta potential, dynamic light scattering (DLS) was employed using the Zetasizer Ultra machine and following the standard DLS procedure outlined in the Zetasizer manual.

3.3.2 UV-Vis Spectroscopy

Three samples of diluted MNP stock solution were prepared at concentrations of 5 mg/mL, 0.5 mg/mL, and 0.05 mg/mL. These samples were added to a cuvette and the UV-Vis spectrum obtained by measuring them using the MicroDrop One Microvolume UV-Vis Spectrophotometer.

3.4 Effects of salt-based dispersants on Nanoparticles

The stock solution of MNPs was vortexed for 2 minutes at 2500 rpm using the digital vortex mixer. The solution was then diluted to a concentration of 2.5 mg/mL using PBS. This sample was vortexed for 2 minutes and sonicated for 10 minutes to ensure thorough mixing. Particle size, concentration and particle count rate data was gathered using DLS.

3.5 Nanoparticle encapsulation

3.5.1 Sample Preparation

Four vials containing 5 mg of lipid solution were prepared by combining DOPC and DOTAP in a 90:10 mol% ratio (respectively). This mixture was then dried utilizing a stream of nitrogen gas, rotating continuously to form a film along the walls of the vial, and placed in a desiccator overnight. After 3 hours, each batch was rehydrated by adding 1 mL of fluidMAG-Amine MNP solution at various dilutions (0.5 mg/mL, 1.0 mg/mL, 2.5 mg/mL, and 5 mg/mL) and the mixture vortexed for 2 minutes and sonicated for 10 minutes. The mixture was then left for another 24 hours [11]. Once rehydration was completed, the samples were analysed using DLS.

3.5.2 Sample Imaging – Transmission Electron Microscopy (TEM)

Samples of concentration 2.5 and 5 mg/mL were prepared as described in section 3.5.1 and imaged using TEM.300-mesh grid size carbon support grids (Agar Scientific) were used, and these were glow discharged for 1 minute. Following this, 2 μ L of UA solution was added to 4 μ L of the prepared MNP samples and thoroughly mixed. 4 μ L of this mixture was swiftly added to the discharged grid and it was stained for 1 minute. The solution was blotted away using a piece of Whatman 1 filter paper and the grid was left in the air until fully dried. This process was repeated using PTA. TEM imaging was carried out under 200 kV.

3.6 Magnetoliposome Size Reduction – Extrusion

Following the encapsulation of the MNPs, the 5 mg/mL solution of magnetoliposomes was vortexed for 2 minutes then sonicated for 15 minutes. The sample was then extruded 21 times through a membrane with pore size of 1 μ m. Following this, the sample was further extruded through a membrane of pore size 0.2 μ m.

3.7 Magnetoliposome Purification

5 mg of 90:10 mol% DOPC: DOTAP lipid film was prepared and dried in the desiccator for 2 hours. A 5 mg/mL solution of MNPs was prepared and this was sonicated for 15 minutes. Following this, the lipid film was

rehydrated with 1 mL of the MNP solution, and this mixture was again sonicated for 15 minutes.

3.7.1 Magnetic separation

400 μ L of the rehydrated lipid mixture was added to an Eppendorf tube and a magnet placed below the tube. The sample was allowed to separate for 25 minutes forming a pellet of magnetic material at the bottom of the tube. Following this, the supernatant was extracted and added to another Eppendorf tube and the remaining pellet was resuspended in 300 μ L of DI water. The fractions were then analysed using UV-Vis's spectroscopy.

3.7.2 Size Exclusion Chromatography (SEC)

The SEC column was first filled with resin and allowed to settle for 1 hour. Following this, 200 μ L of the rehydrated lipid mixture was added to the column and excess fluid removed from the bottom. Once the sample was approaching the bottom of the column, DI water was added to the column in 200 μ L increments and 200 μ L of sample was collected from the bottom of the column. 12 samples were collected and were then analysed using UV-Vis spectroscopy.

3.8 Nanoparticle encapsulation optimization

3.8.1 Low-volume rehydration

A batch of a 90:10 mol% DOPC: DOTAP lipid mixture corresponding to a mass of 5 mg was prepared (with 10 μ L of fluorescent dye for future imaging purposes), dried using a stream of nitrogen gas to form a lipid film on the walls of the vial and placed in a desiccator for 3 hours. This film was then rehydrated with 100 μ L of the 25 mg/mL stock MNP solution. This mixture was vortexed for 4 minutes and sonicated in a warm water bath for 15 minutes, then vortexed for a further 4 minutes. Once thoroughly agitated the sample was stored for 24 hours. Finally, the sample was diluted by the addition of 900 μ L of DI water and analysed using the DLS.

3.8.2 High-concentration hybrid film

A 100 μ L sample of the 25 mg/mL stock MNP solution was dried in a vial via a stream of nitrogen gas. Following this, a 5 mg batch of a 90:10 mol% DOPC: DOTAP lipid mixture was prepared (with 10 μ L of fluorescent dye for future imaging purposes), dried using a stream of nitrogen gas to form a lipid film on the walls of the vial and on top of the MNP film and placed in a desiccator for 3 hours. To the hybrid film, 100 μ L of the 25 mg/mL stock MNP solution was added. The mixture was vortexed 4 times for 2 minutes each time and between each round of vortexing, the mixture was sonicated (15 minutes then two rounds of 5 minutes each). This mixture was then diluted by first adding 900 μ L of DI water to a MNP concentration of 5 mg/mL, then diluted further by adding

1 mL to a concentration of 2.5 mg/mL. Analysis using DLS was done after each dilution.

3.9 Bacteria Preparation

3.9.1 Bacteria fermentation

The bacteria cells used in this process are stored at -80°C. These cells were thawed on ice and then inoculated. The inoculation medium consisted of Lysogeny broth (LB) supplemented with ampicillin (1000x diluted). The cells were inoculated in 5 mL of this medium then incubated at 37°C and shaken at 200 rpm for 2 hours. Following this, the cells were inoculated on an LB-Ampicillin agar and incubated overnight. These plates were stored at 4°C for up to four months.

When needed, a single colony was inoculated from the plate and added to 5 mL of medium in a 50 mL Falcon tube. This was left in a shaking incubator at 37°C and 200 rpm overnight. The following day, 50 µL of this culture was added to 5 mL of medium in a 50 mL Falcon tube and left in a shaking incubator at 37°C and 200 rpm until the desired optical density (OD₆₀₀ or OD) was achieved.

3.9.2 Bacteria purification

500 µL of bacteria stock was added to an Eppendorf tube. This sample was then centrifuged at 5 RCF for 5 minutes. Once the separation was complete, the supernatant was extracted and discarded leaving the bacteria pellet behind. The pellet was then initially redistributed in 500 µL of DI water and the OD measured. Further dilution was done, if necessary, to achieve the desired OD.

3.10 Biohybrid Formation – via Encapsulated Nanoparticles

3.10.1 Interfacing Bacteria with Magnetoliposomes

A 100 µL sample of the purified bacteria was added to a new Eppendorf tube. To this tube, 100 µL of magnetoliposomes, produced using the Low-volume rehydration method, were added and the contents agitated for 10 seconds using a vortex mixer to achieve homogeneity. The Eppendorf tubes containing bacteria and magnetoliposomes were centrifuged for 5 minutes at 5 RCF twice, discarding the supernatant and rehydrating with 200 µL of DI water each time.

3.10.2 Control Preparation

A 100 µL sample of purified bacteria was added to a new Eppendorf tube. To this tube 100 µL of fluidMAG-Amine MNPs at a concentration of 2.5 mg/mL was added and the mixture agitated via vortex for 5 seconds to ensure mixing. The Eppendorf tube containing the bacteria and the free MNPs was centrifuged for 5 minutes at 5 RCF twice, discarding the supernatant and rehydrating with 200 µL of DI water each time.

3.11 Biohybrid Formation – via Carbodiimide Facilitated Coupling

3.11.1 Bacteria encapsulation

Three batches of 5 mg/mL lipid vesicle solutions with compositions of 85:10:5 mol % DOPC: DOTAP: DOPE, 80:10:10 mol % DOPC: DOTAP: DOPE and 50:50 mol% DOPC: DOTAP: DOPE were prepared. This was done by adding the required volume of each lipid to a vial with 7.5 µL of Rhodamine B and vortexing the mixture for 2 minutes. The mixture was then dried to create a lipid film on the walls of the vial. These vials were then placed in the desiccator for 3 hours to allow for complete drying. After drying the films, they were rehydrated using 1 mL of deionized water, vortexed for 2 minutes, sonicated for 15 minutes and finally extruded through a 100 nm membrane.

200 µL of bacteria stock were extracted and added to an Eppendorf tube. This was centrifuged for 5 minutes at 5 rcf. The supernatant was extracted and 800 µL of deionized water was added to the Eppendorf tube. The mixture was agitated to dissolve the bacteria bead and ensure homogeneity. The OD of the redistributed bacteria was measured, and the solution was diluted to the required OD (1.1 for OD experiment and 0.7 for light microscopy). Following this process, three Eppendorf tubes were prepared with 200 µL of purified bacteria. To these tubes, 200 µL of each liposome formulation was added and they were labelled accordingly.

3.11.2 FluidMAG-PAS Nanoparticle activation

10 mg of EDC was added to a vial and 150 µL of deionized water was added. This mixture was vortexed for 2 minutes to ensure homogeneity. Following this, the fluidMAG-PAS MNP stock was vortexed as recommended by the manufacturers and 400 µL were added to the vial containing EDC. This mixture was then vortexed for 2 minutes. [12]

3.11.3 Interfacing encapsulated bacteria with activated nanoparticles

For each lipid composition, 200 µL of encapsulated bacteria was added to an Eppendorf tube with 200 µL of activated MNPs. The mixture was vortexed for 5 seconds. Following this, the samples were centrifuged for 5 minutes at 5 rcf to help remove any free MNPs. The supernatant was then extracted, and the sample was redistributed in 400 µL of DI water.

3.11.4 Control preparation

200 µL of encapsulated bacteria with 200 µL of non-activated MNP solution (the MNP solution was vortexed for 2 minutes before use). The mixture was vortexed for 10 second then centrifuged for 5 minutes at 5 rcf. The supernatant was extracted and discarded, and the sample redistributed in 400 µL of DI water. This was done for each lipid solution composition.

3.12 Biohybrid Confirmation

3.12.1 Light microscopy

Light microscopy was employed to confirm the presence of bacterial biohybrids. A microfluidic assay was filled with DI water and 20 μL of a sample was added to one of the wells. Following this, a magnet was placed downstream of the added sample, the response of the sample was observed, and a video recorded. This procedure was used for both the nanoparticle encapsulation and carbodiimide facilitated coupling approaches and all corresponding controls.

3.12.2 Optical Density experiment

This method was utilized for confirmation of biohybrids formed through the cation exchange approach. Prior to encapsulation, the bacteria was diluted to an OD of 1.1. 200 μL of encapsulated bacteria plus 150 μL of activated MNPs was added to an Eppendorf tube. The mixture was vortexed for 10 seconds then a magnet was placed below the sample for 15 minutes allowing for a pellet of magnetic material to be formed. Following this, the supernatant was extracted and to this 175 μL of DI water was added. The OD of the supernatant was then measured. This procedure was conducted for each of the lipid compositions and corresponding controls.

4 Results

4.1 Nanoparticle characterization

4.1.1 FluidMAG-Amine

The MNPs in solution had a red-brown colour which decreased in intensity with dilution. It was observed that the size measurements obtained for each sample varied with sizes ranging from $200\text{ nm} \pm 30\text{ nm}$ to $5000\text{ nm} \pm 30\text{ nm}$. The measured size for each concentration is presented in figure 1 below.

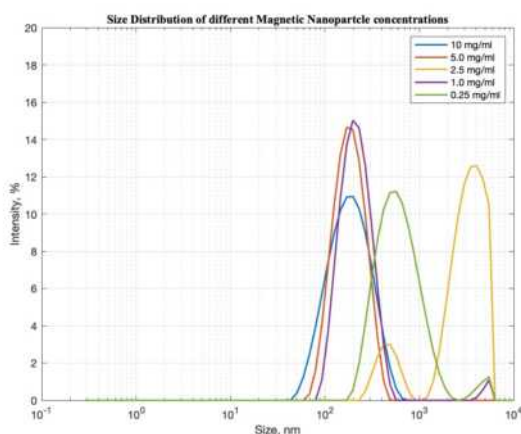


Figure 1: Size distribution of MNPs at different concentrations

The focus when collecting concentration and count rate data was whether at a given concentration, DLS could

give a reliable result. It was clear the higher the concentration, the less likely it was that DLS would give any results – made apparent by the equipment not being able to provide concentration or count rate data for the 10 mg/mL concentration or higher. As a result, it was determined that concentrations of 5 mg/mL and below would be used to ensure the reliability of data. The zeta potential was found to be $\cong -2.54\text{ mV}$.

Using UV-Vis spectrometry, it was determined that these nanoparticles give a characteristic peak between 290 nm and 390 nm. At a concentration of 5 mg/mL, there was saturation observed indicating that UV-Vis measurements for these nanoparticles should be done at lower concentrations. The spectrum obtained can be found in the supplementary information provided.

4.1.2 FluidMAG-PAS

Size measurements of the activated MNPs were conducted at concentrations of 1 mg/mL and 0.5 mg/mL. In both cases, a size measurement of 218 nm which was above the manufacturer specified size of 100 nm was obtained. At both concentrations, it was possible to obtain count rate and concentration data for these nanoparticles. The zeta potential was found to be $\cong -32.01\text{ mV}$.

4.2 Effects of salt-based dispersants on Nanoparticles

4.2.1 FluidMAG-Amine

The results obtained after carrying out DLS measurements showed a significant increase in particle size with the average size being 4600 nm. The change in particle size when dispersed in PBS in place of DI water is shown in figure 2. There was also significant aggregation and settling observed when looking at the sample. Concentration measurements for this sample did not give meaningful data as the particle count and concentration measurements both gave a result of 0 particles/mL.

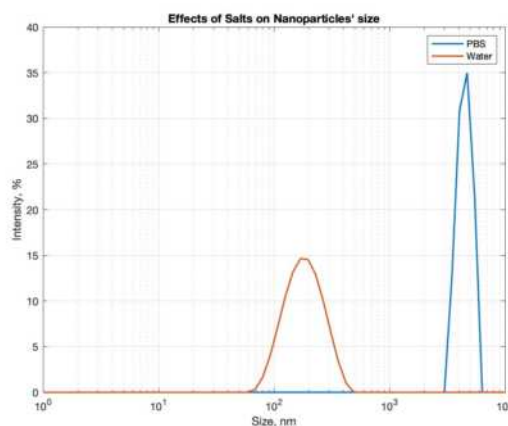


Figure 2: Comparison of measured particle size when dispersed in water and PBS.

4.2.2 FluidMAG-PAS

The effect of PBS on this particle was less prominent than that observed with the fluidMAG-Amine MNPs. There were two size populations observed after DLS with average sizes of 270 nm and 4600 nm. This shows that there is aggregation within the sample. However, this was time dependent, and became more significant over time.

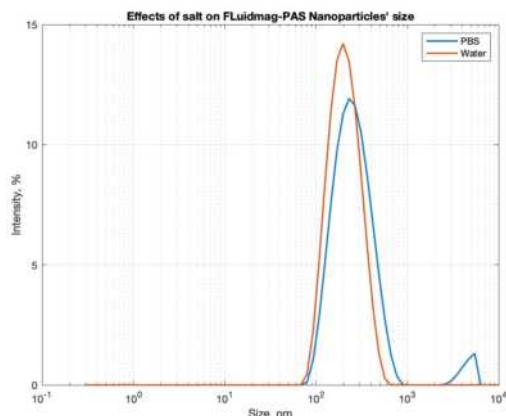


Figure 3: Effect of PBS on fluidMAG-PAS MNPs

4.3 Nanoparticle encapsulation

There was limited initial success observed with the method of MNP encapsulation utilized. The average sizes of the encapsulated MNPs can be seen in table 1 below.

Table 1: Average size of lipid encapsulated MNPs.

MNP Concentration (mg/mL)	Encapsulated Size (nm)
5	2200
2.5	3500
1	420 and 2990
0.5	890

Through imaging via transmission electron microscopy of 2.5 and 5 mg/mL samples, it was possible to confirm that there was some encapsulation taking place, however, the efficiency of encapsulation was notably low. Most of the structures visible upon imaging were either empty vesicles or free nanoparticles with encapsulated nanoparticles being a rarity. From the images obtained, the size of the successfully encapsulated MNPs is within the desired size range. Figure 4 below shows the images obtained using a 2.5 mg/mL PTA-stained sample. Images obtained using other samples are included in supplementary information.

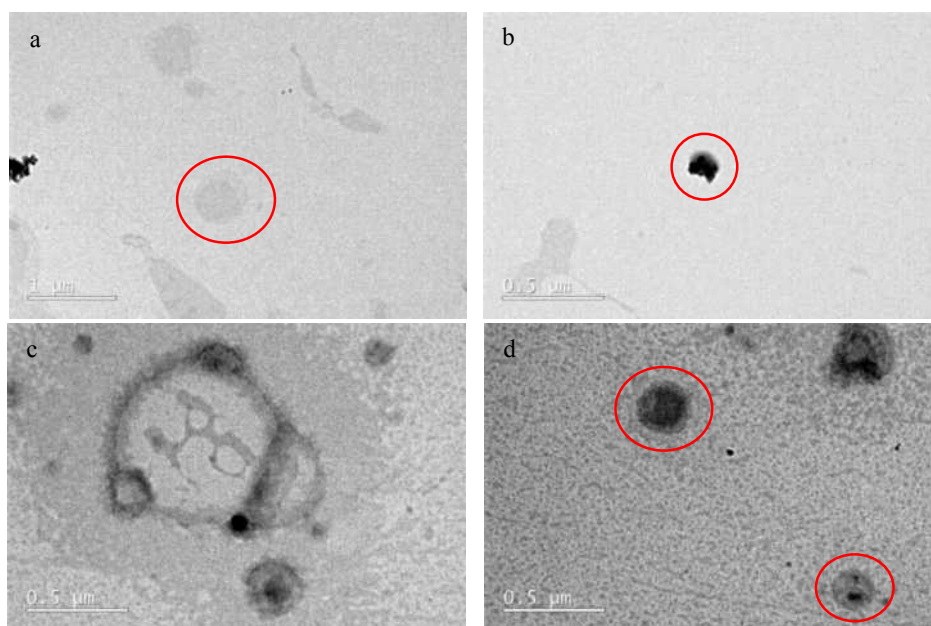


Figure 4: a. Free Liposome, b. free MNP, c. cluster of liposomes, d. MNPs, Encapsulated MNPs

4.4 Magnetoliposome size reduction – Extrusion

The particle size of the magnetoliposomes was measured using DLS and was found to be ≈ 2300 nm. This sample was extruded through a $1\ \mu\text{m}$ extrusion membrane and the obtained particle size was ≈ 870 nm. As the desired size was ≈ 200 nm, the sample was then extruded through a 200 nm extrusion membrane.

After just 4 passes through the extrusion membrane, the characteristic red-brown colour of the MNPs was removed from the sample leaving behind the bright pink colour of the Rhodamine B fluorescent dye. The size of the sample was measured using DLS and there were two size populations with averages sizes of 244 nm and 21 nm. This confirmed that when considered

with the colour of the sample indicated that most of the MNPs had been removed from the sample.

4.5 Magnetoliposome purification

4.5.1 Magnetic separation

The spectra obtained gave a characteristic peak for the MNP fraction at a wavelength of 350 nm (Figure 5). A characteristic peak for the liposome fraction could also be seen at approximately 575 nm.

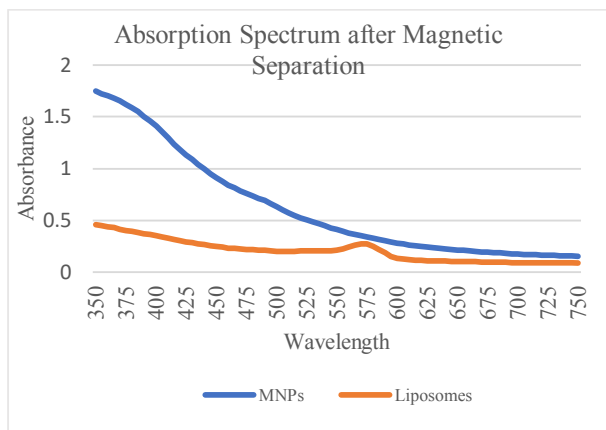


Figure 5: Absorption spectrum obtained after magnetic separation.

4.5.2 Size Exclusion Chromatography

The spectra obtained after conducting UV-Vis spectroscopy on the SEC fractions showed that for the first collected fraction (A1), there were MNPs present but there was no characteristic peak for the liposomes. However, the intensity of the spectrum for both MNPs and liposomes increased through fractions 2-4. Following this, there was a notable decrease in intensity for both components with the final fraction again having no liposomes with MNPs present. The obtained spectrum is presented in figure 6 however, some fractions are excluded to improve the clarity of the illustration. A complete spectrum could be found in the supplementary information provided.

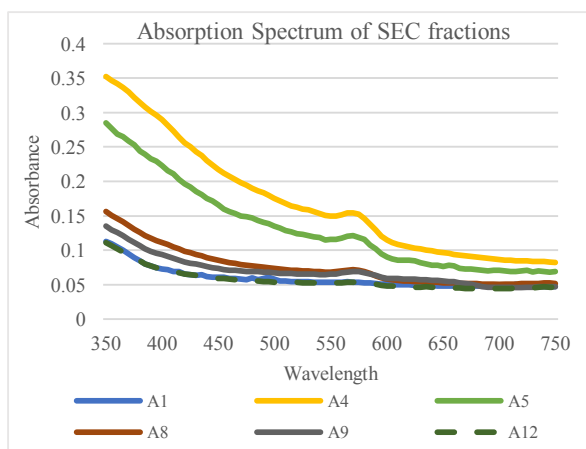


Figure 6: Absorption spectrum obtained for SEC fraction.

4.6 Nanoparticle encapsulation optimization

4.6.1 Low-volume rehydration

DLS analysis of this sample gave an average particle size of 260 nm which suggests that there was successful encapsulation of the MNPs as the size was similar to that obtained from the sample which was visualised using TEM. Based on this positive result, this method of magnetoliposome preparation was deemed suitable.

4.6.2 High-concentration hybrid film

After being left overnight for rehydration, this sample was diluted using 1 mL of DI water. It was then vortexed for 2 minutes and sonicated for 10 minutes however, significant sedimentation was observed. DLS analysis of this sample gave a notably large particle size of approximately 1200 nm. To reduce sedimentation and therefore reduce the particle size, the sample was further diluted, vortexed for 2 minutes and sonicated for 5 minutes however, the size obtained was of the same magnitude. This suggests that this method gives rise to significant and irreversible aggregation and as such, it is not a viable method for magnetoliposome preparation.

4.7 Biohybrid Confirmation

4.7.1 Light Microscopy

For the biohybrid cells formed using nanoparticle encapsulation, it was observed that there was successful biohybrid formation as clear motion of the rod-shaped bacteria was observed after the purification of the sample (removing excess nanoparticles) by centrifugation. This when compared to uncentrifuged sample and the control which displayed a cloud of motion gave a clear indication of successful biohybrid formation. Videos of the observed motion can be found in the supplementary information.

For the cells prepared via carbodiimide facilitated coupling, similar results were obtained for the 85:10:5 and 80:10:10 DOPC: DOTAP: DOPE lipid films. In both cases there was clear and distinct movement of the bacterial cells in response to the applied magnetic field. When compared to the control there was a noticeable difference - in these samples, as with the sample prepared with the first method, there was again a cloud of motion as opposed to clear motion of individual cells. For the 50:50 DOTAP: DOPE lipid film, an anomalous result was obtained as there was far less fluorescence observed in the sample and the control than expected. Barring this, there was still some apparent movement in response to the applied magnetic field.

4.7.2 Optical Density

Measuring the OD of the supernatants obtained after magnetic separation showed that when compared to the prepared controls, the OD of the samples was consistently lower. This result is favourable suggesting that there is less bacteria present in the supernatant of the samples when

compared to their corresponding controls. This points to successful formation of bacterial biohybrid cells when in

the presence of activated MNPs. These results are presented in figure 7 below.

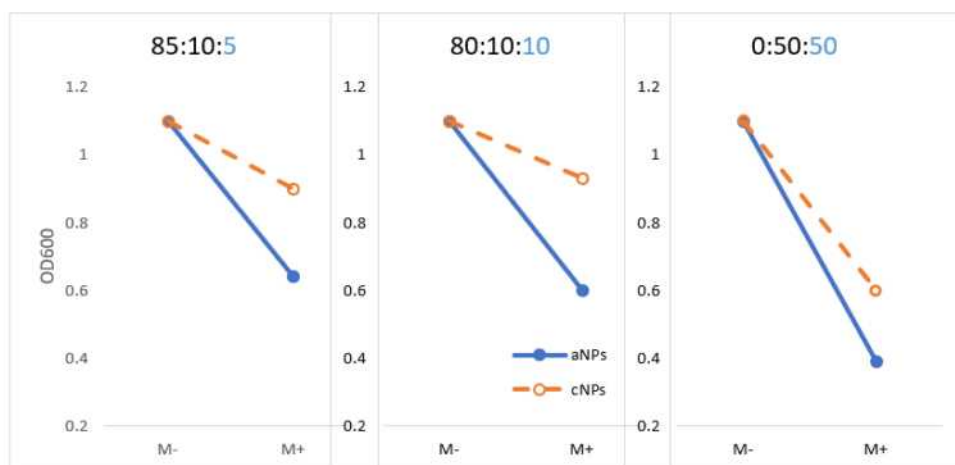


Figure 7: Graph comparing the measured optical densities of the tested supernatants.

5 Discussion

The goal of the project is to form magnetically steerable biohybrids via the attachment of magnetic material to the surface of bacteria. To achieve this, two routes were taken into consideration: nanoparticle encapsulation and carbodiimide facilitated coupling. The first route involved the use of fluidMAG-Amine MNPs. Upon carrying out characterization of these nanoparticles, we observed a time dependent nature of the size measurement – when measured after dilution we saw that the size was larger than expected and this was believed to reflect the aggregation of the nanoparticles. Considering this observation, it was imperative that for each experiment, the dilution of nanoparticles was done immediately prior to use. Characterization of the fluidMAG-Amine nanoparticles also illuminated a zeta potential that, though slightly negative, corresponds to a neutral charge. This, neutrality provides some explanation for the tendency of the nanoparticles to exhibit aggregation as particles with such zeta potentials tend to aggregate faster. Additionally, low zeta potential values (<5 mV and >-5 mV) can lead to agglomeration [13]. However, we hoped this characteristic would prevent any adverse interactions between the MNPs and any other components.

To bind synthetic structures to the surface of the *E. coli* bacteria, due to their gram-negative character [14], the structure must possess a positive charge. Therefore, for fluidMAG-Amine MNPs to achieve binding with the bacteria, they were encapsulated within cationic lipid vesicles. The vesicles used for encapsulation were composed of 90 mol% DOPC and 10 mol% DOTAP, with DOPC being needed to establish the structure of the vesicles and DOTAP being the component that provides the cationic nature needed to facilitate the interaction with the bacteria [15]. Due to the size of the bacteria, a metric

used for evaluating the success of the encapsulation was whether the average size of the encapsulated MNPs produced was ≈ 200 nm. Using the initial method of encapsulation outlined in section 3.5, we found that the size of the encapsulated MNPs was significantly larger than desired – as per the results given in section 4.3. We believe this could be attributed to aggregation of the nanoparticles in the time between dilution to desired concentrations and using these solutions for rehydration of the lipid film. As a result of this observation in addition to the images acquired using transmission electron microscopy, which showed limited encapsulation efficiency, it was deemed necessary that the method of encapsulation be reviewed.

The methods proposed for encapsulation optimization were intended to encourage an increase in the interactions between the MNPs and the lipid films. Furthermore, using stock MNPs for the lipid film rehydration in each case would have eliminated the issue of a time delay between dilution and use of the MNPs which was speculated to be the cause of the aggregation of the MNPs. With limited access to TEM, it wasn't possible to confirm whether the encapsulation efficiency was improved, however, when quantified by size, using the low volume rehydration method resulted in an increase in the quality of the MNP encapsulation. As such, all subsequent MNP encapsulations were done using this method.

With MNP encapsulation completed, the next stage of the process was the interfacing of the bacteria with the produced magnetoliposomes. Using light microscopy, movement of the bacteria in response to an applied magnetic field could be confirmed. Initially, the exact cause of the observed movement was indistinguishable as there were two possible causes – movement due to the formation of biohybrids or due to unbound bacteria being pulled towards the magnet by the drag force created

through the bulk motion of free MNPs. This uncertainty was further enforced by observing that the control samples, which should not have been capable of forming biohybrids, displayed the same response to the presence of a magnetic field. On the other hand, after centrifuging it was observed that the motion still occurred even when it was apparent that the number of free MNPs present in solution had been reduced. Since it was not possible to remove all the free MNPs from the solution, it may be fair to consider the motion observed as being a combination of biohybrids being formed as well as bacterial cells being dragged along by the motion of the free MNPs.

The second method of biohybrid formation that was attempted involved the use of fluidMAG-PAS nanoparticles. Upon conducting characterisation of these MNPs, a similar time dependence regarding the particle size was observed - which indicated that, much like the fluidMAG-Amine MNPs, any dilutions must be done immediately before use. From the zeta potential measurement, it was observed that these MNPs were highly anionic. With a zeta potential value of -32.01 mV, the solution is not expected to aggregate as solutions with zeta potential values < -30 mV or > 30 mV are considered stable [16]. In this case, the MNPs would be likely to repel each other and aggregation to the extent observed with the fluidMAG-Amine nanoparticles would not be expected. However, as these MNPs have a density of 1.25 g/cm^3 [18] which is slightly larger than that of the dispersant (water) they will eventually begin to settle therefore leading to aggregation [16].

The activation of these nanoparticles involved the incorporation of a carbodiimide group, leading to the formation of a highly reactive intermediate O-acylisourea group on the surface of the MNP [18]. Once activated, the MNPs can now very thermodynamically favourable reactions with amine groups. As such, in this method, the bacteria was encapsulated as opposed to the nanoparticles. The vesicles used for encapsulation of the bacteria were composed of DOPC, necessary for vesicle formation, DOTAP, to provide the cationic charge needed for interaction between the vesicles and the bacteria, and DOPE, which provided the amine group needed for interaction between the encapsulated bacteria and the MNPs.

Without imaging via electron microscopy or any other super-resolution microscopy, prior to interfacing with the MNPs it is difficult to evaluate the degree of success of the encapsulation. It was possible to infer success, however, through light microscopy, as it was observed that bacteria responded to the presence of the magnetic field. From visual observation, the characteristics and behaviour of bacteria encapsulated in vesicles with the composition 85:10:5 DOPC: DOTAP: DOPE and bacteria encapsulated in vesicles with the composition 80:10:10 DOPC: DOTAP: DOPE appeared to be indistinguishable. Bacteria encapsulated in the fusogenic, 50:50 DOTAP: DOPE vesicles, however, exhibited clearly different behaviour from the other compositions. Specifically, it was clear that although the

bacteria were present and responded to the presence of a magnetic field, they were no longer GFP expressing. Within the limits of this project, it was difficult to infer with 100% certainty the reason for this occurrence. However, we speculate that the fusogenic nature of the vesicles resulted in the destabilization of the outer membrane of the bacteria which resulted in the formation of a pore. This either led to the leaking of GFP from the bacteria, or the dilution of GFP within the bacteria due to the release of the contents of the vesicle (deionised water) into the cell upon fusion [15].

On the other hand, it was possible to observe a correlation between the OD measurements and the concentration of DOPE present in each vesicle composition. This observation further helps to confirm that the encapsulation was a success. Additionally, it was observed that with greater concentrations of DOPE, the OD was lower. This implied that more biohybrids were formed and upon the introduction of the magnetic field, more bacteria were removed from the supernatant. Based on this, it can be inferred that higher DOPE - though potentially detrimental to the survival of the bacteria due to destabilization of the membrane - results in a greater efficiency of biohybrid formation. This is expected as DOPE is the lipid that provides the amine functional group that is necessary for the interfacing of the bacteria and the MNPs to take place.

6 Conclusion

It was determined that low volume rehydration allowed for the best control of the size of encapsulated nanoparticles when measured using DLS. This gives a feasible approach to MNP encapsulation and provides a basis for future encapsulation protocols.

Through light microscopy and OD measurements, it was confirmed that the employed methods lead to the formation of bacterial biohybrids which are capable of being guided by a magnetic field. However, the efficiency of formation could not be determined using these methods and further confirmation should be obtained using a viable form of super-resolution microscopy such as TEM.

When encapsulating bacteria with lipids, it was found that, although increasing the percentage of DOPE present in the lipid film led to increased interaction between the encapsulated bacteria and the MNPs, high concentrations of DOPE have potentially detrimental effects on the bacteria. As such, it is important to determine the optimal liposome composition for encapsulating bacterial cells.

In conclusion, the methods employed in this study for forming bacterial biohybrids were successful in producing magnetically steerable bacterial cells. Further work is needed to confirm the efficiency of biohybrid formation and confirm that motion in the samples is due to the effect of the magnetic field on the biohybrid cells as opposed to bulk motion caused by the movement of the

MNPs. Following this, the behaviour of the bacteria when guided through a more complex system should be studied.

7 Acknowledgements

We would like to thank Juan Ivars Miñana for his guidance and support throughout the duration of this research. We would also like to thank Yeyang Sun for supporting us to carry out TEM. We also wish to extend our gratitude to the PhD students in the Membrane Biophysics and Elani Groups of Imperial College London for supporting us and providing us with all necessary training before using the laboratory.

8 References

- [1] M. B. Akolpoglu *et al.*, “Magnetically steerable bacterial microrobots moving in 3D biological matrices for stimuli-responsive cargo delivery,” *Science Advances*, vol. 8, no. 28, Jul. 2022, doi: <https://doi.org/10.1126/sciadv.abo6163>.
- [2] F. Wang *et al.*, “Magnetically targeted photothermal cancer therapy in vivo with bacterial magnetic nanoparticles,” *Colloids and Surfaces B: Biointerfaces*, vol. 172, pp. 308–314, Dec. 2018, doi: <https://doi.org/10.1016/j.colsurfb.2018.08.051>.
- [3] C. Chen, P. Wang, and L. Li, “Applications of Bacterial Magnetic Nanoparticles in Nanobiotechnology,” *Journal of Nanoscience and Nanotechnology*, vol. 16, no. 3, pp. 2164–2171, Mar. 2016, doi: <https://doi.org/10.1166/jnn.2016.10954>.
- [4] B.-W. Park, J. Zhuang, O. Yasa, and M. Sitti, “Multifunctional Bacteria-Driven Microswimmers for Targeted Active Drug Delivery,” *ACS Nano*, vol. 11, no. 9, pp. 8910–8923, Sep. 2017, doi: <https://doi.org/10.1021/acsnano.7b03207>.
- [5] C. Piñero-Lambea, D. Ruano-Gallego, and L. Á. Fernández, “Engineered bacteria as therapeutic agents,” *Current Opinion in Biotechnology*, vol. 35, pp. 94–102, Dec. 2015, doi: <https://doi.org/10.1016/j.copbio.2015.05.004>.
- [6] M. R. Benoit *et al.*, “Visualizing implanted tumors in mice with magnetic resonance imaging using magnetotactic bacteria,” *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, vol. 15, no. 16, pp. 5170–5177, Aug. 2009, doi: <https://doi.org/10.1158/1078-0432.CCR-08-3206>.
- [7] S. Taherkhani, M. Mohammadi, J. Daoud, S. Martel, and M. Tabrizian, “Covalent Binding of Nanoliposomes to the Surface of Magnetotactic Bacteria for the Synthesis of Self-Propelled Therapeutic Agents,” *ACS Nano*, vol. 8, no. 5, pp. 5049–5060, Apr. 2014, doi: <https://doi.org/10.1021/nn5011304>.
- [8] O. Felfoul *et al.*, “Magneto-aerotactic bacteria deliver drug-containing nanoliposomes to tumour hypoxic regions,” *Nature nanotechnology*, vol. 11, no. 11, pp. 941–947, Nov. 2016, doi: <https://doi.org/10.1038/nnano.2016.137>.
- [9] D. T. Riglar and P. A. Silver, “Engineering bacteria for diagnostic and therapeutic applications,” *Nature Reviews Microbiology*, vol. 16, no. 4, pp. 214–225, Feb. 2018, doi: <https://doi.org/10.1038/nrmicro.2017.172>.
- [10] Z. Cao, X. Wang, Y. Pang, S. Cheng, and J. Liu, “Biointerfacial self-assembly generates lipid membrane coated bacteria for enhanced oral delivery and treatment,” *Nature Communications*, vol. 10, no. 1, Dec. 2019, doi: <https://doi.org/10.1038/s41467-019-13727-9>.
- [11] R. Sabaté, R. Barnadas-Rodríguez, J. Callejas-Fernández, R. Hidalgo-Álvarez, and J. Estelrich, “Preparation and characterization of extruded magnetoliposomes,” *International Journal of Pharmaceutics*, vol. 347, no. 1–2, pp. 156–162, Jan. 2008, doi: <https://doi.org/10.1016/j.ijpharm.2007.06.047>.
- [12] Chemicell, “Covalent Coupling Procedure on fluidMAG-PAS by Carbodiimide Method.” Accessed: Nov. 22, 2023. [Online]. Available: <http://www.chemicell.com/products/protocols/docs/fluidMAG-PAS.pdf>
- [13] M. Gumustas, C. T. Sengel-Turk, A. Gumustas, S. A. Ozkan, and B. Uslu, “Chapter 5 - Effect of Polymer-Based Nanoparticles on the Assay of Antimicrobial Drug Delivery Systems,” *ScienceDirect*, Jan. 01, 2017. <https://www.sciencedirect.com/science/article/abs/pii/B9780323527255000058>
- [14] J. Y. Lim, J. W. Yoon, and C. J. Hovde, “A Brief Overview of Escherichia coli O157:H7 and Its Plasmid O157,” *Journal of microbiology and biotechnology*, vol. 20, no. 1, pp. 5–14, Jan. 2010, Available: [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3645889/#:~:text=Escherichia%20coli%20\(E](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3645889/#:~:text=Escherichia%20coli%20(E)
- [15] A. Scheeder, M. Brockhoff, E. N. Ward, G. S. Kaminski Schierle, I. Mela, and C. F. Kaminski, “Molecular mechanisms of liposome interactions with bacterial envelopes,” *Research Gate*, Oct. 2023, Accessed: Dec. 14, 2023. [Online]. Available: https://www.researchgate.net/publication/374611119_Molecular_mechanisms_of_liposome_interactions_with_bacterial_envelopes
- [16] Malven Instruments Wolrdwide, “Zeta Potential - An introduction in 30 minutes.” Available: <https://www.research.colostate.edu/wp-content/uploads/2018/11/ZetaPotential-Introduction-in-30min-Malvern.pdf>
- [17] Chemicell, “Product Information - fluidMAG-Amine,” *Chemicell*. http://www.chemicell.com/products/nanoparticles/docs/P_I_fluidMAG-Amine_4121.pdf (accessed Oct. 19, 2023).
- [18] Chemicell, “Product Information - fluidMAG-PAS,” *Chemicell*. http://www.chemicell.com/products/nanoparticles/docs/P_I_fluidMAG-PAS_4110.pdf (accessed Nov. 22, 2023).

Data driven modelling using time series recurrent neural networks (RNN) for glycosylation prediction in mAbs

Final year research project, Group 79: Minqi Wang and Putian Yao
Department of Chemical Engineering, Imperial College London, U.K.

Abstract

N-linked glycosylation is a critical post-translational process that greatly affects the efficacy and efficiency of monoclonal antibodies (mAbs). Previous researchers have published a hybrid modeling system combining a mechanistic kinetic model and a static Artificial Neural Network (ANN), to predict the glycan distribution through intracellular states. However, as glycosylation is a sequential process, Recurrent Neural Network (RNN) seems more suitable as past information is also taken into account in future prediction. Hence, this project focuses on replacing the ANN model with an RNN to predict glycan distribution through time-series nucleotide sugar donor (NSD) concentrations in Chinese hamster ovary (CHO) cells. Two RNN units were connected, forming a new RNN-in-series system. The first unit was responsible for predicting time-series NSD concentration, the result of which was provided as the input of the second unit to predict glycan distribution. The model was trained and validated by the *Keras tuner* library, with Long Short-Term Memory (LSTM) selected in specific for RNN. The model accurately predicts the glycan distribution which closely followed the trend of the test data, with an average absolute error of 1.19%. The RNN-in-series model's success offers a new depth of predicting in the dynamic glycosylation process, with significant benefits for optimizing mAb bioprocessing feeding strategies, and hence meeting the growing demand for high-quality therapeutic proteins.

Keywords: N-linked Glycosylation, monoclonal antibodies (mAbs), Nucleotide sugar donor (NSD), Artificial Neural Network (ANN), Recurrent Neural Network (RNN), Deep-learning, Machine learning, Multiscale, Dynamic modelling

1. Introduction

Over the past few years, monoclonal antibodies (mAbs) are the most successful type of drug used in the biopharmaceutical industry (Kontoravdi & Jimenez del Val, 2018). They can be used to treat a large variety of diseases, such as cancer, autoimmune and infectious diseases. Glycosylation in the constant fragment, as shown in Figure 1, is the key product quality of mAbs. The presence of terminal galactose residues on mAb Fc glycans plays an important role in complement-dependent cytotoxicity (CDC) and antibody-dependent cell-mediated cytotoxicity (ADCC) functions (Kotidis et al., 2019). High mannose N-glycans produced can also influence immunogenicity and biological activity and stability (Buettner et al., 2018). As its critical role, achieving optimal glycosylation level is of great interest to the biopharmaceutical industry.

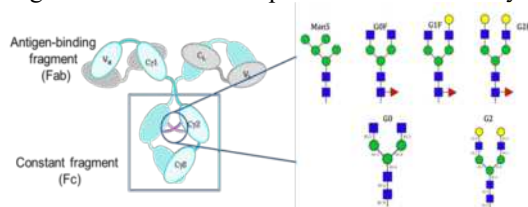


Figure 1: mAbs structure and 6 major N-linked glycans (del Val et al., 2012)

Glycosylation is a post-translational modification process, happening in the endoplasmic reticulum (ER) and Golgi apparatus within cells. Instead of template-driven, glycans produced are the result of a complex network of metabolic and enzymatic reactions influenced by a variety of factors, including nucleotide sugar donors (NSDs) (Harnish Mukesh Naik, Majewska & Betenbaugh,

2018) and the extracellular environment (Hossler, Khattak & Zheng Jian Li, 2009). Figure 2 shows how the glycosylation profile links to the extracellular metabolites through NSDs.

Considering that biological experiments are expensive and time-consuming, researchers in the past have built a mechanistic kinetic model and succeeded in predicting the experimental glycosylation level (Kotidis et al., 2019) with many kinetic parameters. However, background information and knowledge of computational tools and bioprocess are required. To tackle this problem, researchers proposed the use of machine learning to describe the glycosylation process, which requires less bioprocess knowledge (Lancashire, Christophe Lemetre & Ball, 2008). A static artificial neural network (ANN) was proposed to accurately predict glycan distributions in the products with an average absolute error of 1.1% (Pavlos Kotidis & Kontoravdi, 2020), but with no dynamic information provided.

Given that glycosylation is defined as a sequential process, it is critical to consider historical data for accurate predicting of future glycan profiles. Therefore, the research in this project aims at advancing the existing model by replacing the static ANN with a Recurrent Neural Network (RNN). By considering previous metabolites, NSD levels in Golgi, and previous glycan data, the RNN can predict the next day or next few days of glycans. Following this improvement, the research may be beneficial in designing an optimal feeding strategy that aims to maximise the final glycosylation contents through dynamic optimisation.

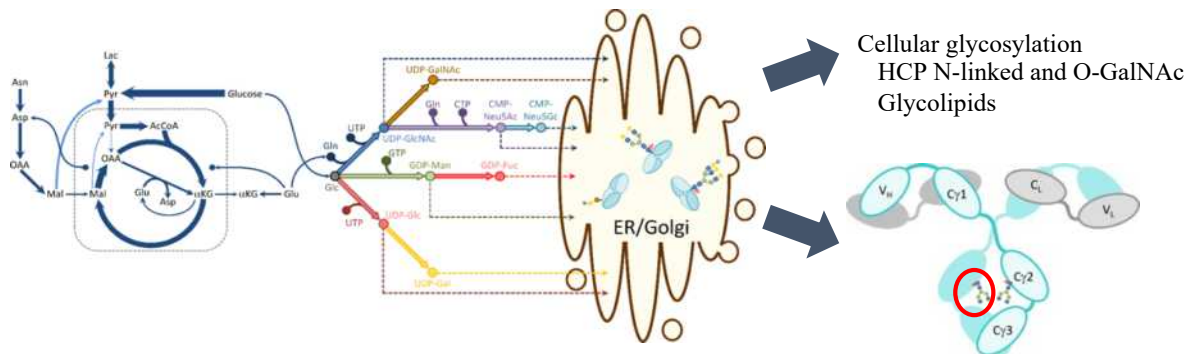


Figure 2: The relationship between extracellular metabolites, NSDs in Golgi and the glycans produced (del Val et al., 2012)

2. Methodology

2.1 System description

The biological system involved in this project is Chinese Hamster Ovary (CHO) cells producing an IgG antibody. The cell culture is fed with glucose and amino acid nutrients every 2 days (Kotidis et al., 2019). To improve product quality, i.e. glycosylation, additional galactose and uridine are fed on day 4, day 6, day 8, and day 10. They are the metabolic precursors needed for uridine diphosphate galactose (UDP-Gal) synthesis, which is the specific type of NSD required for galactosylation (Grainger & James, 2013).

2.2 Machine learning (ANN) model

Data used for machine learning was generated using the mechanistic kinetic model proposed by Kotidis et al. (2019). 10 groups of data were simulated, each containing extracellular metabolites concentrations, NSDs concentrations, and glycan distributions from Day 0 to Day 12.

For the ANN model, the input variables containing 7 types of NSDs, namely CMPNeu5Ac, UDPGal, UDPGlc, UDPGalNAc, UDPGlcNAc, GDPMan, GDPFuc, while the output are the 6 major N-linked glycans: Man5, G0, G0F, G1F, G2, G2F. The structure of ANN is shown below in Figure 3.

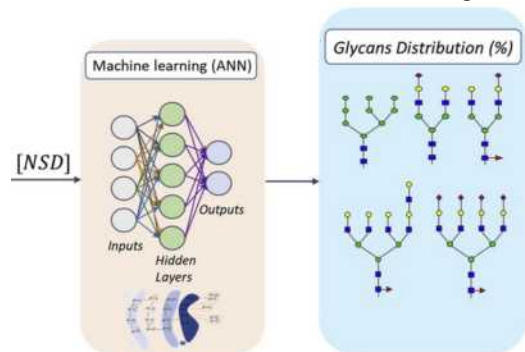


Figure 3: Schematic representation of ANN used for predicting N-linked glycosylation using NSDs from (Pavlos Kotidis & Kontoravdi, 2020)

The whole dataset was split into three portions. 70% of the data was used to train the model, while 23% was used for validation and the remaining 7% for testing of the model. The ANN model was

trained and validated by the *Keras tuner* library, which determines the number of hidden layers and the number of neurons inside each hidden layer. The best model was selected based on minimum average absolute error while ensuring the total parameters used were within acceptable range.

When using Keras in determining the optimal model for the neural network, usually challenges arise due to noise and issues related to the size of the dataset. As shown in Figure 4, the validation error tends to increase after a point, whereas the training error continues to decrease. This phenomenon is known as overfitting. An approach called ‘early-stopping’ was implemented, which is suggested as one of the most effective solutions to prevent overfitting (Xue, 2019).

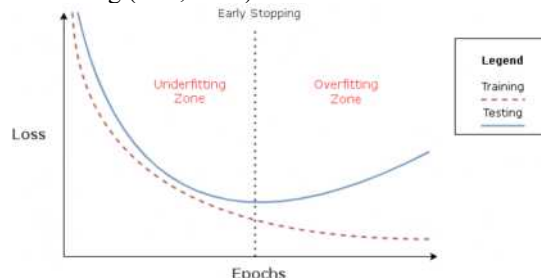


Figure 4: Training example illustrating the underfitting and overfitting regions and the optimal epoch to stop (Igarata, 2021).

2.3 Recurrent neural network model

Recurrent neural network (RNN) is a type of neural network used to process sequential data. By replacing ANN with RNN models, the glycosylation results predicted are presented in the time series.

Various architectural forms can be employed in recurrent neural networks, such as gated recurrent units, bi-directional RNNs, long short-term memory (LSTM), etc. For a better performance, LSTM is selected for the RNN models due to its advantage in addressing the challenges of gradient vanishing or exploding during training (Trupti Katte-Bangayya, 2018). Additionally, compared to standard RNN, LSTMs have better performance on longer sequences as it has long-term memory ability to store past data and proceed with fresh input and working memory. Basically, the repeating block in standard RNN consists of a very simple structure,

like a single activation function layer. The LSTM has a much more complicated structure within the block (Olah, 2015), which is shown in Figure 5.

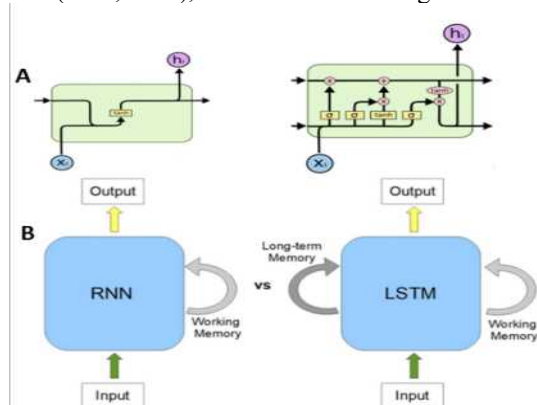


Figure 5: A.) The core idea of Standard RNN (left) and LSTM (right) (Olah, 2015). B.) compare RNN and LSTM (Robail Yasrab & Pound, 2020)

2.4 RNN-in-series system

To predict glycans through time-series NSDs, an RNN-in-series system was implemented. The system consists of two distinct RNN units (RNN unit 1 and RNN unit 2), each specialized in handling a different aspect of the glycosylation prediction process. As shown in Figure 6, two RNN units were connected. The first one is to predict NSD in time series and the second one is responsible for predicting the glycan distributions using the prediction results from the previous unit.



Figure 6: RNN-in-series system

The details of the two units are shown in Figure 7. The RNN unit 1 uses the NSD data from day 0 to day 11 as the input, while the output is from day 1 to day 12. The daily NSD concentration is predicted using concentration from the previous day. The output of RNN unit 1 is stored and serves as the input for the second RNN unit. Glycan information memorized and passed down through hidden states are used together with the inputs to predict the glycan distribution outputs on Days 1 to 12.

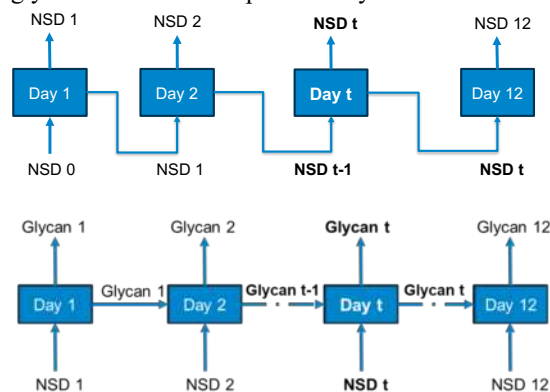


Figure 7: Detailed scheme of RNN unit 1 (above) and RNN unit 2 (below)

3. Results and discussion

3.1 ANN & RNN comparison

The primary comparison consisted of evaluating the prediction result of the Artificial Neural Network (ANN) with that of RNN Unit 2, both of which are implemented for predicting glycan distribution profile from Day 0 to Day 12 using NSD data.

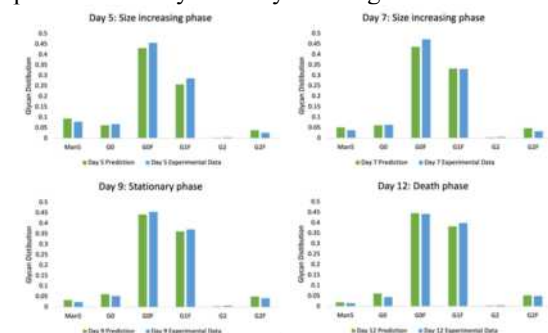


Figure 8: ANN prediction results in different phases

To ensure the robustness and predictability of machine learning, the ANN was reconstructed based on the architectural principles presented in the previous paper (Pavlos Kotidis & Kontoravdi, 2020). In the 12-day period of the biology experiment, CHO cell culture can be divided into three major different culture phases: an exponential growth phase, a stationary phase, and a death phase with the two transition days being day 8 and day 10. The exponential phase is divided into the number increasing phase (the number of cells increases exponentially from day 0 to 4), and size increasing phase (Pan et al., 2017). Considering this and feeding strategies, 4 days after number increasing phase (day 5, day 7, day 9, and day 12) were chosen for ANN model results presenting, to evaluate ANN's predictive performance at each phase after feeding. The prediction results for these days are compared with experimental data, as shown in Figure 8. The average absolute errors are calculated in Table 1. As clearly displayed, the results are on independent days and not connected to previous results. This is indicated by the structure of the ANN. As characterized by its direct feedforward topology, ANN is adept at processing static inputs to generate only present outputs.

Table 1: Error analysis of ANN on selective days

ANN	Average Absolute error
Day 5 (Size increasing phase)	1.48%
Day 7 (Size increasing phase)	1.17%
Day 9 (Stationary Phase)	0.85%
Day 12 (Death phase)	0.76%
Overall	0.97%

Table 2: Error analysis for RNN unit 2 for each glycan

Glycans	Average absolute error
Man5	0.74%
G0	0.54%
G0F	1.34%
G1F	1.26%
G2	0.03%
G2F	0.42%
Overall	0.72%

In contrast, RNN Unit 2 is adept at temporal data integration, which allows it to predict glycan distributions over the entire experimental period (From day 0 to day 12). Glycan distribution predictions are not only based on the current NSD values but also enriched by the historical glycan information provided by the hidden states. The results are shown in Figure 9. Despite some minor differences, RNN can predict the overall trend of glycan distributions accurately.

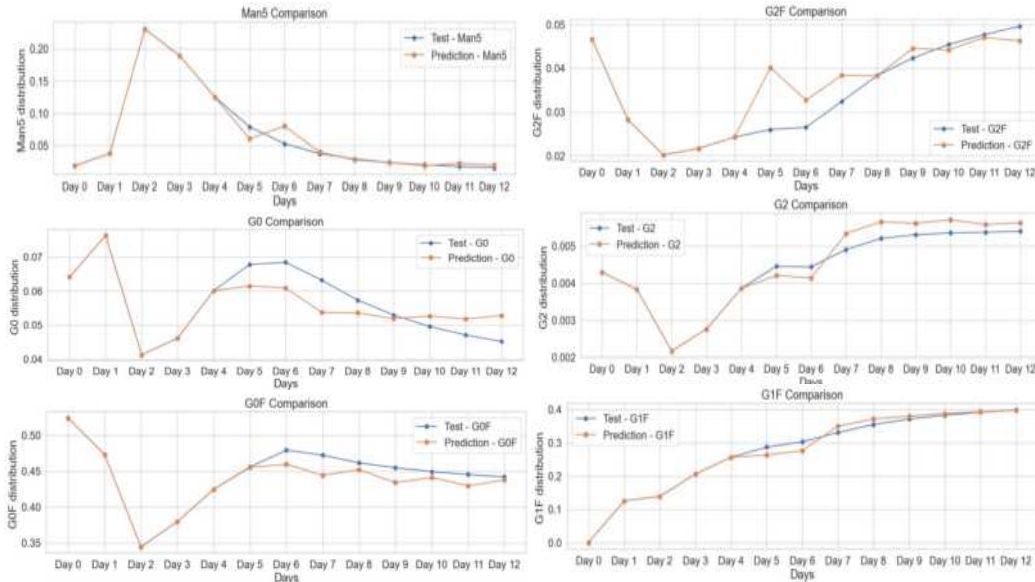


Figure 9: Example RNN results for glycans distribution.

ANN summary

Layer (type)	Output Shape	Param #	Layer (type)	Output Shape	Param #
dense_3 (Dense)	(None, 10)	90	lstm_4 (LSTM)	(None, 8, 3)	60
dense_4 (Dense)	(None, 2)	22	lstm_5 (LSTM)	(None, 4)	128
dense_5 (Dense)	(None, 7)	21	dense_2 (Dense)	(None, 7)	35
Total params: 133 (532.00 Byte)			Total params: 223 (892.00 Byte)		
Trainable params: 133 (532.00 Byte)			Trainable params: 223 (892.00 Byte)		
Non-trainable params: 0 (0.00 Byte)			Non-trainable params: 0 (0.00 Byte)		

Figure 10: Hyperparameters Comparison for ANN model and RNN-unit-2

RNN summary

The mean absolute error of RNN unit 2 for each glycan was calculated and shown in Table 2. Ranging from 0.03% for G2 to 1.34% for G0F, the overall average absolute error is 0.72%, which is lower than that of ANN (0.97%). Beyond the error analysis, it is essential to recognize the proficiency of RNNs in time-series modelling. This attribute is particularly beneficial for dynamic glycosylation prediction tasks where temporal dependencies are crucial. Consequently, the RNN is identified as having a distinctly superior performance in modelling glycosylation dynamics.

The hyperparameters summary in Figure 10 shows that RNN has a higher number of parameters. This is justified by the complexity and richness of the datasets used for RNN training. The expansive parameter set of the RNN considers the breadth of data points and incorporates the necessary recurrent connections that are critical for capturing the temporal sequences inherent in the glycosylation process.

3.2 RNN Unit 1 prediction

RNN Unit 1 is responsible for predicting time series NSD concentrations. The prediction result is compared with experimental data (test data) as shown in Figure 11. Levels of metabolites and

several important indicators relative to the glycosylation process are also predicted. These include, but are not limited to, glucose and ammonia concentrations, because these two are important nutrients needed for cell growth. The viable cell density (Xv) and concentration of monoclonal

antibodies (mAb) are also included as they indicate the product quantity. The results are shown in Figure 12.

In general, the NSD predictions align well with the trend observed in the experimental test data, indicating a good foundation for the glycosylation prediction by RNN unit 1 in the next step. However, gaps between the two lines are also significant. The percentage error of RNN unit 1 for each NSD was calculated and shown in Table 3. Even though the overall error is within an acceptable range (3.70%), the average percentage error for each NSD varies a lot, ranging from 0.33% for GDPFuc to 9.67% for UDPGal. This large difference in error implies further optimization of the RNN model. The complexity of the NSD synthesis is also evidenced by the fact that the error varies so much.

For this RNN unit 1, the hyperparameters summary is also displayed in Figure 13. RNN unit 1 has 547 parameters, which is more than twice as much as RNN unit 2. This outcome is anticipated considering there are nearly 2000 data points used for RNN unit 1 training, whereas only around 900 data points are involved in unit 2.

Table 3: Error analysis for RNN unit 1 for each NSD

NSD	Average percentage error
UDPGal	9.67%
UDPGlc	2.06%
UDPGlcNAc	5.91%
GDPMan	0.55%
GDPFuc	0.33%
Overall	3.70%

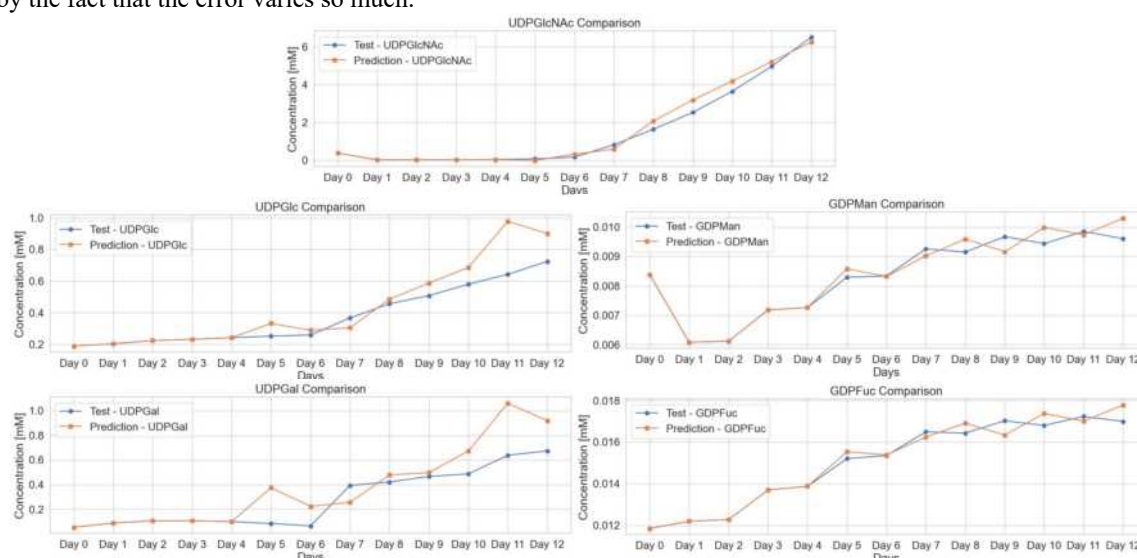


Figure 11: Significant NSD prediction

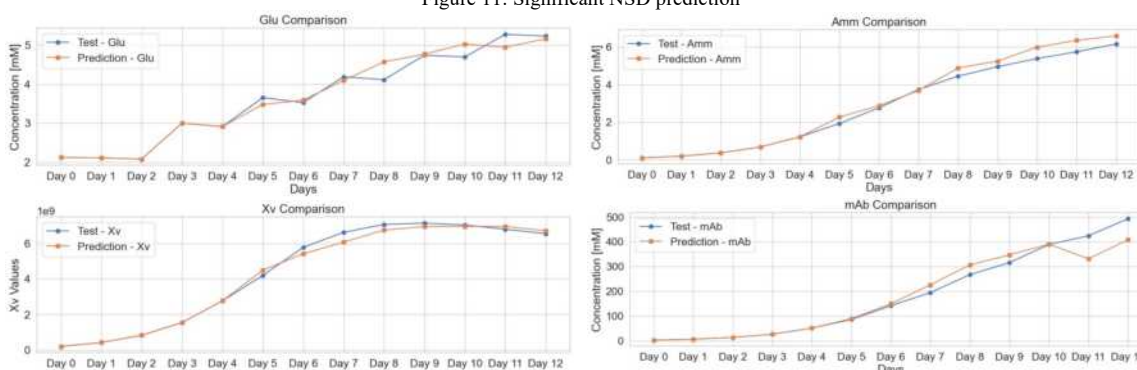


Figure 12: Example results for important parameters and metabolites.

Layer (type)	Output Shape	Param #
lstm_4 (LSTM)	(None, 17, 5)	140
lstm_5 (LSTM)	(None, 6)	288
dense_2 (Dense)	(None, 17)	119
Total params: 547 (2.14 KB)		
Trainable params: 547 (2.14 KB)		
Non-trainable params: 0 (0.00 Byte)		

Figure 13: RNN unit 1 Hyperparameters Summary

3.3 RNN-in-series performance

As mentioned before, in the RNN-in-series model, the predicted NSD output from Unit 1 is stored and fed continuously into RNN Unit 2 as input for the glycosylation predictions. The result of this system is compared with the result from RNN unit 2 only, as presented in Figure 14. Analyzing these two sets of graphs, it is evidenced that the predictions are in

good agreement with the test data, reflecting high accuracy in predicting the glycan distribution and reflecting the fact that the RNN-in-series model performs well across a wide range of variables and conditions.

By comparing the glycosylation predictions side-by-side between the results of RNN unit 2 alone and RNN-in-series, it is worth noticing that the gap between prediction results and test data becomes larger. The error analysis is carried out, with the average absolute error for each glycan calculated in Table 4. Ranging from 0.09% for G2 to 2.65% for

G0F, the overall error is 1.19%, which is indeed higher than that of RNN unit 2 alone (0.72%). The reason for this phenomenon is error accumulation (error propagation) from RNN unit 1 to RNN unit 2. According to the error analysis, where errors in Unit 1's results are carried over to Unit 2, impacts the precision of Unit 2's glycosylation predictions. The increase in final glycosylation prediction error indicated the need for more comprehensive optimization for the model to improve the predictive accuracy for future glycosylation.

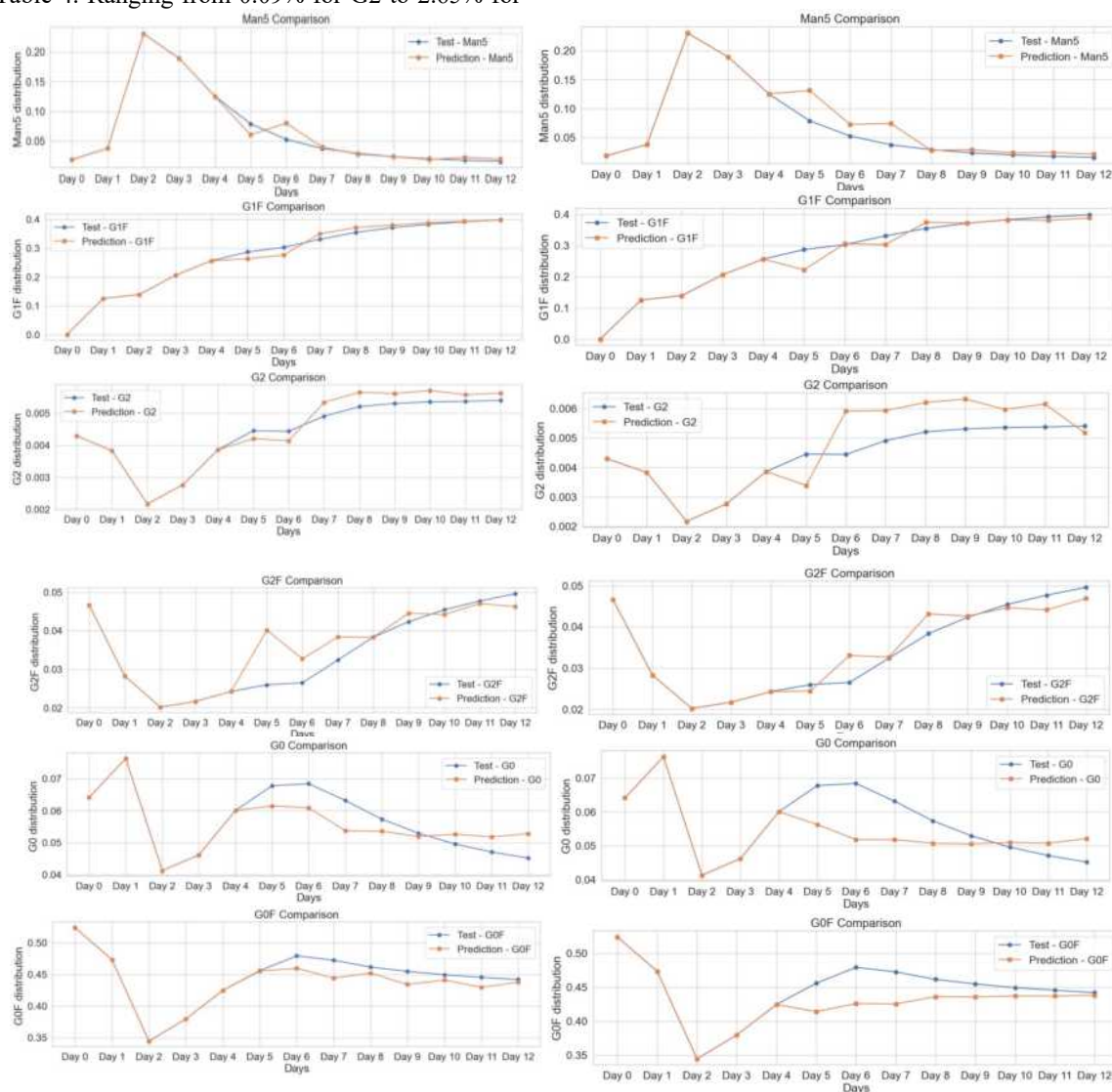


Figure 14: Glycosylation prediction from RNN unit 2 only (left) and RNN-in-series (Right)

Table 4: Error analysis for RNN-in-series system

Glycan	Average absolute error
Man5	1.64%
G0	0.75%
G0F	2.65%
G1F	1.73%
G2	0.09%
G2F	0.26%
Overall	1.19%

4. Outlook

By comparing the RNN-in-series predictions against the experimental data, it is agreed that the system can predict the overall trend of glycosylation. However, the model could be improved to reduce the error propagation issue. Referring to the approach suggested by (Liu, Nathan & Brunton, 2020), the error propagation in the RNN models

could be handled by applying a multiscale system. As shown in Figure 15, The principle behind a multiscale system is that instead of training one neural network with one time step, more neural networks with different time steps could be trained and combined. The smaller the time step, the more accurate the local prediction is. Combining models with different time steps is expected to predict both overall trends and more precise local points.

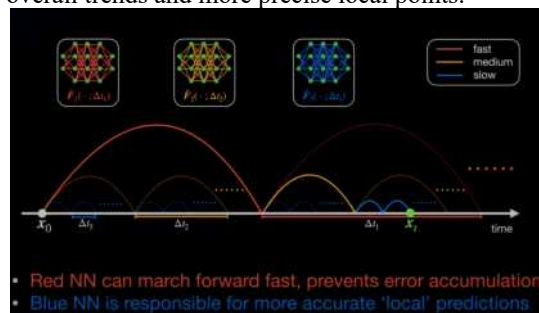


Figure 15: Basic idea about how multiscale works (Brunton, 2021)

In addition, it has been verified that unit 1 of this RNN-in-series model is also able to predict metabolites and other important indicators of glycosylation with high accuracy. Those predictions could be used in the future. For example, viable cell density is highly dependent on specific culture conditions for the glycosylation process and is an indicator of overall health and growth (Jong Hyun Nam et al., 2008). Maintaining an optimal viable cell density ensures that cells are in proper physiological state for efficient glycosylation. Apart from that, the concentration of metabolites is one of the few possible steps that could be controlled by human beings in the glycosylation process. Therefore, these concentrations can be monitored and predicted to optimize feeding strategies in the near future, thereby influencing the overall glycosylation process. The ability to predict and regulate glycosylation contributes to the efficiency and cost-effectiveness of biomanufacturing.

5. Conclusion

In conclusion, due to the importance of glycosylation in influencing the efficacy and efficiency of monoclonal antibodies, it is important to predict the glycan profile in mAbs. The ability to trace glycosylation trajectory is critical for understanding and optimizing the bioprocessing conditions. This project aims at predicting glycan levels through time series NSDs. Recurrent neural network, which requires minimal biological knowledge and specializes in handling sequential data, is applied. Two RNN units are implemented and connected, forming a novel RNN-in-series system, with the first one to predict time-series NSD levels and the second one to predict glycans profile. Experimental data containing NSD concentrations and glycan distributions in CHO cells are provided for training and testing the data-driven model. Keras library is used to tune the parameters, with the

LSTM method selected specifically for the RNN units. The glycan predictions of the system follow the experimental trend very well, with an average absolute error of 1.19%. The significance of this research is not limited to the replacement of ANNs with RNNs in time-series glycosylation, but also to the potential of integrating dynamic modelling approaches into various aspects of bioprocessing and biomanufacturing. In the future, the model will be improved and applied to relevant industries by applying multi-scale algorithms for careful optimisation and manual control of metabolic levels. For the pharmaceutical industry in particular, this dynamic prediction approach will help meet the growing demand for high-quality therapeutic proteins.

References

- A. Sherstinsky (2020) Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. *Physica D: Nonlinear Phenomena*. 404, 132306–132306. doi:<https://doi.org/10.1016/j.physd.2019.132306>.
- Bhat, H. (2023) *Recurrent Neural Network: Applications and Advancements*. August 2023. AlmaBetter. <https://www.almabetter.com/bytes/articles/recurrent-neural-network> [Accessed: 13 December 2023].
- Brunton, S. (2021) Deep Learning of Hierarchical Multiscale Differential Equation Time Steppers. *YouTube*. <https://www.youtube.com/watch?v=JfI3dIISTrU>.
- Buettner, M.J., Shah, S.R., Saeui, C.T., Ariss, R. & Yarema, K.J. (2018) Improving Immunotherapy Through Glycodesign. *Frontiers in Immunology*. 9. doi:<https://doi.org/10.3389/fimmu.2018.02485>.
- Coral Fung Shek, Pavlos Kotidis & Betenbaugh, M.J. (2021) Mechanistic and data-driven modeling of protein glycosylation. *Current Opinion in Chemical Engineering*. 32, 100690–100690. doi:<https://doi.org/10.1016/j.coche.2021.100690>.
- del Val, I.J., Jedrzejewski, P.M., Exley, K., Sou, S.N., Kyriakopoulos, S., Polizzi, K.M. & Kontoravdi, C. (2012) Application of Quality by Design Paradigm to the Manufacture of Protein Therapeutics. In: *Glycosylation*. <https://www.intechopen.com/chapters/39460>.
- Ginestoux, C. (2015) *IMGT Lexique*. 2015. <https://www.imgt.org/IMGTeducation/IMGTlexique/G/Glycosylation.html> [Accessed: 30 November 2023].
- Gordan Lauc, Huffman, J.E., Maja Pučić, Zgaga, L.,

- Adamczyk, B., et al. (2013) Loci Associated with N-Glycosylation of Human Immunoglobulin G Show Pleiotropy with Autoimmune Diseases and Haematological Cancers. *PLOS Genetics*. 9 (1), e1003225–e1003225. doi:<https://doi.org/10.1371/journal.pgen.1003225>.
- Grainger, R.K. & James, D.C. (2013) CHO cell line specific prediction and control of recombinant monoclonal antibody N-glycosylation. *Biotechnology and Bioengineering*. 110 (11), 2970–2983. doi:<https://doi.org/10.1002/bit.24959>.
- Harnish Mukesh Naik, Majewska, N.I. & Betenbaugh, M.J. (2018) Impact of nucleotide sugar metabolism on protein N-glycosylation in Chinese Hamster Ovary (CHO) cell culture. *Current Opinion in Chemical Engineering*. 22, 167–176. doi:<https://doi.org/10.1016/j.coche.2018.10.002>.
- Hossler, P., Khattak, S.F. & Zheng Jian Li (2009) Optimal and consistent protein glycosylation in mammalian cell culture. *Glycobiology*. 19 (9), 936–949. doi:<https://doi.org/10.1093/glycob/cwp079>.
- Igareta, A. (2021) *The Million-Dollar Question: When to Stop Training your Deep Learning Model*. 24 June 2021. Medium. <https://towardsdatascience.com/the-million-dollar-question-when-to-stop-training-deep-learning-models-fa9b488ac04d> [Accessed: 12 December 2023].
- Jong Hyun Nam, Zhang, F., Ermonval, M., Linhardt, R.J. & Sharfstein, S.T. (2008) The effects of culture conditions on the glycosylation of secreted human placental alkaline phosphatase produced in Chinese hamster ovary cells. *Biotechnology and Bioengineering*. 100 (6), 1178–1192. doi:<https://doi.org/10.1002/bit.21853>.
- Kontoravdi, C. & Jimenez del Val, I. (2018) Computational tools for predicting and controlling the glycosylation of biopharmaceuticals. *Current Opinion in Chemical Engineering*. 22, 89–97. doi:<https://doi.org/10.1016/j.coche.2018.08.007>.
- Kotidis, P., Jedrzejewski, P., Sou, S.N., Sellick, C., Polizzi, K., del Val, I.J. & Kontoravdi, C. (2019) Model-based optimization of antibody galactosylation in CHO cell culture. *Biotechnology and Bioengineering*. 116 (7), 1612–1626. doi:<https://doi.org/10.1002/bit.26960>.
- Lancashire, L., Christophe Lemetre & Ball, G. (2008) An introduction to artificial neural networks in bioinformatics--application to complex microarray and mass spectrometry datasets in cancer studies. *Briefings in Bioinformatics*. 10 (3), 315–329. doi:<https://doi.org/10.1093/bib/bbp012>.
- Liu, Y., Nathan, K.J. & Brunton, S.L. (2020) *Hierarchical Deep Learning of Multiscale Differential Equation Time-Steppers*. 2020. arXiv.org. <https://arxiv.org/abs/2008.09768> [Accessed: 10 December 2023].
- Madhuresh Sumit, Sepideh Dolatshahi, An-Hsiang Adam Chu, Cote, K., Scarcelli, J.J., Marshall, J., Cornell, R.J., Weiss, R., Lauffenburger, D.A., Bhanu Chandra Mulukutla & Figueroa, B. (2019) Dissecting N-Glycosylation Dynamics in Chinese Hamster Ovary Cells Fed-batch Cultures using Time Course Omics Analyses. *iScience*. 12, 102–120. doi:<https://doi.org/10.1016/j.isci.2019.01.006>.
- Olah, C. (2015) *Understanding LSTM Networks -- colah's blog*. 2015. Github.io. <https://colah.github.io/posts/2015-08-Understanding-LSTMs/> [Accessed: 1 December 2023].
- Pai, A. (2020) *Analyzing Types of Neural Networks in Deep Learning*. 17 February 2020. Analytics Vidhya. [https://www.analyticsvidhya.com/blog/2020/02/cnn-vs-rnn-vs-mlp-analyzing-3-types-of-neural-networks-in-deep-learning/#Recurrent_Neural_Network_\(RNN\)_-_What_is_an_RNN_and_why_should_you_use_it?](https://www.analyticsvidhya.com/blog/2020/02/cnn-vs-rnn-vs-mlp-analyzing-3-types-of-neural-networks-in-deep-learning/#Recurrent_Neural_Network_(RNN)_-_What_is_an_RNN_and_why_should_you_use_it?) [Accessed: 1 December 2023].
- Pan, X., Ciska Dalm, Wijffels, R.H. & Martens, D.E. (2017) Metabolic characterization of a CHO cell size increase phase in fed-batch cultures. *Applied Microbiology and Biotechnology*. 101 (22), 8101–8113. doi:<https://doi.org/10.1007/s00253-017-8531-y>.
- Pavlos Kotidis & Kontoravdi, C. (2020) Harnessing the potential of artificial neural networks for predicting protein glycosylation. *Metabolic Engineering Communications*. 10, e00131–e00131. doi:<https://doi.org/10.1016/j.mec.2020.e00131>.
- Robail Yasrab & Pound, M. (2020) *PhenomNet: Bridging Phenotype-Genotype Gap: A CNN-LSTM Based Automatic Plant Root Anatomization System*. 3 May 2020. ResearchGate. https://www.researchgate.net/publication/341131167_PhenomNet_Bridging_Phenotype-Genotype_Gap_A_CNN-LSTM_Based_Automatic_Plant_Root_Anatomization_System [Accessed: 2 December 2023].
- Taki Hasan Rafi & Karim, R. (2020) *TIME SERIES ANALYSIS -A COMPARATIVE ANALYSIS BETWEEN ANN AND RNN*. 10 September 2020. ResearchGate. https://www.researchgate.net/publication/344189801_TIME_SERIES_ANALYSIS_-

A_COMPARATIVE_ANALYSIS_BETWEEN_A
NN_AND_RNN [Accessed: 12 December 2023].

Trupti Katte-Bangayya (2018) *Recurrent Neural Network and its Various Architecture Types*. 2018. RSIS.
https://www.academia.edu/39509903/Recurrent_Neural_Network_and_its_Various_Architecture_Types [Accessed: 1 December 2023].

Xue, Y. (2019) An Overview of Overfitting and its Solutions. *Journal of physics*. 1168, 022022–022022. doi:<https://doi.org/10.1088/1742-6596/1168/2/022022>.

Zhang, P., Susanto Woen, Wang, T., Liao, B., Zhao, S., Chen, C., Yang, Y., Song, Z., Wormald, M.R., Yu, C. & Rudd, P.M. (2016) Challenges of glycosylation analysis and control: an integrated approach to producing optimal and consistent therapeutic drugs. *Drug Discovery Today*. 21 (5), 740–765.
doi:<https://doi.org/10.1016/j.drudis.2016.01.006>.

Zhou, H., Zhang, Y., Yang, L. & Du, Y. (2019) *Short-Term Photovoltaic Power Forecasting Based on Long Short Term Memory Neural Network and Attention...* 14 June 2019. ResearchGate.
https://www.researchgate.net/publication/333791434_Short-Term_Photovoltaic_Power_Forecasting_Based_on_Long_Short_Term_Memory_Neural_Network_and_Attention_Mechanism [Accessed: 5 December 2023].

The Role of Ground Source Heat Pumps in UK Domestic Heat Decarbonisation

Aaron Suthagar, Jonathan Wright

Chemical Engineering Department, Imperial College London, UK

Abstract

Decarbonising the UK's residential buildings is a critical challenge due to the substantial contribution of domestic heating to greenhouse gas emissions. Current reliance on fossil fuel-based technologies highlights the need for low-emission alternatives. Among these alternatives, the adoption of heat pumps, particularly Ground Source Heat Pumps (GSHPs), has emerged as a promising alternative over natural gas boilers in households. This study evaluated the suitability of GSHPs as a low-carbon technology within the UK's residential sector. Utilising two bespoke models, a household model and a Heat energy System Optimisation model (HeSO), this research analysed GSHP penetration and the influential factors that characterise this deployment. The household model constructed a comprehensive database of Ground Source Energy Systems (GSESs), including GSHPs and thermal batteries, serving as a foundation for the HeSO model to perform a total system cost minimisation to select cost-effective technologies. By adjusting parameters such as the natural gas price and GSES capital expenditure (CAPEX) subsidy level, reflecting the extent of governmental incentives, findings indicated that optimal GSHP penetration occurred with higher natural gas prices and lower CAPEX for the majority of buildings. Despite some degree of adoption within the domestic sector, this study exposed a discrepancy between current market and policy conditions and those required for widespread GSES adoption. Addressing this gap is pivotal for achieving substantial reductions in household carbon emissions and advancing towards national net-zero objectives. Further study into measures to ensure the economically feasible mass deployment of low-carbon technologies is essential for making tangible progress towards net zero and fulfilling long-term environmental targets.

1 Introduction

Mitigating climate change and transitioning to sustainable energy systems have become crucial global objectives. In 2015, parties from across the world formed the Paris Agreement which set a global warming limit to below 2°C compared to pre-industrial levels^[1] and stated that above 1.5°C there are risks of hitting climatic tipping points such as melting arctic permafrost which would release stored greenhouse gases^[2]. The UK has set its own internal requirements such as: reducing greenhouse gas emissions by 100% from 1990 levels by 2050^[3], ensuring that by 2035 the UK will be powered entirely by clean electricity, and the deployment of new flexibility measures through resource storage to smoothen future price spikes^[2]. Current heating technologies rely on the combustion of fossil fuels and have a high carbon footprint, therefore for improvements to be made here, new and novel technologies, with low carbon footprints, will need to be adopted.

Residential and commercial heating contributes to a substantial portion of global greenhouse gas emissions. In 2017, the combustion of natural gases for space heating and hot water contributed 14.3% of the total greenhouse emissions in the UK^[4]. The UK relies on fossil fuels with 95% of homes being centrally heated with 81% of heat demand being met by gas

boilers on natural gas networks^[5]. Homes are either heated with a natural gas boiler, which on average requires 78% of energy consumption to be used on heating, or with electric heaters, which only require 12%.^[6] The UK government has set ambitious targets to achieve net-zero emissions through the 6th carbon budget, which highlights the immediate requirement for a transition to low-carbon technologies and a reduction in demand for carbon-intensive activities. This would facilitate a 78% reduction in UK territorial emissions between 1990 and 2035 and another goal is to achieve between 65-125GW of low-carbon electricity by 2050^[7].

In the residential sector, initiatives and policies have been laid out to reduce the emissions from homes. Aims such as ensuring all homes have an Energy Performance Rating (EPC) of band C or higher by 2035^[8], and the gas boiler ban in new homes to start as early as 2025^[6]. Further policies include the removal of VAT on new solar panels until 2027 and on the installation of heat pumps and insulation until March 2027^[9]. Other government schemes that have been implemented include the Heat Pump Ready program, a funding pot of £60 million for the installation of 600,000 heat pumps by 2028, and the Boiler Upgrade Scheme - a UK government grant which offers £7,500 towards the cost and installation of a heat pump or a biomass boiler in homes^[10]. One scheme al-

ready in place is the Green Deal scheme, where homeowners can obtain a loan for certain energy efficiency measures and pay off the loan later via their energy bills. The main initiative that will be considered in this study is the replacement of domestic gas boilers with ground source heat pumps.

Heat pumps coupled with the use of clean renewable energy offer a compelling low carbon solution. There are two main forms of heat pumps: electrically and thermally driven heat pumps. This paper focuses on electrically driven heat pumps, specifically Ground Source Heat Pumps (GSHPs). GSHPs make use of thermal energy stored in the ground. They do this by circulating a refrigerant fluid through the ground collectors which absorb thermal energy from the ground. The compressor inside the heat pump increases the temperature and passes the fluid through a heat exchanger. This transfers heat to the hot water cylinders and radiators to provide space heating. Once the fluid has been delivered heat to the distribution system it is passed through an expansion valve and then the cycle restarts. The heat pumps are connected to a series of pipes, known as the ground collectors, either laid out horizontally (often referred to as slinkies) or vertically (through the drilling of boreholes). These pipes are located underground where the average temperature is typically between 10-12°C^[11]. The implementation of heat pumps has many complications that may make them inaccessible to the whole population. These complications include the high initial capital investment, the complexities of government incentives, and extra capital costs that are incurred from installation. Understanding the appropriate size of heat pump is complex due to the regional climate and EPC rating variations, this demands tailored strategies for optimising heat pump performance.

Several studies assessed various aspects of heat pump implementation and the related challenges. Petrovic and Karlsson assessed the penetration of heat pumps into the Danish heating network^[12]. In this study, a constraint on the available ground area was considered. GIS tools were used to determine if buildings had sufficient area to place the required horizontal pipes. If this constraint was ignored it resulted in a 0.3% lower system cost and a 2.7 times higher uptake of GSHPs. The study was then extended using the TIMES-DK model, which optimised both investment and operation for electrically driven heat pumps. This predicted that heat pumps would be responsible for 66-70% of heat production from individual heat sources. This corresponds to 24-28% of total heat demand after 2035 and without residential heat pumps, total system costs would increase by 16% and biomass usage by 70%^[12].

Brenn et al. supported electrically driven heat pumps for low CO₂ intensity electricity sources, while advocating for natural gas-driven heat pumps' eco-

nomic efficiency for space heating^[13]. Another study by Wang et al. found that the use of GSHPs achieved 70% fuel saving and reduced greenhouse gases by 45% when compared to natural gas furnaces. Wang et al. found that the use of heat pumps in Iran was low due to the low natural gas price and the high electricity price, therefore favouring gas boilers^[14]. This was opposed by the paper by Mersch et al. which found that the required installed grid capacity can be more than twice as high in scenarios with high gas prices when compared to low gas prices. Although the size of the power sector increases significantly with increasing gas prices, the optimal deployed technology portfolio does not change fundamentally^[15]. This indicated that with suitable policies and subsidies, heat pumps could become a favourable technology.

The aim of this study was the examination of heat pump adoption within the UK domestic heating sector. This was done using two bespoke optimisation models, one focusing on individual households and the other taking a comprehensive whole-system perspective, which forms the core of this study.

As of the current writing, the integration of GSHPs within these model types remains unexplored. This represents a research gap in the existing literature. Consequently, this study evaluated the necessary conditions for widespread heat pump deployment across the UK, understanding sensitivity to various influencing factors.

The outcomes of these models were rigorously evaluated, both independently and comparatively. The findings emphasise the impracticality of mandating a singular technology for all end-users at a given time, suggesting a phased approach for end-user adoption.

2 Methodology

The novel approach taken in this study was the integration of two bespoke models. The first model considered was the household model and the second was the Heat energy System Optimisation model (HeSO). Parameters including natural gas price and Capital Expenditure (CAPEX), to reflect government subsidies, were varied to map the penetration of heat pumps in the domestic heating sector. To model the UK's domestic heating sector five typical buildings were used. These buildings were chosen using a clustering algorithm on data from the Cambridge housing model^[16] to have varying total heat demands and space heating to hot water ratios. These five buildings give a representative sample of all UK homes.

2.1 Household Model

The household model, as shown in Figure 1, is an optimisation framework that relies on first-law design models and data-driven component costing meth-

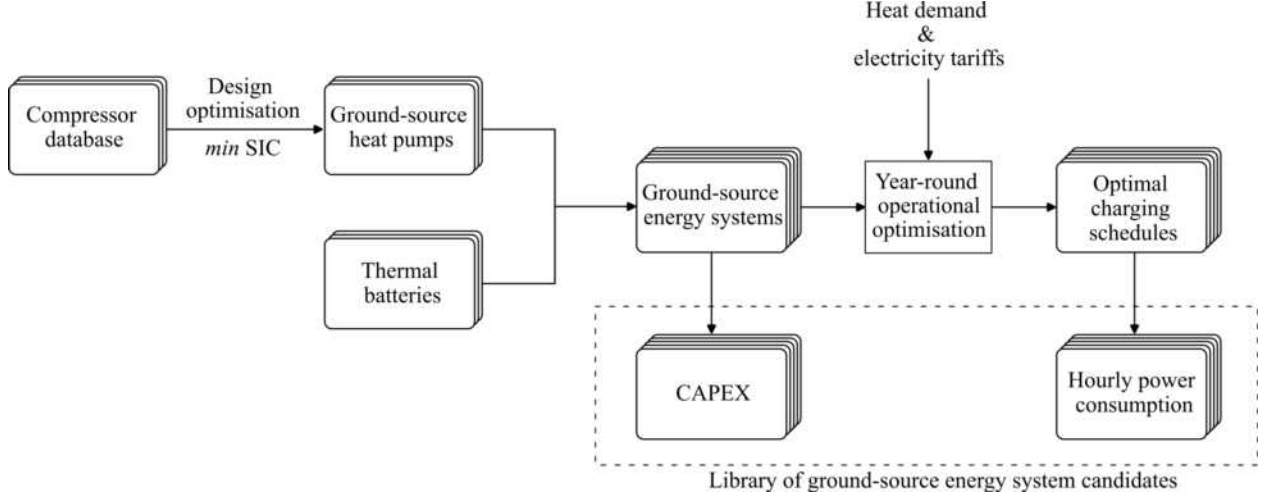


Figure 1: The block diagram of the structure of the HeSO model. The model takes inputs such as total thermal demands, initial renewable and non-renewable resource availability, fuel prices, and availability of current technologies. From this, the model does a simultaneous optimisation for the minimisation of the total system cost while ensuring all demands are met.

ods to optimise: (i) the design of ground-source heat pumps; and (ii) the year-round operation of ground-source energy systems (GSESs) with integrated thermal energy storage.

This optimisation framework is built upon an extensive database collected from various manufacturers of reciprocating-piston, scroll, and rotary-vane compressors. The model assesses the off-design and part-load performance using data obtained from these manufacturers. A GSHP is designed for each compressor held in the database; the condenser and evaporator units are sized to minimise the specific cost of the heat pump. The depth and diameter of the required bore-hole are determined based on the heat pump’s nominal operating conditions (with brine water circulating in underground heat exchangers at 0 °C, hot water supplied at 35 °C and compressor running at 50 Hz). The wide range of GSHPs thus obtained are integrated with thermal batteries of various sizes (ranging from 5 to 50 kWh) to provide a portfolio of over 100 ground-source energy systems (GSESs) for each typical building. These candidates supplied the heat demand for the 5 buildings considered in the whole-energy system model. It is worth noting that not all GSESs will be suitable for all 5 buildings and that only those with powerful enough compressors or large enough thermal stores, that can meet the required heat demand, will be selected.

The optimisation of the year-round operation of each GSES to minimise the operating costs (namely, the cost of electrical energy consumed throughout the year) while supplying the building-specific heat demand is a highly non-linear problem (NLP). The non-linearity is due to the dependency of the heat pump coefficient of performance (COP) not only upon the instantaneous ground temperature but also upon the heating rate. Since the HeSO optimiser is based on a

linear problem solver, it is proposed to optimise the charging schedule of the thermal batteries for each building and each GSES with time-varying electricity tariffs extracted from the HeSO model using the ipopt NLP solver and provide the whole-system model with fully-optimised candidates. In summary, year-round hourly-defined power consumption signals were provided along with the capital cost associated with each GSES candidate. Another output of the model was the Levelised Cost Of Heating (LCOH) which consists of the CAPEX of the GSES combined with the summation of the operation cost over the GSES’s lifetime, this is all divided by the summation of the specific heat demands over the system’s lifetime. This information is displayed mathematically in Equation 1.

$$LCOH = \frac{CAPEX + \int_{t_i}^{t_f} OPEX(t) dt}{\int_{t_i}^{t_f} HeatDemand(t) dt} \quad (1)$$

2.2 HeSO Model

The whole system, or HeSO, model is designed as a capacity expansion and unit commitment optimisation tool. Its primary goal is to minimise the overall cost associated with transitioning entire energy systems in alignment with net-zero policies. Based on the ESO model developed by Heuberger et al.^[17], this model pursues the most cost-effective transformation of the energy system up to 2050. It achieves this by optimising technology investments, decommissioning strategies, and yearly adjustments, while also using hourly dispatch profiles to ensure alignment with demand patterns.

The HeSO model includes several parameters that can be changed with each run. The initial parameter that was considered was the number of typical days considered. This can be varied between 2 and 365

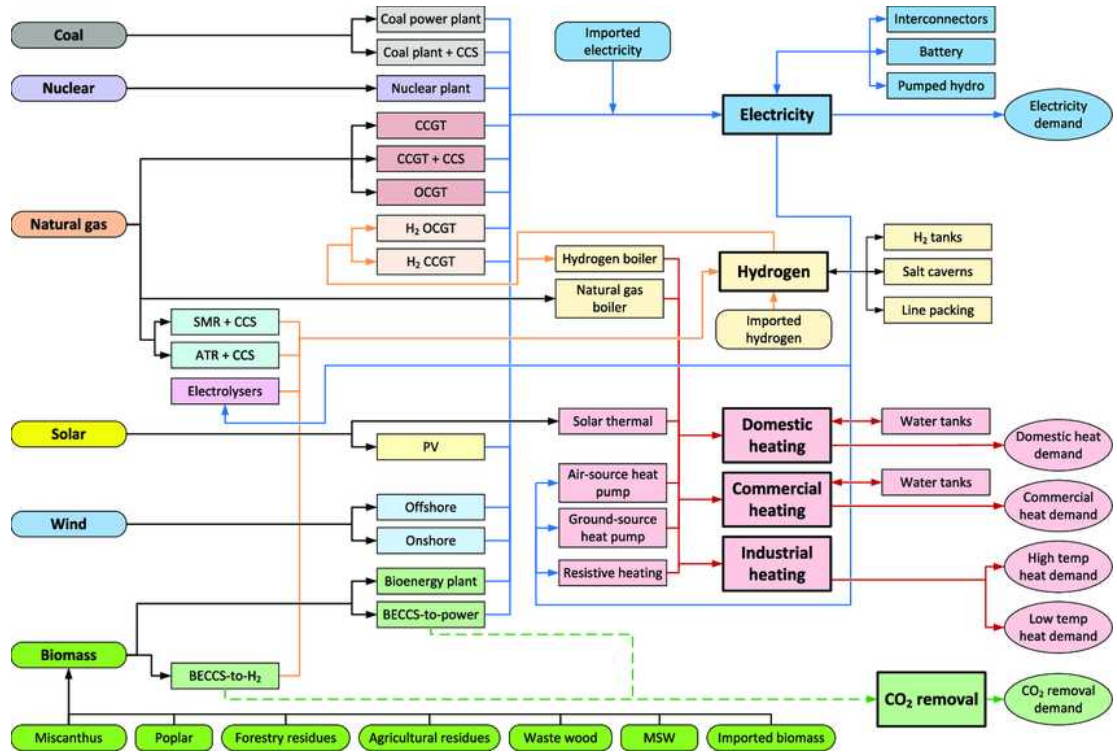


Figure 2: The block diagram of the structure of the HeSO model^[15]. The model takes inputs such as total thermal demands, initial renewable and non-renewable resource availability, fuel prices, and availability of current technologies. From this, the model does a simultaneous optimisation for the minimisation of the total system cost while ensuring all demands are met.

days with a greater number improving accuracy but significantly increasing the computation time. This comes in addition to 4 extreme days which the model always includes, this ensures that any suggested solution is robust enough to deal with the days with even the most significant heat demands and the days where renewable electricity is at the lowest availability. In this report, 4 representative days in addition to the 4 extreme days were used to reduce the required computation time while still maintaining an accurate representation of the different conditions that occur throughout the year.

Some of the considerations that applied in the model were: individuals were assumed to have perfect knowledge and make timely, cost-minimising decisions; investment and decommissioning decisions were to be made every five years; dispatch was optimised over 4 representative days; and technology deployment was limited by build rate constraints. Although this model also considers commercial and industrial heating demands, the focus of this paper is strictly on the heat demand in the residential sector.

The model accounts for carbon emissions within the constraint of pre-set targets. Specifically, a linear reduction in permissible CO₂ emissions occurs every 5 years up until net zero in 2050. These reductions are motivated by the emissions targets outlined on the UK government website to target emissions reductions of

68% by 2030 compared to 1990s levels and ultimately net-zero by 2050^[18]. The constraint on CO₂ emissions acted as one of the driving factors for the increased uptake of low-emission heating technologies within the residential sector.

Within the model, all forms of fuel have a carbon dioxide footprint attributed to them and additionally, electricity has an aggregated CO₂ footprint per unit of power produced. These feed into the aforementioned permissible carbon dioxide emissions constraints. The model also includes several financial metrics with both interest rates and slack penalties for domestic heating values, which can be altered based on changing policies. Slack penalties provide a financial penalty when heating demands are not met.

2.3 Model Integration

In this study, the two models are used in succession starting with the household model and then using its outputs as parameters in the HeSO model. This is necessary as the HeSO model is a linear program optimiser but the optimisation of heat pumps is nonlinear, therefore the heat pump optimisation is done in the household model and the output is used as an input into the HeSO. Figure 3 shows the integration of the two models to get the final result of heat pump penetration.

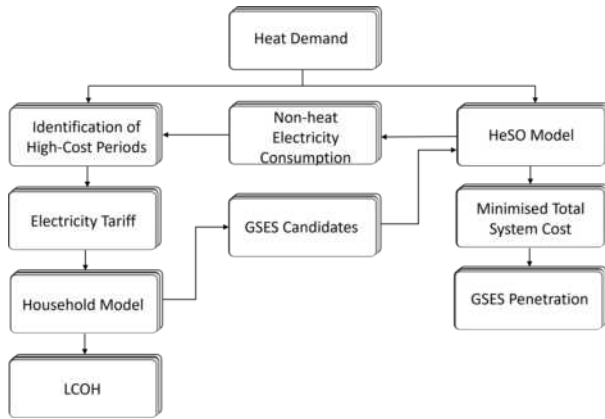


Figure 3: A block diagram showing the integration of both the household and HeSO model, showing the inputs and outputs for each model with heat demand being the primary input and LCOH and GSHP penetration being the primary outputs

The HeSO model was first executed with its preset library of heating technologies to provide the non-heat electricity consumption values which, combined with the heat demand, were used as proxies to calculate the time-resolved electricity tariff. Both heat demand profiles and time-varying electricity tariffs were fed into the household model, which in turn provided a set of year-round optimised GSES for each of the typical buildings. The fully-optimised GSESs were then added back to the HeSO library as an available technology for the HeSO model to choose from. Through successive run iterations of the HeSO model, it successfully selected heat pumps from the available library, alongside the originally provided gas boilers, in order to comply with the specified constraints. The total system cost was minimised and the heat pump penetration was calculated. Other preset technologies in the HeSO library had their CAPEXs set at an arbitrarily high value to ensure they were not picked in the minimisation.

One parameter which was manipulated, to analyse the impact of cost reductions and incentives, is the CAPEX of the GSES. Every optimised heat pump has a representative cost, which was calculated by the household model, and an additional cost which is dominated by the underground heat exchanger cost. Upgrade and installation costs are also included but to a much lesser magnitude. These additional costs typically range between £8000-£12000^[19] depending on the nature of the ground source heat exchanger (slinky versus borehole) and how many home upgrades need to be made (including better radiators and insulation). The sum of these two costs aggregates to the total CAPEX of installing a new ground source heat pump. The total CAPEX, with consideration of government subsidies/incentives, has the potential to become less expensive so by varying the percentage of heat pump CAPEX their uptake in different scenarios can be investigated.

Another parameter that significantly affects the

uptake of GSESs within the model is the natural gas price. Past data has shown that natural gas price is unpredictable and is susceptible to price shocks which can be unexpected and significant. Consequently, modelling some variation in gas prices and the change in uptake gives valuable insight into how the demand for GSESs may change.

To see the effect that the variation of both of these parameters would have on GSES penetration, the penetration of a single high-performance heat pump and thermal battery combination was mapped and then the results were compared from 2020 to 2050 to evaluate the deployment of that GSES. The CAPEX was given a subsidy percentage of 0, 50, or 90 percent resulting in the model being able to select the heat pump at a full, or a variation of a discounted price. Additionally, natural gas prices were varied from 5 to 50 £/MWh at intervals of 5 £/MWh.

Following this, a more rigorous study into the extent of penetration in the UK with a larger library of GSESs for each of the five typical buildings was completed. Here, an optimised set of GSESs, consisting of over 100 units for each building type, from the household model were implemented directly as options to be selected from the whole system model. However, to present the findings more clearly, the library was simplified to showcase only the top 10 most frequently chosen GSESs. The total penetration of GSESs in the domestic heating system was then able to be mapped with a breakdown of specific GSESs. The study was also extended and natural gas prices were varied from 10 to 150 £/MWh at intervals of 10 £/MWh but under the same CAPEX subsidy level.

The whole system HeSO model has several limitations. It operates agnostically of many real-world requirements such as difficulties regarding installation, variability between houses, only selecting representative days to reflect the year, and considering humans as perfectly rational agents. The model assumes that individuals have access to and fully understand the quantity of information required to make the optimal decision on what GSESs to get and make such a decision promptly. When it comes to investigating which GSESs are selected over others, it is difficult to account for such difficulties. Hence, the solutions reached may differ from what is observed in the real world.

3 Results & Discussion

3.1 UK-Wide GSHP Penetration

The initial study, mentioned in Section 2.3, was completed with a single optimal GSES to verify that the parameters chosen had the desired effect on GSES penetration. Building on this a full library of GSESs were then implemented into the HeSO model and the

model was run from 2020 to 2050 at 5-year intervals and 2030, 2040 and 2050 were chosen as representative years. Figure 4 shows the total penetration of the GSESs across the whole UK, done by summing the total uptake per typical building for each representative year.

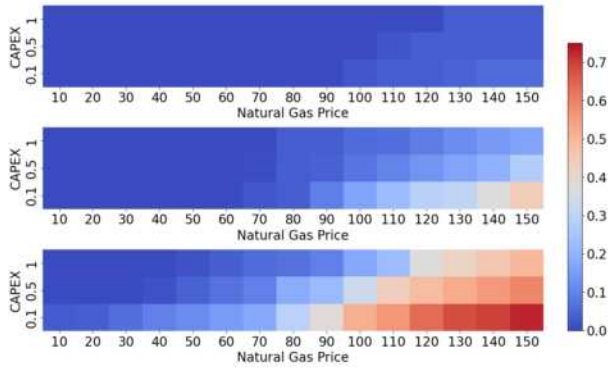


Figure 4: Penetration of all GSESs for the five typical buildings across the 3 representative years of 2030 (Top), 2040 (Middle), and 2050 (Bottom). The x-axis portrays the natural gas price (£/MWh) and the y-axis displays the % of total CAPEX (£). Red represents 75% penetration and blue represents 0% penetration.

Penetration increased predictably with higher natural gas prices and CAPEX subsidy levels, displaying a positive correlation with deployment. In 2030 there was minimal uptake of GSESs in all typical buildings, with adoption only starting at 100 £/MWh at 90% subsidy to a maximum value of around 5% penetration at 150 £/MWh with 90% subsidy. In 2040, there is greater penetration than in 2030. The penetration now starts at a lower natural gas price of 70 £/MWh and reaches a maximum penetration value of 45%, at 150 £/MWh and 90% subsidy. 2050 has the largest penetration across the natural gas price range, with slight penetration starting at 10 £/MWh at 90% subsidy and at 50 £/MWh at the 0% subsidy mark. For the highest subsidy and natural gas price, the penetration of GSESs reaches 70%, with 50% penetration being reached from 100 £/MWh across all subsidy values.

With a higher CAPEX subsidy level, the GSES is more likely to be picked during the total system cost-minimisation. Increasing natural gas prices will increase the operational expenditure (OPEX) of the gas boiler, but since this does not affect the OPEX of the GSES, GSESs become the lower-cost technology. Therefore the trend that penetration increases with natural gas price and CAPEX subsidy was validated. The variations in penetration levels amongst different buildings were not uniform; some of the buildings experienced notably higher penetration levels compared to others.

3.2 Comparison of Penetration Between Two Building Types

To further analyse the penetration of GSESs across the UK, two of the typical buildings were chosen to illustrate their individual penetration. Buildings 2 and 4 were picked as they have a large variation in uptake both over the representative years and in comparison to each other. Buildings 2 and 4 also have different heat demands and hot water to space heating demand ratios, ensuring a notable comparison in the energy requirements and the proportional allocation of energy usage.

In 2030, Building 2 shows minimal penetration over the whole range of natural gas prices as seen by the constant blue penetration map. Penetration in Building 4, on the other hand, started at 100 £/MWh increasing to 100% penetration at 140 £/MWh. The shape seen in the top map of Figure 4 (2030 total) is dictated by the shape of Figure 5a (Building 4) since the other buildings exhibit minimal penetration, similar to Building 2. Therefore the majority of GSES penetration in 2030 is from Building 4. The reason why there was minimal uptake in the majority of buildings is due to the annual emission limit set in 2030. It is at a comparatively relaxed value of 160 MtCO₂-equivalent leading to the preexisting gas boilers being the technology with the lowest cost in houses, showing that there is no strong driving force to remove the gas boiler and replace it with a GSHP. Uptake in Building 2 was equivalent to about 1 heat pump of each type reflecting a very small proportion of the millions of representative buildings in the system. Building 4 has a much greater uptake of GSES than Building 2 due to the larger heat demand in Building 4 and the ability to have a combination of a GSES with an immersion heater. This lowered the total cost of the system and allowed for a significantly improved penetration.

In 2040 there was greater penetration across all of the typical buildings in comparison to 2030. This was expected since the annual CO₂ emission limit has linearly decreased from 2030 to a value of 80MtCO₂, driving the implementation of more heat pumps. In Building 4, there is a more significant threshold of where the GSESs are used. Penetration spikes to 40% penetration at 70 £/MWh for the highest CAPEX subsidy and 50% penetration at 80 £/MWh for the other CAPEX subsidies. There is then 100% penetration from 100 £/MWh onwards. Building 2 shows a significantly smaller uptake in the GSESs with penetration only starting at the 120-130 £/MWh mark for the 50% and 90% subsidy rate, and only reaching around 20% total penetration at the highest natural gas price and subsidy. This lack of uptake is similar to that in 2030 where gas boilers can meet the heat demand for Building 2 more economically than the GSESs therefore the gas boilers are chosen.

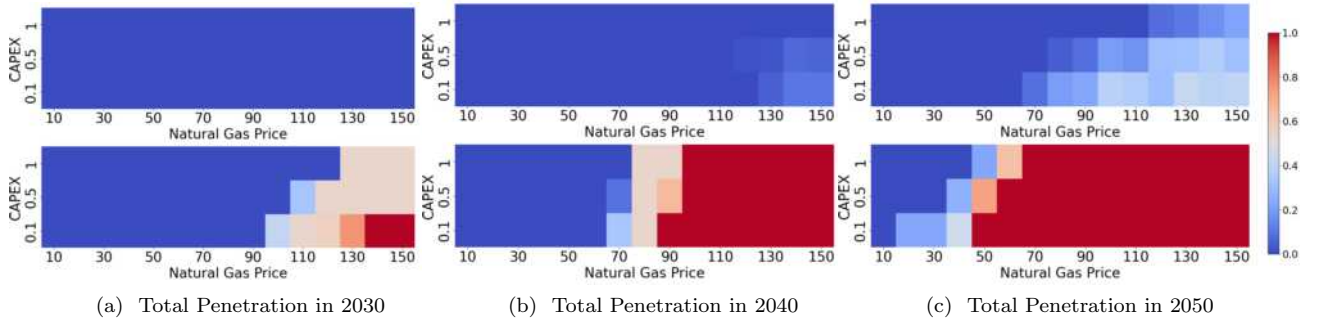


Figure 5: GSES penetration over 2030 (Left), 2040 (Middle), and 2050 (Right) for Building 2 (Top) and Building 4 (Bottom). Natural gas price is varied on the x-axis, and % of total CAPEX is varied on the y-axis, with red representing 100% penetration and blue representing 0% penetration.

The deployment in 2050 shows a similar trend to that in 2040 but to a greater extent. 2050 is the end point of the net zero transition so it is expected that more heat pumps will be implemented. Building 4 has the greatest total uptake, with penetration starting at 20 £/MWh for 90% CAPEX subsidy but only starting at 40 and 50 £/MWh for 50% and 0% CAPEX subsidy respectively. Building 4 reaches 100% penetration onwards from 50 £/MWh for 90% CAPEX subsidy and 70 £/MWh for 0% CAPEX subsidy. Building 2 shows the least penetration but has a significant increase from 2040, with 10% penetration only starting at 50 £/MWh for the 90% subsidy and at 120 £/MWh for the 0% subsidy. Penetration reaches a maximum value of around 40% at the expected high end of natural gas prices. This trend supports the previous discussion that it is simply not economical to use the designed GSESs in Building 2.

3.3 LCOH as a Selection Variable in GSES Uptake

The HeSO model minimises the total system cost therefore it picks the GSESs with a cheaper CAPEX, which is reflected in a smaller thermal battery size and lower heat pump nominal power. This generally aligns with the prediction from Figure ref, the LCOH contour plot produced from the building model but should not be used as a sole factor. The LCOH, as mentioned in Section 2.1, is simply the sum of the CAPEX and the annual OPEX.

Building 2 observes a general trend that all the GSESs that were selected lie around the contours with the smallest LCOH. But none of the selected GSESs lie within the minimum LCOH contour but instead lie mainly within the 2nd and 3rd smallest contours of 9.6 and 9.8 p/kWh respectively. Since the lowest LCOH contour is centred at a nominal heat pump power of around 6 kW, it was expected that pumps in this size would be picked. In contrast to that, the chosen GSESs have a nominal power between 2.9 kW

and 3.4 kW with varying thermal battery sizes from 5-25 kWh. With the most selected GSES, with an adoption ratio of 52%, of pump power 2.9 kW and 15 kWh.

Building 4 has much fewer chosen GSESs, only having three in comparison to Building 2 which had eight. The minimum LCOH contour is much greater than the Building 2 minimum and lies over a large range of nominal power between 7.5 kW and 20 kW and between 30 kWh and 40 kWh thermal battery size. The most selected GSES, with an adoption rate of 81%, has a nominal pump power of 1.6 kW and a thermal battery size of 20 kWh. The other selected heat pumps have the same nominal power but larger thermal battery sizes. The pump power chosen for Building 4 is surprising, considering for a building with a larger heat demand it has a smaller heat pump than Building 2. If the minimum LCOH cannot be found for the required pump size, in the household model, then an immersion heater can be used with the heat pump to meet the heat demand. This option was only available for Buildings 4 and 5 due to the high heat demand. Being able to use the additional heater decreases the total GSES CAPEX in comparison to solely using a heat pump of appropriate size. Therefore the cost minimisation chooses to take the backup heater option preferentially.

This discrepancy between the minimised LCOH and the chosen GSES is due to the minimisation that the HeSO model performs. The HeSO model minimises cost while meeting the required heat demand so as the heat demand is met the model then chooses the GSES with the lowest CAPEX, which is typically the smallest thermal battery size and lowest nominal power. As seen in Building 4, the adoption percentage decreases as thermal battery size increases since the CAPEX is greater. Due to the low uptake of GSESs, Building 2 has a larger variety of selected options. Although the trend can still be seen in Figure 6a, the most selected options are of the smallest nominal power and thermal battery size. This trend shows that LCOH cannot solely be used to determine GSES

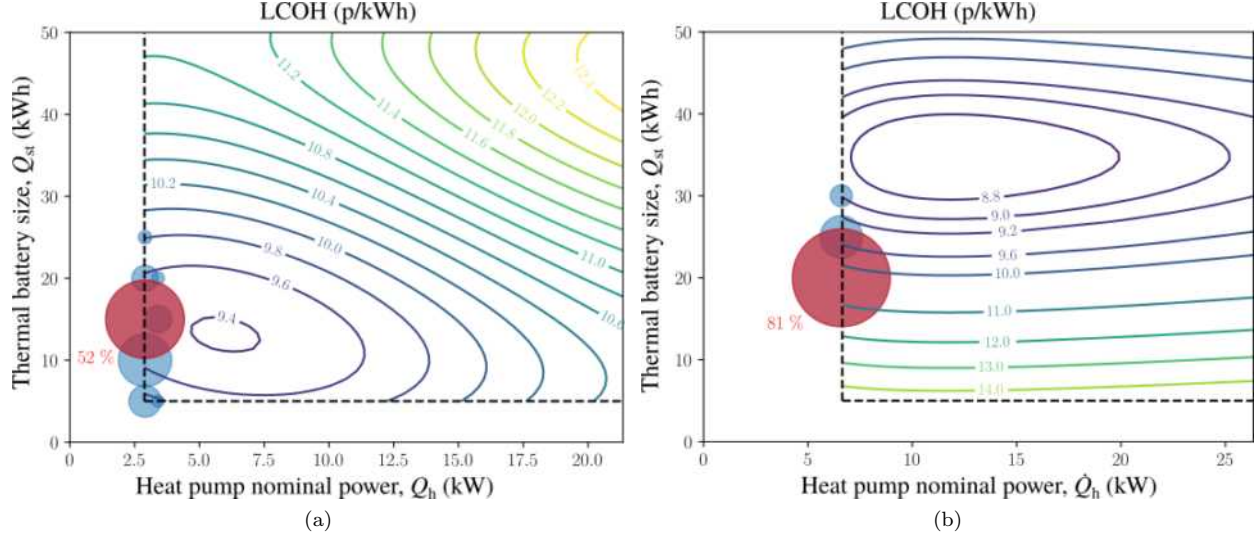


Figure 6: Levelised cost of heat (LCOH) as a function of the heat pump nominal power, \dot{Q}_h (which includes both the mechanical vapour compression heat pump and immersion backup heater) and size of the thermal battery, Q_{st} , reported as contour maps for: (a) Building 2; and (b) Building 4. The contour lines show the iso-LCOH values based on 2023 electricity tariffs. GSES candidates actually picked by the whole-system optimisation model are reported as bubbles, the sizes of which represent their adoption ratio (in %). Red bubbles represent the most likely adopted GSES configuration for each building, while other selected candidates are represented with blue bubbles.

selection. Instead, CAPEX is the driving factor for which GSESs are selected.

In summary, GSES penetration is highest at high CAPEX subsidies and high natural gas prices and increased from 2030 to 2050. Although there is a great amount of penetration in Building 4 there is significantly less penetration in the other buildings showing that even though the GSESs are optimised they are still not chosen over gas boilers. In June 2023 natural gas was priced at 4.56 p/kWh (equivalent to 45.6 £/MWh)^[20], so if this price and the 90% subsidy were maintained until 2050 then there would be varied penetration over the buildings with a high amount in Building 4 but minimal in Building 2. Natural gas prices are near impossible to predict and have been inaccurately forecasted in the past, therefore it would be crass to assume what the price will be in 2050. But what can be said for the GSESs to become viable in the UK, there needs to be either large natural gas price shocks or a reduction in CAPEX through subsidies or simply tech becoming cheaper, or for greater penetration a combination of the two.

4 Conclusions

This study assessed the penetration of Ground Source Energy Systems (GSESs) in the UK domestic heating sector. This was done firstly through the use of the household model. Several GSESs were optimally designed with varying heat pump powers and thermal storage sizes and added to a library. This library was then used as a database, amongst other available tech-

nologies, for the HeSO model to pick from during the minimisation of the total system cost. The findings presented in Section 3 highlight significant insights into how varying fuel prices and CAPEX subsidies significantly influence optimal net-zero energy system transition strategies over the key years of 2030, 2040 and 2050.

It was found that the penetration of the GSESs was greatly affected by both the natural gas price and also the level of CAPEX subsidy, which acted as a government incentive. The general trend seen throughout all the typical buildings was that penetration was the greatest for high natural gas prices combined with a high subsidy percentage and lowest at no subsidy and low natural gas prices. When examined individually, the extent to which natural gas prices would need to increase for greater GSES penetration is considerable and only previously observed during natural gas price shocks, which historically have not resulted in a permanent price increase. But if balanced with strong governmental policies and incentives, the gas price does not need to be as high for the same amount of uptake. This shows that, under the correct conditions, heat pumps can become a viable low-carbon heating option in the domestic sector.

Previous assumptions suggested that the levelised cost of heating would be the primary factor influencing the choice of Ground Source Energy Systems. However, although Building 2 primarily favoured GSES options located in the lowest LCOH range, other buildings, including Building 4, opted for GSES selections that deviated from the lowest LCOH region

illustrated on their respective contour plots. To bring GSES options within the minimum LCOH range, adjustments such as an increase in nominal pump power, a larger thermal battery, or a combination of both adjustments would be necessary. Notwithstanding, implementing these changes would significantly escalate the system's CAPEX.

Initial expectations assumed that GSES within the lowest LCOH range would inevitably possess the lowest lifetime costs for the system. On the contrary, it became evident that the CAPEX held greater significance in the selection of GSES over the LCOH. Consequently, relying solely on LCOH as the decisive factor for choosing GSES is not viable within the limitations of this study. This finding indicates that the heat pumps were already optimised with a high coefficient of performance, efficiently and affordably meeting heat demands from an overall system perspective without the need for larger compressors.

Although heat pumps offer a compelling solution to the decarbonisation of the UK's domestic heating sector, they are not always appropriate. For the buildings with smaller heat demands, even at extremely high natural gas prices and high CAPEX subsidy levels, there was still very little GSES penetration. This is due to the buildings' heat demand being met by another technology at a lower total system cost. This shows that when the GSESs are not optimally priced for the specific heat demand, they will not be blanketly chosen over natural gas boilers.

5 Outlook

Ideally, this study would be repeated with a wider selection of GSESs. The GSESs that were designed for the buildings were not always contained within the lowest LCOH contour. With more configurations of GSESs with nominal powers that are aligned with the lowest LCOH, as well as other alternative test cases, a relationship between which heat pumps are selected for each building and its respective LCOH can be better determined.

Another aspect of the study that could be extended is looking into more CAPEX subsidies. Due to the intensity of the computation only a 0%, 50% and 90% CAPEX subsidy was chosen to model. These subsidies capture the whole range of possible subsidies but may not have been reflective of what the government can offer through schemes or initiatives. With a larger range of subsidy percentages, a more precise prediction of penetration can be made for a given level of government support.

The HeSO model reflects an inherently complicated situation and contains assumptions to permit its function. These assumptions may introduce deviation from the decision-making expected in applied situations. Firstly, an assumption was made about the end

user, assuming perfect foresight and knowledge when picking technologies to provide rational cost minimisation. Further analysis into the end-user's perspective and buyer psychology is necessary to accurately predict the deployment. Another assumption that was made was that the CAPEX of technologies remains the same over time. Some of the constraints set on the model were that investment and decommissioning decisions happen every five years, the dispatch optimisation was over four typical days and the technology deployment was limited by build rate constraints. These may not be reflective of real-world conditions which may either exceed or fall short of these build rate constraints.

The scope of this study was refined to only look at GSHPs. Therefore an extension of this study would be to add a larger library of technologies for the HeSO model to pick from. Many other available heating technologies such as Air Source Heat Pumps (ASHPs) and hydrogen boilers may be well-equipped to cost-effectively meet the housing heat demands within environmental constraints. Doing this would allow for more specialised renewable options that can replace gas boilers to increase the uptake of low-carbon heating solutions in the domestic sector.

Currently, the HeSO model is used to identify high and low electricity demand periods. These high and low periods then get assigned electricity tariffs from Octopus Energy. These tariffs are then used as an input in the household model for the design of the GSESs. The extension in this study would be to integrate the variable electricity tariff from the HeSO model into the household model. This would allow for a potentially more accurate LCOH calculation and the start of a more detailed investigation into the efficacy of LCOH in determining which technologies are more likely to be selected.

6 Acknowledgments

We would like to thank both our supervisors Matthias Mersch and Dr Paul Sapin for their continued support throughout this project.

References

- [1] United Nations Framework Convention on Climate Change (UNFCCC). Paris agreement. <https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement>, 2015. Accessed on: December 1, 2023.
- [2] Department for Business, Energy & Industrial Strategy (BEIS). Net zero strategy. <https://assets.publishing.service.gov.uk/media/6194dfa4d3bf7f0555071b1b/>

- [net-zero-strategy-beis.pdf](#), 2021. Accessed on: December 3, 2023.
- [3] Mission zero: Independent review of net zero. <https://lordslibrary.parliament.uk/mission-zero-independent-review-of-net-zero/#:~:text=The%20%27net%20zero%20target%27%20refers,the%20UK%20from%20the%20environment,2023>, 2023. Accessed on: December 3, 2023.
 - [4] Oliver Broad, Graeme Hawker, and Paul E. Dodds. Decarbonising the uk residential sector: The dependence of national abatement on flexible and local views of the future, 2020.
 - [5] Modassar Chaudry, Muditha Abeysekera, Seyed Hamid Reza Hosseini, Nick Jenkins, and Jianzhong Wu. Uncertainties in decarbonising heat in the UK. *Energy Policy*, 87:623–640, 2015.
 - [6] EDF Energy. Uk boiler ban: What does the future hold for home heating? <https://www.edfenergy.com/heating/advice/uk-boiler-ban>, 2022. Accessed on: December 1, 2023.
 - [7] Committee on Climate Change. Sixth carbon budget, 2020. Accessed on: December 5, 2023.
 - [8] House of Commons Library. Research briefing: Uk climate change targets, 2020.
 - [9] United Kingdom Government. Powering up britain: Net zero growth plan. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1147457/powering-up-britain-net-zero-growth-plan.pdf, 2023. Accessed on: December 5, 2023.
 - [10] Energy Saving Trust. Boiler upgrade scheme. Available at: <https://energysavingtrust.org.uk/grants-and-loans/boiler-upgrade-scheme/>, 2023. Accessed on: December 4, 2023.
 - [11] Thermal Earth. How do ground source heat pumps work? Available at: <https://www.thermalearth.co.uk/how-do-ground-source-heat-pumps-work>, 2023. Accessed on: December 4, 2023.
 - [12] Stefan Petrović and Kenneth Karlsson. Residential heat pumps in the future danish energy system. *Energy*, 114:787–797, 11 2016.
 - [13] J. Brenn, P. Soltic, and Ch. Bach. Comparison of natural gas driven heat pumps and electrically driven heat pumps with conventional systems for building heating purposes. *Energy and Buildings*, 42(6):904–908, 2010.
 - [14] Zheng Wang, Mark B. Luther, Mehdi Amirkhani, Chunlu Liu, and Peter Horan. State of the art on heat pumps for residential buildings. *Buildings*, 11(8):350, August 2021.
 - [15] Matthias Mersch, Christos N Markides, and Niall Mac Dowell. The impact of the energy crisis on the uk’s net-zero transition. *Iscience*, 26(4), 2023.
 - [16] University of Cambridge Cambridge Centre for Sustainable Development. Cambridge housing model. <https://cambridgeenergy.org.uk/project/cambridge-housing-model-decc/>, 2022. Accessed on: December 2, 2023.
 - [17] Clara F. Heuberger. Electricity systems optimisation with capacity expansion and endogenous technology learning (eso-xel). <https://doi.org/10.5281/zenodo.1048943>, 2017.
 - [18] Pm recommits uk to net zero by 2050 and pledges a fairer path to achieving target to ease the financial burden on british families. Accessed on: December 1, 2023.
 - [19] Kensa Heat Pumps. Heat pump costs. Available at: <https://www.kensaheatpumps.com/heat-pump-costs/>, 2023. Accessed on: December 8, 2023.
 - [20] Quarterly energy prices september 2023. Available at: https://assets.publishing.service.gov.uk/media/651d7540e4e658000d59d961/Quarterly_Energy_Prices_September_2023.pdf, 2023. Accessed on: December 9, 2023.

Experimental Design to Investigate Aqueous Amino-Acid Solvents for CO₂ Capture

Ermis Christodoulou and Shannan Huang

Department of Chemical Engineering, Imperial College London, U.K.

Abstract

Recent research has identified aqueous amino-acid salt (AAS) solutions as potential CO₂ absorbents for post-combustion capture (PCC). This study investigates the CO₂ cyclic absorption capacity by potassium glycinate (KGly) over a temperature range of 40°C – 120°C. To achieve this, a bespoke synthetic vapour-liquid equilibrium (VLE) apparatus was employed. Prior to conducting measurements with amino-acids, it was crucial to validate the set-up's operation using monoethanolamine (MEA), a well-studied CO₂ absorbent. Many modifications were made to the standard operating procedure (SOP) and the VLE apparatus to obtain results consistent with literature. The obtained results showed a slight deviation from literature data, which was attributed to poor heat transfer causing the temperature of the CO₂ added to the VLE cell to be overestimated. A sensitivity analysis confirmed this temperature discrepancy. By applying a temperature correction, experiments of 1.930 mol·kg⁻¹ KGly were performed over the temperature range specified. As the set-up is close to validation, it is anticipated that the equipment can be used to investigate AAS as carbon capture solvents in the near future.

Keywords: vapour-liquid equilibrium, monoethanolamine, potassium glycinate, amino-acids, carbon capture

1. Introduction

1.1. Motivation

The past two centuries have seen an unprecedented rise in anthropogenic carbon dioxide greenhouse gas emissions, which has led to escalating global temperatures and detrimental environmental impacts (Wilberforce et al., 2021). The United Nations Intergovernmental Panel on Climate Change has advised to limit global warming to 1.5°C by the end of this century to avoid severe climate change impacts (UNFCCC, 2023).

To achieve the required level of decarbonisation, carbon capture and storage (CCS) is a key technology. Post-combustion capture (PCC) is the most mature method employed to date and involves capturing carbon dioxide from the exhaust gases of fossil fuel power plants or other industrial sources prior to release into the atmosphere. The currently most well-studied technique in PCC is chemical absorption using aqueous amine solvents, such as monoethanolamine (MEA), or blends of aqueous amines, to selectively remove CO₂. This is followed by heating the CO₂-loaded solvent to strip it of CO₂. The lean solvent is recycled back for the absorption step, while the highly concentrated CO₂ may be geologically sequestered or processed for utilisation in industrial processes.

MEA is a well-studied benchmark solvent for PCC, which exhibits favourable properties such as high CO₂ selectivity, high absorption rate and affordability. However, the solvent also possesses significant drawbacks

including high energy requirement, corrosivity, poor stability and environmental issues (Sang Sefidi & Luis, 2019).

Amino-acid salt (AAS) solutions have recently been identified as promising solvents for PCC. They demonstrate low volatility, resistance to thermal and oxidative degradation, and high biodegradability. On the other hand, amino-acids are expensive, as they are currently produced primarily at high purity in the food and pharmaceutical industries. Amino-acid solvents have been commercialised to an extent for acid gas removal, but further development is needed to improve their affordability (Sang Sefidi & Luis, 2019). Therefore, although the performance of AAS for carbon capture is less well-studied, its potential could pave the way towards efficient CCS technologies and its deployment at scale.

1.2. Objectives

The aim of this work is to evaluate the cyclic CO₂ absorption capacity by aqueous AAS. This is achieved by designing a bespoke experimental set-up and validating using 30 mass% aqueous MEA at 40°C, comparing to literature data and modifying the set-up as appropriate.

The equipment can then be used to assess the performance of potassium glycinate (KGly) solution, as an initial AAS of interest. CO₂ absorption over a temperature range of 40°C – 120°C will be investigated as this encompasses both absorption and stripping conditions for PCC. The loading of CO₂ in the solvent is quantified using a thermodynamic model for the gaseous phase.

Glycine was chosen as it is the simplest amino-acid and provides a point of reference from which amino-acids with more complex structures can be studied.

2. Background

2.1. Mechanisms of CO₂ Absorption

CO₂ is captured by the solvent via both physical and chemical absorption. Physical dissolution is dependent on gas solubility within the solvent, temperature and pressure. In chemical absorption, CO₂ undergoes chemical reaction with the solvent (Stewart, 2014). This enables chemical solvents to perform well even at low CO₂ partial pressures, which is important for maximal CO₂ removal from flue gas streams. The overall chemical reaction between CO₂ and an amino-acid salt, or primary or secondary amine, is given by:



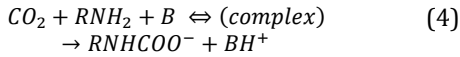
Where, for example, R ≡ CH₂CH₂OH for MEA, and R ≡ CH₂COO⁻ K⁺ for potassium glycinate.

The two main reaction pathways for the absorption of CO₂ by amino-acid salts are the zwitterion and termolecular mechanisms (Vaidya et al., 2010), as described below:

Zwitterion mechanism:



Termolecular mechanism:

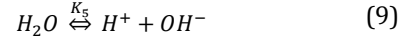
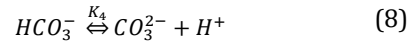
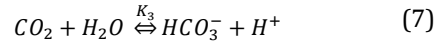
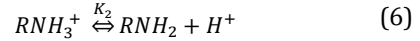


where B denotes a base.

2.2. Thermodynamic Modelling

While the thermodynamic modelling of amine solvents, and specifically aqueous MEA, has been extensively studied, modelling for amino-acid salt solutions is still under development. Important obstacles to this research are the limited range of experimental phase data available and the difficulty in determining the required chemical equilibrium constants (Suleman et al., 2018).

The overall reaction occurring between alkaline amino-acid salts and CO₂ is given by equation (1), however, to develop the framework for a thermodynamic model, the individual chemical reactions occurring in solution must be defined:



The corresponding equilibrium constants are:

$$K_1 = \frac{m_{RNH_2} m_{HCO_3^-} \gamma_{RNH_2} \gamma_{HCO_3^-}}{m_{RNHCOO^-} a_w \gamma_{RNHCOO^-}} \quad (10)$$

$$K_2 = \frac{m_{RNH_2} m_{H^+} \gamma_{RNH_2} \gamma_{H^+}}{m_{RNH_3^+} \gamma_{RNH_3^+}} \quad (11)$$

$$K_3 = \frac{m_{HCO_3^-} m_{H^+} \gamma_{HCO_3^-} \gamma_{H^+}}{m_{CO_2} a_w \gamma_{CO_2}} \quad (12)$$

$$K_4 = \frac{m_{CO_3^{2-}} m_{H^+} \gamma_{CO_3^{2-}} \gamma_{H^+}}{m_{HCO_3^-} \gamma_{HCO_3^-}} \quad (13)$$

$$K_5 = m_{H^+} m_{OH^-} \frac{\gamma_{OH^-} \gamma_{H^+}}{a_w} \quad (14)$$

a_w represents the activity of water, and m_i and γ_i represent the molality and activity coefficient of species i , respectively. An overall equilibrium constant for the reaction given by equation (1), is proposed by (Kumar et al., 2003):

$$K_{ov} = \frac{K_3}{K_1 K_2} \quad (15)$$

To relate the molality of CO₂ with its partial pressure, Henry's law can be employed to describe the vapour-liquid equilibrium:

$$p_{CO_2} = H \cdot m_{CO_2} \quad (16)$$

H is Henry's constant, here defined in terms of molality instead of volume concentration, as volume is temperature-dependent.

Many empirical correlations have been developed to calculate the equilibrium constants for carbamate hydrolysis (K_1) and amine deprotonation (K_2) for different amino-acids as functions of concentration and temperature.

Several models have been developed to model the vapor-liquid equilibrium (VLE) of these systems. Perhaps the simplest, is the Kent-Eisenberg model, which considers the liquid phase as ideal, thus all the activity coefficients are set to unity. This model is widely used on account of its computational simplicity, however, depending on the amino-acid salt, the average deviation of the model can be poor, especially at low CO₂ loadings. Other, more complex models being studied are the UNIQUAC and Deshmukh-Mather models which are based on activity

coefficients and the Electrolyte-NRTL model based on excess Gibbs energy, which relate the activity coefficients to the composition without replacing the overall reaction equation. Additionally, Artificial Neural Network models are also being explored. These complex models often offer a more accurate representation of the absorption phenomena occurring (Ramezani et al., 2022).

2.3. Alkanolamine Solvents

Aqueous alkanolamines have frequently been used in natural gas sweetening: the removal of acidic gases, such as CO₂ and H₂S from gas streams in the refinery industry (Benamor, Aroua & Aroussi, 2012). More recent studies have shown that their applications extend to carbon capture by gas-liquid absorption and is considered the most viable technology for large-scale deployment. The performance of these solvents is evaluated using parameters such as loading, energy consumption and environmental impact. Alkanolamines can be classified by the number of alkyl groups bonded to the amine group's nitrogen atom (Tong et al., 2013).

Monoethanolamine (MEA) is a primary amine with well-documented VLE and thermodynamic data. Aqueous MEA absorbs CO₂ through physical dissolution and chemical reaction. It is used as a benchmark in carbon capture research as it reacts rapidly with CO₂, has high absorption capacity on a mass basis on account of its low molecular weight, and is straightforward to regenerate. Its disadvantages include high energy consumption in the desorption step, corrosivity and loss of solvent at higher vapour pressures (Ma'mun et al., 2005). Here, MEA at 30 mass% and 40°C is used to validate the accuracy and reliability of the experimental set-up.

As reference data for the validation, three sources of literature data in agreement were taken from different time periods with different experimental set-ups. This ensures a high level of accuracy when comparing the MEA absorption results obtained in this study with the literature values (Tong et al., 2012; Aronu et al., 2011; Jou et al., 1995).

Another amine studied for carbon capture purposes is methyldiethanolamine (MDEA), a tertiary amine which captures CO₂ by catalysing the hydrolysis of CO₂. Compared to MEA, its advantages include lower regeneration energy, high resistance to thermal degradation and lower corrosivity (Song et al., 2006). However, reaction rates of tertiary alkanolamines are slower and at low CO₂ partial pressures, absorption capacity is poor (Ma'mun et al., 2005).

2-amino-2-methyl-1-propanol (AMP) is the simplest hindered form of MEA and shows a high potential cyclic absorption capacity (Sartori & Savage, 1983; Xu et al., 1996). However, reaction kinetics are slow with a notably lower extent of reaction (Tong et al., 2013).

Piperazine (PZ) is a diamine with a cyclic structure. A blend of aqueous AMP+PZ was found to have over twice the theoretical CO₂ absorption capacity of MEA at the same mass concentration. The reaction was also over five times faster than AMP alone (Tong et al., 2013). Despite PZ enhancing kinetics by promoting CO₂ mass transfer rates, its drawbacks include high volatility and high viscosity, similar to those of MEA (Freeman et al., 2010). PZ also precipitates at high concentrations and low temperatures, which poses challenges in process operations, flow assurance and safety (Gaspar et al., 2014).

2.4. Aqueous Amino-acid Salt Solvents

Recently, researchers have sought alternative compounds with an amine group for use in CO₂ capture to mitigate against the disadvantages of conventional alkanolamine solvents. One such group of compounds is amino-acid salt solutions, which presents environmental advantages such as synthesis from biological sources and biodegradability, as well as higher surface tension, low volatility and resistance to degradation (Sang Sefidi & Luis, 2019). The performance of these solvents in carbon capture is less studied. The measured VLE parameters from experimentation on a selection of amino-acid salt solutions used for CO₂ absorption is presented in Table 1.

Table 1: Summary of VLE data of amino-acid salt solutions in literature

Source	Amino-acid Salt Solution	VLE Data				Set-up Type
		Temperature Range /K	CO ₂ Pressure Range /MPa	Concentration Range (mol·kg ⁻¹)	Loading Range (mol/mol)	
Song et al., 2006	Sodium glycinate	303.15 – 323.15	0.0001 – 0.2135	1.14 – 4.42	0.170 – 1.075	Analytical
Hamzehie and Najibi, 2016	Potassium glycinate	293.15 – 323.15	0.0051 – 2.5087	0.09 – 0.98	0.885 – 6.948	Synthetic
Chang et al., 2015	Potassium proline	313.15 – 353.15	0.0003 – 0.9286	0.53 – 2.46	0.242 – 1.160	Analytical
Garg et al., 2017	Sodium phenylalaninate	303.15 – 333.15	0.2 – 2.5	0.00 – 0.01	0.164 – 1.750	Synthetic

3. Methods

3.1. Experimental Principle

The set-up can be classified as a synthetic VLE apparatus since the compositions of the co-existing phases were not measured directly. Rather, a mixture of known mass and composition was brought to equilibrium within a vessel of known volume. Using a thermodynamic model for the gas phase, the compositions of both phases were evaluated.

Figure 1 illustrates the experimental set-up used. Vessel E-1 stored compressed CO₂ gas, maintained at 30°C, to reduce effects of fluctuations in the ambient temperature. VLE was established within temperature-controlled vessel E-2. Degassed solvent was added to E-2 using a syringe with a hypodermic needle, through a custom port which accommodated a septum. Vessel E-2 contained a magnetic stirrer, connected to a magnetic stirrer plate on which the vessel rested. After equilibrium was reached in E-2, CO₂ was added stepwise from E-1 by manually controlling flow using V-4.

Using temperature and pressure measurements, as well as the known volume of E-1, the amount of CO₂ contained in the cylinder and how much CO₂ was transferred to E-2, were calculated through the equation of state (Span & Wagner, 1996). Although most of the CO₂ transferred dissolved in the solvent, some remained in the headspace region of E-2, which signified the need for an equation of state for the gas phase. Based on density measurements of the unloaded solvent, pressure and temperature, the equation of state was employed again to calculate the actual quantity of CO₂ dissolved. After each transfer step, the system was allowed to reach equilibrium.

3.2. Apparatus Description

Vessel E-1 was a stainless-steel cylinder (Swagelok 316L-HDF4-150), surrounded by an insulated thick-walled aluminium sleeve. The sleeve was fitted with a low-voltage electric heater pad and a K-type thermocouple to control

the vessel temperature. A platinum resistance thermometer (PRT) was also placed in the sleeve for accurate temperature measurement (TT-1).

Vessel E-2 was a custom-made stainless-steel cell, comprising a main body, a seal-retaining ring and a closure plate. The closure plate was fitted with two threaded fluid ports: 1/8" outer diameter (OD) for the pressure sensor (PT-2) and 1/16" OD to transfer CO₂ into the vessel. The main body had two more ports: one held a plug and the other fitted a custom-made septum holder, through which the degassed solvent was injected into the cell. To control the temperature precisely, E-2 was placed in an insulated, thick-walled aluminium jacket, covering the entire surface of the vessel. The jacket was fitted with four electric cartridge heaters and a PRT for temperature control. The vessel itself housed an axial thermowell, which allowed for an additional PRT to be inserted to measure the vessel temperature (TT-2). The vessel was placed on a magnetic stirrer plate.

Both vessels were fitted with high-precision digital pressure transmitters (Keller, 33X series), with an expanded uncertainty of 0.05%. PT-1 had a range of 0 – 60 bar and PT-2 had a range of 0 – 10 bar. A scroll pump (Edwards XDS35) was used to evacuate the system, which had an ultimate vacuum of 1 mbar.

Other apparatus used included an Anton Paar SVM 3001 density meter and a degassing apparatus comprising an electric heating mantle, a round-bottomed flask, a water-cooled reflux condenser, and a mini-cooler.

3.3. Standard Operating Procedure

Before the solvent can be added to the VLE cell, it must be degassed to remove any CO₂ and other non-condensable gases already dissolved. A 500ml round-bottomed flask was filled with approximately 75ml of solvent and anti-bumping granules. The flask was fitted with a cold-water reflux condenser, and cold water was supplied at 5°C from a mini cooler. At the start of the degassing, a nitrogen line was placed close to the solvent surface, to displace all the air from the flask. After a few minutes the nitrogen line tube was moved to the top of the condenser, to maintain the nitrogen blanket. A heating mantle was used to boil the solvent for 60 minutes, then allowed to cool for 60 – 90 minutes, under continuous nitrogen flow. A 60ml syringe flushed with nitrogen was filled with the degassed solvent. The remaining solvent was used for density measurements.

To begin the VLE experiment, the whole set-up must first be under vacuum conditions. With all the valves open, including V-1 open to the CO₂ supply gas line, the vacuum pump was

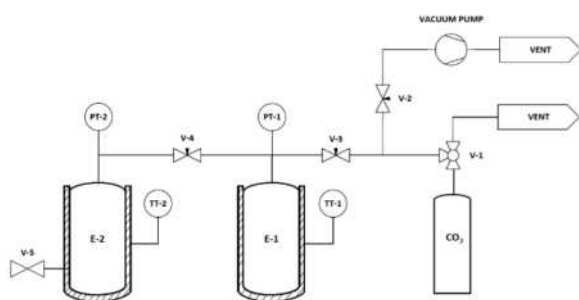


Figure 1: Diagram of VLE apparatus

tuned on until ultimate vacuum was reached. To ensure no contaminants remained in the system, it was first entirely flushed with CO₂. Vessel E-1 was filled with approximately 10 bar of CO₂ from the gas cylinder. Using valve V-4, E-2 was carefully filled with CO₂. Care should be taken to not exceed the maximum allowable pressure reading of 10 bar. Using V-1, the contents of first E-1 and then E-2 were vented to atmosphere until PT-1 and PT-2 read just above atmospheric pressure. The vacuum pump was switched on until ultimate pressure was achieved again. The next step was to fill E-1 with CO₂ to 30-40 bar, then slowly meter CO₂ into E-2 until PT-2 read just above atmospheric pressure. This created a CO₂ blanket in E-2. Using the syringe, fitted with a hypodermic needle, the degassed solvent was transferred into E-2 by piercing the septum. The syringe was weighed before and after the injection to calculate the quantity of solvent added. The amount of CO₂ present in E-2 and which dissolved in the solvent was also calculated. Temperature control was turned on for both E-1 (usually at 30°C) and E-2 (at the temperature under investigation), and the system was allowed to equilibrate. CO₂ was then added in steps of 2 bar from E-1, with the flow manually controlled by V-4. The system was allowed to equilibrate between steps.

3.4. Data Analysis and Uncertainty

The aim of the experiment was to determine how the loading, α , of the solvent varies with the partial pressure of CO₂. The loading is defined as:

$$\alpha = \frac{n_{CO_2}}{n_{S,0}} \quad (17)$$

n_{CO_2} is the moles of dissolved CO₂ and $n_{S,0}$ is the initial moles of unloaded solvent. α can also be reported on a mass basis.

To calculate the amount of CO₂ transferred from E-1 to E-2 at each step, equation (18) is used.

$$n'_{CO_2} = [\rho(T_1, p_{1,i}) - \rho(T_1, p_{1,f})] \left(\frac{V_1 - V_{d,1}}{M_{CO_2}} \right) + [\rho(T_a, p_{1,i}) - \rho(T_a, p_{1,f})] \left(\frac{V_{d,1}}{M_{CO_2}} \right) \quad (18)$$

T_1 , p_1 are the temperature and pressure of the reservoir E-1, respectively. The subscripts i, f are used to denote the initial and final pressure of E-1, before and after each equilibrium step. Here, the volume V_1 is the total volume between valves V-3 and V-4. $V_{d,1}$ is the dead volume of V_1 , i.e., the overhead tubing (including the tubing of pressure transmitter PT-1). This volume was not temperature-controlled, hence why the overhead

temperature associated with E-1, T_a , is defined. This often equals the ambient temperature. To calculate the density of CO₂ at the points defined, the NIST REFPROP function is applied, that uses the equation of state of (Span & Wagner, 1996). M_{CO_2} is the molar mass of CO₂.

The amount of CO₂ transferred to E-2, does not entirely dissolve, with some remaining in the headspace volume. The amount dissolved is given by:

$$n_{CO_2} = n'_{CO_2} - \rho(T_2, p_2) \left(\frac{V_2 - V_s - V_{d,2}}{M_{CO_2}} \right) - \rho(T_a, p_2) \left(\frac{V_{d,2}}{M_{CO_2}} \right) \quad (19)$$

where T_2 , p_2 are the temperature and pressure of the reservoir E-2 respectively. V_2 is the total volume between V-4 and V-5, $V_{d,2}$ is the dead volume (overhead tubing) of E-2 and V_s is the volume of the liquid solvent.

When loading the solvent with CO₂, the density can vary. This change in density is represented by equation (20) as a linear function of a . V_s can be evaluated from equation (21).

$$\rho_s = \rho_{s,0}(1 + b\alpha) \quad (20)$$

$$V_s = \frac{m_{s,0}(1 + M_{CO_2}\alpha)}{\rho_{s,0}(1 + b\alpha)} \quad (21)$$

$\rho_{s,0}$ is the initial solvent density, b is a constant and $m_{s,0}$ is the initial mass of unloaded solvent.

Notably, when the solvent was an aqueous solution of potassium glycinate (KGly), the solution was assumed to be incompressible, hence the value of constant b was set to zero. Density measurement of the loaded solvents at the tested temperatures were not measured, hence b could not be determined experimentally. Aldenkamp et al. (2014) also assumed the incompressibility of KGly solution. Moreover, since the pressure of the VLE cell was kept below 10 bar, the compressibility was neglected.

The partial pressure of CO₂ is given by:

$$p_{CO_2} = p_2 - p_{vap} \quad (22)$$

where p_{vap} is the pressure of the VLE cell at the beginning of the experiment.

To calculate the standard uncertainty for a quantity y , denoted $u(y)$, where y is a function of other measured quantities, the equation for combined variance was used.

$$u^2(y) = \sum_{i=1}^N \left(\frac{\partial f}{\partial x_i} \right)^2 u^2(x_i) \quad (23)$$

$$y = f(x_1, x_2, \dots, x_N) \quad (24)$$

Using equation (23), $u(n_{CO_2})$, $u'(n_{CO_2})$, $u(V_s)$ and finally $u(\alpha)$ can be calculated.

3.5. Validation of VLE apparatus

Before conducting experiments with aqueous amino-acid solutions, it was imperative to first validate the bespoke set-up. CO₂ loading of 30 mass% aqueous MEA at 40°C was chosen, since this solvent is widely studied in literature.

In this work, although originally not the main objective, validating the operation of the system became such. Experimental results did not agree with literature initially (Aronu et al., 2011; Jou et al., 1995; Tong et al., 2012), hence modifications to the set-up and experimental procedure were made as needed. The experiment was repeated for each change.

3.5.1. Experimental Modifications

An early observation was that the vapour pressure of CO₂ was higher than expected according to literature, for all loadings. This suggested that impurities remained in the system. The most significant change was altering the SOP to include the CO₂ blanket. Initially, the SOP detailed injecting the sample into a nitrogen-filled E-2. A vacuum pump was then used to remove this nitrogen after the solvent is introduced. The downside to this was that a small amount of nitrogen remained in the system, but also, some of the solvent would evaporate due to the vacuum conditions. Therefore, filling E-2 with CO₂ eliminates the presence of nitrogen gas as an impurity and enables an accurate recording of the solvent mass.

Other changes to the SOP also aimed to remove impurities. For example, a higher performance scroll pump replaced the original two-stage diaphragm pump, reducing the ultimate vacuum from 4 mbar to 1 mbar. Moreover, during the degassing stage, the solvent was boiled more strongly and for a longer duration (from 30 minutes to 60 minutes) to ensure that any dissolved CO₂ and other non-condensable gases would be removed. The temperature of the water used in the reflux condenser was also reduced to 5°C using a mini cooler, having used ambient temperature water prior to this. Additionally, to remove doubt of a contaminated gas supply, new research-grade CO₂ and N₂ cylinders were tested and used.

All aforementioned modifications had a positive impact that brought the experimental data in closer agreement with literature data on CO₂ absorption by MEA, but some disparity still remained. The purity of the MEA solution was investigated. Through titration, the purity was found to be 28.8 mass%, rather than the assumed

30 mass%. A new solution was prepared, using higher purity MEA (Sigma Aldrich 99.5 mass%), and the concentration was verified through titration to be 29.9 mass%.

The final modifications included altering the experimental apparatus. Poor thermal contact between the reservoir and thermostatic sleeve, and heat loss to the environment due to insufficient insulation, were suspected. This would result in the temperature of the vessel being lower than the controlled temperature of the sleeve. Therefore, temperature control of the gas reservoir E-1 was investigated. Even though the dead volume in the headspace was negligible in comparison to the gas reservoir E-1, insulation was placed to cover all tubing. Moreover, to improve contact between the vessel and the aluminium sleeve, liquid dodecane was added as a thermal compound to improve heat transfer.

Although dead space volume calibration had already been performed, an additional calibration was also carried out to verify the results. High-pressure CO₂ was used to fill the vessels E-1 and E-2 and associated tubing separately. The vessels were weighed before and after filling using a high-precision scale (uncertainty of ± 0.001 g). Using the equation of state (Span & Wagner, 1996), the density was found, and the volume was calculated.

The outcomes of the previous volume calibration and the second calibration described in this work are shown in Table 2.

Table 2: Comparison of apparatus volume calibrations /ml

	Gas reservoir			VLE cell		
	V _{total}	V _{cell}	V _{tubing}	V _{total}	V _{cell}	V _{tubing}
Calibration with water	149.193	146.281	2.912	79.420	77.920	1.500
Calibration with CO ₂	150.015	148.271	1.744	78.554	77.699	0.855

3.5.2. E-1 Temperature Sensitivity Analysis

A sensitivity analysis was performed to investigate discrepancies between the actual temperature of the CO₂ in the gas reservoir and the temperature of the aluminium sleeve. The system was allowed to remain at 30°C for several days, therefore it was assumed that the CO₂ reached thermal equilibrium. The temperature of the sleeve (TT-1) was then increased in steps up to 70°C, and the system was then allowed to equilibrate for 30 minutes between steps, which was representative of the time needed for the VLE cell (E-2) to reach equilibrium. The total amount of CO₂ contained between V-3 and V-4 was calculated. Using the pressure, volume and the density through the equation of state, the actual temperature of the reservoir E-1 could be calculated.

3.6. Potassium Glycinate Experiments

The pressure transmitter PT-2 is rated for only 10 bar with an uncertainty of 0.05%, or 5 mbar. This error is significant for pressures below 10 mbar. In this range, it was initially thought that the results from the MEA experiment agreed with the literature data. Hence the system was thought validated and experiments with KGly were performed.

For these experiments, a batch solution of KGly with a molality of $1.930 \text{ mol} \cdot \text{kg}^{-1}$ was prepared. Three experiments were conducted, using the same experimental apparatus and SOP as the MEA. The temperature of the VLE cell (T_2) was varied, with experiments carried out at 40°C, 90°C and 120°C. The rationale behind the choice of temperatures, was to replicate the conditions of a carbon capture plant. An absorption column would operate at 40°C, and a stripping column at 120°C. To explore the effects of temperature on CO_2 loading, 90°C was chosen as an intermediate temperature.

4. Results

4.1. MEA Validation Results

Figure 2 illustrates the experimental results for the VLE apparatus validation experiments with MEA. The partial pressure of CO_2 is plotted against the solvent loading, for both an initial experiment prior to the modifications detailed in section 3.5.1, and a subsequent experiment after the modifications to the SOP. The results are compared with experimental values from literature. The effects of implementing changes can be clearly observed between the two experiments. Moreover, the size of the error bars demonstrates how the uncertainty decreases with increasing partial pressure of CO_2 .

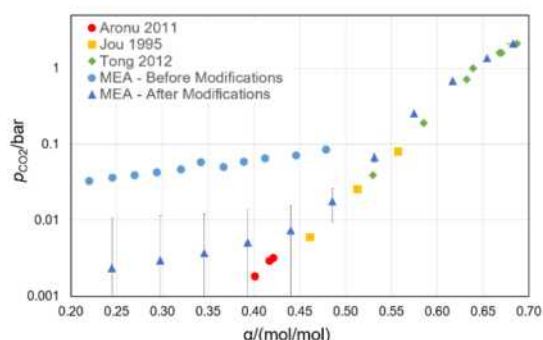


Figure 2: Semi-log plot of the solubility of CO_2 in 30 mass% MEA solution at 40°C, before and after modifications - comparison with literature

Figure 3 again displays the experimental results for MEA, at high partial pressures of CO_2 . The values plotted are the final results obtained, after applying the modifications to the VLE apparatus and SOP.

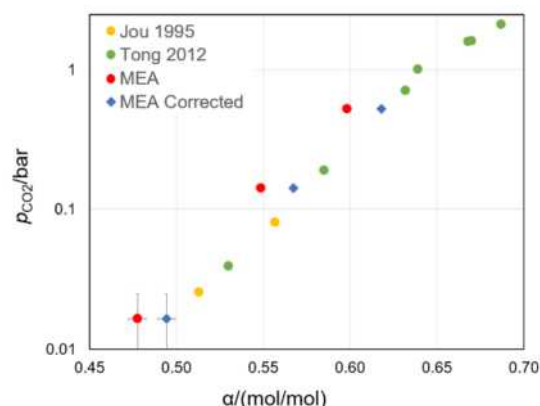


Figure 3: Semi-log plot of the solubility of CO_2 in 30 mass% MEA solution at 40°C, before and after temperature correction - comparison with literature

After the sensitivity analysis on the temperature of the gas reservoir E-1 was performed, it was found that the actual temperature of the CO_2 and the measurement from the PRT in the aluminium sleeve (TT-1) showed a discrepancy at all temperature setpoints investigated. For a 2 bar increase in the gas reservoir pressure (the size of the pressure steps usually made in loading experiments), the temperature discrepancy was approximately 2°C. Hence, CO_2 partial pressure against solvent loading is plotted again in Figure 3, with a correction for temperature correction. Through trial and error, it was found that a temperature reduction of 3°C in T_1 , shifts the data in alignment with the literature.

4.2. Sensitivity Analysis Results

The results of the sensitivity analysis are illustrated in Figure 4. The temperature of the CO_2 , initially at 30°C, was raised in steps, with the measured (assumed) temperature of the vessel E-1 labelled on the x-axis. For each step, the actual temperature of CO_2 was calculated and the difference between the measured and calculated values is plotted on the y-axis.

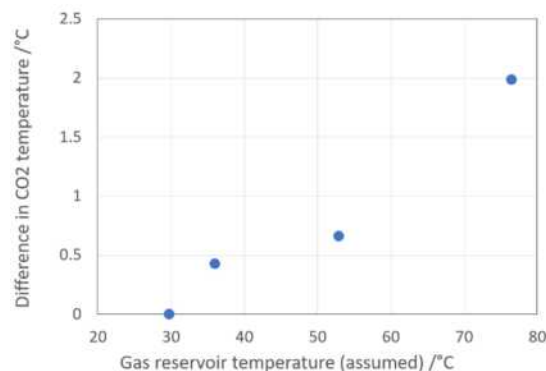


Figure 4: Sensitivity analysis of the CO_2 gas reservoir temperature (T_1)

4.3. Potassium Glycinate Results

Figure 5 illustrates the VLE results for the loading experiments of KGly at 40°C, 90°C and 120°C. The trend of how the partial pressure of CO₂ varies with loading at different temperatures is in line with expectations. At low temperatures, the loading capacity is greater at each partial pressure. To interpolate between data points, a smoothing function was used, given by:

$$\ln(p_{\text{CO}_2}) = A_0 + A_1\alpha + A_2\alpha^2 + A_3\alpha^3 \quad (25)$$

where the parameters, A_i , were found using a least squares regression. A_i and the standard deviation, σ , of the fit are given in Table 6.

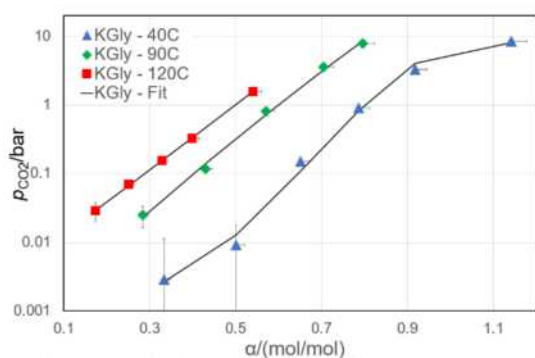


Figure 5: Semi-log plot of the solubility of CO₂ in 1.930 mol · kg⁻¹ aqueous KGly at 40°C, 90°C and 120°C.

It should be noted that there is an uncertainty regarding the temperature of the CO₂ gas reservoir E-1, as was the case for MEA. This will only affect the values of α at each partial pressure measured. To account for this error, the loadings were re-calculated for a temperature correction of 3°C (the same as the MEA temperature correction), and the difference was added as a positive horizontal error bar. The results of the experiments are reported in Tables 3-5.

Table 3: VLE data for 1.930 mol · kg⁻¹ KGly solution at 40°C

p _{CO2}	α	u(p _{CO2})	u(α)
bar	mol/mol	bar	mol/mol
0.00001	0.193	0.00849	0.010
0.00288	0.333	0.00849	0.014
0.00915	0.501	0.00849	0.018
0.15060	0.651	0.00849	0.021
0.90514	0.787	0.00849	0.024
3.29893	0.918	0.00849	0.028
8.45072	1.142	0.00849	0.035

Table 4: VLE data for 1.930 mol · kg⁻¹ KGly solution at 90°C

p _{CO2}	α	u(p _{CO2})	u(α)
bar	mol/mol	bar	mol/mol
0.00001	0.174	0.00849	0.009

0.02479	0.283	0.00849	0.012
0.11628	0.431	0.00849	0.016
0.80640	0.571	0.00849	0.020
3.54932	0.704	0.00849	0.023
7.84032	0.796	0.00849	0.027

Table 5: VLE data for 1.930 mol · kg⁻¹ KGly solution at 120°C

p _{CO2}	α	u(p _{CO2})	u(α)
bar	mol/mol	bar	mol/mol
0.00001	0.096	0.00849	0.007
0.02875	0.174	0.00849	0.009
0.07025	0.252	0.00849	0.011
0.15360	0.329	0.00849	0.013
0.32460	0.400	0.00849	0.015
1.56994	0.541	0.00849	0.019

Table 6: Parameters of equation (25)

T/°C	A ₀	A ₁	A ₂	A ₃	σ /bar
40	-30.228	60.582	-25.161	-3.166	0.351
90	-2.112	13.820	-7.519	-	0.341
120	0.543	10.457	-5.368	-	0.007

5. Discussion

5.1. Apparatus Validation with MEA

The series of experimental modifications undertaken through the course of this investigation were crucial in making the apparatus viable to study CO₂ absorption experimentally.

The most significant change was introducing the solvent into CO₂ rather than N₂. The presence of nitrogen as an impurity had caused an increase in the observed partial pressure of CO₂ with each addition of CO₂ into the VLE cell, which led to the error compounding with an increasing number of data points. This can be observed in the data before modifications being greatly shifted left and in poor agreement with the trend found in literature, shown in Figure 2. The elimination of this source of impurity increased the accuracy of the expected data closer to the results observed in literature.

The investigation of MEA purity by titration showed a notable discrepancy between the target concentration and the actual concentration. Adjusting the MEA purity also led to a significant improvement of the collected experimental data with that of literature.

The additional volume calibration confirmed that the volume of E-2 was close to the original calculation, done by calibration using water, and was useful to ensure that the vessel's volume was validated by a second measurement. The introduction of more rigorous degassing is good practice to rule out imperfect degassing as a source of error. However, these modifications did not make as significant an improvement in

the data's agreement with literature as the two aforementioned changes.

The final modification which is yet to be made is improved temperature control of E-1. As shown by the temperature sensitivity analysis, the temperature of the CO₂ in the reservoir does not match that of the aluminium sleeve around it, and an accurately quantified discrepancy remains unknown.

The discrepancy then leads to an inaccurate calculation of the amount of CO₂ dissolved, based on erroneous temperature and pressure data. This is the likely source of the remaining deviation from the literature values, as the temperature-adjusted data in Figure 3 aligns with the expected trend.

This source of error must first be corrected by further modification to the apparatus or SOP, or by quantifying the temperature discrepancy and accounting for it in calculations. Only when the experimental set-up can be verified as valid can it be used to collect viable data on AAS absorption of CO₂.

5.2. KGly Experiments

In the literature, there are no references for potassium glycinate at the concentration and temperatures tested in this work. While this means it difficult to judge the validity of the measurements obtained, the apparatus is close to satisfactory validation when a plausible adjustment of the reservoir temperature is considered. The VLE data collected on KGly can tentatively be taken as valid with the temperature adjustment applied, but further experimentation should be conducted once sufficient modifications are made to accurately quantify this temperature adjustment and confirm the accuracy of the results.

The results show that the error is greater when the VLE cell is at lower temperatures. The error also increases throughout the progression of the experiment. However, even with a 3°C temperature correction, the measurement uncertainty is relatively small.

6. Conclusions

An experimental set-up has been designed to evaluate the cyclic absorption capacity of amine-based solvents for carbon capture through the collection of high accuracy VLE data. Two such solvents studied in this work are aqueous MEA, an alkanolamine, and aqueous potassium glycinate, an amino-acid.

The design is in its early stages and underwent several modifications to improve the accuracy of the data collected. The apparatus' validity was tested using 30 mass% aqueous

MEA solution, a well-studied CO₂ absorbent, and the results were compared to three literature data sources. The set-up is currently not in sufficient agreement with literature data to be deemed valid for use in investigating the CO₂ absorption capacity of AAS solutions.

However, assessing the impact of temperature discrepancy in the CO₂ reservoir shows strong indication that it is the remaining source of error. This can be corrected for in future modifications to the method. Possible solutions include implementing a more complete thermal enclosure, such as investigating the addition of a more effective thermal contact liquid between vessel E-1 and the sleeve around it and improving insulation over the sleeve, fittings and tubing. An alternative approach may be to quantify the amount of CO₂ in the vessel through calculations independent of temperature, such as by taking mass readings.

Further testing should be carried to confirm that this temperature discrepancy is the key source of error. Once corrected by the modifications suggested and shown to be valid with MEA, the apparatus can be used to achieve the objective of studying the cyclic absorption capacities of amino-acid salt solutions.

The validated set-up will be able to contribute to the understudied area of alternative carbon capture solvents.

7. Acknowledgements

We would like to thank Hossam Qusty for his guidance and support throughout this project.

8. References

- Aldenkamp, N., Huttenhuis, P., Penders-van Elk, N., Steinseth Hamborg, E. & Versteeg, G. F. (2014) Solubility of Carbon Dioxide in Aqueous Potassium Salts of Glycine and Taurine at Absorber and Desorber Conditions. *Journal of Chemical & Engineering Data*. 59 (11), 3397-3406. 10.1021/je500305u.
- Aronu, U. E., Gondal, S., Hessen, E. T., Haug-Warberg, T., Hartono, A., Hoff, K. A. & Svendsen, H. F. (2011) Solubility of CO₂ in 15, 30, 45 and 60 mass% MEA from 40 to 120°C and model representation using the extended UNIQUAC framework. *Chemical Engineering Science*. 66 (24), 6393-6406. 10.1016/j.ces.2011.08.042.
- Benamor, A., Aroua, M. K. & Aroussi, A. (2012) Kinetics of CO₂ Absorption Into Aqueous Blends of Diethanolamine and Methyldiethanolamine. In: Aroussi, A. & Benyahia, F. (eds.). *Proceedings of the 3rd Gas Processing Symposium*. Oxford, Elsevier. pp. 64-70.

- Freeman, S. A., Dugas, R., Van Wagener, D. H., Nguyen, T. & Rochelle, G. T. (2010) Carbon dioxide capture with concentrated, aqueous piperazine. *International Journal of Greenhouse Gas Control*. 4 (2), 119-124. 10.1016/j.ijggc.2009.10.008.
- Gaspar, J., Thomsen, K., von Solms, N. & Fosbøl, P. L. (2014) Solid Formation in Piperazine Rate-based Simulation. *Energy Procedia*. 63 1074-1083. 10.1016/j.egypro.2014.11.115.
- Jou, F., Mather, A. E. & Otto, F. D. (1995) The solubility of CO₂ in a 30 mass percent monoethanolamine solution. *The Canadian Journal of Chemical Engineering*. 73 (1), 140-147. 10.1002/cjce.5450730116.
- Kumar, P. S., Hogendoorn, J. A., Timmer, S. J., Feron, P. H. M. & Versteeg, G. F. (2003) Equilibrium Solubility of CO₂ in Aqueous Potassium Taurate Solutions: Part 2. Experimental VLE Data and Model. *Industrial & Engineering Chemistry Research*. 42 (12), 2841-2852. 10.1021/ie020601u.
- Ma'mun, S., Nilsen, R., Svendsen, H. F. & Juliussen, O. (2005) Solubility of Carbon Dioxide in 30 mass % Monoethanolamine and 50 mass % Methyl-diethanolamine Solutions. *Journal of Chemical & Engineering Data*. 50 (2), 630-634. 10.1021/je0496490.
- Padurean, A., Cormos, C., Cormos, A. & Agachi, P. (2011) Multicriterial analysis of post-combustion carbon dioxide capture using alkanolamines. *International Journal of Greenhouse Gas Control*. 5 (4), 676-685. 10.1016/j.ijggc.2011.02.001.
- Ramezani, R., Mazinani, S. & Di Felice, R. (2022) State-of-the-art of CO₂ capture with amino acid salt solutions. 38 (3), 273-299. 10.1515/revce-2020-0012.
- Sang Sefidi, V. & Luis, P. (2019) Advanced Amino Acid-Based Technologies for CO₂ Capture: A Review. *Industrial & Engineering Chemistry Research*. 58 (44), 20181-20194. 10.1021/acs.iecr.9b01793.
- Sartori, G. & Savage, D. W. (1983) Sterically hindered amines for carbon dioxide removal from gases. *Industrial & Engineering Chemistry Fundamentals*. 22 (2), 239-249. 10.1021/i100010a016.
- Song, H., Lee, S., Maken, S., Park, J. & Park, J. (2006) Solubilities of carbon dioxide in aqueous solutions of sodium glycinate. *Fluid Phase Equilibria*. 246 (1), 1-5. 10.1016/j.fluid.2006.05.012.
- Span, R. & Wagner, W. (1996) A New Equation of State for Carbon Dioxide Covering the Fluid Region from the Triple-Point Temperature to 1100 K at Pressures up to 800 MPa. *Journal of Physical and Chemical Reference Data*. 25 (6), 1509-1596. 10.1063/1.555991.
- Stewart, M. I. (2014) Chapter Nine - Gas Sweetening. In: Stewart, M. I. (ed.). *Surface Production Operations (Third Edition)*. Boston, Gulf Professional Publishing. pp. 433-539.
- Suleman, H., Maulud, A. S. & Syalsabila, A. (2018) Thermodynamic modelling of carbon dioxide solubility in aqueous amino acid salt solutions and their blends with alkanolamines. *Journal of CO₂ Utilization*. 26 336-349. 10.1016/j.jcou.2018.05.014.
- Tong, D., Maitland, G. C., Trusler, M. J. P. & Fennell, P. S. (2013) Solubility of carbon dioxide in aqueous blends of 2-amino-2-methyl-1-propanol and piperazine. *Chemical Engineering Science*. 101 851-864. 10.1016/j.ces.2013.05.034.
- Tong, D., Trusler, J. P. M., Maitland, G. C., Gibbins, J. & Fennell, P. S. (2012) Solubility of carbon dioxide in aqueous solution of monoethanolamine or 2-amino-2-methyl-1-propanol: Experimental measurements and modelling. *International Journal of Greenhouse Gas Control*. 6 37-47. 10.1016/j.ijggc.2011.11.005.
- UNFCCC. (2023) *UNFCCC The Paris Agreement*. <https://unfccc.int/process-and-meetings/the-paris-agreement>.
- Vaidya, P. D., Konduru, P., Vaidyanathan, M. & Kenig, E. Y. (2010) Kinetics of Carbon Dioxide Removal by Aqueous Alkaline Amino Acid Salts. *Industrial & Engineering Chemistry Research*. 49 (21), 11067-11072. 10.1021/ie100224f.
- Wilberforce, T., Olabi, A. G., Sayed, E. T., Elsaid, K. & Abdelkareem, M. A. (2021) Progress in carbon capture technologies. *Science of the Total Environment*. 761 143203. 10.1016/j.scitotenv.2020.143203.
- Xu, S., Wang, Y., Otto, F. D. & Mather, A. E. (1996) Kinetics of the reaction of carbon dioxide with 2-amino-2-methyl-1-propanol solutions. *Chemical Engineering Science*. 51 (6), 841-850. 10.1016/0009-2509(95)00327-4.

Recovery of Materials (Li, Mn, Ni, Co) from Lithium-Ion Battery Cathode

Weng Hin Hong and Li Xun Khor

Department of Chemical Engineering, Imperial College London, London SW7 2AZ, UK

Abstract

The global market of lithium-ion battery (LiB) was fuelled by the growing demand of the consumer electronics and electric vehicle. Therefore, investigation of scaling-up of battery recycling was necessary. In this field of study, researchers aimed to investigate the electrochemical kinetics and thermodynamics of the prevalent cathode materials in LiBs, namely NMC 111, NMC 622 and NMC 811, in the presence of Al^{3+} and Cu^{2+} impurities. This study emphasised the recovery of Ni^{2+} and Co^{2+} through the application of cyclic voltammetry into electrodeposition experiments. From the cyclic voltammetry tests, neutral condition and glassy carbon as the working electrode were found to be able of minimising the impact of hydrogen reduction (side reaction). Additionally, reduction of Ni^{2+} and Co^{2+} to their metallic states both occurred at an electrode potential at ca. -0.7 V vs. Ag/AgCl in the solutions of NMC 111 and NMC 811 containing Al^{3+} and Cu^{2+} impurities, where separation of Co and Ni was not feasible via electrodeposition on a carbon paper. Cu can be separated through electrodeposition at -0.1 V vs. Ag/AgCl whilst Al was concluded to have a negligible effect on Ni and Co recovery.

Keyword: Cyclic Voltammetry, Electrodeposition, Pourbaix diagram, Lithium-ion battery recycling

1. Introduction

The increasing demand for lithium-ion batteries, as one of the irreplaceable components for electric vehicles (EVs) and renewable energy storage, has brought a substantial growth in future demand, sales of electric vehicles (EVs) reached the highest value of 10 million in 2020, and it was estimated that 230 million EVs will be circulating by 2030 [1]. The main components include anode, cathode, electrolyte, current collector and separator, where the anode was typically made of graphite and the most common cathode material was composed of metal oxides, such as lithium cobalt oxide (LiCoO_2), lithium manganese oxide (LiMn_2O_4) and lithium nickel manganese cobalt oxide (LiNiMnCoO_2 or NMC) [2]. NMC made up 30% of the global demand for cathode materials in all applications in 2017. Among the NMC compositions, $\text{LiNi}_{1/3}\text{Mn}_{1/3}\text{Co}_{1/3}\text{O}_2$ (NMC111) was the most in-demand dominant market, accounting for about 50% of the total NMC demand [3]. The recycling of cathode materials in spent lithium-ion batteries, with a particular emphasis on recovering nickel, cobalt and manganese, is a significant area of focus in the renewable battery industry.

Various methods were commonly employed for recycling battery, the common approaches are

pyrometallurgy, hydrometallurgy and direct recycling [4]. Hydrometallurgical leaching was the most used industrial technique to recycle Ni^{2+} and Co^{2+} in LiB due to its efficiency, simplicity, and ability to produce high-quality recycled metals [5]. On the other hand, pyrometallurgy and direct recycling were less discovered due to its high energy intensive and scale-up issue [4][6]. It is noteworthy that the lithium-ion battery recycling market was estimated at \$4.2 billion in 2022, with an expected growth rate of 21.43% by 2030. The collection of 108,000 tonnes of used portable batteries as recyclable waste in 2021 according to the European Union [7], highlighting the significance of implementing and improving the recycling processes for spent lithium-ion batteries.

The primary objective of this study was to investigate the recovery of valuable materials from spent lithium-ion batteries through electrodeposition with a specific focus on widely used cathodic materials such as lithium, nickel, cobalt and manganese. This study involved the use of glassy carbon and platinum as working electrodes, along with different composition solutions such as NMC 111 ($\text{LiNi}_{0.33}\text{Mn}_{0.33}\text{Co}_{0.33}\text{O}_2$), NMC622 ($\text{LiNi}_{0.6}\text{Co}_{0.2}\text{Mn}_{0.2}\text{O}_2$) and NMC811 ($\text{LiNi}_{0.8}\text{Co}_{0.1}\text{Mn}_{0.1}\text{O}_2$) at pH3 and pH7 during cyclic voltammetry (CV) tests. A key objective of this

research is to mitigate the impact of the hydrogen evolution reaction (HER) during the electrodeposition for nickel and cobalt recovery. This strategic approach enables the enhancement of the recovery efficiency of nickel and cobalt metal, in order to obtain a holistic overview of the recycling performance of spent lithium-ion batteries.

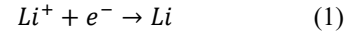
2. Background

Building up upon the prior research led by Ms. Yutong Ji and Dr. Xiaochu Wei, which primarily addressed the recovery of manganese through electrodeposition on the anode, the current investigation extends its scope. This research emphasised broadening the understanding of nickel and cobalt recovery by implementing electrodeposition on the cathode through negative potential. Electrodeposition is a significant technique for recycling and recovering nickel, cobalt coatings from spent lithium-ion batteries (LiBs), it is an environmentally friendly technique due to its high recovery efficiency for valuable metals, selectivity of recovery for specific metals, minimal waste production and low environmental impact [8]. Those ions can be oxidised or reduced on carbon paper, depending on the amount of potential applied to the solutions and the nature of thermodynamic properties shown in a Pourbaix diagram.

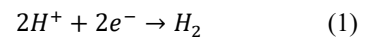
The charge and current response from electrodeposition are influenced by several factors such as temperature, composition of solutions, deposition time and pH [9]. The optimal condition of electrodeposition, where valuable metal ions can be oxidised or reduced in aqueous solutions, can be obtained by implementing Pourbaix diagrams and cyclic voltammetry tests. These contain graphical representation of the thermodynamic stability and kinetics of electrochemical reactions [10], a comprehensive analysis of electrodeposition process can therefore be obtained, this allows for the optimisation of the process parameter, such as the applied potential and pH, to maximise the efficiency of metal ion recovery from spent lithium-ion batteries.

2.1 Electrodeposition on Li

Electrodeposition of lithium involves the reduction of Li^+ ion to metallic lithium, as represented by the following equation:



The standard half-cell potential of Li^+ reduction is -3.3 V vs the standard hydrogen electrode (SHE). On the other hand, the hydrogen evolution reaction (HER), with an electrode potential of -0.6 V [11], is thermodynamically dominant relative to lithium reduction, which is described as the following reaction:



2.2 Electrodeposition on Ni and Co

Electrodeposition of nickel and cobalt mainly depends on the applied potential at which its reduction and oxidation occur. Therefore, Pourbaix diagram is necessary to obtain a holistic electrochemical stability analysis for different redox states as a function of pH. In this study, reduction of Ni^{2+} and Co^{2+} to Ni and Co were investigated, which are given in the following equations:

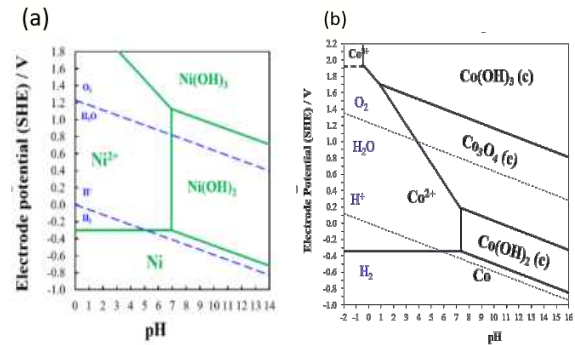
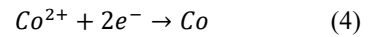
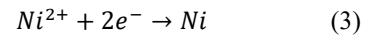


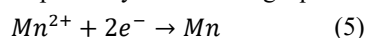
Figure 1: (a) Pourbaix diagram (nickel/concentration: 0.01 mol dm^{-3} , 298.15K, 1.0 bar) (a) nickel-water solution. (b) cobalt-water solution [11]

During cyclic voltammetry measurements, an Ag/AgCl reference electrode was used, and solutions were tested under pH7 and pH3. The Pourbaix diagrams in Figure 1a and 1b indicate the standard half-cell potential against SHE of the two reactions, which are -0.25 V and -0.28 V under neutral condition. However, it is worth noting that an Ag/AgCl reference was used in the experiment, the standard half-cell potential against

Ag/AgCl are corresponding to -0.987 V and -1.017 V respectively. Since the electrodeposition potential of Ni^{2+} and Co^{2+} reductions remain constant approximately against pH, solutions with neutral conditions primarily serve to make the reduction of metal ions more prevalent than HER according to Figure 1, this hypothesis was further validated by the research conducted by Nicolas Dubouis [12].

2.3 Electrodeposition on Mn

The electrodeposition of manganese via a negative potential involves the reduction of Mn^{2+} ion to manganese metal, as depicted by the following equation:



The Pourbaix diagram of manganese reveals the electrode potential associated with the Eq.5 is -1.2 V vs. SHE. under neutral condition [11]. Hence, this electrochemical reduction is thermodynamically unfavourable relative to hydrogen evolution reaction.

3. Methodology

3.1. Materials

Lithium sulphate monohydrate, nickel (II) sulphate hexahydrate, cobalt (II) sulphate heptahydrate, manganese (II) sulphate monohydrate, sulphuric acid (95%), copper (II) sulphate pentahydrate, anhydrous aluminium sulphate was used.

3.2 Cyclic Voltammetry Test

The cyclic voltammetry (CV) test is a vital technique in the study of electrochemical reactions, where the redox processes inherent to materials. This approach provided the characteristic current responses by the material that was investigated when a range of applied potential across an electrochemical cell in a cyclic variation. The purpose of utilising CV test in this study was to identify the redox reaction of nickel, cobalt and manganese at an applied potential.

3.3 Experimental Setup and Solutions Preparation

In this study, different electrolyte solutions were investigated, as shown in Table 1, which contains NMC111, 622, 811 and addition of impurities. The

impact of pH on cyclic voltammetry results was also examined by using solutions at pH7 and pH3, prepared with 0.1 M H_2SO_4 which was diluted from 95% sulphuric acid. All solutions were prepared with deionised water to ensure analytical purity.

Table 1: Solution Composition of Different Solution in Ambient Conditions

Solution category	Solute Concentration
NMC 111	$\text{Li}_2\text{SO}_4 = 0.25 \text{ M}$, $\text{NiSO}_4 = 0.01 \text{ M}$, $\text{MnSO}_4 = 0.01 \text{ M}$, $\text{CoSO}_4 = 0.01 \text{ M}$
NMC 622	$\text{Li}_2\text{SO}_4 = 0.025 \text{ M}$, $\text{NiSO}_4 = 0.03 \text{ M}$, $\text{MnSO}_4 = 0.01 \text{ M}$, $\text{CoSO}_4 = 0.01 \text{ M}$
NMC 811	$\text{Li}_2\text{SO}_4 = 0.05 \text{ M}$, $\text{NiSO}_4 = 0.08 \text{ M}$, $\text{MnSO}_4 = 0.01 \text{ M}$, $\text{CoSO}_4 = 0.01 \text{ M}$
Impurities Test	$\text{Al}_2(\text{SO}_4)_3 = 0.005 \text{ M}$, $\text{CuSO}_4 = 0.000125 \text{ M}$

A 100 mL glass beaker filled with a 50 mL electrolyte solution was used in each measurement, it acted as a conductive medium for redox reactions during the cyclic voltammetry test. The electrochemical cell configuration included glassy carbon (diameter = 3 mm) or platinum (diameter = 1.6 mm) working electrodes, an Ag/AgCl reference electrode and a platinum counter electrode (13 mm x 10 mm). Platinum was selected for its properties of being highly electrically conductive, have a short response time and high chemical stability while resistant to corrosion [13]. However, one drawback could be its relatively small surface area. This resulted in a limitation that suppressed the growth of layer deposition on the electrode [13]. Additionally, the glassy carbon electrode was used with its large surface area and high hardness and smoothness, which contributes to its durability [14,15]. A pre-treatment through polishing with silicon carbide (SiC) paper and polishing powders of alumina (1 μm , 0.3 μm and 0.05 μm) was done on the working electrodes. While the counter electrode was polished using silicon carbide paper with 1 μm polishing powders, and the reference electrode was rinsed with deionised water before each measurement. The potentiostat (Metrohm AUTOLAB

PGSTAT) interfaced with Nova software was utilised for real-time monitoring of the working electrode's current and charge responses. A cyclic voltammetry scan, starting from 0 V to -1.2 V, followed by -1.2 V to 0.2 V, and returning to 0 V in a loop, was executed at a scan rate of 50 mV/s for five consecutive scans.

3.4 Electrodeposition Experiment – High Recovery

Nickel and cobalt, as primary materials being investigated, were recovered via the electrodeposition. Formation of side-products and recovery efficiency of nickel and cobalt were examined during the electrodeposition. Experiments were conducted with NMC 111 and NMC 811, containing aluminium and copper impurities under neutral conditions.

The experiment procedures were performed using 50 mL of solutions in a 100 mL beaker. A reference electrode of Ag/AgCl, working and counter electrode composed of thermally pre-treated carbon papers were used. Carbon paper, configured in a rectangular shape were prepared with an area of 7.5 cm² (2.5 cm x 3 cm) and an attached copper tape. An insulated tape was attached on the interconnecting part of the copper tape and carbon paper, resulting in an effective surface area of 5 cm² (2.5 cm x 2 cm). The electrodeposition experiment was conducted using Nova software with potentiostat (Metrohm AUTOLAB PGSTAT), the electrodeposition was operated under a constant potential setting to achieve a 100% recovery of nickel and cobalt, assuming reduction of nickel and cobalt were the only cathodic products. However, side reactions, such as HER, could be thermodynamic feasible under that applied potential. Hence, an actual recovery efficiency of the Ni and Co were estimated.

Eq.6 and 7 below outlined the charge of the nickel and cobalt reduction and the actual recovery efficiency respectively:

$$Q_{total} = n_e F V ([Ni^{2+}] + [Co^{2+}]) \quad (6)$$

$$\eta = \frac{FV(\Delta[Ni^{2+}]_{t_0 \rightarrow t} + \Delta[Co^{2+}]_{t_0 \rightarrow t})}{Q_{total}} \quad (7)$$

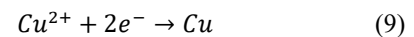
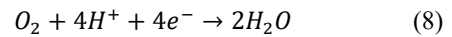
Where Q is the total charge, n_e is the number of electron, F is the Faraday constant, $[Ni^{2+}]$ and $[Co^{2+}]$ are concentration, η is the efficiency of nickel and cobalt recovery.

3.5 Electrodeposition Product Characterisation

Inductively coupled plasma mass spectrometry (ICP-MS) and UV-Vis spectroscopy was the intended method for analysing and quantifying the concentrations of nickel and cobalt recovered on the carbon papers after electrodeposition. However, due to the malfunction of ICP-MS and UV-Vis spectroscopy, further study on the concentration of nickel and cobalt recovered after electrodeposition is required.

4. Results for CV test

Graph of current density against applied potential vs. Ag/AgCl was plotted to visualise when redox reaction occurs. Optimal point at the negative range of current density (the local minimum point) represents a reduction reaction and optimal point at the positive range in current density (the local maximum point) represents an oxidation reaction. CV analysis for cathode in NMC was complicated as Ni^{2+} , Co^{2+} and H^+ reduce at similar range of potential as shown in the Figure 1. This makes the smaller peak in the graph difficult to be distinguished when one reaction dominates the other. Peak in graph were labelled as $P_{A,x}$ for anodic peak and $P_{C,x}$ for cathodic peak. The remaining reaction that was not mentioned above were described as follows:



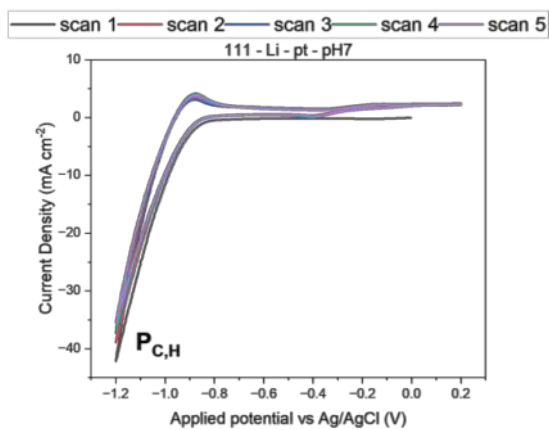


Figure 2

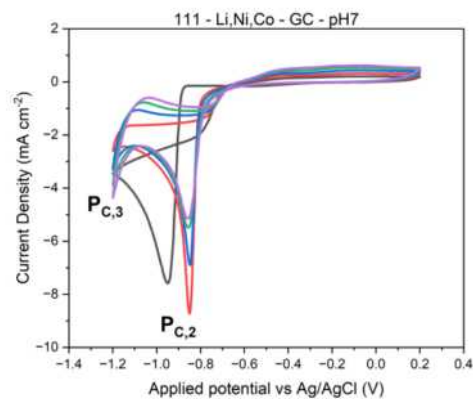


Figure 6

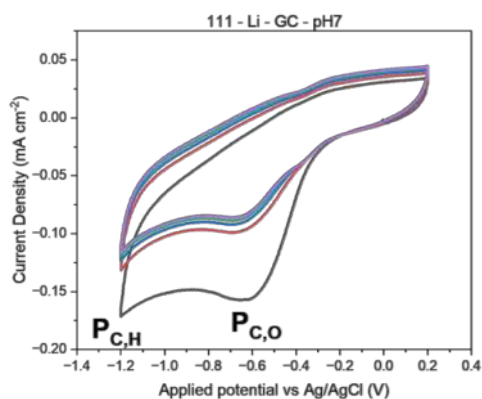


Figure 3

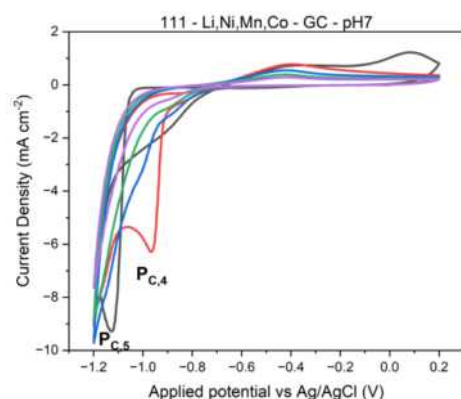


Figure 7

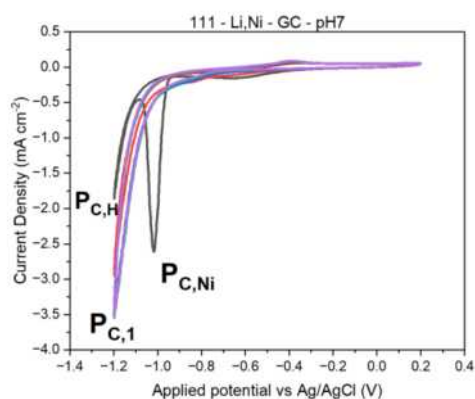


Figure 4

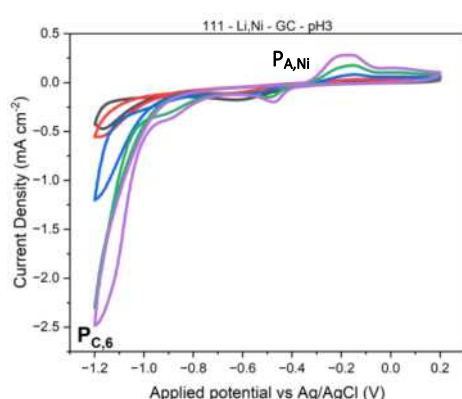


Figure 8

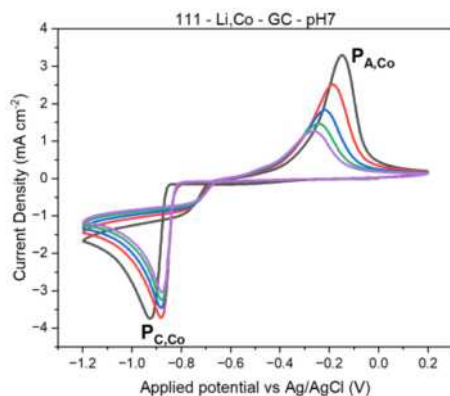


Figure 5

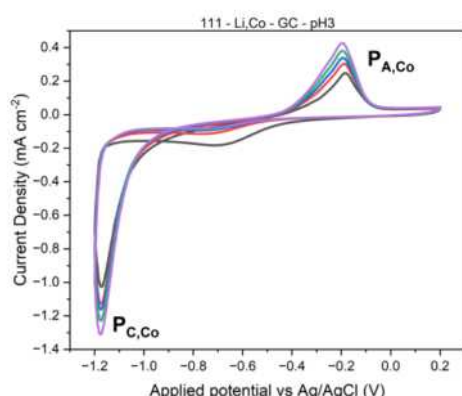


Figure 9

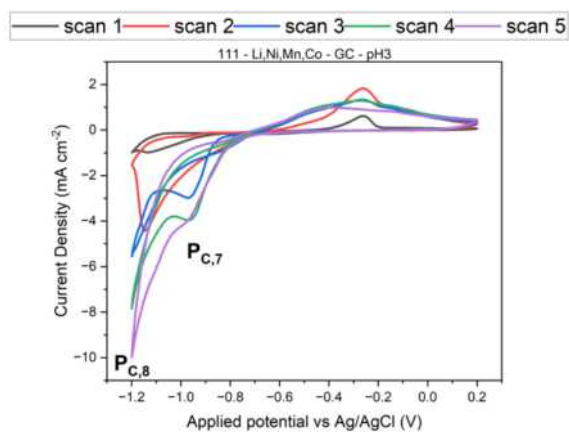


Figure 10

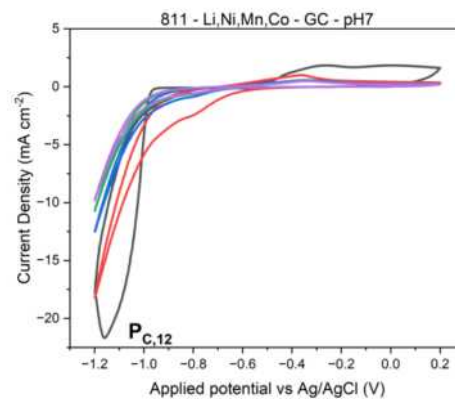


Figure 14

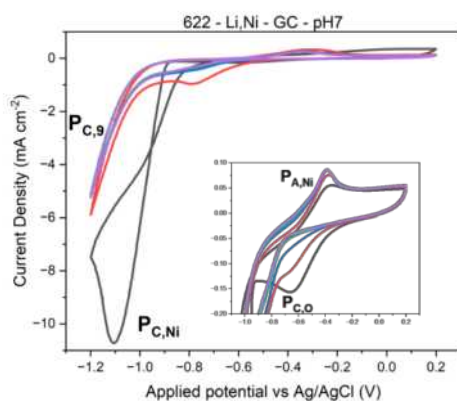


Figure 11

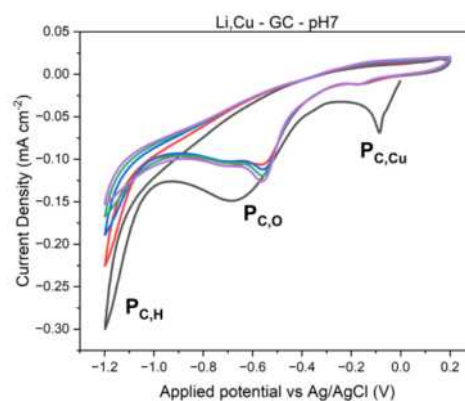


Figure 15

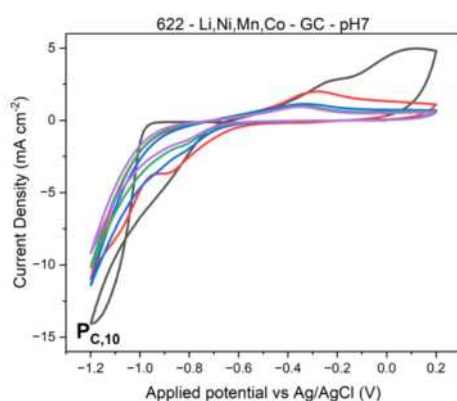


Figure 12

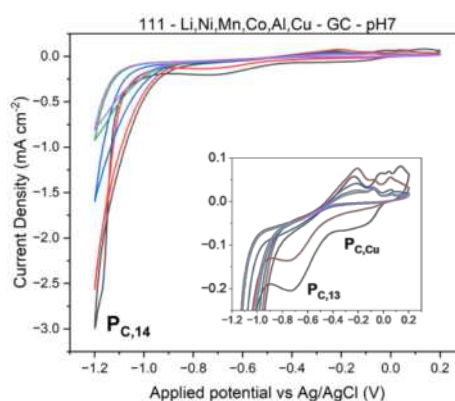


Figure 16

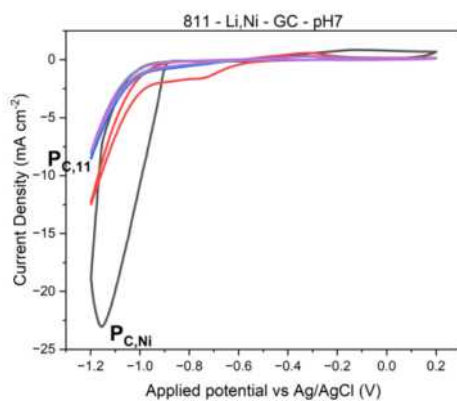


Figure 13

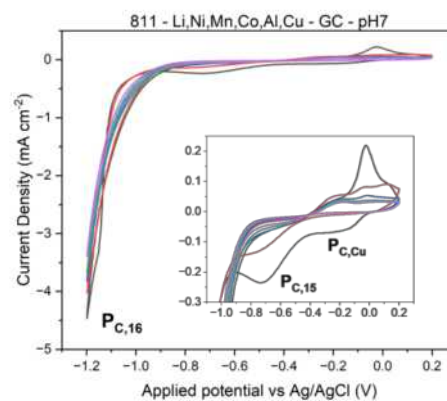


Figure 17

4.1 Discussion on the Effect on Different Type of Electrode Used in Electrolysis (Pt vs. GC)

As shown in Figure 2 and Figure 3, $P_{C,H}$ corresponded to HER. Platinum electrode yielded a much higher value for current density ca. 40 mA cm^{-2} as compared to carbon electrode ca. 0.13 mA cm^{-2} which indicates a higher rate of HER in platinum electrode. This result was reasonable as platinum acts as a catalyst for HER which was unfavourable for the efficiency of recovery for nickel and cobalt metal from NMC solution, due to its excellent electrical conductivity [13]. In addition, glassy carbon is simply a more economical option compared to platinum [14,15]. The cost for electrode is crucial criteria as a large surface area electrode is preferred for higher efficiency of recovery for Ni and Co metal in electrodeposition. Therefore, glassy carbon as an electrode will be a more efficient and economical option for electrodeposition.

4.2 Discussion for NMC111

There was no peak corresponding to the reduction of Li^+ . This can be explained by the Pourbaix diagram for lithium, which shows that lithium ion can have a reduction reaction only when the potential difference exceeds ca. -3.3 V vs. SHE [11].

$P_{A,O}$ is a peak corresponds to oxygen reduction (Eq.8). Although oxygen reduction (Eq.8) existed in all the CV analysis performed, it was negligible as the amount of dissolved oxygen was very low and contributed to negligible current loss.

In Figure 4, $P_{C,Ni}$ only occurred in the first scan which corresponds to the reduction of Ni^{2+} (Eq.3). $P_{C,H}$ corresponds to HER (Eq.2) in the first scan, this increases the local pH and promotes Ni^{2+} to form $\text{Ni}(\text{OH})_2$ via Eq.10. Consequently, in the second to fifth scan, $P_{C,1}$ would be dominated by HER and possibly reduction of $\text{Ni}(\text{OH})_2$, resulting in greater current response at potential of -1.2 V . This hypothesis was made due to the possibility of reduction of $\text{Ni}(\text{OH})_2$ as shown in Figure 1(a). $P_{A,Ni}$ corresponds to oxidation of Ni, the reverse reaction for Eq.3.

Cobalt reduction can be seen clearly on Figure 5 in $P_{C,Co}$. In the first scan $P_{C,Co}$ is at potential ca. -0.95 V and then the peak shifted to ca. -0.9 V . This phenomenon can be commonly seen on deposition of metal in electrolysis

because nucleation of solid requires high activation energy [16], which was observable in the first scan of CV. An anodic peak, $P_{A,Co}$ was also observed at potential ca. -0.2 V , this peak also corresponds to oxidation of Co^{2+} . There was no reaction involving Mn^{2+} within the potential range of the CV analysis, which was similar to Li^+ as discussed before. Therefore, the graph of current density against applied potential vs Ag/AgCl for solution containing Li and Mn had similar trend as Figure 3, no extra peak observed.

When Ni and Co are both present in the solution, the kinetics of the reactions change as both metal ions competes at the anode for electrons [17]. Two peaks, $P_{C,2}$ and $P_{C,3}$ were observed in Figure 6 which were coherent to the peak shown in Figure 4 and Figure 5. The difference between them was the magnitude of the peak which increased significantly. This was caused by the merging of two peaks for reaction Eq.3 and Eq.4.

While recognising the peak for Co and Ni reduction, Figure 7 shows a reasonable trend for both reactions. The addition of Mn changes the kinetic of the reaction by further increasing the conductivity of the solution, increasing the peak intensity for $P_{C,5}$.

$P_{C,4}$ at ca. -0.9 V has lower peak intensity compared to $P_{C,2}$ in Figure 6. Reaction mainly happened at ca. -1.2 V , which involved HER, Eq.3 and Eq.4.

4.3 Discussion for pH3 CV analysis

The main difference in a pH3 and pH7 solution was the increase in H^+ concentration, leading to an increase in concentration gradient of H^+ to the site of reaction leading to an increase in the mass diffusion of H^+ [18]. From Figure 8, solutions with presence of Ni showed an interesting trend, that had a peak at ca. -1.2 V . The peak increases over scan 1 to scan 5. This trend demonstrates the increase in rate of reaction in nickel reduction (Eq.3) when the reaction in Eq.3 was limited by nucleation of Ni metal on the site of reaction. More Ni metal deposited on the reaction site after each scan, increasing the reaction of Eq.3, because the nucleation energy was not needed after nickel being deposited. The anodic peak, $P_{A,Ni}$ in Figure 8 also validates the explanation by having an increasing peak over each scan since oxidation of Ni metal is not

affected by the increase in H^+ concentration, therefore the $P_{A,Ni}$ indicates the amount of Ni metal that is available for oxidation.

In Figure 9, a shift of peak $P_{C,Co}$ in the x-axis was observed. Indicating that higher energy was required to deposit Co metal in higher concentration of H^+ .

The trend for $P_{C,6}$ in Figure 8 also can be observed from $P_{C,8}$ in Figure 10, showing that the increase in H^+ also gave a significant effect in LNMC solution.

4.4 CV discussion for NMC622 and 811

From Figure 11 and Figure 13, $P_{C,Ni}$ showed a similar trend as $P_{C,Ni}$ in Figure 4 for NMC111. However, $P_{C,Ni}$ in NMC622 and 811 had a higher peak intensity. This highlighted that these three peaks correspond to Ni^{2+} reduction (Eq.3) as increase in concentration of Ni^{2+} effectively increases the rate of Eq.3 as discussed in Section 4.3. The peak intensity for $P_{C,9}$ and $P_{C,11}$ also increases when the concentration of Ni^{2+} becomes higher. The reason being more $Ni(OH)_2$ was formed after the first scan and was reduced at potential of -1.2 V in the following scan.

For LNMC solution in NMC 622 and 811 shown in Figure 12 and Figure 14, the trend was similar for Li-Ni solution as the peak intensity for Ni reduction was greater compared to the peak of cobalt reduction.

4.5.1 CV Discussion with Impurities

The Pourbaix diagram for aluminium outlines that the reduction of Al^{3+} ions was thermodynamically feasible at ca. -2 V vs. S.H.E. under neutral condition [20]. On the other hand, the Pourbaix diagram for copper shows that the reduction of Cu^{2+} ions to metallic copper is feasible at ca. 0 V vs. S.H.E [21]. These hypotheses derived from the Pourbaix diagrams, were validated by CV tests in which aluminium or copper were introduced as impurities. The CV test of the solution containing Li^+ and Al^{3+} ions is similar to the one in pure Li_2SO_4 (Figure 3), indicating no further reduction of aluminium. On the other hand, Figure 15 highlighted an additional $P_{C,Cu}$ peak at ca. -0.07 V, corresponding to the reduction of Cu^{2+} ions to copper metal. Subsequent tests were conducted on LNMC 111 and LNMC 811, both containing aluminium and copper impurities. Figures 16 and 17 both outline an additional $P_{C,Cu}$ peak at ca. -0.1 V. However, current density was

getting less negative after each scan was observed at $P_{C,14}$, corresponding to ca. -3.0 mA cm^{-2} for the first scan and ca. -0.75 mA cm^{-2} for the last scan, whereas $P_{C,16}$ was observed at a constant potential of ca. -4.5 V for each scan. This phenomenon can be attributed to the limitation of mass transport in the case of LNMC 111 with impurities [18]. Given that the concentration of Ni^{2+} is smaller compared to LNMC 811 with impurities, resulting to lower concentration gradient between the bulk solution and the electrode surface, leading to a lower diffusion rate. Hence, hydrogen evolution reaction would become predominant, resulting in an increased pH. As described by the Figure 1a, $Ni(OH)_2$ formation is thermodynamically feasible when pH increases from 7. Hence, hypothesis was made that the electrode was not fully covered by glassy carbon electrode, leading to a smaller surface area available for the reduction to take place and small current response. This phenomenon also validated a stronger response in current density of $P_{C,16}$ than $P_{C,14}$ by 50% due to higher nickel concentration. Hence, faster diffusion rate to the electrode surface [18]. Additionally, noting that cathodic peaks of $P_{C,13}$ and $P_{C,15}$ at an electrode potential of ca. -0.7 V vs. Ag/AgCl were primarily due to the reduction of Ni^{2+} and Co^{2+} . This was supported by the CV results of Li-Ni and Li-Co solutions, as discussed in Section 4.2. On the other hand, Figure 16,17 highlights that the cathodic peaks of $P_{C,14}$ and $P_{C,16}$ were resulted from the reduction of nickel, cobalt and hydrogen reduction, with the HER being the predominant reaction.

5. Electrodeposition Results and Discussion

Electrodeposition experiments were done on both NMC111 and NMC 811 at applied potential of 1.0 V and -0.7 V vs. Ag/AgCl, which are for recovery of Mn and Ni/Co respectively. The Mn recovery at the potential of 1.0 V was supported by the research conducted by Dr. Xiaochu [11], while the Ni/Co recovery at the potential of -0.7 V was supported by $P_{C,13}$ and $P_{C,15}$ in Figure 16 and 17. However, the ICP-MS machine was down and was not able to generate any results from it. UV-vis spectrometry was not available to be used within the timeframe designed for characterisation of results as well. ICP-MS

analysis will be carried on in the future to obtain the concentration of nickel and cobalt deposited on the carbon paper after electrodeposition. Hence, recovery efficiency of Ni and Co could be obtained by Eq.7 and to predict the impact of HER on Ni and Co recovery.

6. Conclusion

In this study, the electrochemical kinetics and thermodynamics of Li-Ni-Mn-Co system in water solutions were investigated within the applied potential range of -1.2 V to 0.2 V vs Ag/AgCl. Neutral condition and glassy carbon working electrode were the optimal condition for electrodeposition, in order to maximise the recovery efficiency of Ni and Co metal. In another word, suppressing the effect of HER by decreasing the concentration of H^+ ion and inhibiting the usage of platinum as working electrode could facilitate reduction of Ni^{2+} and Co^{2+} . During the cyclic voltammetry tests for the solutions of NMC 111 and NMC 811 with impurities (Al and Cu), the cathodic peak at ca. -0.7 V vs. Ag/AgCl was concluded to be the optimum reduction potential point of Ni^{2+} and Co^{2+} ions. Being able to separate Cu selectively at electrical potential. ca. -0.1V vs. Ag/AgCl, Al and Cu was concluded to have a negligible effect for nickel and cobalt recovery due to the large potential difference of its cathodic peak from nickel and cobalt cathodic peak. Hence, impurities of Cu could be extracted by applying a much lower potential (ca. -0.1 V vs. Ag/AgCl) compared to the reduction potential of -0.7 V vs. Ag/AgCl, while Co and Ni peak merged at ca. -0.7 V vs. Ag/AgCl and separation of these metal could not be done by electrodeposition. Additionally, an electrode potential of -1.0 V vs Ag/AgCl was also concluded to be a point where Ni and Co reduction occurred the most, as supported by the CVs test of Li-Ni and Li-Co aqueous solution. Consequently, final condition of pH7, working electrode of glassy carbon with the applied potential -0.7 V and -1.0 V vs Ag/AgCl was selected in electrodeposition which would carry on in the future study.

7. Outlook

Further research could be conducted to further enhance the recovery efficiency of Ni and Co. It is crucial to obtain

an optimum potential that minimise the rate of HER. Product characterisation of the solution after electrodeposition was required, which could be done by ICP-MS to obtain the concentration of nickel and cobalt after electrodeposition, and hence to obtain the recovery efficiency. Additionally, sensitivity analysis of electrode potential around -0.7 V and -1.0 V vs. Ag/AgCl during electrodeposition could also be conducted. Hence, optimal recovery efficiency could be obtained. Lastly, the effect of leaching could be investigated to extract nickel and cobalt separately, recovery of Ni and Co would be achieved by extracting metal ions from a solid material by selectively dissolving it in an acid due to solubility differences.

8. Reference

1. Nurdiawati, A., & Agrawal, T. K. (2022). Creating a circular EV battery value chain: End-of-life strategies and future perspective. *Resources, Conservation and Recycling*, 185, 106484.
2. Soge, A. O., et al. (2021). Cathode Materials for Lithium-ion Batteries: A brief review. *Journal of New Materials for Electrochemical Systems*, 24(4).
3. Kelly, J. C., Dai, Q., & Wang, M. (2020). Globally regional life cycle analysis of automotive lithium-ion nickel manganese cobalt batteries. *Mitigation and Adaptation Strategies for Global Change*, 25, 371-396.
4. Zhang, N., Xu, Z., Deng, W., & Wang, X. (2022). Recycling and upcycling spent LIB cathodes: a comprehensive review. *Electrochemical Energy Reviews*, 5(Suppl 1), 33.
5. Schiavi, P. G., et al. (2021). Resynthesis of NMC111 cathodic material from real waste lithium-ion batteries. *Chemical Engineering Transactions*, 86, 463-468.
6. Ji, H., Wang, J., Ma, J., Cheng, H. M., & Zhou, G. (2023). Fundamentals, status and challenges of direct recycling technologies for lithium-ion batteries. *Chemical Society Reviews*.
7. Defalque, C. M., Marins, F. A. S., da Silva, A. F., & Rodríguez, E. Y. A. (2021). A review of wastepaper

recycling networks focusing on quantitative methods and sustainability. *Journal of Material Cycles and Waste Management*, 23, 55-76.

8. Chan, K. H., Malik, M., & Azimi, G. (2021). Recovery of valuable metals from end-of-life lithium-ion battery using electrodialysis. In *Rare Metal Technology 2021* (pp. 11-17). Cham: Springer International Publishing.

9. Zhang, H., Ni, Y., Zhong, Y., Wu, H., & Zhai, M. (2015). Fast electrodeposition, influencing factors and catalytic properties of dendritic Cu-M (M= Ni, Fe, Co) microstructures. *RSC advances*, 5(117), 96639-96648.

10. Reyes-Valderrama, M. I., Salinas-Rodríguez, E., Montiel-Hernández, J. F., Rivera-Landero, I., Cerecedo-Sáenz, E., Hernández-Ávila, J., & Arenas-Flores, A. (2017). Urban mining and electrochemistry: cyclic voltammetry study of acidic solutions from electronic wastes (printed circuit boards) for recovery of Cu, Zn, and Ni. *Metals*, 7(2), 55.

11. Wei, X. (2022). Recovery Materials from End-of-Life Lithium-ion Batteries. Department of Chemical Engineering, Imperial College London.

12. Dubouis, N., & Grimaud, A. (2019). The hydrogen evolution reaction: from material to interfacial descriptors. *Chemical Science*, 10(40), 9165-9181.

13. Chen, D., Tao, Q., Liao, L. W., Liu, S. X., Chen, Y. X., & Ye, S. (2011). Determining the active surface area for various platinum electrodes. *Electrocatalysis*, 2, 207-219.

14. de Boode, B., Phillips, C., Lau, Y. C., Adomkevicius, A., McGettrick, J., & Deganello, D. (2022). Glassy carbon manufacture using rapid photonic curing. *Journal of Materials Science*, 1-12.

15. Shahrokhian, S., Kahnamoui, M. H., & Salimian, R. (No date). Surface modification of glassy carbon electrode with the functionalized carbon nanotube for ultrasensitive electrochemical detection of risperidone.

16. De Yoreo, J. J., Sommerdijk, N. A., & Dove, P. M. (2017). Nucleation pathways in electrolyte solutions. New perspectives on mineral nucleation and growth: from solution precursors to solid materials, 1-24.

17. Sahlman, M., Aromaa, J., & Lundström, M. (2021). Copper cathode contamination by nickel in copper electrorefining. *Metals*, 11(11), 1758.

18. Skoog, D., Holler, F., & Crouch, S. (2007). *Principles of Instrumental Analysis*.

19. Saneie, R., et al. (2022). Recovery of copper and aluminum from spent lithium-ion batteries by froth flotation: a sustainable approach. *Journal of Sustainable Metallurgy*, 8(1), 386-397.

20. Sukiman, N. L., et al. (2012). Durability and corrosion of aluminium and its alloys: overview, property space, techniques and developments. *Aluminium Alloys-New Trends in Fabrication and Applications*, 5, 47-97.

21. Iqbal, A. (2023). Ovarian Leiomyoma Associated with Serous Cystadenoma-A Case Report of an Uncommon En

Derivative-free Optimisation of Neural Networks in Reinforcement Learning for Process Control

Ivan Pchelintsev and Andreas Must

Department of Chemical Engineering, Imperial College London, UK

Abstract

The paper explores methods of training neural networks to control process units as an alternative to standard Proportional-Integral-Derivative (PID) controllers. Neural networks have theoretical potential in handling plant-wide control better than standard PID controllers, due to the complex dynamics of multivariate control. Several derivative-free optimisation algorithms were tested on a model of a Continuous Stirred-Tank Reactor, with the most promising algorithms being selected for further testing against a more complicated Multistage Extraction model with challenging set point changes. The algorithms were susceptible to pitfalls, such as failure to control multiple set-points at once. A hybrid algorithm that introduced more exploration into its decision process was proposed, with limited improvement in performance. However, using a Bayesian Optimisation algorithm to optimise the structure of the neural network along the algorithm itself proved to be a viable method, with the algorithms seeing a stark improvement in performance, managing to consistently adhere to multiple set-points. This demonstrates the potential capabilities of neural networks to learn the dynamics and the methods of controlling a variety of processes.

Keywords: *Reinforcement Learning, Machine Learning, Process Control, Derivative-free Optimisation*

1 Introduction

Control systems are an integral aspect of any chemical manufacturing environment, and have broad applications outside of chemical engineering, such as in the automotive, aerospace and robotics industries [1]. Therefore, new methods of establishing and maintaining quality control are of high interest to the scientific community and beyond.

The most common and widely accepted method of automatic control of systems is PID control, where the controller uses the current error and the history of errors in the past to determine the appropriate action to take. PID controllers have been widely accepted as the most efficient way to control industrial processes due to their simplicity and robustness [2]. However, while PID control equations provide a good framework for quality control, tuning a controller can be a challenging and time-consuming process [3]. There

has been extensive research dedicated to the most efficient ways of tuning a PID controller, such as the Ziegler-Nichols method [4]. More modern methods make use of optimisation techniques [5], wherein the effectiveness of process control is represented as a function of the PID parameter space to be optimised. This is an example of a Reinforcement Learning problem (see Section 2.2).

Neural networks can also be used to control processes, by training them to take appropriate actions given the current state of the system. The advantage of tuning a neural network over a PID controller is that in order to effectively tune a PID controller, the dynamics of the system must be known in advance. This requires experimentation and can be time-consuming. In contrast, a neural network learns the dynamics of the system and how to control the system simultaneously, through exploring the parameter space. This project builds on the idea of using

optimisation techniques to tune a controller, focusing on implementing an Artificial Neural Network into a Reinforcement Learning loop, and optimising the neural network's parameters.

The objective of the project is to develop an algorithm that can be used to train a neural network to control any set-point changes in any process. The project makes use of two models of chemical engineering process units to achieve this - a simple first-order Continuous Stirred-Tank Reactor (CSTR) model as a proof of concept, and a more complex Multistage Extraction model as a benchmark. Six derivative-free optimisation algorithms were tested on a neural network controlling the CSTR model, and compared with the performance of a tuned PID controller. Algorithms that showed promise were further tested against the Multistage Extraction model. Finally, an algorithm that incorporated optimisation of the neural network structure itself was developed, and its performance compared to the standard derivative-free optimisation algorithms that used more conventional neural network structures.

Section 2 focuses on the Background of the projects and the algorithms used, Section 3 outlines the dynamics of the two models that were used, Section 4 outlines the procedures for the tests and their reasoning, and Sections 5 and 6 demonstrate the key outcomes and insights of the project.

2 Background

2.1 PID Controller

A Proportional-Integral-Derivative (PID) controller is a system that monitors and regulates a crucial process state, x , to match desired process set-point, x_{sp} , by taking appropriate control actions u . The action that the controller takes is a sum of three functions of the error:

$$u(t) = K_p e(t) + K_i \int_0^t e(\tau) d\tau + K_d \frac{de(t)}{dt} \quad (1)$$

where error is defined as

$$e(t) = x(t) - x_{sp} \quad (2)$$

and K_p , K_i , K_d are the control system parameters that need to be tuned, or optimised, in order to achieve stable and effective control. In real-world applications, the state is a continuous function of time, but due to the computational nature of the project, the states are simulated at discrete time-steps (see Section 4.1).

2.2 Reinforcement Learning

A Markov Decision Process (MDP) is a control process where a system, or agent, uses a policy

π_θ to select actions at every discrete time step that have a certain probability of influencing the environment. These actions cause a corresponding change in the environment, which is then observed by the agent which calculates a corresponding reward. This relationship can be expressed mathematically as a 4-tuple:

$$(\mathcal{X}, \mathcal{U}, P(x_{t+1} = x' | x_t = x, u_t = u), R_t) \quad (3)$$

where \mathcal{X} is the state space, \mathcal{U} is the action space, $P(x_{t+1} = x' | x_t = x, u_t = u)$ is the probability of action u to alter state x to x' , and R_t is the calculated reward at that iteration. Reinforcement Learning (RL) is a subset of machine learning which makes use of the Markov Decision Process. In RL, the agent is a neural network, the policy is the network's parameters, and crucially, the reward for a particular action is not known prior to taking said action [6]. In every iteration of the reinforcement learning loop, the agent interacts with the environment according to its policy and calculates the total reward for following this policy after t time steps. Then, the agent updates its behaviour, or policy, with the goal of maximising future rewards. This loop is represented visually in Figure 1.

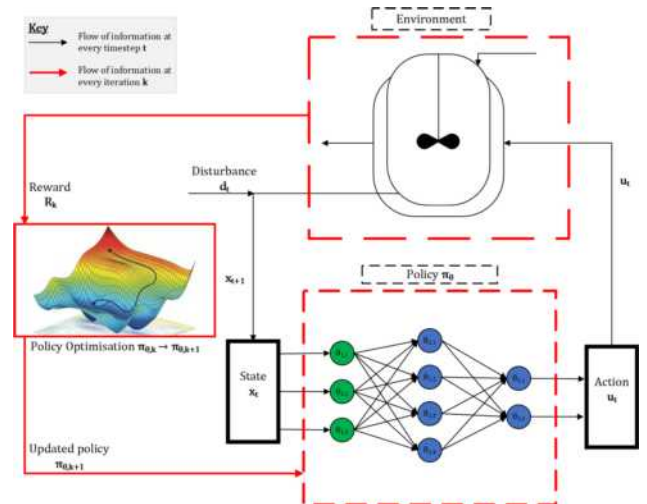


Figure 1: A representation of the Reinforcement Learning loop. 'Policy Optimisation' image source: [7]

2.3 Derivative-free Policy Optimisation

Derivative-free optimisation (DFO) methods are algorithms that do not use derivative information or finite differences to optimise functions. Instead, they usually operate by sampling data from the objective function, and applying metaheuristics to approach an optimal point. This presents several advantages that DFO holds over policy gradients. For instance, since DFO relies solely on objective function values, backpropagation is not required,

which can significantly reduce computational demand for the policy search. Additionally, due to its reliance on gradient information, policy gradients can be prone to getting stuck in local optima (however this is very unlikely, given a high enough number of dimensions). This makes derivative-free optimisation an attractive option to explore for relatively low-dimensional reinforcement learning problems, such as process control for a reactor. There are a number of metaheuristic algorithms developed, this paper included Generalised Policy search, Particle Swarm Optimisation, Simulated Annealing, Genetic Algorithm, Artificial Bee Colony and Firefly algorithms, with the first two explained in detail in Sections 2.3.1 and 2.3.2.

Table 1: Generalised Policy Search pseudocode

Algorithm 1 Generalised Policy search

Input:

shrink ratio (ϕ_{shr}), radius (r), evaluations shrink (e_{shr}), evaluations (e), search ratio, bounds min and max

Output:

θ^* - Best Policy Parameters, R^* - Best Reward

Start:

Initialise R^* , i_{fail} , divide e into e_{rs} and e_{ls}

for $i = 0$ to e_{rs} **do**

 Sample search space and evaluate reward:

$\theta_i \leftarrow U[min, max]$, $R_i \leftarrow \text{Evaluate } \pi_{\theta_i}$

if $R_i < R^*$ **then**

 Update best policy and reward:

$\theta^* \leftarrow \theta_i$, $R^* \leftarrow R_i$

end if

end for

$r_{0,min} \leftarrow r \times min$, $r_{0,max} \leftarrow r \times max$

while $i < e_{ls}$ **do**

if $i_{fail} \geq e_{shr}$ **then**

 Reset i_{fail} and reduce radius:

$i_{fail} \leftarrow 0$, $r \leftarrow r \times \phi_{shr}$

$r_{0,min} \leftarrow r \times min$, $r_{0,max} \leftarrow r \times max$

end if

 Sample local points and evaluate reward:

$\theta_i \leftarrow L[\theta^*, r_{0,min}, r_{0,max}]$, $R_i \leftarrow \text{Evaluate } \pi_{\theta_i}$

if $R_i < R^*$ **then**

$R^* \leftarrow R_i$, $\theta^* \leftarrow \theta_i$, $i_{fail} \leftarrow 0$

else do

$i_{fail} \leftarrow i_{fail} + 1$

end if

 Update counter $i \leftarrow i + 1$

end while

return R^* , θ^*

2.3.1 Generalised Policy Search

Generalised Policy Search (GPS) is a hybrid algorithm that consists of two phases - random search and local search. In the random search phase, the search space is populated with random guesses, and the point with the lowest reward is chosen as the starting point of the local search. In the local search phase, guesses are made randomly within a certain radius of the current best guess. With every guess, the current best guess value updates if there was improvement, and the radius around which the guesses are made is shrunk if there was no improvement for pre-determined number of iterations [8]. The pseudocode for GPS algorithm is given in Algorithm 1.

Table 2: Particle Swarm Optimisation pseudocode

Algorithm 2 Particle Swarm Optimisation

Input:

number of particles (S), inertia (W), cognitive (ϕ_p), social (ϕ_g), iterations (e), bounds min and max

Output:

θ^* - Best Policy Parameters, R^* - Best Reward

Start:

Initialise R^* , R_{best}

for $p = 1$ to S **do**

$x_p \leftarrow U[min, max]$

$v_p \leftarrow U[min/100, max/100]$

end for

for $p = 1$ to S **do**

$p_{best,p} \leftarrow x_p$, $R_{best,p} \leftarrow \text{Evaluate } \pi_{x_p}$

end for

$R^* \leftarrow \min[R_{best}]$, $\theta^* \leftarrow \text{argmin}[p_{best}]$

for $i = 0$ to e **do**

$r_1 \leftarrow U[0, 1]$, $r_2 \leftarrow U[0, 1]$

$v \leftarrow W \cdot v + r_1 \phi_p(p_{best} - x) + r_2 \phi_g(\theta^* - x)$

$x \leftarrow x + v$

$R \leftarrow \text{Evaluate } \pi_x$ for all particles

if $R_p < p_{best,p}$ **then**

$p_{best,p} \leftarrow x_p$, $R_{best,p} \leftarrow R_p$

end if

if $R_p < R^*$ **then**

$\theta^* \leftarrow x_p$, $R^* \leftarrow R_p$

end if

end for

return θ^* , R^*

2.3.2 Particle Swarm Optimisation

Particle Swarm Optimisation (PSO) is a global derivative-free optimisation algorithm [9]. Particles are initialised with a random position and velocity, and after each iteration, each particle updates its position according to its velocity value. Additionally,

the velocity of each particle gets updated every iteration according to the specified hyperparameters. The inertial coefficient makes the particle velocity more resistant to change, the cognitive coefficient directs the particle's velocity to its last recorded best point in the parameter space, and the social coefficient directs the particle's velocity to the best point detected across all particles in the parameter space. The pseudocode for PSO is given in Algorithm 2.

Table 3: Bayesian Optimisation pseudocode

Algorithm 3 Bayesian Optimisation Algorithm

Input:

n_{calls} - number of BO calls, U_n - acquisition function, hyperparameter space

Output:

y^* - Best Reward, θ^* - Best hyperparameters

Start:

Initialise hyperparameters $\theta^* = \theta_0$, current best value $y^* = f(\theta_0)$, $S_0 = \{\theta_0, y_0\}$

for $n = 1$ to n_{calls} **do**

Fit a Gaussian process to sample S_n :

$$f(\theta) \approx \hat{f}(\theta) \sim \mathcal{GP}(\mu_{\hat{f}}(\theta), \sigma_{\hat{f}}(\theta))$$

Select new hyperparameter set optimising U_n :

$$\theta_n = \arg \max U_n(\theta, S_n)$$

$$y_n = f(\theta_n)$$

Update sample set:

$$S_{n+1} = S_n \cup \{\theta_n, y_n\}$$

if $y_n < y^*$ **then**

$$y^* \leftarrow y_n, \theta^* \leftarrow \theta_n$$

end if

end for

return y^*, θ^*

2.3.3 Bayesian Optimisation

Bayesian Optimisation (BO) is global optimisation strategy that is commonly used for optimising computationally expensive functions [10]. It relies on building a Gaussian processes regression from the current sampled data points, and using a certain acquisition function to choose a next point to sample. The computational demand for building a Gaussian process scales rapidly with every additional point, which is why this algorithm is only typically used for problems with expensive functions which require less than 100 iterations. The pseudocode for BO is given in Algorithm 3.

3 Case Studies

Two case studies were used for testing purposes - a CSTR model for initial testing and a Multistage

Extraction model for more thorough evaluation.

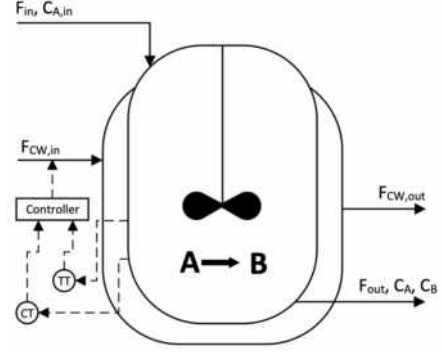


Figure 2: CSTR Model Process Flow Diagram

3.1 First Order CSTR

The first case study was based on a Continuously Stirred-Tank Reactor (CSTR) model taken from the dynamic optimization web course [11]. The simplified process flow diagram is outlined in Figure 2. The reaction taking place in the reactor is a simple first order reaction:



With the reaction rate given as:

$$r_A = k_0 e^{\frac{-E}{RT}} C_A \quad (5)$$

Where r_A is the reaction rate in [kmol/s], k_0 denotes a pre-exponential factor in [1/s], E is the activation energy in [J], T refers to reactor temperature in [K], and C_A is the concentration of substance A in CSTR [mol/m³]. The equations for mass (6) and energy (7) balance equations are given below:

$$\frac{dC_A}{dt} = \frac{F}{V(C_{A,in} - C_A)} - r_A + \delta_{C_A} \quad (6)$$

$$\begin{aligned} \frac{dT}{dt} = & \frac{F}{V(T_{in} - T)} + \frac{\Delta H}{\rho_{AB} C_{p,AB}} r_A \\ & + \frac{UA}{V \rho_{AB} C_{p,AB} (T_c - T)} + \delta_T \end{aligned} \quad (7)$$

Where t denotes time in [s], F is the volumetric flow rate in [m³/s], V refers to reactor volume in [m³], $C_{A,in}$ is the feed concentration of substance A in [mol/m³]. T and T_{in} refer to reactor and feed temperatures in [K] respectively, ΔH denotes heat of reaction 4 in [J/mol], ρ_{AB} and $C_{p,AB}$ are the density and heat capacity of components A and B mixture in [kg/m³] and [J/(kgK)], U and A are the heat transfer coefficient area in [W/(m²K)] and [m²], and T_c denotes the cooling jacket temperature in [K]. The noise for C_A and T are denoted as δ_{C_A} and δ_T , which both are distributed uniformly with $\delta_{C_A} \sim U[-0.01, 0.01]$ and $\delta_T \sim U[-0.5, 0.5]$.

Table 4: Nominal values of the key states and actions of the CSTR model

Variable x	x_{LB}	x_{UB}	Setpoint x_{sp}
T	315	330	$325 \rightarrow 320 \rightarrow 330$
C_A	0.7	0.9	$0.85 \rightarrow 0.9 \rightarrow 0.8$
T_C	295	305	-

Table 5: Parameters for the CSTR model

Parameter	Value
T_f	350
q	100
$C_{A,in}$	1
V	100
ρ_{AB}	1000
$C_{p,AB}$	0.239
ΔH	50000
$\frac{E}{R}$	8750
k_0	7.2×10^{10}
UA	50000

In this model, the key states to be controlled were the temperature (T) and concentration of A (C_A) in the reactor, and the action taken to control these states was the temperature of the cooling fluid (T_C). The bounds of these variables, as well as the set-point for the desired states are outlined in Table 4. The process parameters that were chosen for this model are shown in Table 5.

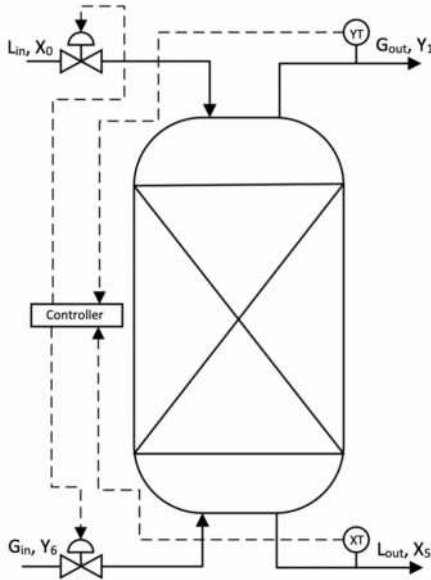


Figure 3: Multistage Extraction Process Flow Diagram

3.2 Multistage Extraction

The model for multistage extraction is based on page 471 of J. Ingham's book "Backmixing in a Multi Stage, Multi-mixer Liquid-liquid Extraction Colum" [12]. The model has five stages. A simplified diagram is shown in Figure 3. For a stage n ,

Table 6: Table showing the nominal values of the key states and actions of the Multistage Extraction model

Variable x	x_{LB}	x_{UB}	Setpoint x_{sp}
X_5	0.05	0.55	$0.3 \rightarrow 0.4 \rightarrow 0.4 \rightarrow 0.3$
Y_1	0.05	0.75	$0.3 \rightarrow 0.3 \rightarrow 0.35 \rightarrow 0.35$
L	5	500	-
G	10	1000	-

the equilibrium solute concentration in liquid $X_{n,eq}$ [kg/m³] is calculated as:

$$X_{n,eq} = \frac{(Y_n)^{exponent}}{m} \quad (8)$$

Where Y_n is the concentration of solute in gas phase of stage n in [kg/m³], *exponent* is the change of nonlinearity of the equilibrium equation [-], and m is the equilibrium constant [-]. The mass transfer rate Q_n [kg/hr] for stage n is calculated as:

$$Q_n = k_{la}(X_n - X_{n,eq})V_l \quad (9)$$

Where k_{la} is the mass transfer capacity constant in [1/hr], X_n is the concentration of solute in liquid phase of stage n in [kg/m³], and V_l denotes the liquid volume in each stage in [m³]. The differential equation used to calculate concentration of solute in liquid phase at stage n is given as:

$$\frac{dX_n}{dt} = \frac{1}{V_l} (L(X_{n-1} - X_n) - Q_n) + \delta_{X_n} \quad (10)$$

And for a solute concentration in gas phase at stage n :

$$\frac{dY_n}{dt} = \frac{1}{V_g} (G(Y_{n+1} - Y_n) + Q_n) + \delta_{Y_n} \quad (11)$$

Where V_l and V_g refer to liquid and gas volumes at each stage in [m³], L and G are solute-free liquid and gas flow rates in [m³/hr] respectively, and subscripts $n-1$ and $n+1$ denote the previous and the next stage. $\delta_{X_n} \sim U[-0.01, 0.01]$ and $\delta_{Y_n} \sim U[-0.01, 0.01]$ are the added noise to liquid and gas phase at stage n . As per mass balance, $L_{in} = L_{out} = L$ and $G_{in} = G_{out} = G$.

For this model, the outlet concentration of solute in liquid (X_5) and gas (Y_1) phases are controlled variables with desired set points, with both liquid (L) and gas (G) flow rates being manipulated variables. The controlled and manipulated variable ranges with a tested set point changes are given in Table 6, The numerical values used in simulations are outlined in Table 7.

4 Methodology

4.1 Problem Statement

The reinforcement learning problem in question was the control of a process governed by certain model

Table 7: Parameters for the Multistage Extraction model

Parameter	Value
V_l	5
V_g	5
m	1
k_{la}	5
$exponent$	2
X_0	0.6
Y_6	0.05

equations. Given certain key process states and a set-point to adhere to for each of those states, the neural network agent was required to learn how to optimally control the process by outputting actions that would keep the process states steadily at the set-point, as well as follow any desired set-point change. Optimal control was approached by minimising the reward function through optimising the policy parameters. The optimisation problem is formulated as following:

$$\begin{aligned}
& \min R(\mathbf{x}_t, \mathbf{u}_t) \\
& \text{s.t. } \mathbf{x}_0 = \mathbf{x}_{t_0} \\
& \mathbf{u}_t = \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}_t) \\
& \mathbf{x}_{t+1} = f(\mathbf{x}_t, \mathbf{u}_t, d_t) \\
& 0 \leq t \leq 100
\end{aligned} \tag{12}$$

Where total reward associated with that set of policy parameters was calculated as:

$$R = \sum_{t=1}^{t_{max}} \left(\sum_i \frac{\|x_{i,t} - x_{i,sp}\|}{w_i} + \sum_j \frac{\|u_{j,t} - u_{j,LB}\|}{w_{j,mag}} + \sum_j \frac{\|u_{j,t} - u_{j,t-1}\|}{w_{j,ch}} \right) \tag{13}$$

Where the first term records the error accumulated over the course of the simulation, the second term penalises actions of high magnitude (using up a lot of cooling water or fuel to achieve a desired set-point is expensive), and the third term penalises rapidly changing actions, as that could damage whatever actuators are simulated in the system. Each of these costs is weighted with their respective coefficients for each state i and action j . These coefficients varied across the two case studies in the project, as shown in Tables 8 and 9.

4.2 Algorithm Test on CSTR Case Study

The optimisation algorithms from Section 2.3, as well as the Simulated Annealing, Artificial Bee Colony, Genetic, and Firefly algorithms were implemented into a reinforcement learning loop similar to the one illustrated visually in Figure 1. Initially, a two-layer neural network with the configurations specified in Table 10 was used as the agent in the reinforcement

Table 9: Reward split coefficients for multistage extraction column case study

Coefficient	Value
w_{X_5}	0.2
w_{Y_1}	0.2
$w_{L,mag}$	20000
$w_{G,mag}$	20000
$w_{L,ch}$	20000
$w_{G,ch}$	20000

Table 8: Reward split coefficients for CSTR case study

Coefficient	Value
w_{C_A}	0.2
w_T	15
$w_{T_C,mag}$	10
$w_{T_C,ch}$	10

Table 10: Neural network architecture for CSTR model

Layer	Layer size	Activation	Activation argument
Input	4	-	-
Hidden 1	8	LeakyReLU	0.1
Hidden 2	2	LeakyReLU	0.1
Output	1	ReLU6	-

loop. The agent neural network took in as inputs the state \mathbf{x} and set-point error $\mathbf{x} - \mathbf{x}_{sp}$ for the two variables it had to control, namely reactor temperature and reactant concentration, and returned an action u in response. This was done according to the neural network's set of policy parameters $\boldsymbol{\theta}$, and after every simulation, a cost was calculated using equation 13, which was used by the optimisation algorithm to update to its next set of parameters. This process would repeat for a specified number of iterations, at which point the best recorded set of parameters and the corresponding costs were returned.

In order to appropriately compare the performance and effectiveness of different algorithms, each algorithm was allocated an equal amount of computational time to optimise the neural network. Furthermore, due to the stochastic nature of the models and the algorithms, single results may not be reliable enough for comparison, as they are not reproducible. Hence, an average optimal cost across 10 runs was used to compare the performance of each algorithm. Finally, the algorithms optimising the Neural Network were compared with a tuned PID controller to assess their viability in industry, as PID is a widely accepted method for process control.

4.3 Algorithm Test on Multistage Extraction Case Study

Two more promising algorithms - GPS and PSO - were also benchmarked against a more complex Multistage Extraction model with 10 states and 2 actions, the dynamics of which are outlined in Section 3.2. The neural network structure used for initial evaluation is outlined in Table 11. Out of 10 states, only two (X_5 and Y_1) were controlled and used as

Table 11: Neural network architecture for Multistage Extraction model

Layer	Layer size	Activation	Activation argument
Input	4	-	-
Hidden 1	8	ELU	0.1
Hidden 2	4	LeakyReLU	0.1
Output	2	ReLU6	-

neural network inputs with the corresponding set point error $\mathbf{x} - \mathbf{x}_{sp}$. To account for nonlinear system, the first hidden layer used exponential linear unit (ELU) as activation function instead of leaky ReLU. The network had two actions, L and G , as outputs.

The performance of selected algorithms were tested with equal computational time. The parameters for Generalised Policy Search were selected as $\phi_{shr} = 0.9$, $r = 0.1$, $e_{shr} = e/30$ and search ratio of 0.1. Particle swarm used 50 particles, $W = 0.6$, $\phi_p = 0.2$ and $\phi_g = 0.2$. Due to inherent stochastic behaviour of the models and algorithms, an average of 10 runs were used to compare the performance of algorithms.

4.3.1 Hybrid algorithm - Random PSO

The performance of PSO algorithm was unsatisfactory, which lead to a hybrid RandPSO algorithm, that combined the explorative nature of Random Search with the exploitative behaviour of the PSO algorithm. The algorithm initialises as Random Search, up to a certain amount of iterations, after which it populates a radius around the best guess with several particles, and activates the PSO algorithm. This algorithm was also tested on the Multistage Extraction model.

4.4 Neural Network Architecture Optimisation

The number of layers in the neural network, the size of each layer and the activation function, as well as the DFO algorithm hyperparameters used for the agent in the Multistage Extraction case study were reconsidered as a set of hyperparameters to be optimised. Algorithm 4 combines DFO methods explored thus far for parameter optimisation with Bayesian Optimisation for hyperparameter optimisation. The input and output layer structure were fixed at $n_x + n_{x,sp}$ neurons for input and n_u neurons with ReLU6 activation function for output layer, where n_x , $n_{x,sp}$ and n_u is the number of states, set points and actions, respectively. The number of neurons per layer varied between 1 and 16, number of layers between 1 and 4, and the activation functions available were tanh, sigmoid, ELU, ReLU, and leakyReLU. In order to ensure the reliability of each sample in the BO

Table 12: Hyperparameter optimised policy search pseudocode

Algorithm 4 Hyperparameter Optimisation Algorithm

Input:

n_{calls} - number of BO calls,
 n_{iter} - number of DFO iterations,
 n_{rep} - number of DFO epoch repetitions
 U_n - acquisition function,
DFO algorithm

Output:

y^* - Best Average Reward, θ^* - Best hyperparameters

Start:

Initialise hyperparameters $\theta^* = \theta_0$, current best value $y^* = \text{Algorithm}(\theta_0)$, $S_0 = \{\theta_0, y_0\}$

for $n = 1$ to n_{calls} **do**

 Use BO algorithm (Algorithm 3) to select new hyperparameters to sample

 Initialise set of results, $\mathcal{R} = \{\}$

for $n_r = 1$ to n_{rep} **do**

 Evaluate algorithm with chosen θ :

$R_{n_r} \leftarrow \text{Algorithm}(\theta_{n_r})$

$\mathcal{R} \leftarrow \mathcal{R} \cup R_{n_r}$

 Evaluate average performance across repetitions:

$y_n \leftarrow \text{mean}(\mathcal{R})$

$S_n \leftarrow S_{n-1} \cup \{\theta_n, y_n\}$

if $y_n < y^*$ **then**

$y^* \leftarrow y_n$, $\theta^* \leftarrow \theta_n$

end if
end for

return y^*, θ^*

loop, each set of hyperparameters was tested on the algorithm of choice n_{rep} times.

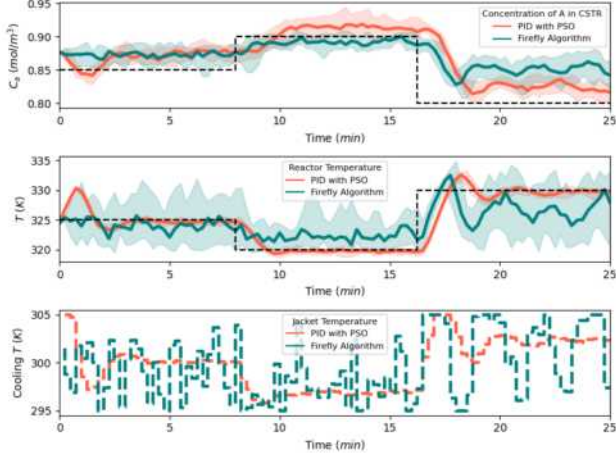
5 Results

5.1 CSTR Case Study Results

The mean rewards and standard deviations for the initial screening of the algorithms are presented in Table 13. Figure 4 demonstrates the performance difference of a good control algorithm with a sub-optimal algorithm. In this case, the good control is represented by a PID controller with PSO parameter optimisation, the sub-optimal control by Firefly algorithm. The first two subplots show the mean trajectory of concentration of A (C_A) and reactor temperature (T) with a distribution of runs. The subplot below demonstrates the trajectory of reactor cooling jacket (T_c) in time.

Table 13: CSTR Case Study Results

Algorithm	Mean Reward	Std Dev
PID with PSO	18.9	0.28
Artificial Bee Colony	23.8	1.00
Particle Swarm	21.4	1.72
Genetic	22.4	2.84
Firefly	52.9	14.56
Simulated Annealing	45.8	25.40
General Policy Search	18.2	0.60

**Figure 4:** Comparison of PID controller and Firefly performances

5.2 Multistage Extraction Case Study Results

The mean rewards and standard deviations of results of Multistage Extraction case study are presented in Table 14.

Table 14: Multistage Extraction Case Study Results

Algorithm	Mean Reward	Std Dev
Particle Swarm	44.9	21.8
General Policy Search	12.2	4.90
Random PSO	26.8	4.55

5.3 Bayesian Optimisation of Neural Network Architecture

Table 15 shows the results for determining the optimal number of layers with Bayesian optimisation. Table 16 demonstrates the variance of optimal set of neural network and GPS hyperparameters for two-layer neural network optimised with Bayesian Optimisation. r is radius, ϕ is search ratio, R denotes reward, and Std is standard deviation. Figure 5 shows the improvement of BO optimised GPS compared to GPS without hyperparameter optimisation. The upper two subplots depict the trajectory of controlled variables (X_5 and Y_1), and lower two of the actions L and G .

Table 15: Screening Results for Optimal Number of Layers

Hidden Layers	Reward
1	23.4
2	14.7
3	27.1
4	23.0

Table 16: Variance of Bayesian Optimisation Results (r = radius, ϕ = search ratio, R = reward, Std = standard deviation)

Layer 1 Layer 2	r	ϕ	R	Std
$16 \times \tanh$ $4 \times \text{sigmoid}$	0.208	0.388	7.62	1.97
$10 \times \tanh$ $4 \times \text{leakyReLU}$	0.273	0.298	9.1	2.62
$8 \times \text{leakyReLU}$ $5 \times \text{leakyReLU}$	0.5	0.279	11.7	2.58
$16 \times \text{ReLU}$ $8 \times \text{sigmoid}$	0.417	0.173	7.9	1.17

6 Discussion

6.1 CSTR Case Study Insights

The CSTR Case Study served as a preliminary low-computational demand measure to disregard unsuitable algorithms from being explored. As Table 13 shows, Simulated Annealing and Firefly algorithms had a high standard deviation and reward, making them a poor fit for this model. In the case of Simulated Annealing, its main disadvantage to most other algorithms is in the fact that it is a single-point algorithm, and one point is insufficient to explore a high-dimensional parameter space. Furthermore, there are insufficient mechanisms built in to exploring the parameter space in Simulated Annealing, as the algorithm only explores the parameter space through small perturbations in each dimension.

The best performing algorithms were the standard PID controller tuned using the PSO algorithm, and the GPS algorithm, both of which had a low standard deviation, indicating the consistency of their performance. This again demonstrates the robustness of PID control, and why it should be used when applicable. However, this data doesn't take into account the time required to establish an appropriate search space for all of the parameters in PID control. In the case of multivariate control, where multiple states and multiple actions are involved, PID control becomes significantly harder to tune; as establishing approximate starting guesses for all of the parameters in PID control requires thorough experimentation with process dynamics. On the other hand, the neural network approach automates the exploration of the process dynamics by incorporating the model

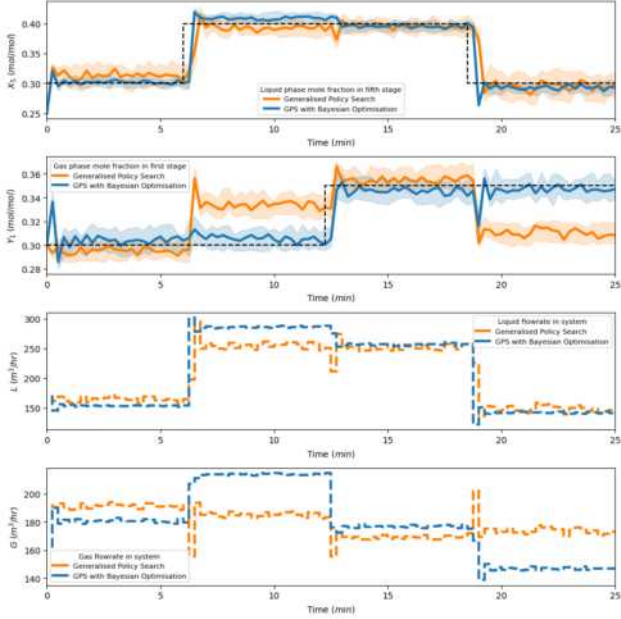


Figure 5: Comparison of Generalised Policy Search performance with and without Bayesian optimisation performances

equations into the parameter search space.

From the data shown in Table 13, the PSO and GPS algorithms were chosen for further testing on the multistage extraction model. The main advantage of these algorithms is their simplicity, which is especially the case for the GPS algorithm. The PSO algorithm is easy to vectorise, making it more computationally efficient, enabling it to run more iterations within the given time frame. Likewise, GPS involves a simple Random Search followed by a Local Search, which is easy to run, enabling it to carry out many function evaluations which allows the algorithm to explore the search space further.

6.2 Multistage Extraction Case Study Insights

Table 14 shows the results of testing the aforementioned algorithms on the multistage extraction model. Due to the higher computational demand, the runtime for each search was set to 300 seconds. It can be observed that the PSO algorithm's performance suffers significantly from this increase in complexity, especially given its impressive performance in the CSTR case study, as shown in Section 5.1. A likely cause of this is the nature of the model that the neural network interacts with. Given 2 set-points to adhere to in the multistage extraction model and 2 actions to use, the agent has enough degrees of freedom to achieve any configuration of set-points and set-point changes. However, during its exploration, the agent has a high probability of coming across a sub-optimal solution that consists of using both actions to control the 'easier' set-point, and ignore the existence of another one. In order to find a

solution that improves upon this, the agent would have to first give up control of the 'easier' set-point, and adjust the actions that look to control the second state, while simultaneously keeping the first state satisfied. In that sense, this sub-optimal solution is a local optimum, and the PSO algorithm is under-equipped to handle local optima when compared to the GPS or Genetic algorithms, which have random co-ordinate generation at every step. On the other hand, the GPS algorithm performed much better: as the algorithm is essentially a simple Random Search followed by a Local Search, there is a lot of randomness involved, which helps the algorithm avoid the pitfalls of sub-optimal solutions. It must be noted that due to different weights being used to calculate the reward for the multistage extraction model seen in Table 9, the rewards of algorithms cannot be compared across case studies. Therefore, while the mean reward for the GPS algorithm was lower in the Multistage Extraction model compared to the CSTR model, this does not indicate a better performance in the more complex model. These results inspired the testing of a hybrid 'Random PSO' algorithm that incorporated Random Search in an attempt to initially set the algorithm on the correct path, by initially exploring the search space through a Random Search before activating the more exploitative PSO algorithm. This gave the algorithm a marginal boost in performance and removed some cases where the PSO algorithm was getting 'stuck' - both the mean and the standard deviation of the reward decreased. However, the algorithm was still unable to consistently control both set-points, which required further investigation.

6.3 Neural Network Bayesian Optimisation Insights

Due to the superiority of the GPS algorithm compared to the other algorithms in the Multistage Extraction case study, it was further improved on by incorporating it into an outer hyperparameter optimisation loop, as described in Section 4.4. The result was a significant improvement in control, as seen in Figure 5. Qualitatively, the algorithm is able to handle conflicting set-point changes, and keep track of both state variables simultaneously. In contrast, the GPS algorithm can handle controlling one of the variables very well, but struggles to keep the second one under control at the same time. Table 16 also shows optimal hyperparameters yielded by the BO algorithm, and it is evident from the rewards that there is a large improvement in performance. It should be noted that the optimal initial radii r and shrink ratios ϕ are all relatively low, meaning a more exploitative variation of the local search stage

is desirable for this model. Furthermore, the best performing neural network architecture uses the tanh and sigmoid activation functions. This is justified by the core equations that the multistage extraction model is governed by - the phase equilibria for each stage are based on an exponential correlation, and both tanh and sigmoid activation functions make use of the exponential function. This facilitates the learning process of the neural network by adjusting the parametrisation of the neural network to fit the process dynamics.

7 Conclusions and Future Work

This project explores the novel concept of using trained neural networks, as opposed to tuned PID controllers, to control chemical process units. The main goal was to develop an algorithm that is sophisticated enough to train a neural network to control a range of process units and models. Various algorithms were used for training a neural network that acted as an agent in a reinforcement learning loop of controlling a process. The most prominent algorithms were the General Policy Search, due to low computational cost, and Particle Swarm Optimisation, due to it being parallelisable. These two algorithms were further tested against a more challenging model, where the General Policy Search algorithm outperformed the Particle Swarm algorithm significantly. One possible cause of this was the PSO's tendency to converge on local optima, and two main approaches were used to improve the performance of the algorithms. For the case of the PSO algorithm, it was merged with a Random Search to form a hybrid algorithm with more exploring traits. This yielded marginal improvements to the performance of the algorithm, but still resulted in unsatisfactory control of the process. Therefore, the GPS algorithm, optimising the neural network parameters, was combined with a Bayesian Optimisation algorithm, which changed the GPS algorithm hyperparameters to improve its performance. The Bayesian Optimisation algorithm yielded hyperparameters for the GPS algorithm and the neural network that led to a significant and sustained improvement of the GPS algorithm, showing capability of controlling multiple states at once with multiple actions, thereby successfully creating an algorithm that can control a range of processes. Further research could be done on the Bayesian hyperparameter optimisation algorithm, such as looking at ways of making it more efficient. Furthermore, further testing could be done on this algorithm with even more convoluted models, such as plant-wide models that include controlling multiple process units at once.

8 Acknowledgements

We would like to thank the Ph.D. student Maximilian Bloor for his continued support in implementing the models used to benchmark the algorithms.

9 Supplementary Information

The code used for this project can be found [here](#).

References

1. Levine, W. S. *Control System Applications* (CRC Press, 2000).
2. *The PID Controller and Theory Explained* <https://www.ni.com/en/shop/labview/pid-theory-explained.html>. (accessed: 27.11.2023).
3. Skogestad, S. *Simple analytic rules for model reduction and PID controller tuning* tech. rep. (Norwegian University of Science and Technology, 2003).
4. Ziegler, J. G. & Nichols, N. B. Optimum settings for automatic controllers. *Transactions of the ASME* (1942).
5. Bloor, M. *Hierarchical Reinforcement Learning for Plant-Wide Control* tech. rep. (Imperial College London, 2022).
6. Shoham, Y., Powers, R. & Genager, T. *Multi-agent reinforcement learning: a critical survey* tech. rep. (Stanford University, 2003).
7. Amini, A., Soleimany, A., Karaman, S. & Rus, D. *Spatial Uncertainty Sampling for End-to-End Control* 2019. arXiv: [1805.04829](https://arxiv.org/abs/1805.04829) [cs.AI].
8. Luus, R. & Jaakola, T. H. I. Optimization by direct search and systematic reduction of the size of search region. *AIChE Journal* **19**, 760–766 (1973).
9. Kennedy, J. & Eberhart, R. *Particle swarm optimization* in *Proceedings of ICNN'95 - International Conference on Neural Networks* **4** (1995), 1942–1948 vol.4.
10. Močkus, J. *Bayesian Approach to Global Optimisation* (Dordrecht: Kluwer Academic, 1989).
11. *Nonlinear Model Predictive Control* <http://apmonitor.com/do/index.php/Main/NonlinearControl>. (accessed: 21.11.2023).
12. Ingham, J. *Backmixing in a Multi Stage, Multi-mixer Liquid-liquid Extraction Column* <https://books.google.co.uk/books?id=Mki4zQEACAAJ> (The author, 1970).

Characterisation of a Complex Mixture of Tri-, Di- and Monoacylglycerols from Ethanolysis of Sunflower Oil by NMR Spectroscopic Techniques

Ana Derqui Serrano and Mia Sophie Jones

Department of Chemical Engineering, Imperial College London, U.K.

Abstract In this study, several characterisation techniques were employed to analyse the composition of a glyceride mixture obtained by mechanochemical ethanolysis of triacylglycerols (TAGs) contained in sunflower oil. The desired products for this reaction were mono- and diacylglycerols (MAGs and DAGs). The focus was placed on 1D and 2D NMR techniques, specifically ^1H -NMR, DEPT-135 ^{13}C -NMR and HSQC. The efficacy of GPC and FTIR for characterisation was also tested. GPC analysis proved to be inefficient due to co-elution of species. FTIR analysis identified the hydroxyl group characteristic of MAGs and DAGs in the purified product. It also revealed the presence of water and absence of glycerol. However, it did not enable complete characterisation of the product mixture. ^1H -NMR confirmed the absence of glycerol and the presence of 1,2-DAG. More detailed characterisation was not possible with this technique due to significant signal overlap which arises from complex mixtures. The 2D NMR technique HSQC was therefore performed which successfully separated the signals overlapped in 1D NMR. This allowed the identification of the remaining species present in the mixture and the determination of the product composition, given on a water-free basis, by semi-quantitative analysis. It was found that the purified product contained unreacted triglycerides (6 mol%), 1,2-diglycerides (6 mol%), 1,3-diglycerides (40 mol%), fatty acid ethyl esters (33 mol%) and unevaporated ethanol (15 mol%).

Keywords: Ethanolysis, MAGs, DAGs, TAGs, ^1H -NMR, DEPT-135, HSQC, GPC, FTIR

1. Introduction

Mono and di-acylglycerols (MAGs and DAGs), also known as mono- and diglycerides, are compounds classified as non-ionic surfactants which are widely used in many industries such as the food and pharmaceutical sectors due to their stabilising, emulsifying and thickening properties. Mixtures of food grade MAGs and DAGs are valued at around £1.0/kg^[1], whereas those of analytical grade are valued at £189/kg^[2]. The global market for these surfactants is expected to grow significantly in the next decade, at a CAGR (Compound Annual Growth Rate) of 7.20% between 2024 and 2032^[3]. Thus, further research into the synthesis of these products is valuable from an economic standpoint. These compounds can be produced by transesterification of vegetable oil containing triacylglycerols (TAGs) with a short chain alcohol - such as methanol, ethanol or glycerol - in the presence of a suitable catalyst. This is a stepwise reaction mechanism and is shown in Figure 1 below, where the reactants and products involved at each reaction step are identified.

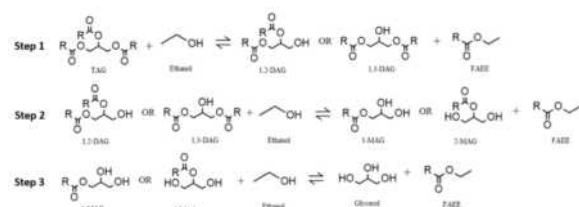


Figure 1. The stepwise ethanolysis reaction mechanism adapted from Yusoff et al.^[4]

Ethanol (EtOH) was used in this study and thus the fatty acid alkyl ester produced at each step is fatty acid ethyl ester (FAEE). Ethanol was chosen over methanol due to the toxicity of the latter since MAGs and DAGs are used in cosmetics and edible products. Sodium hydroxide (NaOH) was used as the catalyst because it allows for

shorter reaction times, milder reaction conditions and smaller alcohol volumes compared to an acid catalyst^[5].

MAGs and DAGs present in two positional isomeric forms depending on the location of the ester bonds. Therefore, from the reaction mechanism shown in figure 1 and the sample work-up, a total of up to 9 species could be present in the product mixture: 1-MAG, 2-MAG, 1,2-DAG, 1,3-DAG, TAG, FAEE, glycerol, ethanol, and water. This results in a very complex mixture that must be characterised. The structures of these molecules are given in figure 2 below.



Figure 2. Reactant and product structures for the ethanolysis of TAGs. The carbons of interest for the NMR analysis are numbered. The substituents labelled “R” refer to the fatty acid chains.

The initial intention of our research was to study the feasibility of mechanochemical catalysis for DAG and MAG synthesis by surveying various reaction conditions and comparing results to those obtained by the conventional heating and stirring methods^[6]. However, the complexity of our product mixture and thus of the spectroscopic data obtained did not yield rapid conclusions regarding the product composition. Therefore, the research focus was switched to NMR (Nuclear Magnetic Resonance) spectroscopic analysis of the product mixture. Gel Permeation Chromatography (GPC) and Fourier Transform Infrared Spectroscopy (FTIR) experiments were also conducted, but to a lesser extent. It was deemed useful to compare the efficacy of each technique for the characterisation of

our mixture. The characterisation by NMR involved ^1H -NMR, DEPT-135 (Distortionless Enhancement by Polarisation Transfer) ^{13}C -NMR and HSQC (Heteronuclear Single Quantum Coherence) analyses. Data processing was carried out with the aid of predicted and experimental chemical shifts from ChemDraw[®] and literature, respectively. Using semi-quantitative HSQC analysis, the purified product was found to contain 40 mol% 1,3-DAG and 6 mol% 1,2-DAG on a water-free basis. 4.05g of purified product was collected. No presence of monoglycerides was detected.

The aim of our work was to help speed up this characterisation process for future research into the ethanolysis of TAGs by highlighting some of the key issues that may be encountered.

2. Background

Spectroscopic techniques, such as high resolution ^1H - and ^{13}C -NMR, paired with computational methods, are key experiments that have been employed to study the composition of vegetable oils and other complex lipid mixtures^[7]. ^1H -NMR is a particularly useful technique due to its non-invasiveness, rapidity, and sensitivity^[8].

Numerous papers^{[9],[10]} have been published on the characterisation of glyceride mixtures. Notably, Hatzakis et al.^[8] offers a detailed comparison of spectroscopic techniques useful for characterisation including ^1H -NMR and ^{13}C -NMR data. However, the use of 1D ^1H -NMR spectra alone is often insufficient for unambiguous identification of complex molecules^[11]. This difficulty is enhanced for mixtures, specifically mixtures of structurally related molecules^[12] such as mono- and diglycerides because significant signal overlap can be observed.

Advances have been made in the automation of NMR spectrum processing to facilitate these issues, such as the use of deep neural network (DNN)-based approaches for peak picking and spectral deconvolution^{[13],[14]}. However, these are still in early stages of development. Therefore, the ability to characterise mixtures manually and to understand which factors may affect the spectroscopic results is still of great importance.

To overcome the limitations of ^1H -NMR experiments, multiple NMR analysis techniques can be used in conjunction for complete characterisation of complex mixtures^[8]. For example, ^{13}C -NMR experiments are known for their characterisation efficacy as they enable the identification of the carbon atoms of each species in the mixture. The wider spectral range of ^{13}C -NMR experiments^[15] compared to that of ^1H -NMR greatly reduces the incidence of peak overlap. However, ^{13}C -NMR experiments are lengthy because long relaxation delays are necessary to obtain a satisfactory signal to noise ratio^[8]. Therefore, it is useful to analyse the performance of faster 1D and 2D-NMR experiments, such as DEPT (Distortionless Enhancement by Polarisation Transfer) ^{13}C -NMR and HSQC (Heteronuclear Single Quantum Coherence). DEPT ^{13}C -NMR shows signals for protonated carbons and is faster than traditional ^{13}C -NMR because it is a double resonance program. It transfers the polarisation

from one excited nucleus to another, specifically from ^1H to ^{13}C , which results in sensitivity enhancement compared to the decoupled 1D ^{13}C -NMR experiment^[16]. Specifically, ^{13}C DEPT-135 is a useful experiment as it uses a proton pulse angle of 135° and thus results in a spectrum in which CH_2 signals have an opposite phase orientation to those of CH and CH_3 . This allows for easier identification of groups.

HSQC is a 2D NMR technique based on the magnetisation transfer from a ^1H nucleus to a neighbouring ^{13}C nucleus and back to the former after a time delay t_1 . This last transfer is detected and creates the signal that is recorded on the spectrum^[17]. Specifically, it determines single bonded C-H correlations by plotting the ^1H - and DEPT-135 ^{13}C -NMR spectra on separate axes. In this way, cross-peaks are determined which enables the separation of signals that appear overlapped in 1D ^1H -NMR spectra. Thus, species characterisation is greatly facilitated for complex mixtures.

3. Methods

3.1. Chemicals

Ethanol (96.3 v/v%, $\text{C}_2\text{H}_5\text{OH}$), sodium hydroxide pellets (99.1%, NaOH), anhydrous sodium sulphate (100%, $\text{Na}_2\text{O}_4\text{S}$), deuterated chloroform (99.8% D, CDCl_3), and sodium chloride (99.5%, NaCl) were purchased from VWR Chemicals[®]. Acetone ($\geq 99.5\%$, $\text{C}_3\text{H}_6\text{O}$), Ion exchanger Amberlite[®] IR-120 (100%), and *N,N*-Dimethylformamide ($\geq 99.9\%$, $\text{C}_3\text{H}_7\text{NO}$) were purchased from Sigma-Aldrich[®]. Vegetable glycerol (99.9%, $\text{C}_3\text{H}_8\text{O}_3$) was purchased from Onyx[®]. Commercial grade sunflower oil was purchased from KTC[®].

3.2. Reaction Conditions

47.51 g of sunflower oil, 2.49 g of EtOH and 0.13 g of NaOH ^[18] were introduced into each of the two milling beakers. The oil was weighed using an A&D[®] GX-6000 balance. The NaOH catalyst was crushed to a fine powder using a mortar and pestle. The EtOH and NaOH were weighed using an A&D[®] BM-252 microbalance. The catalyst was dissolved in the ethanol in a glass vial by intermittent manual shaking and gentle heating using an IKA[®] RH-Digital magnetic hotplate where the temperature was controlled with an IKA[®] ETS-D5 thermocouple.

Five stainless steel milling balls of 10mm diameter were used in each beaker. The shaker mill (Retsch[®] Mixer Mill MM 500 nano) was set to a frequency of 30Hz. The reaction duration was 5 minutes.

3.3. Purification

To quench the reaction, the crude product was stirred with a 10% excess of Amberlite[®] IR 120 with respect to the stoichiometric amount required on an IKA[®] RH-Digital magnetic hotplate at 70°C for 10 minutes. The temperature was controlled with an IKA[®] ETS-D5 thermocouple. The mixture was then separated from the Amberlite[®] IR 120. The following methodology was adapted from Hobuss et al.^[19]. The product was washed twice with a 5 w/w% aqueous NaCl solution to break the

emulsion and remove any potential glycerol. The mass of solution used equalled the mass of the product. This step involved heating the mixture to 70 °C and stirring it for 5 minutes. It was then cooled in an ice bath to facilitate phase separation.

The upper phase was collected and centrifuged in a Sigma® 3-18 centrifuge at 8000 rpm for 5 minutes. Then, the post centrifugation upper phase was weighed and stirred with an equal mass of ethanol for 5 minutes at 2000 rpm. The mixture was then added to a 100 mL decanter and the bottom phase, predominantly oil, was removed. It was washed with ethanol again in the same manner to extract any remaining product.

The upper phase, containing ethanol and the product, was collected each time from the decanter. The solution was dried using anhydrous sodium sulphate until it was free-flowing. It was then collected by pipetting to avoid transfer of the sodium sulphate and placed in a Heidolph® Rotacool rotary evaporator at 170 mbar, 50.5 °C and 120 rpm for 50 minutes to remove the ethanol. Finally, the product was filtered using a syringe filter tip.

3.4. Analytical Procedures

3.4.1. FTIR Experiments

To prepare the samples for FTIR, 1 mL of analyte was transferred to Eppendorf tubes. A droplet of each sample was used for analysis by an Agilent Technologies® Cary 630 FT-IR spectrometer. The results were viewed using the MicroLab FT-IR software.

3.4.2. GPC Experiments

To analyse the samples by GPC, 10-20 mg of sample were introduced into Eppendorf tubes containing 1 mL of the mobile phase (N,N -Dimethylformamide). The analysis was performed by a Shimadzu® High Performance Liquid Chromatography Workstation with an autosampler (SIL-20AHT), oven (CTO-20A), PDA (SPD-M20A), and RI detector (RID-20A). The injection volume was 5 µL, the separation was performed on a set of M and L Polar Gel columns (300mm×7.8mm, Agilent) operating at 60°C with a mobile phase flow of 1 mL/min. The results were processed using the LabSolutions GPC software.

3.4.3. NMR Experiments

The samples for NMR analysis were prepared by introducing 30 µL of analyte into a 5 mm diameter tube containing 1000 µL of CDCl₃. The techniques employed were ¹H, ¹H-¹³C HSQC and DEPT-135 ¹³C-NMR spectroscopy. The analyses were performed in a JEOL® 400 MHz spectrometer operating at 400 and 100 MHz for proton and carbon-13 nuclei, respectively. All experiments were performed at 25 °C.

¹H-NMR spectra were recorded with the following acquisition parameters: 2 scans; relaxation delay 4 s; pulse width 3.2 µs; acquisition time 5.46 s; spectral width 15 ppm; acquired data points 40960.

¹³C-NMR spectra for the DEPT-135 experiments were recorded with the following acquisition

parameters: 512 scans; relaxation delay 2 s; pulse width 11.1 µs; acquisition time 1.30 s; spectral width 200 ppm; acquired data points 40 960.

HSQC spectra were recorded with the following acquisition parameters: 16 scans; relaxation delay 1.5 s; pulse width 6.4 µs; acquisition time 0.29 s; spectral width 9.5 ppm for ¹H-NMR and 170 ppm for ¹³C-NMR; acquired data points 1280 for ¹H-NMR and 32 for ¹³C-NMR.

3.4.4. Spectrum Processing using MestReNova®

The obtained spectra were processed using MestReNova® by Mestrelab Research®. The chloroform signal was used as the reference and set at 7.26 ppm and 77.16 ppm for ¹H- and DEPT-135 ¹³C-NMR spectra, respectively. For the ¹H- and DEPT-135 ¹³C-NMR experiment, spectral data points were increased to 52 430 and 104 858, respectively, using zero filling. For the HSQC experiment, spectral data points were increased to 1640 for ¹H-NMR and 1024 for ¹³C-NMR using zero filling. Baseline correction was performed by applying a polynomial third order function for ¹H-NMR spectra and a Bernstein Polynomial third order function for DEPT-135 ¹³C-NMR spectra. The latter were denoised to a 0.3 noise factor. Automatic phase correction was sufficient for ¹H-NMR spectra, but manual phase correction was required for DEPT-135 ¹³C-NMR spectra.

For HQSC spectra, the DEPT-135 ¹³C- and ¹H-NMR files must be open and adequately processed in the same document, so the HSQC spectrum shows the correct horizontal and vertical traces which allows for phase adjustment and accurate cross-peak to species assignment. Exponential apodization of 3 Hz and 5 Hz was performed along axes f1 and f2 respectively. Phase correction was applied manually.

3.4.5. ChemDraw® Predictions

Predictions of the ¹H- and ¹³C-NMR spectra in CDCl₃ for an applied field of 400 MHz and 100 MHz, respectively, were performed with ChemDraw®. For ¹H-NMR, it estimates approximately 90% of C-H_x groups with a standard deviation of between 0.2 and 0.3 ppm. For ¹³C NMR, it estimates 95% of the shifts with a standard deviation of 2.8 ppm^[20]. The simulations were carried out for linoleic glycerides since linoleic acid is the predominant fatty acid present in sunflower oil (71.73%)^[21].

4. Results and Discussion

The crude product mixture was very viscous and had a fruity smell characteristic of FAEE^[22] which qualitatively indicated product formation.

4.1. GPC Analysis

GPC analysis was conducted on the crude product prior to oil separation. Further GPC analysis could not be conducted due to solubility issues of the purified product in DMF and tetrahydrofuran (THF).

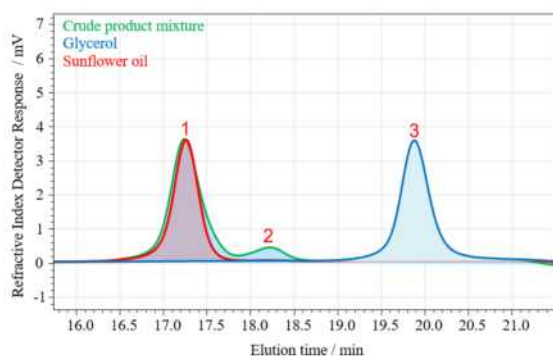


Figure 3. Superposition of the GPC spectra of the crude product mixture, pure glycerol and pure sunflower oil.

One high and one low intensity peak at the retention times of 17.2 (peak 1) and 18.2 minutes (peak 2) are determined, respectively. Peak 1 corresponds to an overlap of the pure sunflower oil and product signals. However, the product signal is broader than that of the sunflower oil. This suggests co-elution of species in the column.

GPC is a technique that identifies compounds by molecular weight separation. Larger molecules have shorter retention times than smaller molecules because they cannot penetrate the pore size of the column packing^[23]. This means that the hydrodynamic radius of the species is a key factor in the separation efficacy, not only its molecular weight. The hydrodynamic radius of a molecule in solution, also known as Stokes radius, is the radius of a sphere whose diffusion coefficient is equivalent to that of the molecule^[24].

Given this, and by contrasting with our NMR results discussed in section 4.3, we expect the co-elution of TAG and 1,3-DAG, whose hydrodynamic radii are reasonably similar, to result in peak 1. Similarly, peak 2 is expected to correspond to the co-elution of FAEE and 1,2-DAG. Their hydrodynamic radii are smaller than those of TAG and 1,3-DAG so are retained longer in the column packing. The higher intensity of peak 1 compared to peak 2 can be explained by the large amounts of unreacted oil present in the crude product, as discussed in section 4.4.

Co-elution occurs when the difference in the retention times of the analytes is inferior to the resolution of the analysis^[25]. It is therefore a severe limitation of GPC analysis which does not allow for accurate identification of species in the product mixture. Co-elution could be overcome by using a higher resolution column^[26]. However, the variation of column resolution to suit a new type of mixture is both time and cost-intensive^[27]. Therefore, GPC is not recommended as a characterisation technique for complex glyceride mixtures.

4.2. FTIR Analysis

From the analysis of figure 4, some of the signals in the glycerol spectrum do not match those of the purified product. Therefore, as detected by GPC analysis, glycerol is absent in the product.

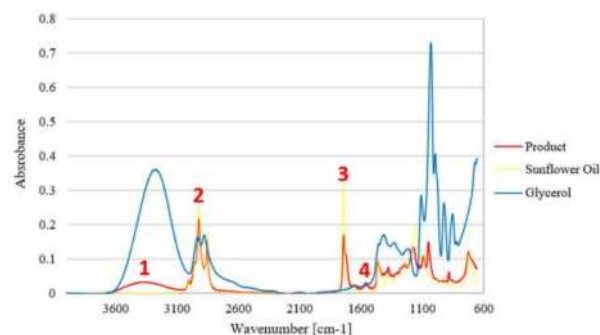


Figure 4. Superposition of the FTIR spectra of the purified product, pure glycerol and pure sunflower oil.

Peak 1 of the product spectrum is a broad signal at 3550–3200 cm^{-1} which corresponds to the intermolecular bonded alcohol group^[28]. This suggests the presence of MAG and/or DAG in the product because TAG does not contain hydroxyl groups. Peaks 2 and 3 correspond to exact superpositions of the signals in the sunflower oil and the product spectra. This indicates shared moieties between compounds in the mixture and in the pure sunflower oil. However, it does not confirm the presence of oil in the product as all glycerides share similar structures, namely the glycerol backbone and fatty acid chains^[29]. Finally, peak 4 is a weak signal at 1645 cm^{-1} and indicates the presence of water as it corresponds to its bending mode of vibration^[30].

We can conclude that FTIR is an effective characterisation technique but is not detailed enough for complete characterisation of complex glyceride mixtures. It mainly enables the determination of functional groups and does not inform on the total hydrogen and carbon content in the mixture^[31]. Therefore, the use of NMR techniques is required.

4.3. NMR Analysis

The structures of the species that may be present in our product are shown in figure 2, which includes the carbon numbering convention used in this report. These labelled groups will be referred to in the following sections. For example, when referring to group C3-H₂ of 1,2-DAG, C3 corresponds to the terminal carbon on the glycerol backbone that is bonded to the hydroxyl group.

4.3.1. ¹H-NMR Analysis

In figure 5, the signals from glycerol are not observed on the product spectrum, for example in the 3.3–3.6 ppm region. Therefore, ¹H-NMR results confirm the absence of glycerol in the product mixture, as also found from the GPC and FTIR results. Two signals at 3.72 ppm and 5.08 ppm, which do not appear on the sunflower oil or glycerol spectra, are observed. These are specific to 1,2-DAG so its presence is identified in the product.

Nevertheless, determining the presence or absence of unreacted sunflower oil as well as of the different product species is not as unequivocal. There is a severe overlap of the signals from the oil and product spectra in the 4.1–4.3 ppm and 5.2–5.4 ppm regions. In the former region, there is an overlap between the signals from the protons bound to the terminal glyceryl carbons (C1/C3-H₂) of both TAG and the products. In the latter, the

signals from the protons bound to the central glyceryl carbon (C2-H) of TAG overlap with those from the fatty acid chain unsaturations, which are present in all glycerides.

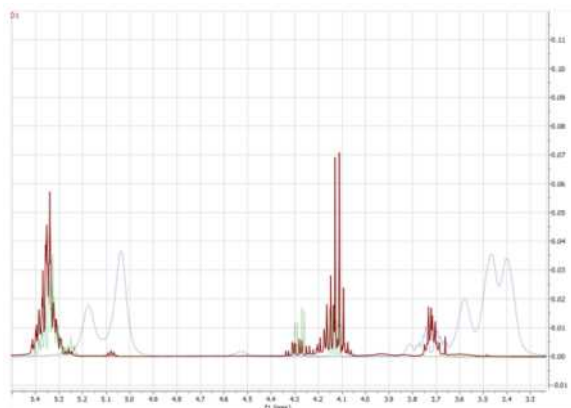


Figure 5. Superposition of the ^1H -NMR spectra of the purified product (red), pure glycerol (blue) and pure sunflower oil (green).

To facilitate the analysis of overlapping peaks, the in-built GSD (Global Spectral Deconvolution) method in MestReNova[®] was employed. It successfully revealed distinct peaks hidden under apparent envelopes. However, given the significant overlap between the signals, clear multiplet patterns could not be inferred. Moreover, the singlet corresponding to H_2O at 1.6 ppm is overlapped with signals from the methylene groups in the fatty acid chains present in all glycerides and is thus difficult to identify from ^1H -NMR. The presence of residual water in the product mixture is known from the FTIR results. Therefore, limiting the analysis of complex mixtures to solely ^1H -NMR experiments does not allow for accurate product characterisation. The

overlap of signals is a significant limitation of this technique which justifies the need for complementary techniques such as DEPT-135 ^{13}C -NMR or 2D NMR to accurately assign the signals to the correct species. HSQC (Heteronuclear Single Quantum Coherence) is a 2D method which is particularly useful, since it indicates which protons are attached to which carbons via $^1\text{J}_{\text{C-H}}$ coupling.

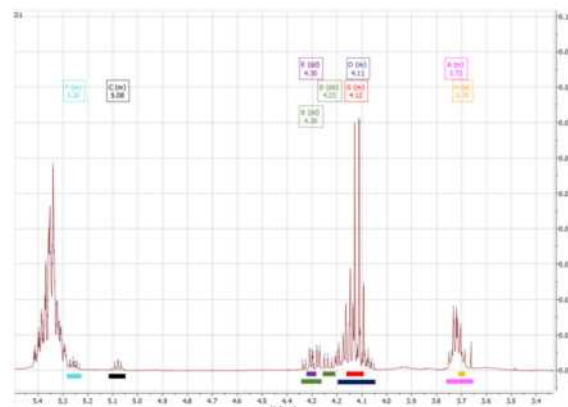


Figure 6. Assignment of ^1H -NMR signals of the purified product mixture to species. The signal boundaries are shown by the colour bands.

The ^1H -NMR spectrum signals were assigned to each species after contrasting the ^1H , DEPT-135 ^{13}C and HSQC spectra and comparing their data with literature. These assignments are shown in figure 6. Table 1 correlates the labels in the figure to the corresponding protons in each species. It also summarises the chemical shift data obtained from ^1H -NMR experiments, ChemDraw[®] predictions and literature.

Table 1 Experimental, Predicted and Literature Results for the ^1H -NMR Analysis of the Purified Product.

Species	Group	Label	^1H -NMR Chemical Shift (ppm)		
			Experimental	ChemDraw [®]	Literature
1,2-DAG	C3-H ₂	A	3.72	3.68;3.62	3.72 ^[8]
	C1-H ₂	B	4.30;4.23	4.42;4.17	4.31;4.23 ^[8]
	C2-H	C	5.08	5.26	5.08 ^[8]
1,3-DAG	C2-H and C1/C3-H ₂	D	4.04-4.20	4.21-4.78	4.07-4.18 ^[8]
TAG	C1/C3-H ₂	E	4.30	4.30	4.29 ^[8]
	C2-H	F	5.26	5.85	5.26 ^[8]
FAEE	C1-H ₂	G	4.12	4.01	4.13 ^[32]
Ethanol	C1-H ₂	H	3.70	3.59	3.63 ^[33]

The groups labelled C1/C3-H₂ correspond to the two identical CH₂ groups of the species. *Two signals are observed for the C1-H₂ group of 1,2-DAG. This is because the protons of this group are diastereotopic^[34].

4.3.2. DEPT-135 ^{13}C -NMR Analysis

The signals in figure 7 are assigned using ChemDraw[®] predictions and literature^[8]. On DEPT-135 ^{13}C spectra, methine group signals appear downfield whereas those of methyl appear upfield. Quaternary carbons do not produce a signal as they are not bonded to protons.

^{13}C -NMR has a much larger spectral width than that of ^1H -NMR, 200 ppm compared to 15ppm^[15]. This is an interesting feature of ^{13}C -NMR since it greatly reduces the incidence of overlaps. DEPT-135 ^{13}C -NMR is faster than decoupled broadband ^{13}C -NMR and presents the

additional advantage of differentiating the signals from methylene groups from those of methine and methyl groups since they are shown in opposite orientations on the spectrum, as seen in figure 7. However, the study of DEPT-135 ^{13}C -NMR on its own did not yield insight on which carbon groups correspond to which glyceride species given their structural similarity. Therefore, 2D NMR was required for a more complete analysis. Nevertheless, the DEPT-135 ^{13}C -NMR spectrum was useful for accurate processing of the HSQC spectrum, as mentioned in section 3.4.4.

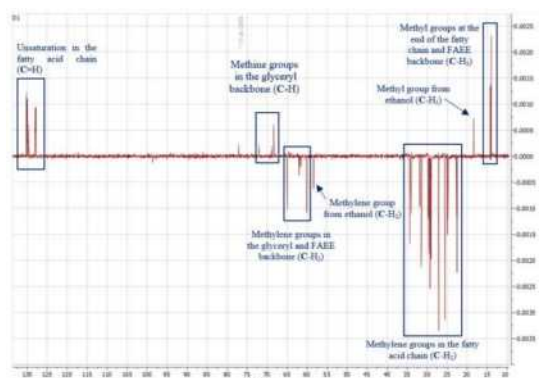


Figure 7. Assignment of DEPT-135 ^{13}C -NMR spectrum signals. Negative phase signals correspond to CH_2 groups. Positive phase signals correspond to CH and CH_3 groups.

4.3.3. ^1H - ^{13}C HSQC Analysis

The HSQC spectrum presents cross peaks, where the vertical axis corresponds to the ^{13}C -NMR chemical shifts and the horizontal axis to those of ^1H -NMR. Thus, identification of which protons are bonded to which carbons through a single bond is possible. Specifically, this spread of signals along two dimensions allows the separation of signals which appear overlapped on 1D NMR spectra^[35].

From figure 8, we see that there is significant overlap of ^1H -NMR signals between 4.05 ppm and 4.35 ppm which HSQC has been able to resolve through the plotting of cross peaks. Specifically, these overlaps occur between 4.09 ppm and 4.16 ppm for FAEE and 1,3-DAG; between 4.2 ppm and 4.35 ppm for 1,2-DAG and TAG; and between 3.68 ppm and 3.75 ppm for ethanol and 1,2-DAG.

Indeed, the terminal glyceryl methylene groups of TAG and the terminal glyceryl methylene group in position 1 of 1,2-DAG present signals which are overlapped on the ^1H -NMR spectrum around 4.30 ppm (signals E and B in figure 6 and table 1). The overlap between the signals from 1,3-DAG and the methylene group from the ethyl chain of FAEE between 4.09 ppm and 4.16 ppm is also evidenced (signals D and G). Furthermore, HSQC analysis allowed the separation of the single signal for 1,3-DAG obtained from ^1H -NMR analysis (signal D) into individual cross peaks at (4.15 ppm; 65.07 ppm) for the terminal glyceryl methylene groups and at (4.07 ppm; 68.39 ppm) for the central glyceryl methine group.

HSQC analysis also revealed the presence of ethanol whose methylene group signal overlaps with that of terminal glyceryl methylene group in position 3 of 1,2-DAG at 3.72 ppm (signals H and A). These findings confirm the necessity of 2D NMR techniques to achieve more accurate signal identification. The signals corresponding to MAGs, at 3.65 ppm, 3.94 ppm and 4.18 ppm for 1-MAG^[9] and 3.84 ppm and 4.28 ppm for 2-MAG^[9], are absent on figure 8. These species are thus

not present in the product mixture. The reasons for this will be explained in section 4.4. Table 2 summarises the chemical shift data obtained from HSQC experiments, ChemDraw[®] predictions and literature.

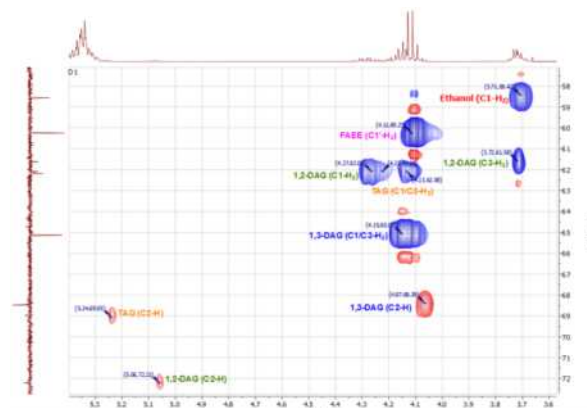


Figure 8. ^1H - ^{13}C HSQC spectrum of purified product. Red signals correspond to positive phase (CH or CH_3). Blue signals correspond to negative phase (CH_2). Unlabelled red signals correspond to residual noise after apodization was applied.

Therefore, HSQC analysis gives conclusive evidence of the presence of 1,2-DAG, 1,3-DAG, FAEE, TAG and ethanol in the final product.

It is worth noting that there is still overlap of ^{13}C -NMR signals between 1,2-DAG and TAG (signals B and E) on the HSQC spectrum at 62 ppm. This overlap can be resolved by using a higher frequency spectrometer, as was done by Hatzakis et al.^[8] In their experiments, a 600 MHz spectrometer was employed whereas a 400 MHz machine was used in our study. Thus, it is evidenced that the use of a spectrometer of sufficiently high frequency is of great importance when dealing with complex mixtures. This results in higher resolution spectra, as can be seen in figure 9^[8], which greatly facilitates analysis.

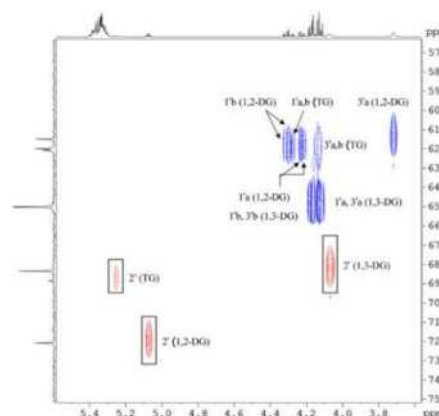


Figure 9. 600 MHz HSQC-DEPT spectrum of DAG olive oil in CDCl_3 solution taken from Hatzakis et al.^[8]

Table 2 Experimental, Predicted and Literature Results for the ^1H - ^{13}C HSQC Analysis of the Purified Product.

Species	Group	^1H -NMR Chemical Shift (ppm)	^{13}C -NMR Chemical Shift (ppm)		
		Experimental	Experimental	ChemDraw®	Literature
1,2-DAG	C3-H ₂	3.72	61.58	61.5	61.49 ^[8]
	C1-H ₂	4.27	62.08	62.5	61.98 ^[8]
	C2-H	5.06	72.21	72.3	72.09 ^[8]
1,3-DAG	C1/C3-H ₂	4.15	65.07	65.4	65.01 ^[8]
	C2-H	4.07	68.39	67.6	68.34 ^[8]
TAG	C1/C3-H ₂	4.13;4.22 ^a	62.08	62.7	62.07 ^[8]
	C2-H	5.24	69.05	69.0	68.85 ^[8]
FAEE	C1-H ₂	4.11	60.25	61.3	60.20 ^[32]
Ethanol	C1-H ₂	3.71	58.42	57.9	58.2 ^[33]

^aTwo signals are observed for these CH₂ groups of TAG. This is because the protons of this group are diastereotopic^[34].

4.3.4. Semi-Quantitative HSQC Analysis

In ^1H - ^{13}C HSQC experiments, the $^1\text{J}_{\text{CH}}$ coupling constant of a group relates to its chemical environment and is affected by atoms up to one bond away from the carbon. The $^1\text{J}_{\text{CH}}$ value of each group affects the correlation peak volume V_c of its cross peak on the HSQC spectrum, as seen in figure 10. This is due to a strong dependence between V_c and $^1\text{J}_{\text{CH}}$ in conventional HSQC where an average $^1\text{J}_{\text{CH}}$ value, $^1\text{J}_{\text{CHtune}}$, is used for analysis. This usually allows for a sufficient qualitative analysis where all correlation peaks can be detected on the spectrum. However, it significantly impacts the quantitative analysis. In figure 10, strong deviations in V_c are observed at $^1\text{J}_{\text{CH}}$ values different to $^1\text{J}_{\text{CHtune}}$. Therefore, to achieve a semi-quantitative analysis of our mixture, it is necessary to compare peak volumes between similar groups of different species. The assumption is that groups in a similar chemical environment will have similar $^1\text{J}_{\text{CH}}$ constants and therefore their correlation peak volumes will be reduced by comparable amounts.

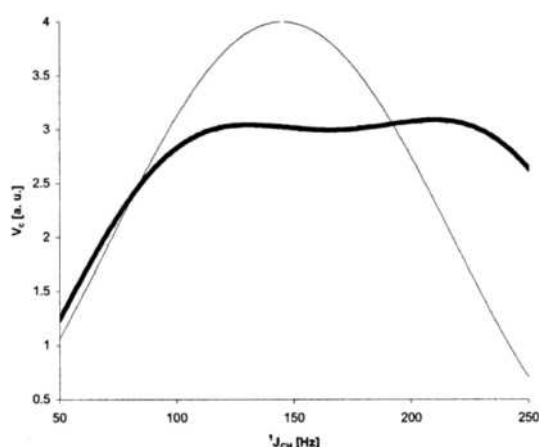


Figure 10. Simulated J-dependence of conventional (thin line) HSQC and Q-HSQC (thick line) taken from Heikinnen et al.^[36]. The conventional HSQC method uses $^1\text{J}_{\text{CHtune}}=145\text{Hz}$. The Q-HSQC method uses iterative optimisation in the $^1\text{J}_{\text{CH}}$ range of 115 – 190 Hz.

For our analysis, we compared integrals between groups C1-H₂ of FAEE and 1,2-DAG; C2-H of TAG, 1,2-DAG and 1,2-DAG; C3-H₂ of 1,2-DAG and C1-H₂ of ethanol. This allowed us to determine integral ratios of each species with respect to 1,2-DAG which is shown in equation 1.

$$\text{Integral ratio} = \frac{I_{\text{species } i, \text{ group } j}}{I_{1,2\text{-DAG, group } j}} \quad \text{Eq. 1}$$

These were used to calculate the final molar composition of the mixture, presented in table 3. The results are given on a water-free basis as non-carbon containing compounds, such as H₂O, do not appear on HSQC spectra. The results are reported to their integer values due to the semi-quantitative nature of the analysis.

Table 3 Molar composition of the purified product on a water-free basis, obtained by semi-quantitative HSQC analysis.

Species	Integral Ratio	Content (mol%)
1,2-DAG	1.00	6
1,3-DAG	6.39	40
TAG	1.02	6
FAEE	5.20	33
Ethanol	2.39	15

From this table we see that the purified product mixture is composed of 46 mol% DAG and 33 mol% FAEE. Since no MAGs were produced, we know that the reaction, shown in figure 1, did not exceed the first step. Therefore, we would expect to find FAEE and the DAG isomers in stoichiometric proportions. 1.4 times more DAG than FAEE was determined but this deviation can be explained by the approximations made for the semi-quantitative analysis.

It should be noted that Heikinnen et al.^[36] achieved a quantitative-HSQC (Q-HSQC) method which greatly minimises the dependence of the correlation peak volume on $^1\text{J}_{\text{CH}}$, as can be seen in figure 10. This was achieved by an iterative optimisation method to obtain a

uniform intensity response for V_c for a given $^1J_{CH}$ range. In this way, all the peak integrals of species in this $^1J_{CH}$ range can be directly compared, avoiding the drawbacks of our semi-quantitative method.

Another method to achieve better quantification could be to conduct ^{13}P -NMR experiments. These have proven to be the most effective at determining the concentration of glycerides^[8]. Drawbacks include long experiment durations since phosphorylation of the sample is required. Additionally, phosphorylation destroys the sample. However, it is highly accurate due to its signals being singlets which are spread over a wide spectral range. Moreover, they offer reduced analysis costs since they can be conducted using cheaper lower frequency spectrometers^[8].

4.4. Insight into the Reaction Mechanism

The characterisation of the mixture provides some insight into the ethanolysis reaction mechanism. Due to the absence of MAGs and glycerol in the product, we were able to confirm that the reaction did not exceed the first step. This could be due to the use of insufficient ethanol and/or the short reaction time of 5 minutes tested. These conditions likely resulted in insufficient formation of DAGs to push the formation of MAGs in the second step. Specifically, 37.4g of unreacted oil was collected during purification and 4.05g of purified product, consisting of 46 mol% DAGs. Therefore, a much higher number of TAGs compared to DAGs would have been available for reaction with ethanol at the time the reaction was quenched, leading to the preferential formation of DAGs instead of MAGs.

5. Conclusions

The characterisation of the glyceride mixture was effective and showed preferential synthesis of the 1,3-DAG isomer compared to the 1,2-DAG isomer. No MAGs or glycerol were produced so the reaction did not exceed the first step. Therefore, the characterisation enabled clearer insight into the reaction mechanism. The purified product contained significant quantities of FAEE and residual amounts of unreacted oil, ethanol, and water. Improvements in the purification procedure would be beneficial to increase the purity of DAGs in the mixture.

The key difficulties encountered during characterisation were highlighted. Specifically, the comparison of analytical techniques demonstrated that GPC was ineffective for the characterisation of the glyceride mixture due to the co-elution of species. FTIR was an effective technique but did not give detailed enough information because it does not inform on the hydrogen and carbon contents of the mixture. Thus, NMR proved to be the most effective technique but the combination of 1D and 2D methods is necessary. 1H -NMR alone was ineffective due to the complexity of the mixture resulting in many peak overlaps. This complexity was not only due to the number of species present in the mixture but was enhanced by their structural similarity. The use of 2D techniques, specifically 1H - ^{13}C HSQC, was successful at resolving

most of these overlaps and enabled semi-quantitative analysis.

6. Outlook

More accurate quantification could be achieved using the Q-HSQC method developed by Heikinnen et al.^[36] or by conducting ^{13}P -NMR experiments^[8]. Alternatively, ^{13}C -NMR experiments could be performed which, similarly to ^{13}P -NMR, have a wider spectral range compared to 1H -NMR. This enables unequivocal signal to species assignment and thus better quantification. If the 1H -NMR data for each of the pure components in the mixture is available, the MestReNova[®] software SMA (Simple Mixtures Analysis) plug-in allows for easier characterisation and quantification^[37].

To overcome the remaining peak overlaps in the HSQC spectrum, a higher frequency spectrometer could be employed to improve spectral resolution. From literature, a 600 MHz frequency seems to be sufficient^[8].

After achieving accurate quantification, the study of mechanocatalysis as a feasible method for ethanolysis could be addressed. Specifically, the reaction conditions for this process could be optimised and its energy efficiency compared to that of conventional methods.

Acknowledgements

The authors would like to extend their gratitude to Matheus dos Santos and Siyuan Gao for their support and guidance throughout this research project.

References

- [1] www.alibaba.com. (n.d.). High Quality Food Grade Mono-and Diglycerides Price - *Mono-and Diglycerides Price, High Quality Mono-and Diglycerides*. [online] Available at: https://www.alibaba.com/product-detail/High-quality-food-grade-Mono-and_62523328922.html?s=p [Accessed 11 Dec. 2023].
- [2]: Cambridge Bioscience Limited. (n.d.). *Mono-and diglycerides - MedChem Express*. [online] Available at: <https://www.bioscience.co.uk/product~1033672> [Accessed 13 Dec. 2023].
- [3]: Expert Market Research.com (2023). *Global Mono Diglycerides Market Report and Forecast 2024-2032*. [online] Available at: <https://www.expertmarketresearch.com/reports/mono-diglycerides-market> [Accessed 11 Dec. 2023].
- [4] Yusoff, M.F.M., Xu, X. and Guo, Z. (2014). Comparison of Fatty Acid Methyl and Ethyl Esters as Biodiesel Base Stock: a Review on Processing and Production Requirements. *Journal of the American Oil*
- [5] Wang, H., Wang, M., Zhao, N., Wei, W. and Sun, Y. (2006). *Preparation and Utilization of CaF₂-ZrO₂ as a Novel Solid base for the Synthesis of Dimethyl Carbonate from Methanol and Propylene Carbonate*.

[online] ScienceDirect. Available at: <https://www.sciencedirect.com/science/article/abs/pii/S0167299106809997>.

[6] Murhadi, Hidayati, S. and Sugiharto, R. (2019). Profile of Monoglyceride and Diglyceride Compounds of the Ethanolysis Products from Palm Kernel Oil (PKO). *IOP Conference Series: Earth and Environmental Science*, 292(1), p.012002. doi:<https://doi.org/10.1088/1755-1315/292/1/012002>.

[7] Sacchi, R., Addeo, F. and Paolillo, L. (1997). ¹H and ¹³C NMR of virgin olive oil. An overview. *Magnetic Resonance in Chemistry*, 35(13), pp.S133–S145. doi:[https://doi.org/10.1002/\(sici\)1097-458x\(199712\)35:13%3Cs133::aid-omr213%3E3.0.co;2-k](https://doi.org/10.1002/(sici)1097-458x(199712)35:13%3Cs133::aid-omr213%3E3.0.co;2-k).

[8] Hatzakis, E., Agiomyrgianaki, A., Kostidis, S. and Dais, P. (2011). High-Resolution NMR Spectroscopy: An Alternative Fast Tool for Qualitative and Quantitative Analysis of Diacylglycerol (DAG) Oil. *Journal of the American Oil Chemists' Society*, 88(11), pp.1695–1708. doi:<https://doi.org/10.1007/s11746-011-1848-2>.

[9] Nieva-Echevarría, B., Goicoechea, E., Manzanos, M.J. and Guillén, M.D. (2014). A method based on ¹H NMR spectral data useful to evaluate the hydrolysis level in complex lipid mixtures. *Food Research International*, 66, pp.379–387. doi:<https://doi.org/10.1016/j.foodres.2014.09.031>.

[10] Kumar, R., Bansal, V., Tiwari, A., Sharma, M., Puri, S.K., Patel, M.B. and Sarpal, A.S. (2011). Estimation of Glycerides and Free Fatty Acid in Oils Extracted From Various Seeds from the Indian Region by NMR Spectroscopy. *Journal of the American Oil Chemists' Society*, 88(11), pp.1675–1685. doi:<https://doi.org/10.1007/s11746-011-1846-4>.

[11] Wu, Y.-S., Li, B.-X. and Long, Y.-Y. (2022). Rapid quantitative ¹H–¹³C two-dimensional NMR with high precision. *RSC Advances*, 12(9), pp.5349–5356. doi:<https://doi.org/10.1039/d1ra08423b>.

[12] Otte, Douglas A.L., Borchmann, D.E., Lin, C., Weck, M. and Woerpel, K.A. (2014). ¹³C NMR Spectroscopy for the Quantitative Determination of Compound Ratios and Polymer End Groups. *Organic Letters*, [online] 16(6), pp.1566–1569. doi:<https://doi.org/10.1021/ol403776k>.

[13] Li, D.-W., Hansen, A.L., Yuan, C., Bruschweiler-Li, L. and Bruschweiler, R. (2021). DEEP picker is a deep neural network for accurate deconvolution of complex two-dimensional NMR spectra. *Nature Communications*, 12(1). doi:<https://doi.org/10.1038/s41467-021-25496-5>.

[14] Schmid, N., Bruderer, S., Paruzzo, F.M., Fischetti, G., Toscano, G., Graf, D., Fey, M., Henrici, A., Ziebart,

V., Heitmann, B., Helmut Gräbner, Jan Dirk Wegner, Sigel, R.K.O. and Wilhelm, D. (2023). Deconvolution of 1D NMR spectra: A deep learning-based approach. *Journal of Magnetic Resonance*, 347, pp.107357–107357. doi:<https://doi.org/10.1016/j.jmr.2022.107357>.

[15] Al-Aasmi, Z.H., Shchukina, A. and Butts, C.P. (2022). Accelerating quantitative ¹³C NMR spectra using an EXtended ACquisition Time (EXACT) method. *Chemical Communications*, [online] 58(56), pp.7781–7784. doi:<https://doi.org/10.1039/d2cc01768g>.

[16] Riegel, S. (2015). *DEPT: A tool for 13C peak assignments*. [online] Available at: <https://www.nanalysis.com/nmready-blog/2015/11/19/dept-a-tool-for-13c-peak-assignments>.

[17] Chemistry LibreTexts. (2023). 7.4: Two Dimensional Heteronuclear NMR Spectroscopy. [online] Available at: https://chem.libretexts.org/Bookshelves/Organic_Chemistry/Introduction_to_Organic_Spectroscopy/07%3A_Two-Dimensional_NMR_Spectroscopy/7.04%3A_Two-Dimensional_Heteronuclear_NMR_Spectroscopy [Accessed 14 Dec. 2023].

[18] Fregolente, L.V., Batistella, Cés.B., Filho, R.Ma. and Wolf Maciel, M.R. (2006). Optimization of Distilled Monoglycerides Production. *Applied Biochemistry and Biotechnology*, 131(1-3), pp.680–693. doi:<https://doi.org/10.1385/abab:131:1:680>.

[19] Hobuss, C.B., Silva, F.A. da, Santos, M.A.Z. dos, Pereira, C.M.P. de, Schulz, G.A.S. and Bianchini, D. (2020). Synthesis and characterization of monoacylglycerols through glycerolysis of ethyl esters derived from linseed oil by green processes. *RSC Advances*, [online] 10(4), pp.2327–2336. doi:<https://doi.org/10.1039/C9RA07834G>.

[20] Perkin Elmer (2022). *Chem Draw 22.0 User Guide*.

[21] Onaneye-Babajide, Omotola & Petrik, Leslie & Musyoka, Nicholas & Bamikole, Amigun & Farouk, Ameer. (2010). Use of coal fly ash as a catalyst in the production of biodiesel. *Petroleum and Coal*, 52.

[22] Shi, W., Li, J., Chen, Y., Chen, Y., Guo, X. and Xiao, D. (2021). Enhancement of C6–C10 fatty acid ethyl esters production in *Saccharomyces cerevisiae* CA by metabolic engineering. *LWT*, 145, p.111496. doi:<https://doi.org/10.1016/j.lwt.2021.111496>.

[23] Jain, A.V. (2006). *CHAPTER 47 - Analysis of Organophosphate and Carbamate Pesticides and Anticholinesterase Therapeutic Agents*. [online] ScienceDirect. Available at: <https://www.sciencedirect.com/science/article/abs/pii/B978012088523750048X>.

[24] Guialab, (2018.). *Comprendiendo y desafiando los límites de la técnica DLS – Guialab*. [online] Available at: <https://www.guialab.com.ar/notas-tecnicas/comprendiendo-y-desafiando-los-limites-de-la-tecnica-dls/> [Accessed 14 Dec. 2023].

[25] Dworkin, J.P. (2011). Chromatographic Co-elution. *Encyclopedia of Astrobiology*, [online] pp.302–302. doi:https://doi.org/10.1007/978-3-642-11274-4_289.

[26] Schuster S.A., Johnson W.L., DeStefano J.J., Kirkland J.J. (2013). Methods for Changing Peak Resolution in HPLC: Advantages and Limitations. *LCGC Supplements*, Vol. 31, Issue 4, pp. 10-18 [online] Available at: <https://www.chromatographyonline.com/view/methods-changing-peak-resolution-hplc-advantages-and-limitations>.

[27] Otte, Douglas A.L., Borchmann, D.E., Lin, C., Weck, M. and Woerpel, K.A. (2014). ¹³C NMR Spectroscopy for the Quantitative Determination of Compound Ratios and Polymer End Groups. *Organic Letters*, [online] 16(6), pp.1566–1569. doi:<https://doi.org/10.1021/ol403776k>.

[28] Sigma Aldrich (n.d.). *IR Spectrum Table*. [online] www.sigmaaldrich.com. Available at: <https://www.sigmaaldrich.com/GB/en/technical-documents/technical-article/analytical-chemistry/photometry-and-reflectometry/ir-spectrum-table>.

[29] Mao, Y., Lee, Y.-Y., Xie, X., Wang, Y. and Zhang, Z. (2023). Preparation, acyl migration and applications of the acylglycerols and their isomers: A review. *Journal of Functional Foods*, [online] 106, p.105616. doi:<https://doi.org/10.1016/j.jff.2023.105616>.

[30] Ni, Y. and Skinner, J.A. (2015). IR and SFG vibrational spectroscopy of the water bend in the bulk liquid and at the liquid-vapor interface, respectively. *Journal of Chemical Physics*, 143(1), pp.014502–014502. doi:<https://doi.org/10.1063/1.4923462>.

[31] Younis, U., Rahi, A.A., Danish, S., Ali, M.A., Ahmed, N., Datta, R., Fahad, S., Holatko, J., Hammerschmidt, T., Brtnicky, M., Zarei, T., Baazeem, A., Sabagh, A.E. and Glick, B.R. (2021). Fourier Transform Infrared Spectroscopy vibrational bands study of *Spinacia oleracea* and *Trigonella corniculata* under biochar amendment in naturally contaminated soil. *PLOS ONE*, 16(6), p.e0253390. doi:<https://doi.org/10.1371/journal.pone.0253390>.

[32] Uranga, C.C., Beld, J., Mrse, A., Córdova-Guerrero, I., Burkart, M.D. and Hernández-Martínez, R. (2016). Data from mass spectrometry, NMR spectra, GC–MS of fatty acid esters produced by *Lasiodiplodia*

theobromae. *Data in Brief*, [online] 8, pp.31–39. doi:<https://doi.org/10.1016/j.dib.2016.05.003>.

[33] Reich I.L., (2012). *NMR Spectroscopy*. University of Wisconsin, Madison.

[34] Ashenhurst, J. (2022). Diastereotopic Protons in ¹H NMR Spectroscopy: Examples. [online] *Master Organic Chemistry*. Available at: <https://www.masterorganicchemistry.com/2022/02/08/diastereotopic-protons-1h-nmr-examples/> [Accessed 14 Dec. 2023].

[35] Dumez, J.-N. (2022). NMR methods for the analysis of mixtures. *Chemical Communications*, [online] 58(100), pp.13855–13872. doi:<https://doi.org/10.1039/D2CC05053F>.

[36] Heikkinen, S., Toikka, M.M., Karhunen, P.T. and Kilpeläinen, I.A. (2003). Quantitative 2D HSQC (Q-HSQC) via Suppression of *J*-Dependence of Polarization Transfer in NMR Spectroscopy: Application to Wood Lignin. *Journal of the American Chemical Society*, 125(14), pp.4362–4367. doi:<https://doi.org/10.1021/ja029035k>.

[37] Mestrelab (n.d.). *SMA*. [online] Mestrelab. Available at: <https://mestrelab.com/software/mnova/sma/> [Accessed 14 Dec. 2023].

[38] Bashyal, J. (2023). *Sodium Hydroxide (NaOH): Formula, Properties, Preparation, Uses*. [online] The Chemistry Notes. Available at: <https://thechemistrynotes.com/sodium-hydroxide-naoh/>.

[39] Fisher Scientific (2023). *Amberlite IRC-120(H), ion exchange resin, Thermo Scientific Chemicals*. [online] Available at: <https://www.fishersci.co.uk/shop/products/amberlite-irc-120-h-ion-exchange-resin-thermo-scientific/15439009> [Accessed 14 Dec. 2023].

Prediction of Aqueous Solubility of Polycyclic Aromatic Hydrocarbons and their Derivatives by ML-QSPR Modelling

Joseph John Stewart-Tull and Adam John Watts

Department of Chemical Engineering, Imperial College London

Abstract

Polycyclic aromatic hydrocarbons and their derivatives (PAHDs) are a class of organic compounds of which many are toxic, mutagenic and/or carcinogenic. . Given their toxicological pertinence it is of interest to be able to assess their behaviour in water systems; this can be done through an understanding of their aqueous solubilities. However, experimental methods can be expensive in both time and cost, and hence, computational methods are being increasingly used.

An ML-QSPR model to predict the aqueous solubilities of PAHDs from Mordred descriptors was produced. The model utilised an XGBoost regressor, of which the hyperparameters were tuned by a cross-validated grid search. Euclidean distances were also used as a metric to remove outliers. The results of the final model (mean absolute error of 0.633, a root mean squared error of 0.906, and an R^2 of 0.870) are comparable with similar models that predict the aqueous solubilities of other organic compounds. Further work should be done to explore the mitigation of propagated experimental error, as well as the implementation of different models and descriptor packages.

1. Introduction

Polycyclic aromatic hydrocarbons (PAHs) are a class of organic compounds found both naturally occurring in crude oil, coal, and gasoline, as well as being artificially produced for use as chemical precursors [1]. PAHs are composed of multiple fused aromatic rings, and hence, are relatively stable, however, given the correct conditions they can be reactive [2]. Such reactivity can lead to the formation of reactive metabolites which have the ability to covalently, and hence irreversibly, bind to DNA, proteins and other macromolecules [3]. Their planar structure also allows for intercalation with DNA [4]. Both of these processes are known to have adverse effects on humans and thus many PAHs have received the classification of being toxic, mutagenic and/or carcinogenic [5].

PAHs are primarily formed as a result of incomplete combustion during both natural and anthropogenic processes; this includes the burning of fossil fuels, vehicle emissions, industrial processes, and volcanic eruptions [5]. Such activities contribute significantly to the pollution of the atmosphere, pollution which includes PAHs and PAH derivatives (PAHDs). PAH derivatives, also known as substituted PAHs, arise as a result of the presence of impurities during the combustion process [6]. PAH derivatives include

base PAH structures with additional groups such as acids and amines.

These derivatives have been found to be equally, if not more, toxic than unsubstituted PAHs [7] and yet little research has been conducted to better understand them. Pollutant particulates are subject to deposition, whereby they are returned to the ground either by gravity or precipitation, thus leading to the direct contamination of bodies of water, or indirect via surface runoff. A quantitative measure of the deposition of PAHs and their derivatives, relative to their airborne concentration, is hard to ascertain as it varies widely depending on environmental conditions and other local factors [8]. Regardless, given their toxicological pertinence it is of interest to be able to assess their behaviour in water systems.

In order to understand PAHs' and their derivatives' relationships with water, a fundamental property to consider is their aqueous solubility. The ability to estimate their dissolution in water allows for their transportation as well as their bioavailability in waterways to be better understood. Generally, PAHs and their derivatives have low aqueous solubilities [5], resulting in relatively low levels of direct water contamination. However, due to their lipophilic nature, small quantities can bioaccumulate in the fatty tissue of

organisms inhabiting aquatic ecosystems [5]. This poses increasing risk to each additional stage of the trophic system, with humans susceptible to the highest accumulated concentrations. Therefore, due to their aforementioned toxic nature, even small quantities have the capacity for damage and hence should be accounted for to adequately assess the risk they impose and facilitate the exploration of remediation techniques.

The determination of aqueous solubility by experimental means has the potential to be expensive, both in time and cost. It requires pure compounds that may be difficult to synthesize and given the tendency for PAHs and their derivatives to form naturally, in mixed compositions, this is particularly prevalent [5]. Thus, the implementation and utilisation of predictive models has become increasingly popular in the estimation of thermophysical properties for the majority of molecules and compounds, with increasing degrees of accuracy as technologies and datasets are continually developed. Predicting the aqueous solubilities of many classes of organic compounds has been explored using both theoretical and data-driven approaches, however, there is limited research with regards to PAHs, especially substituted ones. In terms of theoretical approaches, group contribution models such as UNIFAC and SAFT have proven highly successful in their abilities to predict thermophysical properties, however, their current ability to predict the aqueous solubilities of substituted PAHs is limited [9].

An established data-driven method being increasingly utilised as advancements in the capabilities and accessibility of machine learning (ML) are being made are ML-QSPR (machine learning–quantitative structure-property relationship) models. QSPRs mathematically link thermophysical properties of molecules and compounds with descriptors based on intrinsic structural information. Descriptors range from basic properties such as molecular weight and number of functional groups to more complex variables concerning the fragmentation and fingerprints of given compounds; the calculation of each, insofar with regards to computation, relatively simple. The implementation of QSPR within ML models enables complex relationships between descriptors and target variables, as well as the significance of descriptors, to be found. As a result of this, aqueous solubility predictions of

organic compounds using these models have been extensively explored. However, relatively few have been developed with the objective of predicting the solubilities of PAHs, and even fewer for PAH derivatives. One pioneering model utilized partial least squares (PLS) to predict the solubilities of PAHs with reasonable accuracy, however, the sample size was relatively small with little variance, and contained no substituted PAHs [10]. Nonetheless, this provided valuable insight into the capabilities of such models and provided a foundation on which to develop. More recent solubility prediction models [11] [12] [13], not developed exclusively for PAHs or their derivatives, provide benchmark values of expected accuracies of predictions given the current available technologies. However, considerations must be made when comparing the accuracies of models given inconsistencies between the number of compounds and their variance within datasets, along with the different ML models used.

Taking all of this into account, this study aims to implement an ML-QSPR model that is capable of correlating the aqueous solubilities of both substituted and unsubstituted polycyclic aromatic hydrocarbons to their descriptors. This, in an effort to understand the relationships between such, and hence, be able to accurately predict the aqueous solubilities of unseen PAHDs in an effort to better understand their potential impacts as pollutants.

2. Methodology

2.1. Data Collection

In this work, the AqSolDB dataset was employed – a curation of experimental aqueous solubilities of 9982 organic compounds ranging from alkanes to aromatics. It was formed with the intention for use in data-driven model development. The aqueous solubility values in the dataset are standardised to $\log[S/(\text{molL}^{-1})]$ units for consistency, as well as allowing for better visualisation and comparison. AqSolDB also includes the Simplified Molecular Input Line Entry System (SMILES) string for each of the compounds, easing the generation of molecular descriptors.

During the curation of AqSolDB efforts were made to reduce the errors in the database, utilising an algorithm to select the most accurate experimental solubility value when duplicates were found. However, regardless of its effectiveness, there is still an error associated with experimental results which must be considered in any subsequent

predicted values using the dataset. Furthermore, in an effort to maximise the sample space, solubility values measured at $25 \pm 5^\circ\text{C}$ were included, despite solubility being a function of temperature, as the range was considered acceptably small.

2.2. Molecular descriptor calculation

The molecular descriptor values of each compound were to act as their respective feature values during the machine learning process. It was therefore important to generate relevant descriptors, that adequately represented the individuality of each compound. Multiple descriptor packages are available, the two most widely used being Mordred and PaDEL. They are both open source and can generate large numbers of both 2D and 3D descriptors. Of the other available packages many often utilize proprietary software, are unavailable or deprecated in Python, or calculate significantly fewer, less varied descriptors.

In the interest of computational simplicity, given the size of the dataset, only one of the aforementioned packages was to be used. The selected package was Mordred due to its ability to calculate descriptors for larger molecules than PaDEL is capable of, furthermore, each descriptor calculation requires under half the computational time to that of PaDEL. Both factors were significant given the size and variance of the dataset.

The descriptors were generated from the SMILES of each compound, using the Mordred package in Python generating 1,613 two-dimensional (2D) descriptors with 3D descriptors being disregarded to increase prediction speed and avoid repeatability problems regarding 3D descriptor values. Compounds for which descriptors could not be calculated were removed leaving 8149 compounds, of which 436 were identified as PAHDs. These PAHDs were then used as the basis for processing the data. The removal of the PAHDs left a remaining set of 7,713 non-PAHD compounds.

2.3. Data Processing

Following this, categorical variables and low variance features were removed, utilising a threshold of 0.1, as these features do not vary enough to effect the solubility and would increase computational intensity. This left 603 descriptors. The descriptor pair correlation matrix was then calculated and descriptors with an absolute

Pearson correlation coefficient value greater than 0.8 were removed. This prevented particular descriptors from dominating the model and causing the model to overfit. Following this, 130 descriptors remained for use in the model. An additional step was taken to remove descriptors that were uncorrelated to solubility, with a threshold of less than 0.2 absolute Pearson correlation coefficient being set. This created a dataset of 43 descriptors. These two datasets were tested against each other to find the best performing set to continue with, as some important descriptors could have been removed.

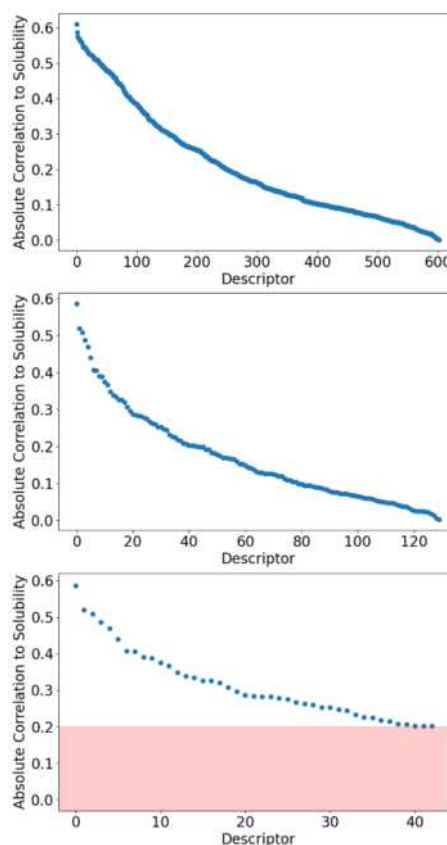


Figure 1: Absolute correlations to solubility for each descriptor with different numbers of descriptors.

2.4. Machine Learning Overview

The machine learning algorithm used in this study was a regressor of gradient-boosted decision trees (XGBRegressor) from the XGBoost library. XGBoost is an implementation of gradient boosting, which is an ensemble learning technique where models are added to correct the errors of previous models. It uses decision trees as base learners and combines these to create a strong predictive model, choosing how to build a more powerful model by using the gradient of a loss function, which captures the performance of the model. XGBoost was chosen in this study as it

gives state-of-the art results on a wide range of problems . The XGBoost regressor has several hyperparameters which can be tuned to optimise performance on a given dataset.

During this study 4 different hyperparameters were studied to optimise the model, those being, the number of estimators, the learning rate, the maximum depth and the minimum child weight.

The number of estimators is the number of trees which are fit during training. The more estimators, the more complex patterns can be found in the data, however if the number of estimators is set too high it can lead to overfitting, reducing the predictive capability of the model.

The learning rate determines the step size at each iteration while moving towards the minimum of the loss function. It is used to prevent overfitting in the model and determines how quickly the model will train.

The maximum depth hyperparameter determines the maximum depth of each tree. Deeper trees can find more intricate relationships in the data but can lead to the model overfitting to the training data and fail to find the underlying patterns.

The minimum child weight hyperparameter controls the minimum sum of instance weights needed in a child (nodes formed when a tree is split). This stops the creation of nodes with very few instances, and therefore acts as a form of regularisation to prevent overfitting .

The model also included early stopping, which stops the model training once the performance of the test dataset has stopped improving for a certain number of boosting rounds. In this study a value of 5 was set in order to prevent the model from overfitting and to reduce the computational complexity by stopping the training once the training loss begins to increase .

In order to optimise the hyperparameters for each dataset used in this study, two grid searches were conducted which tested different combinations of hyperparameters using 5 cross validation folds, to provide a more robust measure of the model's performance. The performance of the model during the grid search was scored by the negative mean absolute error. The first grid search ranged over a wide range for the hyperparameters with the second one using smaller increments to attempt to get closer to an optimum.

2.5. Machine Learning Application

From the 436 PAHDs found in AqSolDB 25% (109) were completely removed from the dataset to create a test set, to determine the predictive capability of the model on completely unseen data. It was ensured that the test set would have a representative range of solubilities, to demonstrate predictive capabilities across a wide range of PAHDs. This left a training set of 327 PAHDs.

Initially, only the set of 327 PAHDs was used to train the model, testing with both 43 descriptors and 130 descriptors. The number of descriptors which gave the best performance at this stage would be used moving forwards.

Following this, the 327 PAHDs were combined back into the non-PAHD dataset and this dataset would be run with the previously best performing features. This was done as some features of the PAHDs, especially the additional functional groups, may have been poorly represented within the PAHD dataset, so the non-PAHDs were included to provide a more complete set of features for the model to train on. After training this model, it was tested against the test set of PAHDs again.

After each training of the model the model was evaluated with the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and R².

Table 1: Equations for the different metrics used in this work. Prediction_i is the ith model prediction and true_i is the corresponding experimental value

Metric	Equation
Mean Absolute Error	$\frac{\sum_i^n prediction_i - true_i }{n}$
Root Mean Squared Error	$\sqrt{\frac{\sum_i^n (prediction_i - true_i)^2}{n}}$
R ²	$1 - \frac{\sum_i^n (true_i - prediction_i)^2}{\sum_i^n (true_i - mean\ of\ true)^2}$

The MAE was the primary metric used in this study, as it allowed the best comparisons between this model and others from literature and provided the best real-world interpretation of the predictions.

The RMSE was also utilised as larger errors have a larger impact on the RMSE, making it more sensitive to outliers. This allows for identification of models that have several large errors, which may not be visible from MAE alone.

R^2 was employed to measure the goodness of fit of the XGBoost regressor. A higher value meaning a higher proportion of the variance in the dependent variable, in this case solubility, is explained by the model.

The feature importance was visualised using SHapley Additive exPlanation (SHAP) values. SHAP values are based on game theory to assign the importance of each feature in the model, with the value showing the contribution of that feature to the specific prediction. The average SHAP value is therefore a measure of each features average contribution to the model’s predictions, with larger average SHAP values meaning that the feature is more important .

2.6. Euclidean Distances

Given the variety within the dataset, it was hypothesized that some metric to categorize the data could be used to better represent the PAHDs in the training set by adding similar non-PAHD compounds and excluding dissimilar ones. To do this, the Euclidean distances between each of the PAHDs and each of the non-PAHDs from the dataset were calculated. The Euclidean distances considered the difference between each value of the 130 descriptors for every combination of PAHD and non-PAHD compound, outputting a single value for each pair.

These values could then be used to determine a non-PAHD’s similarity to each of the PAHDs, with a lower Euclidean distance indicative of a more similar compound. Increasing threshold values of Euclidean distance were iterated through, increasing the ‘radius’ around each PAHD, encompassing increasing amounts of decreasingly similar non-PAHDS. This was done over 10 percent increments, with the first iteration adding the most similar 10 percent of non-PAHDs, and the last adding most similar 90 percent of non-PAHDs, to the training set. May it be noted that this was carried out in the absence of the test set, to ensure that the non-PAHDs selected had no predetermined relationship with them.

Taking advantage of this enabled exploration into how formulating the dataset, in an effort to reinforce the characteristics of PAHDs, affected predictions.

2.7. y-Randomisation

After the best model was found, a y-Randomisation was carried out to validate the

robustness of the model. y-Randomisation works by randomly shuffling the target variable, while keeping the feature variables unchanged, and is used to determine whether the model is capturing actual relationships between the features and the target variable, or the predictions are due to chance. To show that the model is capturing actual relationships, the performance of the model should be poor with the randomised sets. The target variable was randomised 50 times with the average MAE, RMSE and R^2 of the models being reported.

3. Results and Discussion

3.1. Feature selection

The results of using 130 features and 43 features, showed that the set of 43 features performed worse, as shown in Table 2: Comparison of the datasets using different numbers of features. MAE and RMSE values in units of $\log[S]$. This is due to some of the most important features being removed which are not linearly correlated with solubility, such as the ECIndex and ABC. This demonstrates a limitation in the calculation of the correlation coefficient as it only accounts for linear relationships, and it is likely that many of the features are not linearly related to solubility. Due to this, the 130 features were chosen to move forwards with as it allowed for better PAHD prediction.

Table 2: Comparison of the datasets using different numbers of features. MAE and RMSE values in units of $\log[S/(\text{mol}\cdot\text{L}^{-1})]$

Number of features	MAE (Test)	RMSE (Test)	R^2 (Test)
43	0.743	1.01	0.838
130	0.663	0.892	0.874

3.2. Comparison with whole dataset

The results shown in Table 2 demonstrate that training the model on the whole dataset provides better predictions for the PAHDs. This is due to the non-PAHD dataset having a larger range of different structures and functional groups, allowing the model to learn more complex patterns within the dataset. This allowed for better predictions for the test set as some of the compounds in it show features that are not found in the rest of the PAHD dataset.

Table 2: Comparison of utilising the whole dataset in training versus using just the PAHD set. MAE and RMSE in units of $\log[S/(\text{mol}^{-1})]$

Dataset	MAE (Test)	RMSE (Test)	R ² (Test)
PAHD	0.663	0.899	0.868
Whole set	0.655	0.890	0.871

3.3. Euclidean distances

Analysing the MAEs at increasing threshold increments showed little consistency in either reinforcing the characteristics of the PAHDs or diluting the training set, increasing, or decreasing the MAE respectively. The values fluctuated displaying no discernible trend, which upon inquiry could be a result of the following.

The calculation of Euclidean distances was done irrespective of feature importance, and hence, a compound could be classified as ‘similar’ based on descriptors that had little influence on the model. Therefore, the addition of ‘similar’ compounds to the training set, with the intention of reinforcing the characteristics of PAHDs, could have added compounds with little similarity with regards to the features pertinent to the model, reducing the effectiveness of the training set.

This possibly explains the lack of any trend found using this approach as, given the lack of explicit relationships between descriptors, a compound could be deemed ‘similar’ to a PAHD whilst having contradicting values for the most influential descriptors to the model. Therefore, given the current method, there is no reliable correlation between ‘similarity’ and individual descriptors, resulting in the possibility for influential descriptor values to arise at any threshold of ‘similar’ compounds.

Upon plotting the Euclidean distances by Principal Component Analysis (PCA) it becomes increasingly apparent how little distinction there is between most compounds, further helping to explain the absence of any distinct trends using this method in its current configuration. However, upon visualization of the data, clear outlying non-PAHDs can be identified.

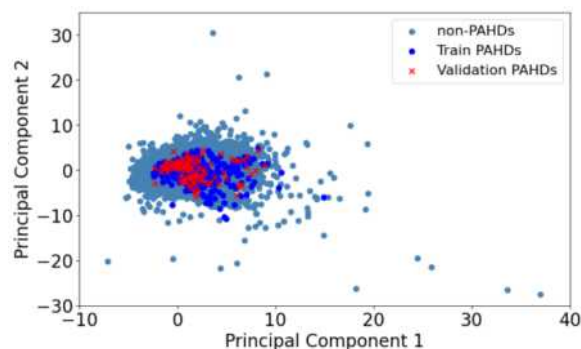


Figure 2: PCA plot displaying 130 descriptors in 2 dimensions

Euclidean distances were therefore employed once again but this time as a means to remove such. Non-PAHDs in the training set with the greatest Euclidean distances were found and new datasets were formed by iteratively removing an additional percentage of the remaining farthest values. The model was retuned and retrained for each iteration.

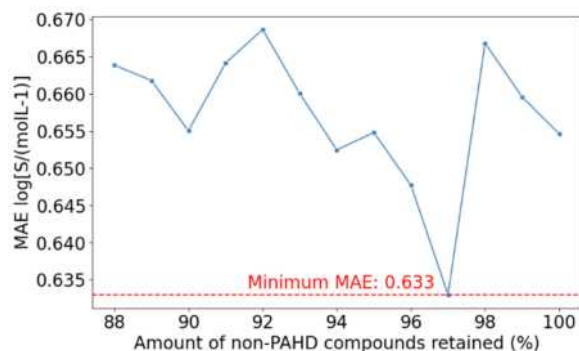


Figure 3: MAE from each dataset produced during outlier removal. Minimum MAE in units of $\log[S/(\text{mol}^{-1})]$

Figure 3 demonstrates the that best model found was when 3% of the outlying non-PAHDs were removed. The performance of this model is detailed in Table 3.

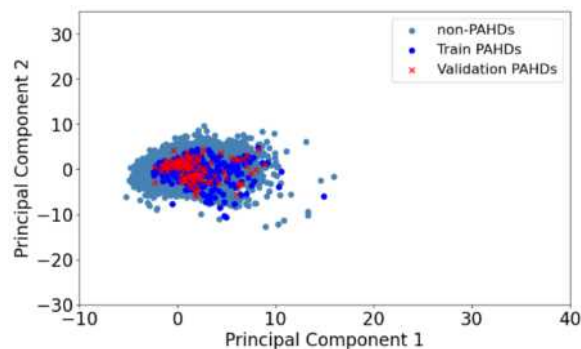


Figure 4: PCA plot with 3% greatest outliers removed

Figure 4 shows the distribution of compounds when the 3% of compounds with the greatest Euclidean distances were removed. It likely gives

the best result as all PAHDs are encompassed in the applicability domain and unrelated non-PAHDs that add unnecessary noise are removed.

3.4. Best Model Analysis

Table 3: Summary of performance of the best model produced in this study. MAE and RMSE in units of $\log[S/(\text{molL}^{-1})]$

Model	MAE (Test)	RMSE (Test)	R ² (Test)
97% retained	0.633	0.906	0.870

The hyperparameters of the best model were, 5000 estimators, a learning rate of 0.01, a maximum depth of 11, and a minimum child weight of 15.

The performance of the model is comparable with other machine learning models in literature which predict aqueous solubility of organic compounds, with the MAEs of some models shown in Table 4.

Table 4: Comparison of this work with other literature machine learning models. MAE in units of $\log[S/(\text{molL}^{-1})]$, MLR-Multilinear regression, Ensemble-Combination of artificial neural network, random forest and XGBoost

Developer	ML Method	MAE (Test)
Yan	MLR	0.68
Sorkun	Ensemble	0.40
Ali	MLR	0.72
This work	XGBoost	0.63

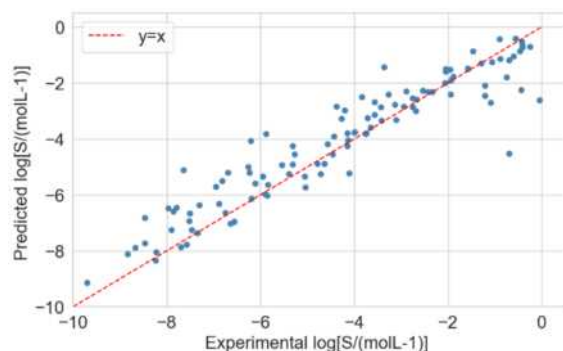


Figure 5: Parity plot showing the correlation of the predicted and experimental solubility values

Figure 5 and the R² value in Table 3 demonstrate that the predictions have a good fit to the actual aqueous solubility values in the test set. The plot shows that at lower solubility values the model is more likely to underpredict the solubility. This is likely due to the lack of compounds in the training set that have positive solubility values. This may

mean that the model fails to learn relationships between the descriptors and aqueous solubility that led to higher solubilities, causing an underprediction at low experimental solubilities.

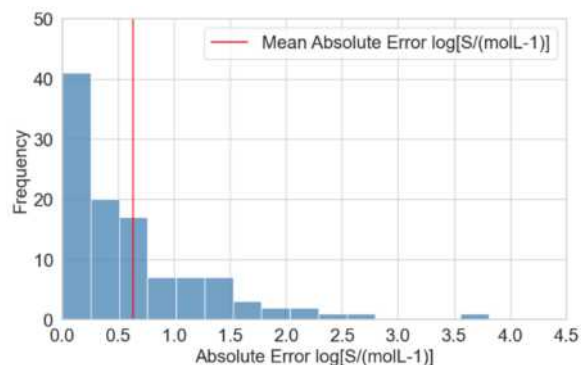


Figure 6: Distribution of the errors between the experimental values and predicted values for the test set using the best model

Figure 6 shows the distribution of the errors in the test set when using the best model. The errors are concentrated towards the lower errors with 56% of errors being less than 0.5 $\log[S/(\text{molL}^{-1})]$. There is one error of 3.81 $\log[S/(\text{molL}^{-1})]$ which shows a very poor prediction. This point represents 1-bis[4-(diethylamino)phenyl]methyl]naphthalene-2,7-disulfonic acid. A reason for the poor prediction of this molecule is that it contains groups that are poorly represented in the training dataset. For example, it contains 2 sulfonic acid groups which are not well represented in the training dataset, appearing only 5 times. Presence of sulfonic acid groups increases aqueous solubility of compounds, due to their large polarity which allows for stronger interactions between the compound and water [22]. The model may not have been able to capture this relationship, and therefore it underpredicted the solubility, predicting a value of -4.51 $\log[S/(\text{molL}^{-1})]$ when the actual value was -0.698 $\log[S/(\text{molL}^{-1})]$.

3.5. y-Randomisation

Table 5: Results of y-Randomisation. MAE and RMSE in units of $\log[S/(\text{molL}^{-1})]$

Model	MAE (Test)	RMSE (Test)	R ² (Test)
Original Model	0.633	0.906	0.870
Average from Randomisation	2.28	3.00	0.224

The results of the y-Randomisation showed that by randomising the target variable the model performed much worse than the original model. This demonstrates that the model is finding actual relationships between the features and aqueous solubility, rather than just being by chance. This is evidence that the model is robust.

3.6. Feature analysis

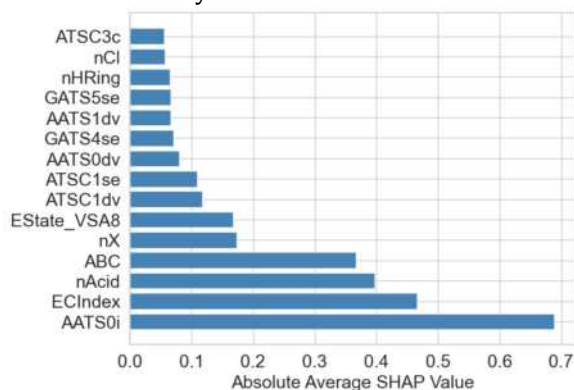


Figure 7: Absolute Average SHAP value for the 10 most important descriptors of the XGBoost model

Figure 7 shows the descriptors that have the largest influence on the predictions from the best model found. It is important to understand the most important descriptors, not just to gain a better understanding of the model, but to also verify that each of these features makes physical sense. The meanings of the most important descriptors are detailed below.

AATS0i corresponds to an Autocorrelation descriptor which represents the average Bronto-Moreau autocorrelation with lag 0, weighted by the first ionisation potential [23]. It captures the distribution of the first ionisation energy across the molecular structure. This influences how the compound will interact with water, as water is a polar molecule. Furthermore, AATS0i provides information about the compounds shape, size, and symmetry, which influences how the compound interacts with water, and therefore its solubility.

The ECIndex is the Eccentric Connectivity Index and is a topological descriptor [23]. It considers the connectivity of the atoms and their eccentricity, which is a measure of how far an atom in the molecule is from other atoms in the molecule [24]. It captures information about the size and shape of the molecule, which influences how it will interact with water, depending on if it is large and complex, or small and simple, with simpler molecules tending to be more soluble [25]. The

connectivity also impacts the polarity of the compound which will affect its interaction with water.

The nAcid descriptor is a count of the number of acidic groups in a molecule, meaning those that can donate a proton in a reaction [23]. Acidic groups can form hydrogen bonds with the water molecules, which allows the compound and water to interact more strongly and increases solubility [26].

3.7. Comparison with SAFT- γ -mie

The SAFT- γ -Mie equation of state is a group contribution method that allows for the calculation of thermodynamic properties, including aqueous solubility. It is based on the Statistical Associating Fluid Theory (SAFT) and utilises the Mie potential to describe the interactions between the Mie segments of the molecules [27]. Mie segments are the individual units that make up a molecule, with the molecule being represented as chains of these segments. The Mie potential is a generalised form of the Lennard-Jones potential, with variable exponents for the repulsive and attractive terms, allowing for a more accurate representation of intermolecular forces for a wider range of substances [27].

SAFT- γ -Mie is considered the “Opus Magna” of the SAFT equation of state and can be used in the prediction of the solubility of PAHs and was therefore compared with the best model found in this work [28]. An MAE for the aqueous solubility of PAHs of 0.42 was found across a set of 22 PAHs, clearly demonstrating superior accuracy to the data-driven model [9]. However, the SAFT- γ -Mie model is more complex than the machine learning model and it is challenging to introduce new groups due to the need to estimate the interaction parameters, which is a very intensive and requires large amounts of data. This makes it difficult to predict the aqueous solubility of the more complex PAHDs as they have much more complex interactions than the simpler PAHDs. Due to this, the machine learning model still has applicability when it comes to more complex molecules.

4. Conclusion

To conclude, an ML-QSPR model to predict the aqueous solubilities of PAHs and PAH derivatives (PAHDs) has been produced, with comparable accuracy to other aqueous solubility models . Thus, aiding in the understanding of their

behaviour in water. Furthermore, the use of SMILES as a means of descriptor generation allows for the descriptor values of novel PAHDs to be found, and therefore, an estimation of their aqueous solubility to be made. This is useful as it bypasses the need to synthesise such compounds to find their aqueous solubility experimentally, a costly and time intensive process.

The final model utilised an XGBoost regressor, a tree-based algorithm, to predict aqueous solubilities of PAHDs, using 130 Mordred descriptors [16]. An additional dataset of non-PAHD organic compounds was also available to supplement the training set [14], of which the closest 97% by Euclidean distance to the PAHDs, were employed. This resulted in a mean absolute error (MAE) of 0.633, a root mean squared error (RMSE) of 0.906, and an R^2 of 0.870. These metrics indicated a better performance than using solely PAHDs in the training set (MAE:0.663, RMSE:0.899, R^2 :0.868), and hence, supports the idea of using similar compounds to aid the model's predictive capabilities. However, the use of Euclidean distances to execute this did not behave exactly as expected and would potentially be improved in future works by accounting for the varying feature importance by assigning representative weights in the calculations.

The model in its current form is capable of producing results on a par with similar models, however, the error of predictions combined with the experimental error propagated from the training data leads to limited application where precise values are required. The model therefore serves better as a tool to gain an initial understanding of the aqueous solubility of a given PAHD, and if within a range of interest, further research should be conducted to verify the value.

That being said, with the continuous development of the capabilities of machine learning, the use of data-driven models will most likely remain at the forefront of thermophysical property prediction. As technology evolves, refinements to model architectures and training methodologies may contribute to mitigating errors, thereby expanding the scope of applications where precise values are achievable. The ongoing integration of advanced techniques and increased data availability holds promise for further enhancing the accuracy and versatility of such models in the realm of thermophysical property prediction.

With regards to expanding the applicability of the PAHD prediction, further models should be investigated. For instance, an ensemble of an artificial neural network, a Random Forest and an XGBoost regressor could be used, as this has shown the best performance in prediction of the aqueous solubility of general organic compounds. Furthermore, different descriptor packages could be investigated, such as PaDEL, as they may generate descriptors that better represent the structure and properties of PAHDs with respect to their aqueous solubility. This would allow the model to better relate the solubility of PAHDs to these descriptors and therefore improve predictions.

References

- [1] C. f. D. C. a. Prevention, "Polycyclic Aromatic Hydrocarbons (PAHs) Factsheet | National Biomonitoring Program | CDC," 24 5 2019. [Online]. Available: https://www.cdc.gov/biomonitoring/PAHs_FactSheet.html. [Accessed 10 11 23].
- [2] B. Moorthy, "Polycyclic Aromatic Hydrocarbons: from Metabolism to Lung Cancer," *Toxicological Sciences*, no. 1, pp. 5-15, 2015.
- [3] U. A. Boelsterli, "Covalent binding of reactive metabolites to cellular macromolecules," in *Mechanistic Toxicology : the Molecular Basis of How Chemicals Disrupt Biological Targets*, Boca Raton, Crc Press, 2007.
- [4] W. D. S. Arnab Mukherjee, "Drug-DNA Intercalation," *Dynamics of Proteins and Nucleic Acids*, pp. 1-62, 2013.
- [5] M. S. M. Hussein I. Abdel-Shafy, "A Review on Polycyclic Aromatic hydrocarbons: Source, Environmental impact, Effect on Human Health and Remediation," *Egyptian Journal of Petroleum*, vol. 25, no. 1, pp. 107-123, 2016.
- [6] M. N. et. al., "Polycyclic Aromatic Hydrocarbons (PAHs) and Their Derivatives (O-PAHs, N-PAHs, OH-PAHs): Determination in Suspended Particulate Matter (SPM) – a Review," *Environmental Processes*, vol. 9, no. 1, 2021.
- [7] B. C. Agnieszka Krzyszczak, "Occurrence and Toxicity of Polycyclic Aromatic Hydrocarbons Derivatives in Environmental Matrices," *Science of The Total Environment*, vol. 788, 2021.
- [8] P. Siudek, "Atmospheric Deposition of Polycyclic Aromatic Hydrocarbons (PAHs) in the Coastal Urban Environment of Poland: Sources and Transport Patterns," *International Journal of Environmental Research and Public Health*, vol. 19, no. 21, pp. 14183-14183, 2022.

- [9] T. Bernet et al, *Solubility prediction of polycyclic aromatic hydrocarbons in water using SAFT-gamma Mie, in preparation*, 2023.
- [10] G.-N. L. e. al., "Estimation of Water Solubility of Polycyclic Aromatic Hydrocarbons Using Quantum Chemical Descriptors and Partial Least Squares," *QSAR & Combinatorial Science*, vol. 27, no. 5, pp. 531-670, 2007.
- [11] M. C. Sorkun, J. V. A. Koelman and S. Er, "Pushing the limits of solubility prediction via quality-oriented data selection," *iScience*, vol. 24, no. 1, 2021.
- [12] A. Yan and J. Gasteiger, "Prediction of Aqueous Solubility of Organic Compounds Based on a 3D Structure Representation," *Journal of Chemical Information and Computer Sciences*, vol. 43, no. 2, pp. 429-434, 2002.
- [13] J. Ali, P. Camilleri, M. B. Brown, A. J. Hutt and S. B. Kirton, "In Silico Prediction of Aqueous Solubility Using Simple QSPR Models: The Importance of Phenol and Phenol-like Moieties," *Journal of Chemical Information and Modeling*, vol. 52, no. 11, pp. 2950-2957, 2012.
- [14] M. C. Sorkun, A. Khetan and S. Er, "AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds," *Scientific Data*, vol. 6, no. 1, 2019.
- [15] D. Weininger, "SMILES, a Chemical Language and Information system. 1. Introduction to Methodology and Encoding Rules," *Journal of Chemical Information and Modeling*, vol. 28, no. 1, pp. 31-36, 1988.
- [16] M. Hirotomo, T. Yu-Shi, K. Norihito and T. Tatsuya, "Mordred: a Molecular Descriptor Calculator," *Journal of Cheminformatics*, vol. 10, no. 1, 2018.
- [17] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, 2016.
- [18] "XGBoost Parameters — xgboost 2.0.2 documentation," 2022. [Online]. Available: <https://xgboost.readthedocs.io/en/stable/parameter.html#parameters-for-tree-boost>. [Accessed 9 12 2023].
- [19] D. Leitão, "Gradient Boosting: To Early Stop or Not To Early Stop," *Medium*, 25 3 2023. [Online]. Available: <https://towardsdatascience.com/gradient-boosting-to-early-stop-or-not-to-early-stop-5ea67ac09d83>. [Accessed 5 11 2023].
- [20] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," 2017.
- [21] C. Rücker, G. Rücker and M. Meringer, "y-Randomization and Its Variants in QSPR/QSAR," *Journal of Chemical Information and Modeling*, vol. 47, no. 6, pp. 2345-2357, 2007.
- [22] E. Block, "Sulfonic acid | chemical compound | Britannica," *Encyclopædia Britannica*, 2019. [Online]. Available: <https://www.britannica.com/science/sulfonic-acid>. [Accessed 24 11 2023].
- [23] H. Moriwaki, "Descriptor List — mordred 1.2.1a1 documentation," 2016. [Online]. Available: <https://mordred-descriptor.github.io/documentation/master/descriptors.html>. [Accessed 2 12 2023].
- [24] V. Sharma, R. Goswami and A. K. Madan, "Eccentric Connectivity Index: A Novel Highly Discriminating Topological Descriptor for Structure-Property and Structure-Activity Studies," *Journal of Chemical Information and Computer Sciences*, vol. 37, no. 2, pp. 273-282, 1997.
- [25] "What factors affect solubility?," AAT Bioquest, 18 4 2022. [Online]. Available: <https://www.aatbio.com/resources/faq-frequently-asked-questions/what-factors-affect-solubility>. [Accessed 10 12 2023].
- [26] J. Clark, "Carboxylic Acids Background," *Chemistry LibreTexts*, 3 10 2013. [Online]. Available: [https://chem.libretexts.org/Bookshelves/Organic_Chemistry/Supplemental_Modules_\(Organic_Chemistry\)/Carboxylic_Acids/Properties_of_Carboxylic_Acids/Carboxylic_Acids_Background#:~:text=high%20boiling%20point-,Solubility%20in%20water,with%20water%20in%20any](https://chem.libretexts.org/Bookshelves/Organic_Chemistry/Supplemental_Modules_(Organic_Chemistry)/Carboxylic_Acids/Properties_of_Carboxylic_Acids/Carboxylic_Acids_Background#:~:text=high%20boiling%20point-,Solubility%20in%20water,with%20water%20in%20any). [Accessed 1 12 2023].
- [27] V. Papaioannou, T. Lafitte, C. Avendaño, C. S. Adjiman, G. Jackson, E. A. Müller and A. Galindo, "Group contribution methodology based on the statistical associating fluid theory for heteronuclear molecules formed from Mie segments," *The Journal of Chemical Physics*, vol. 140, no. 5, 2014.
- [28] "SAFT," [Online]. Available: <http://www.molecularsystemsengineering.org/saft.html>. [Accessed 3 12 2023].

Development of a Unified Kinetic Model for the Hydrothermal Carbonisation (HTC) of Microalgal Biomass

Christos Evripidou and Yunxiang Zheng

Department of Chemical Engineering, Imperial College London, U.K.

Abstract

Hydrothermal Carbonisation (HTC) has gained interest due to its mild operating conditions. It allows energy densification of wet biomass through formation of hydrochar, a solid product valuable for soil remediation, biofuel production and adsorption. HTC is a complex mix of competing parallel reactions that is influenced by factors such as feed species, residence time, temperature, and solids loading (SL). This research develops a unified lumped kinetic model for HTC of microalgae, aiming for applicability across a wide range of microalgal species, providing initial estimates for key results such as hydrochar yield, carbon (C) content, higher heating value (HHV), and energy trade-offs. It considers the dissolution of macromolecules into monomers, and the formation of gas and secondary char. Two modelling approaches were used, resulting in the Multi-Step Optimisation (MSO) model and Direct Optimisation (DO) model. The former used experimental dissolutions to obtain kinetic parameters for the hydrolysis steps of the biomolecules through global optimisation. These were then used, in subsequent optimisations, to attain the rest of the kinetic parameters for hydrochar and gas formations. On the contrary, the DO Model directly obtained all kinetic parameters at once. Due to the limited species included in the hydrolysis set, research was focused on the DO Model to allow for wider applicability. The model provided good initial estimates for the hydrochar yield but tended to overpredict across microalgal species. Further model iterations are expected to improve the current restricted ability in predicting C content and HHV accurately. To reach energy neutrality quicker, higher temperatures, higher SL, and algae with higher total content of carbohydrates and proteins are suggested to be used.

Keywords – Hydrothermal Carbonisation (HTC), Hydrochar, Microalgal biomass, Lumped kinetic model, Carbon (C) Content, Higher heating value (HHV)

1. Introduction

There currently exist several industrial processes that can be used to produce char from biomass. Biochar is a solid product derived from biomass that is rich in carbon content. It has large surface area, high porosity high cation exchange capacity, and high stability. These properties have led to its wide applications such as soil remediation, activated carbon preparation, biofuel production, and materials synthesis (Yang, et al., 2023). Two of the most-promising processes are pyrolysis and hydrothermal carbonisation (HTC). HTC is a relatively new method using sub-critical water as a reaction medium, typically operating at lower temperatures between 180-250°C with pressures between 10-40 bar to produce hydrochar (Bevan, et al., 2021). Compared to pyrolysis, HTC stands out as it allows the utilisation of wet biomass, resulting in improved process efficiency and cost-effectiveness. In addition, hydrochar from HTC exhibits benefits over biochar from pyrolysis, as it has a reduced metal content and higher calorific value when formed from similar operating conditions (Liu, et al., 2018) (Kambo & Dutta, 2015). Between algal and lignocellulosic biomass, algal biomass stands out for its negligible recalcitrant biomass fraction, higher growth rate (Zhang, et al., 2023), and higher energy content with a greater photosynthetic efficiency. (Chen, et al., 2022). Although use of biomass char would be an eco-friendly approach for sustainable energy production, scaling up these processes is largely hindered by varieties in biomass (Cao, et al., 2021). Previous research has

investigated methods such as pre-treatments to homogenise the system to process feedstocks composed of materials from different sources (Adam, et al., 2023).

This research paper focuses on the development of a unified kinetic model for the HTC process, that would be applicable across a wide range of microalgal biomass species. This is a novel model in which lumped kinetics were used to provide a simple tool to obtain sound initial estimates of key parameters such as hydrochar yield, carbon content & energy neutrality points. Two modelling approaches were used to construct the predictive model: A Direct Optimisation (DO) model that involved the direct optimisation for the hydrochar mass and gas yield and a Multi-Step Optimisation (MSO) model which employed a more nuanced approach. This considered an initial kinetic parameter estimation of the hydrolysis steps of the biomolecules followed by an estimation for the hydrochar mass and gas yield.

2. Background

The HTC process involves several complex chain reactions which include hydrolysis, dehydration, decarboxylation, polymerisation, and aromatisation (Czerwińska, et al., 2022) and form several intermediate products.

The process is influenced by many factors with the predominant being the feedstock species, solid loading (SL), temperature, and residence time. Differences in algal biomass species used for HTC affect the formation of biochar including the yield and properties. This is

Page 1 of 10

largely due to their difference in biochemical compositions. For example, for the same operational conditions of HTC, under the same temperature, residence time and SL, the hydrochar yield for *Chlorella Spp.* is 9.7% higher than that of *Dunaliella* (Heilmann, et al., 2010). When comparing the average compositions of carbohydrates, proteins, and lipids of these two algae, it is observed that the carbohydrate content in *Chlorella Spp.* is almost 4 times that of the *Dunaliella*, whilst the protein and lipid content is lower than the *Dunaliella*. The relationship between the hydrochar yield and the species is challenging to determine due to the complexity of the reactions. There are studies on the effects of the removal of lipids (Broch, et al., 2014) or ash (Liu, et al., 2019) from the biomass on the qualities of hydrochar produced. These were used as insights on how the hydrochar is affected by the compositions of biomass used. An increase in SL and residence time increases the hydrochar yield (Aragón-Briceño, et al., 2020). In contrast, an increase in temperature decreases the hydrochar yield but creates a more stable product (Chen, et al., 2021).

Different approaches have been made on modelling the HTC process with only lignocellulosic biomass. The most common methods include statistical modelling with response surface methodology, computational modelling with software such as COMSOL, and mathematical optimisation modelling with lumped reaction networks (Alberto Galifucoco, 2019). Most mathematical models are based on pseudo-first-order reaction kinetics, neglecting higher-order reactions like polymerisation. Some have found that first-order models fit better than second-order models, and give the same square error as the higher-order models to 2 significant figures (Jung, et al., 2018). Despite the complex modelling technique, most research simplifies the process into a reaction scheme as shown in Figure 1 (Ischia & Fiori, 2021).

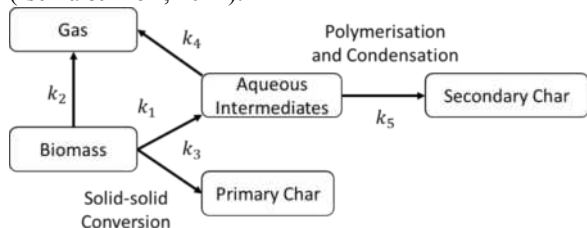


Figure 1: Simplified HTC reactions scheme.

3. Methods

3.1 Data Collection

Experimental data was collected from literature for various algae species. The data gathered was in the temperature range of 453 - 573K, residence time range of 15 -240 minutes and SL range of 1 - 25%. Table 1 shows the number of data points each species is contributing to the three models. A large range of biomass compositions were covered. Data on gas yields were also collected for 9 different macroalgae species across a range of temperatures, residence time and SL.

Table 1: Summary of the calibration data set used. (For detailed calibration & validation data sets, see supplementary materials.)

Species	Hydrochar C	HHV
<i>Dunaliella Salina</i>	13	13
<i>Aphanizomenon Flos-aquae</i>	1	1
<i>Synechocystic Spp.</i>	1	1
<i>Spirulina Spp.</i>	3	3
<i>Chlorella Spp.</i>	49	3
<i>Nannochloropsis Oculata</i>	7	7
<i>Chlamydomonas Reinhardtii</i>	1	1
<i>Gelidium Sesquipedale</i>	4	4
<i>Mixed Species (Polyculture)</i>	18	18

3.2 Model Assumptions

During HTC, the reactions are carried out in a heterogenous system. However, this was simplified into a model with constant volume due to the presence of a large volume of water (low SL). To reduce the model complexity, no mass transfer limitations were considered, and simplifications described in the following sections were made.

3.2.1 Hydrolysis Simplifications

The algal biomass mainly consists of three components: carbohydrates, lipids, and proteins. The first step of HTC reactions is the hydrolysis of these components, where carbohydrates convert into glucose, proteins convert into amino acids, and lipids convert into fatty acids and glycerol.

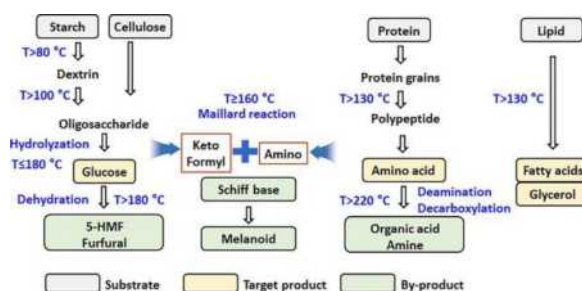


Figure 2: Reaction pathways for hydrothermal hydrolysis of algal biomass. Taken from (Chen, et al., 2022).

In practice, the hydrolysis process would also result in by-products such as 5-HMF as seen in Figure 2. When proteins hydrolyse into the liquid phase, they undergo polymerisation to form insoluble grains. As the temperature increases the peptide bonds break and further decompose into amino acids. At temperatures above 220°C, by-products form from deamination and decarboxylation reactions. When oligosaccharides hydrolyse, the main product is glucose with the possibility of bio-methane production. At temperatures above 180°C, glucose may hydrolyse further into by-products, which limit the methane potential. Similarly, amino acids can perform Maillard reactions with saccharides, which happen at temperatures over 160°C (Chen, et al., 2022). However, implementing these side reactions in the model would heavily increase its complexity. Therefore, no side reactions or by-products were considered.

3.2.2 Lipids Simplifications.

The lipids contained in microalgae were modelled as entirely being retained on the solid primary hydrochar. During the HTC of microalgae, most lipids adsorb onto the hydrochar surface while some remain in the aqueous phase (Valentas, 2011). Experimental work from various groups have reported lipid retention for *N. Oculuta* for the residence time of 30 minutes between 80-100% in the hydrochar (Levine, et al., 2013) (Friedl, et al., 2005). Given the above findings, the assumption of no lipids being lost to the aqueous phase was imposed.

3.2.2 Ash Simplifications

Ash composition varies significantly between the different microalgal species, and there are limited studies on ash dissolution in HTC. However, it is known that HTC reduces the ash content in hydrochar as the inorganics in biomass dissolve in the aqueous product (Akbari, et al., 2019) (Yoshimoto, et al., 2023). It is also known that temperature and residence time are the main factors affecting the percentage of ash remaining in the hydrochar. Another crucial factor for the fate of ash found in the raw biomass is whether it consists more of water-soluble or insoluble chemicals (Mäkelä, et al., 2015). In addition, as there are limited studies on ash, it would be difficult to generate a model which includes a mechanism for its dissolution. As a result, it was deemed appropriate to assume that ash remains in the aqueous phase since a more complex model would be susceptible to overfitting.

3.2.3 Gas formation & gas yield approximations

The gas was characterised as pure CO₂. This is because the concentration of CO₂ was substantially higher than of any other species (Bevan, et al., 2021). Literature suggests that the gas fraction resulting from the HTC of microalgae is small (Lachos-Perez, et al., 2022). However, there was a notable absence in literature of quantified experimental gas yields, especially for microalgae. Consequently, it was necessary to resort to gas yields from macroalgae (multicellular organisms) for the purpose of modelling the gas formation. However, this approach introduces uncertainty and results in gas formation being less reliable in comparison to hydrochar formation.

3.3 Reaction Kinetics

3.3.1 Reaction Schemes

Various reaction schemes were investigated as potential candidates for representing the HTC process. The primary scheme considered is portrayed by Figure 3. Five elementary reactions were adopted to build up the lumped model, which all followed Arrhenius kinetics (Equation 1). Glucose and Amino acids were used as the precursors to describe a series of parallel and competing reactions with various intermediate products that eventually form hydrochar alongside CO₂.

$$k_i = A_{oi} e^{-\frac{E_{ai}}{RT}} \quad (1)$$

The reactions are as follows:

- Reaction 1: Gas evolved through decarboxylation and decarbonylation reactions to produce CO₂.
- Reaction 2: Polymerization reactions (of intermediate products) which utilise glucose to form secondary char.
- Reaction 3: Polymerization reactions (of intermediate products) which utilise amino acids to form secondary char.
- Reaction 4: Hydrolysis of carbohydrates to glucose monomers.
- Reaction 5: Hydrolysis of proteins to amino acids monomers.

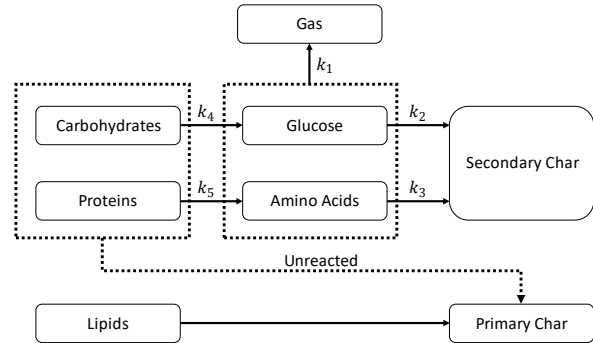


Figure 3: Model reaction scheme with rate constants labelled for each reaction.

An alternative model of greater complexity was also explored. The main distinction was that the gas evolution resulted from both the hydrochar and the dissolved species, which is likely be more representative of the realistic process of HTC. However, its adoption was hindered due to concerns of overfitting given the limited availability of microalgae data for HTC.

3.3.2 Material Balances

Most experimental work found in literature pertains exclusively to batch reactors, especially in the case of microalgal feedstocks. Therefore, the following material balances were formulated to describe HTC as per the primary reaction scheme investigated. Note that because gas was formed from both amino acids and glucose, it was assumed that they contribute equally to the gas production.

$$r_1 = \frac{dM_{gas}}{dt} = k_1(M_a^{n_3} + M_{gl}^{n_4}) \quad (2)$$

$$r_2 = \frac{dM_{gl}}{dt} = k_4(M_{cs}^{n_4} - k_2(M_{gl})^{n_2} - 0.5k_1(M_{gl})^{n_1}) \quad (3)$$

$$r_3 = \frac{dM_a}{dt} = k_5(M_p^{n_5} - k_3(M_a)^{n_3} - 0.5k_1(M_a)^{n_1}) \quad (4)$$

$$r_4 = \frac{dM_{cs}}{dt} = -k_4(M_{cs})^{n_4} \quad (5)$$

$$r_5 = \frac{dM_p}{dt} = -k_5(M_p)^{n_5} \quad (6)$$

$$\frac{dM_{so}}{dt} = +k_4(M_{gl})^{n_4} + k_5(M_a)^{n_5} \quad (7)$$

$$M_{s1} = M_{so} + M_{cs} + M_p + M_{lip} \quad (8)$$

3.4 Optimisation Algorithms

3.4.1 Algorithm & Problem Formulation

To carry out the estimation of process parameters an optimisation problem was set up with the objective being formulated as a cost function which minimises the Mean Square Error (MSE). The formulation utilizes the difference between predicted and experimental hydrochar masses rather than yield to allow for the optimiser to have inherent information regarding SL. The choice of MSE was due to the significantly different number of experimental results available for the mass and gas yields. This aided in placing equal importance to the cost contributions. In subsequent optimisations, weightings were further placed to alter the influence of the contributions and encourage gas evolution. Finally, the mean experimental values for the hydrochar mass and gas yield appear in the objective function for normalisation purposes.

Objective function:

$$MSE = \frac{1}{n} \sum_1^n \frac{(M_{s1}^p - M_{s1}^e)^2}{\mu_{M_{s1}}} + \frac{1}{m} \sum_1^m \frac{(Y_g^p - Y_g^e)^2}{\mu_g} \quad (9)$$

The 'GlobalSearch' optimiser from MATLAB's Global optimisation toolbox was implemented to identify a global minimum. This optimiser falls under the asymptotically complete search methods which can only guarantee global optimality under infinite completion time. The local algorithm of choice used to solve the optimisation problem was the interior point and the feasible search space was set up as:

$$10 \leq A_{oi} \leq 10^{12} \quad \forall i \in \{1, \dots, 5\} \quad (10)$$

$$0.5 \leq n_i \leq 5 \quad \forall i \in \{1, \dots, 5\} \quad (11)$$

$$10^4 \leq E_{ai} \leq 10^7 \quad \forall i \in \{1, \dots, 5\} \quad (12)$$

3.4.2. Evaluation of Model Performance

To evaluate the model performance the coefficient of determination was used which describes the goodness of model fit (Hukkerikar, et al., 2012) as well as the RMSE.

$$R^2 = 1 - \frac{\sum_i (x_i^e - x_i^p)^2}{\sum_i (x_i^e - \mu)^2} = 1 - \frac{RSS}{TSS} \quad (13)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_i (x_i^e - x_i^p)^2} \quad (14)$$

3.5 Carbon Content and HHV Predictions

After the development of the model, predicted mass outputs for glucose, amino acids, hydrochar, gas, carbohydrates, and protein were obtained. The proportions of carbohydrates and proteins that reacted to form char were calculated using Equations 15 and 16. The percentage of char formed from those two biomolecules, was determined by Equation 17.

$$M_{Cs} = M_{Cs,0} - M_{gl} - 0.5M_{gas} \quad (15)$$

$$M_p = M_{p,0} - M_a - 0.5M_{gas} \quad (16)$$

$$y_{i,char} = \frac{M_i}{M_{s1}} \text{ where } i = \{Cs, p, lip\} \quad (17)$$

Next, two different approaches were taken to build a model for C content predictions. The initial approach was to find an average C content for carbohydrates, proteins, and lipids individually, and then use their contributions to find out how much of the carbon was accumulated in the char. However, due to the large variation in the types of carbohydrates, proteins, and lipids in different algal species, this model was found to be very difficult to balance regarding the accuracy and model applicability across a wide variety of species, even for the calibration data set.

The alternative method for modelling the carbon content of char was performed via Equation 18, where the artificial parameters $A_{i,j}$ and $n_{i,j}$ were optimised with the objective function of minimising the Sum of Squared Residuals (SSR) shown in Equation 20. This optimisation was carried using the GRG nonlinear solver with 84 data points. The same approach was also applied to find the predicted hydrogen content, minimising the SSR.

$$y_{j,char}^p = \sum_i A_{i,j} y_{i,char}^{n_{i,j}} \quad (18)$$

$$\text{for } i = \{Cs, p, lip\}, j = \{C, H, O\} \quad (19)$$

$$SSR_j = \sum_1^n (y_{j,char}^p - y_{j,char}^e)^2 \text{ for } j = \{C, H\} \quad (20)$$

Finally, HHV was formulated as a function of the predicted carbon, hydrogen, and oxygen content using Dulong's empirical equation (Smitha, et al., 2016). From this, the oxygen content parameters were optimised by minimising the SSR shown in Equation 22.

$$HHV_{char}^p = 100 \times \left(0.338 y_{C,char}^p + 1.428 \left(y_{H,char}^p + \frac{y_{O,char}^p}{8} \right) \right) \quad (21)$$

$$SSR_{HHV} = \sum_1^n (HHV_{char}^p - HHV_{char}^e)^2 \quad (22)$$

3.6 Energy Balance

An energy balance was performed to calculate the total energy input required for the system. The system was assumed to be under isothermal steady state operation with no heat losses to the surroundings. The total energy supplied to heat up the slurry from ambient to the set operating temperature was modelled as a constant:

$$Q_{heating} = M_{biomass} \Delta T \bar{c}_p \quad (22)$$

$$\Delta T = T_{set} - T_{ambient} \quad (23)$$

$$\bar{c}_p = [(1 - SL)c_{p,water} + (SL)c_{p,algae}] \quad (24)$$

The overall enthalpy of reaction was calculated from Hess's law using the heat of combustions of the components, where γ_i was defined as the extent of reaction i .

$$H_{reaction} = \sum_i H_i \text{ for } i = \{1a, 1g, 2, 3, 4, 5, 6\} \quad (25)$$

$$H_i = \gamma_i (HHV_{product,i} - HHV_{reactant,i}) \quad (26)$$

The amount of energy in debt was then calculated using Equation 27. When E_{debt} is positive, the system is in energy deficit, meaning energy input is higher than energy produced by the reactions; when E_{debt} is negative, the system is in energy surplus.

$$E_{debt} = Q_{heating} + H_{reaction} \quad (27)$$

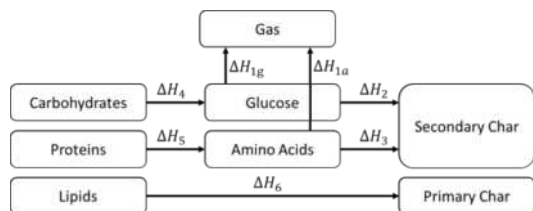


Figure 4: Enthalpy changes map with enthalpy changes labelled for each reaction.

4. Results & Discussion

4.1 Multi Step Optimisation (MSO) Model

Encouraging yet modest results were observed for the integration of the Multi Step Optimisation model (Figure 5). The overall coefficient of determination (R^2) for the hydrochar mass was evaluated to be 0.245 while the RMSE was 0.041. As expected, these were worse in comparison to the DO model that involved direct optimisation. Further, it was apparent that the trained MSO model resulted in systematic overpredictions.

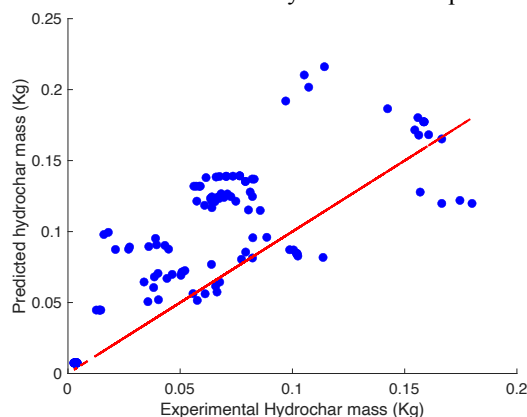


Figure 5: MSO model calibration parity plot.

To gain insight to the outlying behaviour of the system, examining the respective parity plots of the two biomolecules is necessary. The R^2 values obtained, suggest that the estimated kinetic parameters (see Supplementary Section S1) provide a good fit to the experimental hydrolysis data (Figure 6). Additionally, it is evident that carbohydrates are predicted to undergo a more rapid dissolution comparatively to proteins as also documented in literature (Qian, et al., 2021). Thus, the overprediction was understood to stem from the assumption of isothermal conditions and the skewed temperature range that the hydrolysis training data spanned (393K–453K). The typical operating envelope for the HTC of microalgae is between 453–523K and because no heating up stage was assumed; the hydrolysis of the macromolecules was systematically underestimated. As the model considered that the

unreacted biomolecules constitute to the formation of primary char, inevitably overpredictions would be observed. Further, the hydrolysis training data originated from only two microalgal species (*Chlorella Vulgaris* and *Scenedesmus Spp.*). Therefore, efforts shifted towards the development of the DO model to allow for wider applicability.

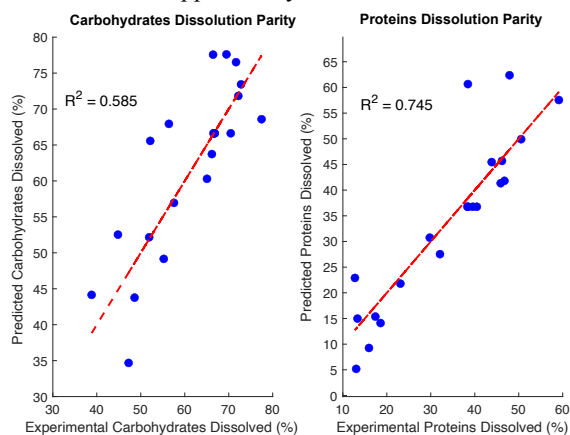


Figure 6: DO model hydrolysis calibration parity plot for (Left) carbohydrates (Right) proteins.

4.2 Direct Optimisation (DO) Model

4.2.1 Calibration Model Performance.

Moving to the DO model there were still hints of overprediction present (Figure 7). However, the extent of it was limited, which was reinforced by the RMSE of 0.023 and the much-improved R^2 coefficient of 0.774 compared to the MSO model. This highlighted that the DO model was successful in providing sound predictions across the microalgae species that it was calibrated against.

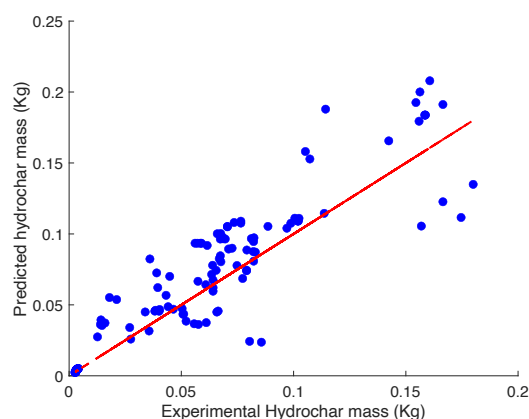


Figure 7: DO model calibration parity plot.

4.2.2 Hydrochar yield Calibration Profiles

Examinations of the DO model concluded that the direct optimisation for the hydrochar conformed with the expectation set from literature that the hydrolysis of carbohydrates would be faster than that of proteins (Figure 8). However, the model also predicted a more delayed hydrolysis of proteins than anticipated. This revealed some of the limitations of not using hydrolysis data for training purposes for the DO model.

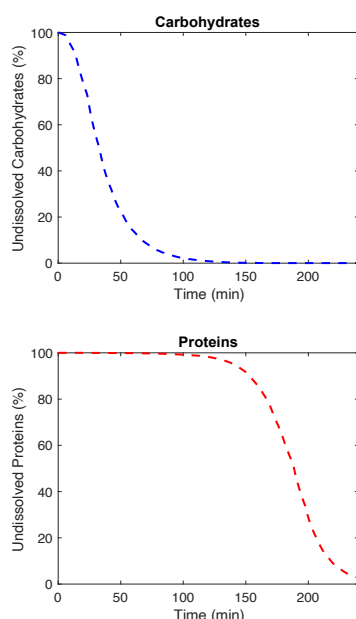


Figure 8: DO model hydrolysis profile of polyculture for (Top) carbohydrates at SL=0.16, $\tau=240$ min (Bottom) proteins SL=0.16, $\tau=240$ min.

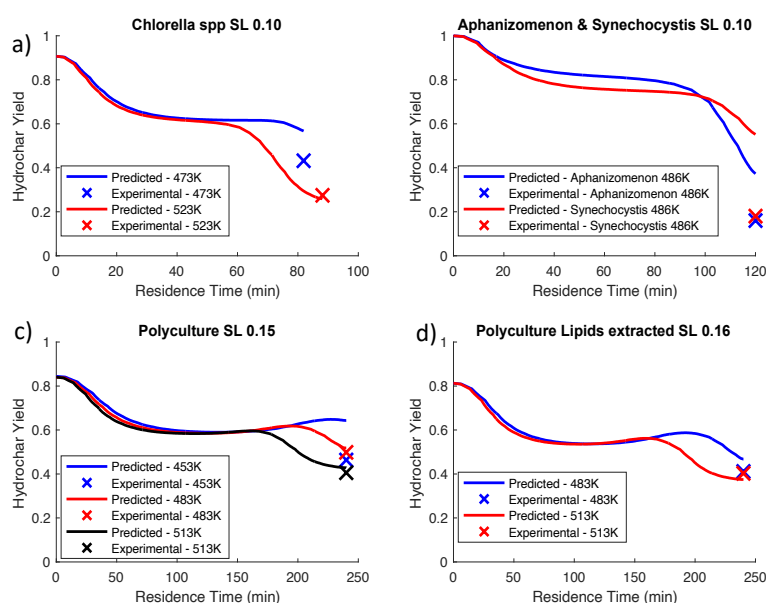


Figure 9: DO model hydrochar yield profiles for (Top Left) *Chlorella* Spp. at SL=0.1 (Top Right) *Aphanizomenon* & *Synechocystis* at SL=0.1, $T=486$ K (Bottom Left) Microalgae Polyculture at SL=0.15 (Bottom Right) lipids extracted Microalgae Polyculture at SL=0.16. Yields do not begin from one due to the assumption of immediate transfer of ash onto the liquid phase.

Figure 9 depicts 4 hydrochar yield profiles obtained from the DO model (further profiles can be viewed in the supplementary section S5). The model predicts that different microalgae species undergo carbonization in similar but distinct ways. This is due to the variations in composition as described previously. In addition, it illustrates that the performance of the model is improved for higher temperatures and that the hydrolysis of carbohydrates is less temperature dependent in comparison to protein hydrolysis and secondary char formation.

The initial drop in hydrochar yield can be traced to the more rapid dissolution of the carbohydrates whilst the second drop can be attributed to the delayed dissolution of proteins. Notably, there were periods of time that the hydrochar yield observed for the microalgae polycultures increased. This can be explained by the secondary char formation beginning to dominate for polycultures that were high in carbohydrate content at longer residence times. This is promising because mixed cultivating ponds are expected to be used in industry. Limited secondary char formation was observed from high protein microalgal species. This was because gas evolution depleted the glucose formed from the more rapid dissolution of carbohydrates, while the residence times were not sufficient for protein dissolution to take place. The approach of using macroalgae that are high in carbohydrate content for the gas formation may have exacerbated this behaviour (Cai, et al., 2013).

4.2.3 Validation of hydrochar yield

Analysis on the predictive capabilities of the DO model, reveals that the tendency to overestimate the hydrochar

yields carries forward to the validation sets as depicted in Figure 10. The coefficient of determination R^2 was evaluated to 0.37. This was considered satisfactory considering the inherent variability of microalgal species, generalised nature of the model and simplifications introduced.

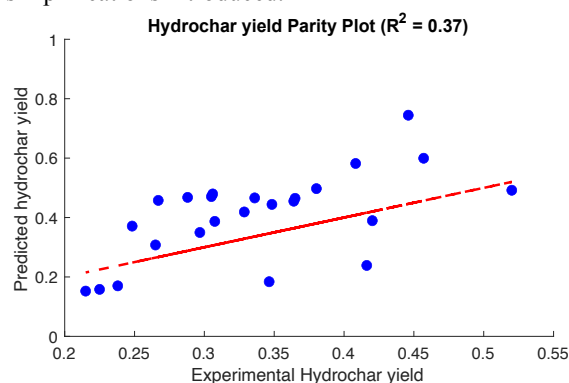


Figure 10: DO model validation parity plot.

Figure 11 suggests that the model predicts an increase in hydrochar yield with SL, as has previously been observed by (Aragón-Briceño, et al., 2020). In contrast, an increase in operating temperature leads to a decrease in the hydrochar yield owing to further degradation arising from gas formation and protein dissolution. Notably, the DO model performs well for deashed *Scenedesmus* across the temperatures examined, while also being able to capture SL variations for *Nannochloropsis*. On the other hand, for the same microalgal species the model poorly encapsulates the effect of temperature on hydrochar yield at lower operating temperatures. Improvements in predictive accuracy emerge at elevated thermal conditions. This

may be an indication of the limited capabilities of the model at accurately predicting the protein hydrolysis for certain microalgae at lower temperatures.

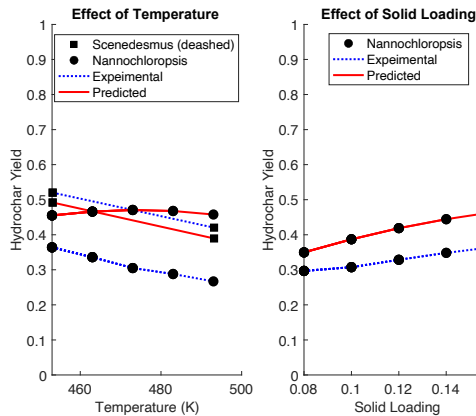


Figure 11: Effects of temperature and solid loading individually on hydrochar yield.

Validation under simultaneous variations of thermal and SL conditions was performed as shown in Figure 12. *Chlorella Vulgaris* is evidently one of the species that the DO model overpredicts the hydrochar yield for. The largest deviation is depicted to occur at experimental conditions of 453K and 5% SL. This once again highlights that at lower temperatures the performance of the model suffers compared to higher temperatures. However, this is not significant because HTC typically operates at higher temperatures to benefit from energy neutrality. Overall, the DO model captures well the effect of temperature and SL variations on the hydrochar yield.

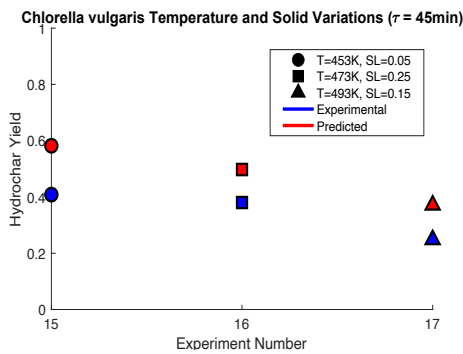


Figure 12: Validation on effect of hydrochar yield by varying temperature and solid loading simultaneously.

4.3.1 Carbon Content & HHV Model

Ensuring model consistency, the carbon (C) content and HHV models were both exclusively trained using microalgae experimental data. The C content model achieved an R^2 value of 0.75 and a $RMSE$ of 0.04, which is half of the $RMSE$ when using the average of the experimental data (0.08). The hydrochar HHV model on the other hand exhibited a lower R^2 value of 0.50, and a $RMSE$ of 2.6 that was 30% lower than when comparing with the use of the average. This disparity in model performance was deduced to be caused by the limited availability of experimental data for the HHV model

calibration, which were even fewer than that for the carbon content.

An examination of the parity plots reveals distinct patterns (Figure 13). The carbon content model tends to overpredict when values are low, and underpredict when values are high. This can originate by the absence of ash content in the solid fraction. The HHV plot indicates to an underfitting of the data set, which could be attributed to both the low number of data available, and the absence of ash.

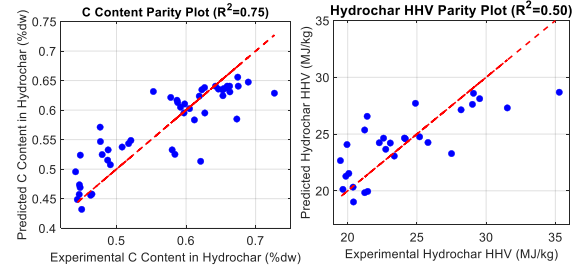


Figure 13: Parity plots for (Left) carbon content (Right) hydrochar HHV.

4.3.2 Validation of C Content & HHV

The obtained R^2 values of -18 and -3.8 respectively suggested a significant degree of model overfitting. To comprehend the shortcomings of the model, a detailed analysis was conducted to assess its performance across different variations. The findings, as illustrated in Figure 14, reveal notable inadequacies, particularly when attempting to account for variances attributed to temperature.

The model has diametrically opposite trends when explaining temperature-induced variations. Interestingly when considering the joint influence of temperature and residence times the model closely aligns with the trend in C content, albeit with an inclination to slightly overpredict. Nevertheless, the model struggles in predicting HHV. The absence of experimental data specifically isolating the effects of residence time, SL, and algal species complicates diagnosing the reasons hindering the model performance.

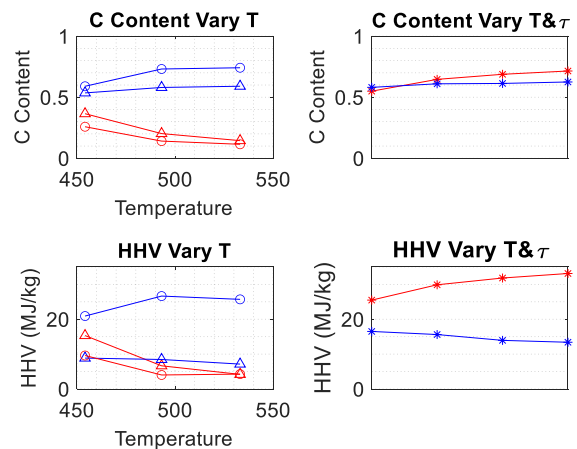


Figure 14: Validation effect of temperature and residence time on C content and HHV. Red stands for predicted values, blue stands for experimental data. Operating conditions: (Triangles) Scenedesmus, (Circles) Deashed Scenedesmus, $t=240\text{min}$, $SL=0.06$. (*) *Spirulina Platensis*, $SL = 0.09$.

Some of the hindrances of the model can be attributed to three main factors. Firstly, local optimisation was employed for the determination of C content, and hence better solutions might be available. Secondly, both the calibration and validation data sets are small, and there are many varieties of algae with different types of carbohydrates, proteins, and lipids each containing different compositions of C, H and O. Hence, larger data sets would help provide a better explanation of the variance. Finally, the ash content was not considered, which is expected to worsen the HHV, meaning overpredictions are expected especially for those with high ash content in the biomass feed (Akbari, et al., 2019).

4.2.3 Energy Debt

The energy balance profiles over the HTC reactor implies that there are endothermic hydrolysis reactions initially, resulting in an energy deficit within the system. Then, as the exothermic reactions take precedence, energy is generated by the system. The residence time necessary to reach energy neutrality varies and is influenced by the algae species in the feed, the operating temperature, and SL.

In figure 15, three ED profiles are depicted for different algae species, all conducted at 240°C with a SL of 15% over a total residence time of 2 hours. Notably, *Nannochloropsis Oculata* demonstrates the earliest attainment of energy neutrality, followed by *Chlorella Spp*. This trend in neutrality times may be attributed to their different compositions, particularly the highest total content of proteins and carbohydrates in *Nannochloropsis*, and the second most in *Chlorella Spp*. This implies a more abundant availability of amino acids and glucose, expediting the formation of char and gas, which are both exothermic reactions that help the system payback the energy debt.

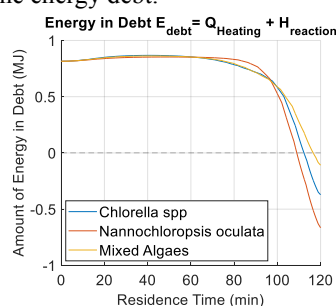


Figure 15: Energy debt profile for different microalgae species.

Figure 16 illustrates on the left the ED profiles for *Chlorella* across five distinct temperatures, maintaining a SL of 15% over a 2-hour period. The graphical representation underscores a relationship between temperature variation and the speed at which energy neutrality is achieved. Higher temperatures result in a more rapid attainment of energy neutrality. Notably, any temperature below 230°C results in a continuous energy deficit throughout the entire process.

On the right of Figure 16 are the ED profiles for *Chlorella* at a constant temperature of 240°C across

various SL over 2 hours. This reveals a nuanced relationship between SL and the time required to reach energy neutrality. Higher SL was found to result in a more rapid realization of neutrality, primarily attributed to the increased in biomass available for reactions. However, this trend diminishes over time.

It is interesting to observe that at a SL of 0.3, the system exhibits an endothermic domination for residence times between 75-95 minutes. This is inherited from the kinetic model, where the proteins hydrolysis predictions are delayed, the secondary hydrolysis region creating a second peak in the ED profile. The starting energy debt is larger at higher temperatures as a larger amount of energy is needed to heat the reactants to the required temperature.

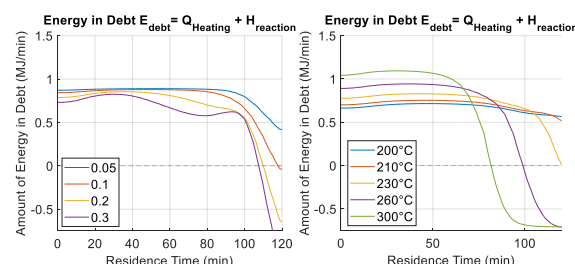


Figure 16: Energy debt profile for (Left) different temperatures (Right) different solid loadings.

4.4.1 Limitations of the Direct Optimisation Model

The current iteration of the unified model is limited to only hydrochar yield predictions. For the HHV and C content, predictions were only made for the set of data for *Spirulina* in the validation data set. Furthermore, the ash simplifications introduced may hinder the use of the unified models for environmental and life cycle assessments, as there is limited information regarding the end state of the inorganic ash components. This would indicate that these environmental assessments would over-estimate the environmental impact of the overall process, as all the ash is assumed to be dissolved in water.

4.4.2 Reasons for Model Limitations

Scarce availability of experimental dissolution yields hindered the performance of the MSO model. Therefore, the hydrolysis of the macromolecules was entirely governed from the best performing model. This impacted the performance of the model at lower temperatures as the hydrolysis of proteins was underestimated. Furthermore, the assumption that lipids remain entirely in the hydrochar begins to falter at higher temperatures. This may have contributed to the systematic overestimation of the hydrochar yields. Moreover, as the ash content would contribute negatively to the HHV value (Timilsina, et al., 2024) the model predicts a better-quality product char than would be expected and may have hindered the optimisation algorithm from converging to a better solution. Finally, utilizing macroalgae that contain higher carbohydrates content for the representation of the gas yields hindered

the accuracy of the predictions on microalgae species that conversely have higher protein content.

5. Conclusion

Unified kinetic models have been demonstrated to reasonably model the dynamics of the HTC of algae, which may prove useful in avoiding the need to develop a separate model for each individual microalgal species. Instead, unified models for groups of microalgal species could be developed based on higher carbohydrate or protein content, saving time and resources.

The results shown suggest that the DO model is a promising approach, but future work should revisit the Multistep Optimisation model to consider a wider variety of data. The models develop perform best for polycultures at higher temperatures, which is where HTC reactors are expected to be operated in industry, using multiple microalgal species at once.

Further work could also consider the collection of in-house data with a particular focus on, obtaining gas and hydrolysis yields for a wide range of microalgae to aid wider applicability. The standardisation of data collection could limit the underlying uncertainty arising from utilising experimental data from several different researchers with different experimental set-ups. Additional model complexity could also be considered by removing the simplifications made on lipids and ash.

5. Nomenclature

Symbol	Definition
A_{oi}	Pre exponential factor for reaction i
\bar{c}_p	Specific heat capacity of slurry
$c_{p,water}$	Specific heat capacity of water
$c_{p,algae}$	Specific heat capacity of algae
E_{ai}	Activation energy of reaction i
E_{debt}	Energy in debt
H_i	Enthalpy of reaction i
HHV_{char}	Higher heating value of hydrochar
k_i	Rate constant of reaction i
k_o	Reference rate constant
m	Number of gas yield experiment data points
M_i	Mass of component i
MSE	Mean square error
n	Number of hydrochar yield experiment data points
n_i	Order of reaction i
N	Number of experiments for property i
$Q_{heating}$	Energy supplied via heating
r_i	Reaction rate of reaction i
R	Gas constant
R^2	Coefficient of determination
SL	Solid Loading

SSR_i	Sum of Squared Residuals of i
τ	Residence Time
TSS	Total Sum of Residuals
$y_{i,char}$	Mass fraction of i in char
y_i^j	Yield of i resulted from method j
γ_i	Extent of reaction i
ΔT	Temperature change of reactor by heating
μ	Experimental mean

Subscripts:

Symbol	Definition
x_a	Property of amino acid
x_c	Property of carbon
x_{cs}	Property of carbohydrates
x_{gas}	Property of gas
x_H	Property of hydrogen
x_{gl}	Property of glucose
$x_{i,0}$	Initial value of the property of i
x_{lip}	Property of lipids
x_O	Property of oxygen
x_p	Property of proteins
x_{so}	Property of secondary hydrochar
x_{s1}	Property of total hydrochar

Superscripts:

Symbol	Definition
x^e	Value obtained from literature experiments
x^p	Value predicted from model

6. Acknowledgements

The authors would like to express their sincere gratitude to Mr. Suhaib Nisar for his insightful guidance and support throughout the entire project.

7. References

- Adam, R. et al., 2023. Systematic homogenization of heterogenous biomass batches – Industrial-scale production of solid biofuels in two case studies. *Biomass and Bioenergy*, 173(106808).
- Akbari, M., Oyedun, A. O. & Kumar, A., 2019. Comparative energy and techno-economic analyses of two different configurations for hydrothermal carbonization of yard waste. *Bioresource Technology Reports*, Volume 7.
- Alberto Galifucoco, 2019. A New Approach to Kinetic Modeling of Biomass Hydrothermal Carbonization. *ACS Sustainable Chemistry & Engineering*, Issue 13073-13080.
- Aragón-Briceño, C., Grasham, O., Ross c, A. & Dupont, V., 2020. Hydrothermal carbonization of sewage digestate at wastewater treatment works: Influence of solid loading on characteristics of

- hydrochar, process water and plant energetics. *Renewable Energy*, Volume 157, pp. 959-973.
- Bevan, E., Fuab, J., Lubertia, M. & Zheng, Y., 2021. Challenges and opportunities of hydrothermal carbonisation in the UK; case study in Chirnside. *Royal Society of Chemistry*, Volume 11.
- Bevan, E., Fu, J., Luberti, M. & Zheng, Y., 2021. Challenges and opportunities of hydrothermal carbonisation in the UK; case study in Chirnside. *Royal Society of Chemistry*, Volume 11.
- Broch, A., Jena, U., Hoekman, S. & Langford, J., 2014. Analysis of Solid and Aqueous Phase Products from Hydrothermal Carbonization of Whole and Lipid-Extracted Algae. *Energies*, 7(1).
- Cai, T., Park, S. Y. & Li, Y., 2013. Nutrient recovery from wastewater streams by microalgae: Status and prospects. *Renewable and Sustainable Energy Reviews*, Volume 19, pp. 360-369.
- Cao, Y. et al., 2021. Hydrothermal carbonization and liquefaction for sustainable production of hydrochar and aromatics. *Renewable and Sustainable Energy Reviews*, 152(111722).
- Chen, C., Liang, W., Fan, F. & Wang, C., 2021. The Effect of Temperature on the Properties of Hydrochars Obtained by Hydrothermal Carbonization of Waste *Camellia oleifera* Shells. *ACS Omega*, 6(25).
- Chen, H. et al., 2022. Hydrothermal hydrolysis of algal biomass for biofuels production: A review. *Bioresource Technology*, 334(126213).
- Czerwińska, K., Śliz, M. & Wilk, M., 2022. Hydrothermal carbonization process: Fundamentals, main parameter characteristics and possible applications including an effective method of SARS-CoV-2 mitigation in sewage sludge. A review. *Renewable and Sustainable Energy Reviews*, 154(111873).
- Friedl, A., Padouvas, E., Rotter, H. & Varmuza, K., 2005. Prediction of heating values of biomass fuel from elemental composition. *Analytica Chimica Acta*, Volume 544, pp. 191-198.
- Heilmann, S. M. et al., 2010. Hydrothermal carbonization of microalgae. *Science Direct*, 34(6), pp. 875-882.
- Hukkerikar, A. S. et al., 2012. Group-contribution+ (GC+) based estimation of properties of pure components: Improved property estimation and uncertainty analysis. *Fluid Phase Equilibria*, Volume 321, pp. 25-43.
- Ischia, G. & Fiori, L., 2021. Hydrothermal Carbonization of Organic Waste and Biomass: A Review. *Waste and Biomass Valorization*, 12(2797-2824).
- Jung, D., Zimmermann, M. & Kruse, A., 2018. Hydrothermal Carbonization of Fructose: Growth Mechanism and Kinetic Model. *ACS Sustainable Chemistry & Engineering*, Issue 13877-13887.
- Kambo, H. S. & Dutta, A., 2015. A comparative review of biochar and hydrochar in terms of production, physico-chemical properties and applications. *Renewable and Sustainable Energy Reviews*, Volume 45, pp. 359-378.
- Lachos-Perez, D. et al., 2022. Hydrothermal carbonization and Liquefaction: differences, progress, challenges, and opportunities. *Bioresource Technology*, 343(126084).
- Levine, R. B. et al., 2013. The Use of Hydrothermal Carbonization to Recycle Nutrients in Algal Biofuel Production. *Environmental Progress & Sustainable Energy*, 32(4), pp. 962-975.
- Liu, H., Chen, Y., Yang, H. & Gentili, F. G., 2019. Hydrothermal carbonization of natural microalgae containing a high ash content. *Fuel*, Volume 249, pp. 441-448.
- Liu, T. et al., 2018. Effect of hydrothermal carbonization on migration and environmental risk of heavy metals in sewage sludge during pyrolysis. *Bioresource Technology*, Volume 247, pp. 282-290.
- Mäkelä, M., Benavente, V. & Fullana, A., 2015. Hydrothermal carbonization of lignocellulosic biomass: Effect of process conditions on hydrochar properties. *Applied Energy*, Volume 155, pp. 576-584.
- Qian, F. et al., 2021. Kinetics of hydrolysis of microalgae biomass during hydrothermal pretreatment. *Biomass and Bioenergy*, Volume 149.
- Smitha, A. M., Singh, S. & Ross, A. B., 2016. Fate of inorganic material during hydrothermal carbonisation of biomass: influence of. *Fuel*, Issue 169, pp. 135-145.
- Timilsina, M. S. et al., 2024. Prediction of HHV of fuel by Machine learning Algorithm: Interpretability analysis using Shapley Additive Explanations (SHAP). *Fuel*, Volume 357.
- Valentas, N., 2011. *Hydrothermal Carbonization: An Innovative New Process for the Extraction of Algal Oil*. s.l., AIChE Annual Meeting.
- Yang, J. et al., 2023. Pyrolysis and hydrothermal carbonization of biowaste: A comparative review on the conversion pathways and potential applications of char product. *Sustainable Chemistry and Pharmacy*, Volume 33.
- Yoshimoto, S. et al., 2023. Effects of potassium on hydrothermal carbonization of sorghum bagasse. *Bioresources and Bioprocessing*.
- Zhang, Y. et al., 2023. Recent advances in lignocellulosic and algal biomass pretreatment and its biorefinery approaches for biochemicals and bioenergy conversion. *Bioresource Technology*, 367(128281).

Machine Learned Equation of State for the 2D Lennard-Jones Fluid

Daniel McGlinchey, Sanjay Kumaran

Department of Chemical Engineering, Imperial College London, U.K.

Abstract: A thermodynamically consistent machine learned equation of state (EoS) is developed for the two-dimensional Lennard-Jones (2D-LJ) fluid. A database of 9116 data points of thermophysical properties for the 2D-LJ fluid is created using molecular dynamics simulations. A physics informed neural network (PINN) is then trained on the database of: compressibility factor, internal energy, heat capacities, thermal expansion coefficient, isothermal compressibility, thermal pressure coefficient and Joule-Thomson coefficient. The EoS property prediction performance is in agreement with the simulation data. The phase equilibria prediction of the EoS is compared to previously published EoS, where it outperforms them around the critical region.

Introduction

Fluid property prediction is essential to chemical engineering. Modelling heat transfer, separation systems, and reactors rely on accurately predicting fluid properties.

Thus, centuries of research have occurred in this field, starting with the ideal gas relation, formed by amalgamating Boyle's law, Charles' law, and Gay-Lussac's law (1). Further developments were made to produce the van der Waals equation of state (EoS) (2) which is the basis for advancements today by introducing intermolecular force contributions. A notable ongoing development is that of Statistical Association Fluid Theory (SAFT)(3), which adds association theory to these models (4).

Many methods of fluid property prediction exist, empirical EoS' such as the Redlich-Kwong EoS (RK EoS) (5), theoretical EoS' such as SAFT- γ -Mie (6), molecular dynamics (MD) simulations (7) and machine-learning approaches.

Empirical approaches can accurately predict fluid properties within their range of validity, provided that experimental data exists to calculate the required parameters. However, there is a trade-off between critical point estimation and density estimation between the methods used to calculate these parameters (8).

MD simulations attain total system properties by averaging a sum of microstates (individual particle arrangements), encompassing total system behaviour (9). Given that an accurate model of the forces between molecules in the system exists, simulations can produce more accurate datapoints than experiments due to the lack of human error. A downside to MD simulations is that only properties at discrete temperatures and densities can be calculated; thus, to model a system with a gradient in temperature or density, the properties must be calculated at intervals along the temperature profile, and properties between these intervals must be interpolated. Unphysical behaviour may be

observed if thermodynamically consistent interpolation isn't used (10).

Previous studies (11), propose implementing a feed-forward artificial neural network (ANN) as an equation of state. Whilst Zhu's approach proved the viability of ANNs and Gaussian Process Regression (GPR), it fitted generated data without a theoretical foundation. Chaparro resolved this using MD and calculating the various Helmholtz derivatives (12). This methodology ensures thermodynamic consistency of the predicted properties since the relations used to predict the properties from the Helmholtz free energy satisfy Maxwell's relations. Given this approach worked well for the Mie fluid, testing this approach on other fluids is the next logical step.

This paper uses this methodology on the two-dimensional Lennard Jones (2D-LJ) fluid. Applications of the 2D-LJ include modelling colloidal systems (13), surfactant adsorption (14) and soft matter applications (15). Utilising the same workflow as Chaparro, this paper aims to produce an EoS for the 2D LJ using molecular dynamics and machine learning for accurate property prediction.

Background

The Lennard Jones (LJ) potential is a theoretical model of spherical particles with a soft-sphere reference state, representing intermolecular interactions. This model is based on van der Waals theory and consists of strong repulsion forces at short distances and attractive forces at longer distances, obeying equation [1].

$$V(r) = 4\epsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^6 \right] \quad [1]$$

$V(r)$ represents the LJ potential, σ is the effective diameter of each particle, r is the distance between centres of the particles, and ϵ is the potential well depth.

In the current literature, the EoS shown in **Table 1** have been proposed for the 2D LJ EoS.

Table 1: Available EoS for 2D-LJ

2D LJ Equation of State	Type	Year of Publication
Percus - Yevick(16)	Theoretical	1977
Reddy and O'Shea(17) (ROS)	Semi-Empirical	1986
CMO(18)	Theoretical	1999
Hypernetted – Chain (HNC3)(19)	Semi-Theoretical	2012
2PT(2D)(20)	Semi-Theoretical	2018

Semi-empirical EoS are correlated to simulation data with no theoretical basis; Semi-theoretical EoS are correlated to simulated data using a theoretical basis (22).

Whilst theoretical models are based on theories, such as perturbation theory and integral equations, forming a theoretical EoS is laborious. An alternative to creating a theoretical equation of state is to use empirical approaches such as least-squares regression.

The Helmholtz free energy is defined by equation [2]:

$$A = U - TS \quad [2]$$

Here, A is the Helmholtz free energy, U is the internal energy, T is the temperature and S is entropy.

The basis for developing the EoS in this paper is to use MD to generate a database of properties defined by the partial derivatives of the Helmholtz free energy. The EoS is then set up such that the density and temperature are inputs to an ANN, which predicts the residual Helmholtz free energy. Automatic differentiation is then used to calculate the partial derivatives of the residual Helmholtz free energy with respect to density and temperature. These partial derivatives are then combined using the thermodynamic relations shown in equations [13]-[24] and their ideal contribution to predict the properties in the database.

With an ANN that has learned the Helmholtz free energy surface, this method leads to a thermodynamically consistent equation of state, as demonstrated by Rosenberger (21).

Artificial Neural Network

An artificial neural network (ANN) uses layers of connected neurons to map inputs to targets. The type of ANN used in this paper is a multilayer perceptron (MLP). This consists of layers of neurons where all the neurons in each layer connect to all the neurons in the next layer. A neuron consists of a weight vector, bias, and activation function. The values sent to the neuron from the previous layer

are multiplied by the weight vector, the sum of the values in this vector is then computed, the bias is then applied, and this value is then input to the activation function, which applies a nonlinear transformation. This output is then sent to every neuron in the next layer. The structure of an ANN is shown in **Figure 1**.

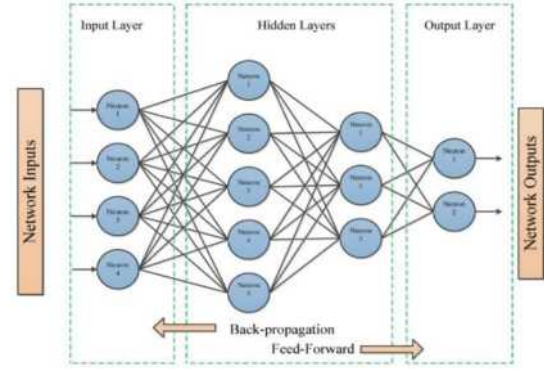


Figure 1: Diagram of ANN structure (23)

According to the Universal Approximation Theorem (24), given a continuous mappable route from input to output, an ANN exists to approximate this relation. A caveat to this lies in the neural network architecture, where the requirement for this can be many neurons per layer and many layers requiring strong computational power. Given that Chaparro's workflow is used in this study, the same ANN architecture is considered with six linear layers, all layers using the Tanh activation function with 45 neurons per layer.

When training, the difference between the desired output and the output from the ANN is computed, this is called a loss. The loss can be calculated in many ways however, it is typical in regression to use a mean squared error (25). When an ANN "learns", or is "trained", it uses the gradient of the loss with respect to the weights and biases in the layers of neurons, along with a gradient descent algorithm to change the weights and biases to decrease the loss.

Methods

Datapoint distribution creation

To fit the ANN on the Helmholtz free energy surface, a database of derivative properties was created at varying reduced temperatures (T^*) and reduced densities (ρ^*). To identify the region to generate data, the vapour-liquid equilibrium (VLE) and fluid-solid equilibrium (FSE) curves were plotted from existing simulation data.

No datasets for the solid, liquid, and vapour phase envelope of the 2D-LJ fluid exist in the literature thus, the envelope was modelled (26). The VLE and critical point data were used to fit

arctangent functions from the critical point to the minimum saturated vapour point, and from the critical point to the minimum saturated liquid point (27,28). FSE data was used to fit an exponential function to represent the FSE curve (29).

The triple point was found at $\rho^*=0.768$, $T^*=0.405$ by solving for where the FSE curve intersects the VLE curve. The phase envelope is shown in **Figure 2**.

A distribution of datapoints was then generated in the ranges $\rho^* \in (0,1)$ and $T^* \in (0.4,10)$. The distribution of the datapoints was generated with a Sobol' sequence. Sobol' sequences are "characterised by low 'discrepancy'"; this maximised the exploration of the space but minimised the number of data points which explored the same region (30). The distribution was then scaled such that the number of data points at a given temperature was inversely proportional to the temperature. This was desired as the low-temperature region is expected to be where "the Helmholtz free energy changes the most due to the presence of phase instability and the transition from subcritical to supercritical behaviour" (31). Isotherms were added to the distribution of points so that the performance of the equation of state could be evaluated against the MD data.

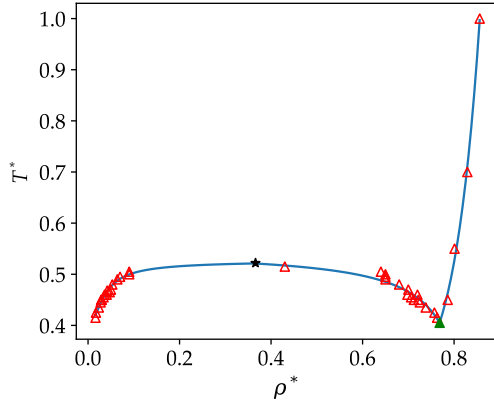


Figure 2: Phase envelope for the 2D-LJ fluid. Red triangles show VLE and SLE data, black star shows critical point, green triangle shows triple point, and blue lines show VLE and FSE curve.

Points were included if they were within 2.5% of the saturation temperature near the critical point, with this percentage linearly increasing to a maximum of 10% at the triple point temperature. This resulted in 13870 discrete points where MD simulations would be performed. The distribution of the data points where MD simulations were performed are shown in **Figure 3**.

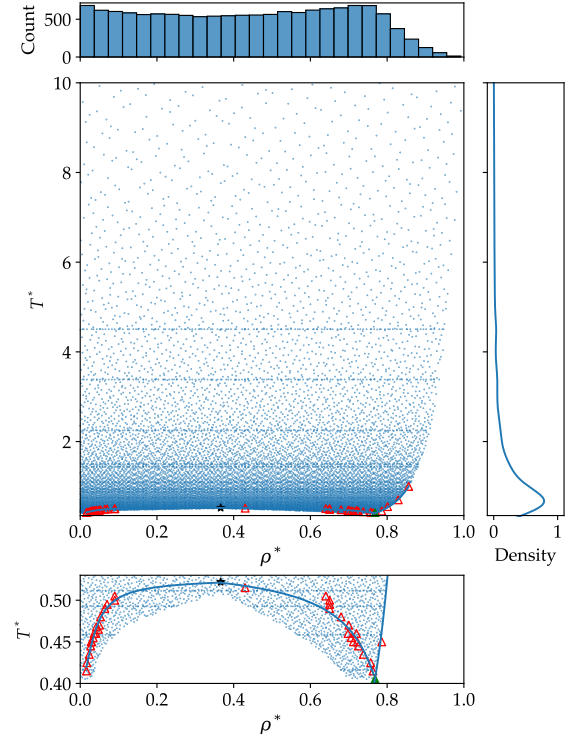


Figure 3: Desired data point distribution. Datapoints are shown in blue dots. X-axis plot shows the distribution of data points with respect to density. Y-axis plot shows kernel density estimate for the distribution of data points with respect to temperature. Lower subplot shows subcritical region and phase envelope from **Figure 2**.

Molecular Dynamics Simulations

The 2D-LJ fluid was simulated using molecular dynamics to obtain the thermophysical properties at each temperature and density. The parameters of the simulations are shown in **Table 2**.

Table 2: Parameters of the molecular dynamics simulations

Parameter	Setting
Number of particles	2048
LJ cut-off radius	5σ
Timestep	0.002
Thermostat damping constant	0.2
Barostat damping constant	2.0
Number of equilibration timesteps	10^6
Number of production timesteps	40^6
Number of timesteps over which average properties were calculated	1000

The simulations were performed using the Large-scale Atomic/Molecular Massively Parallel Simulator code (LAMMPS)(32)

The simulations were first run in the canonical (NVT) ensemble where the internal energy (U^*), isochoric heat capacity (C_v^*), and thermal pressure coefficient (γ_v^*) were calculated using equations [3]-[5] respectively (33).

$$U^* = \langle \mathcal{H} \rangle \quad [3]$$

$$C_v^* = \frac{\langle [\mathcal{E} - \langle \mathcal{E} \rangle]^2 \rangle_{NVT}}{k_b * T^2} \quad [4]$$

$$\gamma_v^* = \frac{\langle [\mathcal{V} - \langle \mathcal{V} \rangle] * [\mathcal{P} - \langle \mathcal{P} \rangle] \rangle_{NVT}}{k_b * T^2} + \rho * k_b \quad [5]$$

Where $\langle X \rangle$ is the mean of property X and $\langle X \rangle_{NVT}$ indicates that the mean of property X was calculated from the results of the NVT ensemble. \mathcal{H} is the Hamiltonian, \mathcal{E} is the instantaneous total energy, T is the mean temperature, k_b is the Boltzmann constant, \mathcal{V} is the instantaneous potential energy, \mathcal{P} is the instantaneous pressure, and ρ is the mean density. The units' style was set to LJ (Lennard Jones) in LAMMPS. This sets the reported properties to be unitless (reduced units) as the mass of the particles, potential well depth, characteristic length scale, and Boltzmann constant are set to one.

The final pressure from the NVT ensemble was then used as the pressure for the isobaric-isothermal (NPT) ensemble. From the NPT ensemble, the isobaric heat capacity (C_p^*), thermal expansion coefficient (α_p^*), isothermal compressibility (β_T^*), Joule–Thomson coefficient (μ_{JT}^*), and compressibility factor (Z^*) were calculated using equations [6]–[10].

$$C_p^* = \frac{\langle \left[\mathcal{E} + \frac{P}{\rho} - \langle \mathcal{E} + \frac{P}{\rho} \rangle \right]^2 \rangle}{k_B * T^2} \quad [6]$$

$$\alpha_p^* = \frac{\langle \left(\mathcal{E} + \frac{P}{\rho} - \langle \mathcal{E} + \frac{P}{\rho} \rangle \right) * (v - \langle v \rangle) \rangle}{\langle v \rangle * T^2} \quad [7]$$

$$\beta_T^* = \frac{\langle [v - \langle v \rangle]^2 \rangle}{\langle v \rangle * T} \quad [8]$$

$$\mu_{JT}^* = \frac{1}{\rho * C_p^*} * (T * \alpha_p^* - 1) \quad [9]$$

$$Z^* = \frac{P}{\rho * T} \quad [10]$$

Where \mathcal{E} is the instantaneous total energy, P is the mean pressure, ρ is the mean density, T is the mean temperature, k_b is the Boltzmann constant, v is the instantaneous volume.

Post-Processing of Simulation Results

Due to unstable NPT ensembles, some simulations produced partial results. Thus, the results of these simulations were discarded.

To remove any data points where the fluid vaporised when changing from the NVT to NPT ensemble, convergence criteria that the mean density and mean temperature of both ensembles must be within $1 * 10^{-3}$ was imposed.

To remove any datapoints where the properties reported from the simulation were not converged, criteria that the absolute difference in relative

standard deviation for every property between the first half and the second half of the production timesteps must be less than 0.05 was imposed.

To remove any non-converged derivative properties for a given datapoint, the properties were calculated over the period shown in equation [11]. Criteria that the relative standard deviation for each calculated property over this period must be less than 0.05 was imposed. However, if the computed property was not converged, the property was replaced with a “NaN”, which was then caught in the loss function of the ANN. This maximised the amount of information which was obtained from each simulation as properties which were converged were retained while properties which weren't converged were rejected.

$$n \in (0,100), \quad \frac{2}{3}t_{max} + \frac{n}{100} * \frac{1}{3}t_{max} \quad [11]$$

Where t_{max} is the maximum timestep in the simulation.

To remove outliers in the derivative properties from the dataset, LocalOutlierFactor from Scikit-Learn was used (34). Outliers were replaced with “NaN” which were caught in the loss function of the ANN. Simulations which produced negative pressures were also discarded.

Artificial Neural Network

As established in previous literature, the ANN was used to predict the residual Helmholtz free energy. Training on the total Helmholtz free energy is difficult for the ANN as the ideal contribution diverges around the zero-density limit (31). In addition, analytical expressions exist for the ideal contribution to the Helmholtz free energy and its derivative properties, thus they need not be learned by the ANN. The residual Helmholtz free energy was obtained using equation [12].

$$A_{res}^* = ANN(\rho^*, T^*) - ANN(\rho^* = 0, T^*) \quad [12]$$

It is important to note that both ANN terms in this equation are the same ANN, evaluated at different conditions.

Once the residual Helmholtz free energy was obtained, automatic differentiation was used to obtain the derivatives and hessian with respect to reduced temperature and reduced density. These derivatives were then used to calculate the derivative properties using equation [9] and equations [13]–[24].

$$S_{res}^* = - \left(\frac{\partial A_{res}^*}{\partial T^*} \right)_{\rho^*} \quad [13]$$

$$U^* = A_{res}^* + T^* S_{res}^* + [T^*] \quad [14]$$

$$\frac{P^*}{\rho^*} = \rho^* \left(\frac{\partial A_{res}^*}{\partial \rho} \right)_{T^*} + [T^*] \quad [15]$$

$$Z^* = \frac{P^*}{\rho^*} \frac{1}{T^*} \quad [16]$$

$$C_v^* = -T^* \left(\frac{\partial^2 A_{res}^*}{\partial T^{*2}} \right)_{\rho^*} \quad [17]$$

$$\left(\frac{\partial P^*}{\partial T^*} \right)_{\rho^*} = \rho^{*2} \left(\frac{\partial^2 A_{res}^*}{\partial T^* \partial \rho^*} \right) + [\rho^*] \quad [18]$$

$$\left(\frac{\partial P^*}{\partial \rho^*} \right)_{T^*} = 2\rho^* \left(\frac{\partial A_{res}^*}{\partial \rho^*} \right)_{T^*} + \rho^{*2} \left(\frac{\partial^2 A_{res}^*}{\partial \rho^{*2}} \right)_{T^*} + [T] \quad [19]$$

$$\alpha_p^* = \frac{\left(\frac{\partial P^*}{\partial T^*} \right)_{\rho^*}}{\rho^* \left(\frac{\partial P^*}{\partial \rho^*} \right)_{T^*}} \quad [20]$$

$$\beta_T^* = \frac{\left(\frac{\partial P^*}{\partial \rho^*} \right)_{T^*}^{-1}}{\rho^*} \quad [21]$$

$$\gamma_v^* = \frac{\alpha_p^*}{\beta_T^*} \quad [22]$$

$$C_p^* = C_v^* + \left(\frac{T^* \alpha_p^{*2}}{\rho^* \beta_T^*} \right) \quad [23]$$

$$\gamma_i^* = \frac{C_p^*}{C_v^*} \quad [24]$$

[...] indicates that this portion of the equation is the ideal contribution to the property at $k_B = 1$.

At low density, the values of β_T^* increased 3-4 orders of magnitude relative to high-density regions. C_p^* also increased by one order of magnitude from low-temperature to high-temperature regions. Using $\rho^* \beta_T^*$ as a target variable instead of β_T^* , and using γ_i^* as a target variable instead of C_p^* , as reported in previous literature, resolved these issues (31). This method acted as batch normalisation and decreased the time to convergence (35).

For each mini-batch of data, several derivative properties were replaced by "NaN" in the pre-processing stage. For each derivative property, if the target value at that density and temperature contained a "NaN", the predicted and target value for that density and temperature were discarded for that property. The individual loss for each property is calculated using equation [25], and the total loss is calculated from the loss of each property using equation [26].

N_x is the length of the tensor containing only non-"NaN" values of property X , $X_{i_{predicted}}$ is the

predicted value of property X , $X_{i_{target}}$ is the target value of property X . σ_X^2 is the variance of property X in the mini-batch w_X is the weight of property X which was set at 1 for $X \in (Z^*, U^*)$ and set to 1/20 when X was any other property. This allows the ANN to accurately predict the first derivatives of the Helmholtz free energy, which are used to recover the chemical potential and pressure. These are then used to predict phase equilibria. However, information from the second derivatives improves the prediction of the first derivatives. Thus, they are added but at a lower weight so to not detrimentally affect the prediction of the first derivatives (31).

$$Loss_X = \frac{1}{N_X} \frac{w_X}{\sigma_X^2} \sum_{i=1}^{N_X} (X_{i_{predicted}} - X_{i_{target}})^2 \quad [25]$$

$$Loss_{Total} = Loss_{Z^*} + Loss_{U^*} + Loss_{C_p^*} + Loss_{\alpha_p^*} + Loss_{\rho^* \beta_T^*} + Loss_{\gamma_v^*} + Loss_{\gamma_i^*} + Loss_{\mu_{JT}} \quad [26]$$

The model was implemented using PyTorch-Lightning and Ray Tune (36,37). The model used one input layer, four hidden layers and one output layer with 45 neurons each. Tanh activation functions were used for all layers except the output layer. This architecture was previously applied to the Mie fluid (31). As the Lennard-Jones potential is a special case of the Mie potential, this model should be more than sufficient to model the 2D Lennard-Jones fluid.

The database was split into 70% training data and 30% validation data using train-test-split from Scikit-Learn (34). The training data was used to calculate the training loss. The validation data was used to calculate the validation loss, which was used to check if the model was overfitting and provided a way to assess the model on data on which it was not trained on.

The final model was trained with a maximum of 200,000 epochs, along with early stopping if the validation loss did not decrease for 500 epochs. The model also terminated if the wall time exceeded 24 hours. During training, the AdamW optimiser was used with a learning rate of 10^{-6} , batch size of 256, and weight decay coefficient of 10^{-2} . AdamW incorporates decoupled weight decay as a form of regularisation into the standard Adam optimiser (38). A batch size of 256 was used as this decreased the probability of a batch containing only NaNs for a property. A learning rate of 10^{-6} is used as the

derivative properties vary significantly when the weights are updated. Models ran with a high learning rate tended to cause the loss to diverge.

Results and Discussion

2D-LJ Fluid Properties Database

Of the 13870 datapoints at which the MD simulations were ran, after postprocessing of the simulation results, 9116 datapoints remained. The distribution of the data points in the database is shown in **Figure 4**.

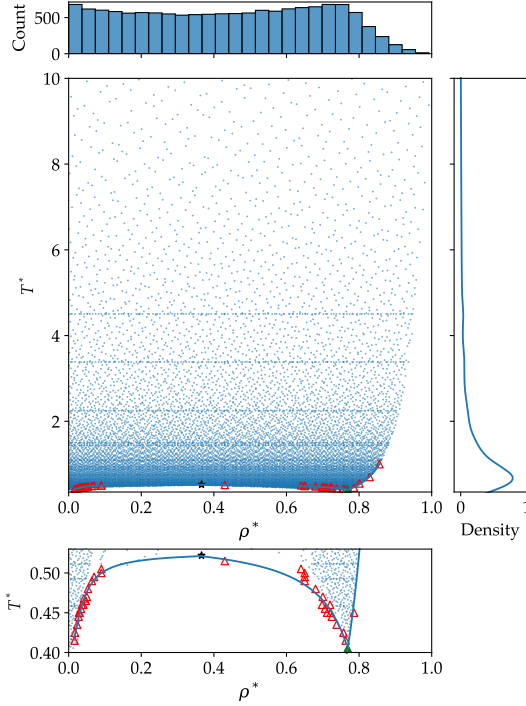


Figure 4: Datapoint distribution in the database after postprocessing. Datapoints shown in blue dots. X-axis plot shows distribution of data points with respect to density. Y-axis plot shows kernel density estimate for the distribution of data points with respect to temperature. Lower subplot shows subcritical region and phase envelope from **Figure 2**.

When comparing the distributions of desired data points to the data points after postprocessing **Figure 3** and **Figure 4** respectively, the distribution of the data points is similar in the supercritical region. The inverse relationship between temperature and number of data points is retained. However, when comparing the area around and inside the VLE region, the postprocessed distribution contains only 14 of the original 948 from the desired data points. The results at 44% at these data points were removed due to failed NPT simulations from unstable ensembles causing unphysical results thus, LAMMPS terminates the simulation. The other 56% of the data points inside the VLE region were removed due to the convergence tolerance that the difference in relative standard deviation of the properties between the

first and second half of the simulations must be less than 0.05. This tolerance was too restrictive, however, we believe that it was best to have fewer total data points but higher quality data points.

The count of non-NaN properties for the 9116 data points in the database are shown in **Table 3**.

Table 3: Count of properties in the property database used for training the EoS.

Property	Count
Internal energy	9116
Compressibility factor	9101
Isochoric heat capacity	9070
Isobaric heat capacity	9070
Thermal pressure coefficient	9092
Thermal expansion coefficient	9093
Isothermal compressibility	9051
Joule–Thomson coefficient	7525

EoS performance

The final model was trained for 32361 epochs and was terminated after being trained for 24 hours due to time constraints.

The performance of the EoS on the 1st derivative properties was assessed by plotting heatmaps of the mean squared error (MSE) between the predicted and target values of internal energy and compressibility factor for the validation dataset in **Figure 6**.

Greater accuracy was seen in the lower temperature region ($T^* < 2$) and higher error was seen in the higher temperature region ($T^* > 2$). This was expected as the lower temperature region had significantly more data due to the inverse scaling; thus the Helmholtz free energy surface was better defined in this region. Both plots showed lower error in the low-density region. This was expected as the fluid behaves as an approximately ideal gas in the low-density region, and the formulation of the model accounts for the ideal contribution to the properties analytically. The highest error in the validation set for both properties was seen in the high-density and high-temperature region.

The ability of the EoS to predict pressure isotherms was assessed by plotting two sub-critical ($T^* = 0.450$ & 0.511), one critical ($T^* = 0.522$) and two supercritical ($T^* = 2.247$ & 4.504) isotherms as shown in **Figure 5**. As discussed in relation to the heatmaps of the error in the first derivative properties, the EoS' predictions were better in the low-density region. This was best shown in the black and magenta lines, which intersected the points at the low-density region, then deviated at the higher-density region.

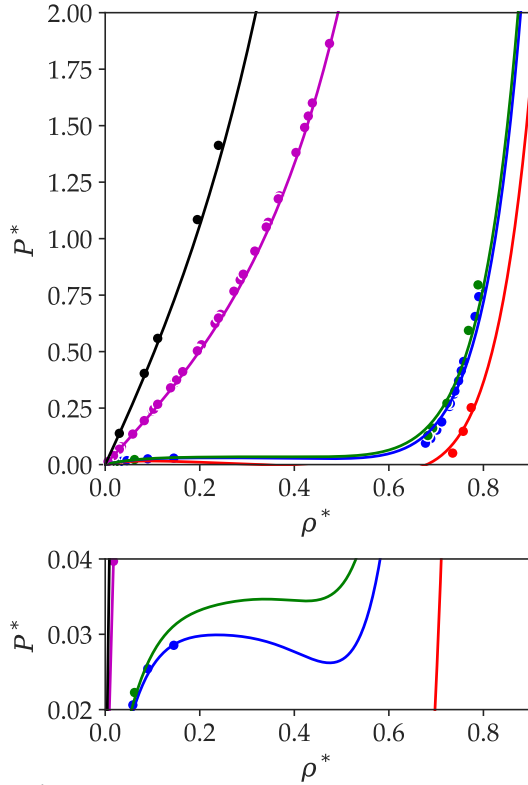


Figure 1: Pressure isotherms. Circles are MD data, lines are EoS prediction. Colours indicate temperature of the isotherm; red: $T^* = 0.450$, blue: $T^* = 0.511$, green: $T^* = 0.522$, magenta: $T^* = 2.247$, black: $T^* = 4.504$. Lower subplot shows region around the critical point.

The EoS was able to predict the existence of van der Waals loops in the subcritical region and a turning point in the critical isotherm, as shown in the lower subplot of **Figure 5**. This was impressive as the database only contained 14 data points in the VLE region, of which none were on the vapour side. This ability to predict the existence of van der Waals loops was understood to be a result of the ANN being trained on the second derivative properties, which encouraged the ANN to predict the curvature of the Helmholtz free energy surface

correctly and thus encourages accurate property predictions in regions which the model wasn't trained.

The predicted turning point at the critical isotherm implied that the model could accurately predict the critical point, and the predicted van der Waals loops implied that the model could predict the VLE envelope. The predicted VLE envelope, VLE envelopes from other 2D-LJ EoS, and predicted critical point from literature are shown in **Figure 7**.

It was seen from **Figure 7** that the EoS from this paper correctly predicts the shape of the VLE envelope. The predicted envelope is in better agreement with the simulated VLE data from literature in the vapour region, with a mean absolute percentage error (MAPE) of 4%, than in the liquid region, with a MAPE of 56%, this was attributed to the fact that all the data inside the VLE envelope in the database was in the liquid region. Therefore, the Helmholtz free energy surface was better defined in the liquid region than the vapour region.

It was also seen from **Figure 7** that the predicted critical point from the EoS from this paper is in good agreement with the critical point from literature with an absolute percentage error of 0.3%.

The predictions of the VLE envelope of other EoS for the 2D-LJ fluid were also shown in **Figure 7**. All the EoS' had good agreement with the simulated data in the low density and low-temperature region ($\rho^* < 0.05, T^* < 0.46$). In the high-temperature region ($T^* > 0.46$) both EoS' from the literature overpredict the liquid and vapour densities, while the EoS from this paper underpredicts both the liquid and vapour density. It was understood that the liquid region was underpredicted to a greater degree due to the lack of data points in the database in this region.

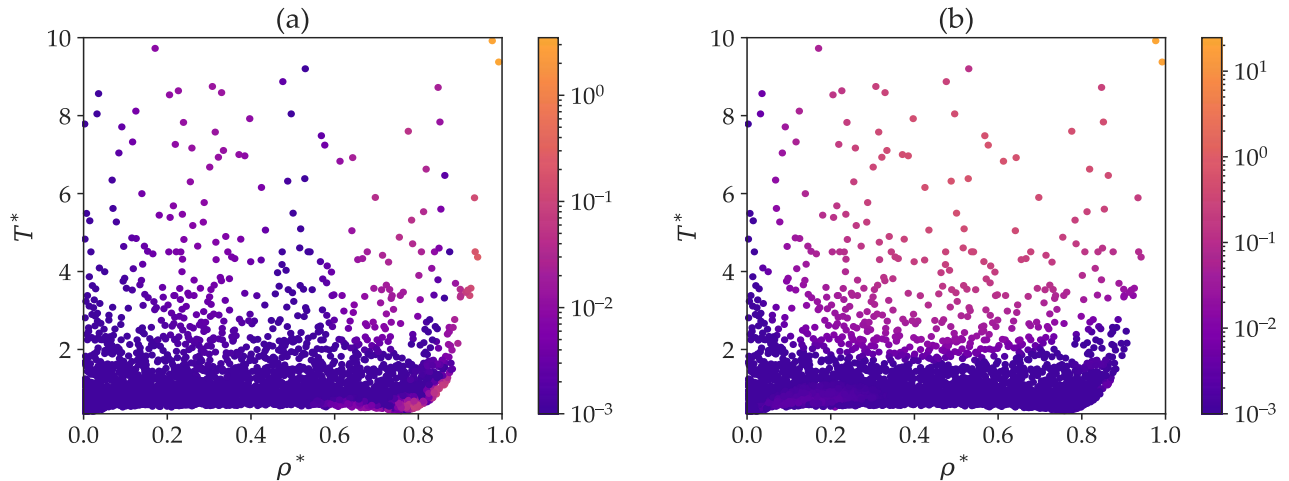


Figure 2: Heatmaps of the MSE of the validation dataset. More orange coloured datapoints are worse predictions, more purple coloured datapoints are better predictions. (a) shows the MSE of the compressibility factor. (b) shows the MSE of the internal energy.

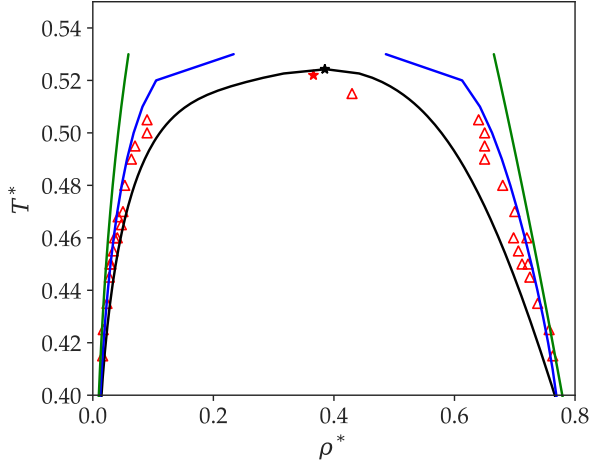


Figure 7: VLE envelope. Black line is predicted VLE envelope from the EoS from this paper. Black star is the predicted critical point from the EoS from this paper. Red star is the critical point from literature.(27) Red triangles are simulated VLE data from literature.(27,28) Blue line shows the RO EoS data from literature.(39) Green line shows CMO EoS data from literature.(39)

Isotherms for the second derivative properties were plotted, these are shown in **Figure 8**. The EoS accurately predicted the thermal expansion coefficient and the isothermal compressibility over the range of densities and temperatures. The

predictions of the thermal pressure coefficient agreed with the MD data until $\rho^* = 0.5$ where the EoS' deviation increased with increasing density however, the EoS captured the correct shape of the isotherms. When predicting the isobaric heat capacity, the EoS showed the same trend, accurate predictions in the low-density region and increasing deviations as density increased. The predictions of the Joule-Thomson coefficient agreed with the MD data in the supercritical and subcritical region however the EoS deviated from the MD data around the critical point. The EoS predicts the isochoric heat capacity poorly at all temperatures above $\rho^* = 0.1$. The turning point at high densities in the subcritical region is not present in the model.

These poor performing properties have not been fitted adequately by the model. This is believed to have occurred due to the low weighting of these properties in the loss function combined with the termination of the training before the loss had stopped decreasing. It is likely that the model may have been able to predict these properties better, given a longer training period and higher weighting in the loss function.

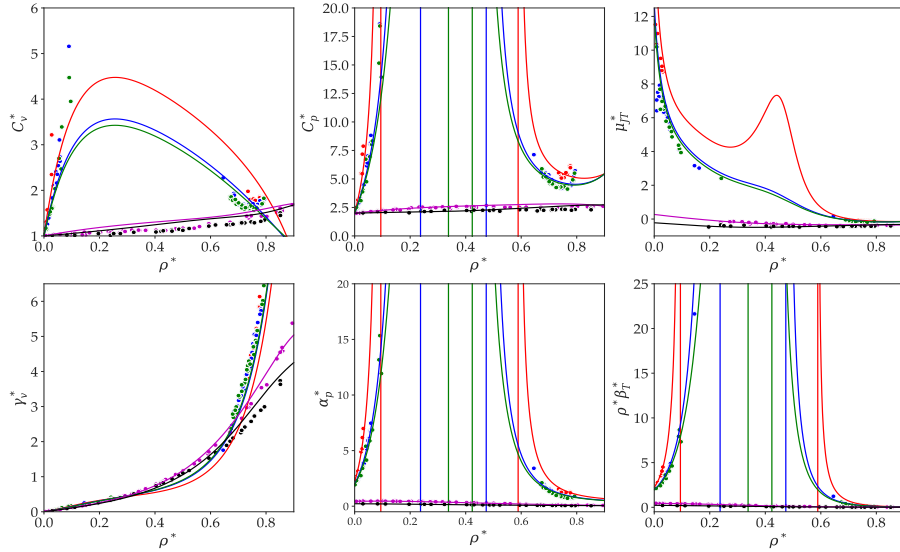


Figure 3: Isotherms of isochoric heat capacity, isobaric heat capacity, Joule-Thomson coefficient, thermal pressure coefficient, thermal expansion coefficient, and isothermal compressibility Circles are MD data, lines are EoS prediction. Colours indicate the temperature of the isotherm; Red: $T^* = 0.450$, Blue: $T^* = 0.511$, Green: $T^* = 0.522$, Magenta: $T^* = 2.247$, Black: $T^* = 4.504$.

Conclusions

This paper implements MD simulations and an ANN to produce an EoS for the 2D-LJ fluid. This approach was previously used in Chaparro's study on the Mie fluid (40). Given the 2D LJ fluid is a specific case of the Mie fluid, this paper investigated the use of this for a 2D-LJ EoS.

This ANN was trained on both the first and second derivatives of the Helmholtz free energy to learn the Helmholtz free energy surface. Training on second derivative properties improved the ability of the EoS to extrapolate the first derivative properties. This was best seen in the ability of the model to predict the existence of van der Waals loops in the subcritical region despite being trained

on only 14 data points in the VLE region. This model accounts for the ideal gas limit analytically, thus has great performance at low densities. The performance of the EoS was good for first derivative property prediction and phase behaviour, as seen in **Figure 7**. For phase prediction, this EoS outperforms existing EoS in the critical region, performs as well in the vapour branch with a MAPE of 4% and performs poorly in the liquid branch with a MAPE of 56%.

Shortcomings of our methodology were identified, such as overcleaning of the data, leading to a lack of data points in the VLE region and weaker performance in the VLE region. Future work should consider the effect of loosening these convergence tolerances. This could result in more subcritical points to aid the prediction of the phase envelope.

To improve property prediction around the VLE envelope, an additional loss could be implemented where the current predicted VLE envelope is compared to the VLE data. This would need to be implemented such that this training only occurs after the model can predict any phase behaviour. Improvements could also be had if hyperparameter optimisation was performed. For future research, developing upon these areas and testing this framework on other fluids is encouraged.

Acknowledgements

We thank Gustavo Chaparro for his advice and kindness in aiding this study. His foundation from previous research and expertise aided our process of discovery.

References

1. Tenny KM, Cooper JS. Ideal Gas Behavior. 2023.
2. Kontogeorgis GM, Privat R, Jaubert JN. Taking Another Look at the van der Waals Equation of State—Almost 150 Years Later. *J Chem Eng Data*. 2019 Nov 14;64(11):4619–37.
3. Chapman WG, Gubbins KE, Jackson G, Radosz M. SAFT: Equation-of-state solution model for associating fluids. *Fluid Phase Equilib*. 1989 Dec;52:31–8.
4. Nezbeda I. On Molecular-Based Equations of State: Perturbation Theories, Simple Models, and SAFT Modeling. *Front Phys*. 2020 Sep 29;8.
5. Redlich Otto, Kwong JNS. On the Thermodynamics of Solutions. V. An Equation of State. Fugacities of Gaseous Solutions. *Chem Rev*. 1949 Feb 1;44(1):233–44.
6. Papaioannou V, Lafitte T, Avendaño C, Adjiman CS, Jackson G, Müller EA, et al. Group contribution methodology based on the statistical associating fluid theory for heteronuclear molecules formed from Mie segments. *J Chem Phys*. 2014 Feb 7;140(5).
7. Thompson AP, Aktulga HM, Berger R, Bolintineanu DS, Brown WM, Crozier PS, et al. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comput Phys Commun*. 2022 Feb;271:108171.
8. Kontogeorgis GM, Folas GK. *Thermodynamic Models for Industrial Applications*. Wiley; 2010. 67–67 p.
9. Coveney P V., Wan S. On the calculation of equilibrium thermodynamic properties from molecular dynamics. *Physical Chemistry Chemical Physics*. 2016;18(44):30236–40.
10. Swesty FD. Thermodynamically Consistent Interpolation for Equation of State Tables. *J Comput Phys*. 1996 Aug;127(1):118–27.
11. Zhu K, Müller EA. Generating a Machine-Learned Equation of State for Fluid Properties. *J Phys Chem B*. 2020 Oct 1;124(39):8628–39.
12. Chaparro G, Müller EA. Development of thermodynamically consistent machine-learning equations of state: Application to the Mie fluid. *J Chem Phys*. 2023 May 14;158(18).
13. Tsiok EN, Fomin YD, Gaiduk EA, Tareyeva EE, Ryzhov VN, Libet PA, et al. The role of attraction in the phase diagrams and melting scenarios of generalized 2D Lennard-Jones systems. *J Chem Phys*. 2022 Mar 21;156(11).
14. Howes AJ, Radke CJ. Monte Carlo Simulations of Lennard-Jones Nonionic Surfactant Adsorption at the Liquid/Vapor Interface. *Langmuir*. 2007 Feb 1;23(4):1835–44.
15. Fernandez-Nieves A, Puertas AM, editors. *Fluids, Colloids and Soft Materials: An Introduction to Soft Matter Physics*. Hoboken, NJ, USA: John Wiley & Sons, Inc; 2016.
16. Glandt ED, Fitts DD. Percus–Yevick equation of state for the two-dimensional Lennard-Jones fluid. *J Chem Phys*. 1977 May 15;66(10):4503–8.
17. Reddy MR, O'Shea SF. The equation of state of the two-dimensional Lennard–Jones fluid. *Can J Phys*. 1986 Jun 1;64(6):677–84.
18. Mulero A, Cuadros F, Faúndez CA. Vapour - Liquid Equilibrium Properties for Two- and Three-dimensional Lennard-Jones Fluids from Equations of State. *Australian Journal of Physics*. 1999;52(1):101.
19. Khordad R. Inhomogeneous 2D Lennard–Jones Fluid: Theory and Computer Simulation. *Commun Theor Phys*. 2012 Nov;58(5):759–64.

20. Sivajothi SSP, Lin ST, Maiti PK. Efficient Computation of Entropy and Other Thermodynamic Properties for Two-Dimensional Systems Using Two-Phase Thermodynamic Model. *J Phys Chem B* [Internet]. 2019 Jan;123(1):180–93. Available from: <http://dx.doi.org/10.1021/acs.jpcc.8b07147>
21. Rosenberger D, Barros K, Germann TC, Lubbers N. Machine learning of consistent thermodynamic models using automatic differentiation. *Phys Rev E*. 2022 Apr 1;105(4):045301.
22. Mulero A, Cuadros F, Faúndez CA. Vapour - Liquid Equilibrium Properties for Two- and Three-dimensional Lennard-Jones Fluids from Equations of State. *Australian Journal of Physics*. 1999;52(1):101.
23. Abdolrasol MGM, Hussain SMS, Ustun TS, Sarker MR, Hannan MA, Mohamed R, et al. Artificial Neural Networks Based Optimization Techniques: A Review. *Electronics (Basel)*. 2021 Nov 3;10(21):2689.
24. Hornik K, Stinchcombe M, White H. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Networks*. 1990 Jan;3(5):551–60.
25. Wallach D, Goffinet B. Mean squared error of prediction as a criterion for evaluating and comparing system models. *Ecol Modell*. 1989 Jan;44(3–4):299–306.
26. Sivajothi SSP, Lin ST, Maiti PK. Efficient Computation of Entropy and Other Thermodynamic Properties for Two-Dimensional Systems Using Two-Phase Thermodynamic Model. *J Phys Chem B* [Internet]. 2019 Jan;123(1):180–93. Available from: <http://dx.doi.org/10.1021/acs.jpcc.8b07147>
27. Smit B, Frenkel D. Vapor–liquid equilibria of the two-dimensional Lennard-Jones fluid(s). *J Chem Phys*. 1991 Apr 15;94(8):5663–8.
28. Singh RR, Pitzer KS, de Pablo JJ, Prausnitz JM. Monte Carlo simulation of phase equilibria for the two-dimensional Lennard-Jones fluid in the Gibbs ensemble. *J Chem Phys*. 1990 May 1;92(9):5463–6.
29. Barker JA, Henderson D, Abraham FF. Phase diagram of the two-dimensional Lennard-Jones system; Evidence for first-order transitions. *Physica A: Statistical Mechanics and its Applications*. 1981 Mar;106(1–2):226–38.
30. Saltelli A, Annoni P, Azzini I, Campolongo F, Ratto M, Tarantola S. Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Comput Phys Commun*. 2010 Feb;181(2):259–70.
31. Chaparro G, Müller EA. Development of thermodynamically consistent machine-learning equations of state: Application to the Mie fluid. *J Chem Phys*. 2023 May 14;158(18).
32. Thompson AP, Aktulga HM, Berger R, Bolintineanu DS, Brown WM, Crozier PS, et al. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comput Phys Commun*. 2022 Feb;271:108171.
33. Allen MP, Tildesley DJ. Statistical mechanics. In: *Computer Simulation of Liquids*. Oxford University Press Oxford; 2017. p. 46–94.
34. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* [Internet]. 2011;12(85):2825–30. Available from: <http://jmlr.org/papers/v12/pedregosa11a.html>
35. Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In: Bach F, Blei D, editors. *Proceedings of the 32nd International Conference on Machine Learning* [Internet]. Lille, France: PMLR; 2015. p. 448–56. (Proceedings of Machine Learning Research; vol. 37). Available from: <https://proceedings.mlr.press/v37/ioffe15.html>
36. Liaw R, Liang E, Nishihara Robert and Moritz P, Gonzalez JE, Stoica I. Tune: A Research Platform for Distributed Model Selection and Training. *arXiv preprint arXiv:180705118*. 2018;
37. Falcon W, The PyTorch Lightning team. PyTorch Lightning [Internet]. zenodo; 2023 [cited 2023 Dec 7]. Available from: <https://zenodo.org/records/10139277>
38. Loshchilov I, Hutter F. Decoupled Weight Decay Regularization. 2017 Nov 14; Available from: <http://arxiv.org/abs/1711.05101>
39. Mulero A, Cuadros F, Faúndez CA. Vapour-Liquid Equilibrium Properties for Two-and Three-dimensional Lennard-Jones Fluids from Equations of State. *J Phys*. 1999;52:101–16.
40. Chaparro G, Müller EA. Development of thermodynamically consistent machine-learning equations of state: Application to the Mie fluid. *J Chem Phys*. 2023 May 14;158(18).

Optimising Detergent Formulation: A Bi-Objective Computer-Aided Molecular Design Approach

David Ke and Esha Langi

Department of Chemical Engineering, Imperial College London, U.K.

Abstract Surfactants, essential in the cleaning industry, face increasing demands for higher cleaning efficiency, as well as more stringent environmental and health regulations. The shift towards synthetic derivation has been significant due to the time-consuming nature of empirical solutions, resulting in continuous innovation becoming imperative. This study presents a bi-objective computer aided molecular design approach for the optimisation of detergent formulations, with particular emphasis placed on a binary mixture of non-ionic and anionic surfactants. The methodology addresses the critical aspect of maximising cleaning efficiency whilst complying with evolving consumer preferences for a reduced environmental impact. Alongside maximising cleaning power, the design of surfactants is tailored for a range of specific industries through consideration of associated soils. Trade-offs between properties have been quantified and their effects minimised so as to allow for the design of a more robust final surfactant. It is found that linear alkyl ethoxylate structures are optimal for achieving these objectives, but three other non-ionic surfactant types are also investigated in this paper. This research not only offers a pragmatic solution to the challenges faced by the detergent industry, but it also significantly accelerates the surfactant development process through the integration of advanced computational algorithms. These findings provide a framework for the future development of effective, industry specific, environmentally sustainable cleaning agents.

1 Introduction

Detergents are a type of surfactant which are specifically catered to the cleaning industry. Their primary purpose is to remove dirt, grease and other impurities from substances and often include additional features ranging from fragrance to fabric softeners.

Cleaning agents serve a vital purpose in society, with the two types of surfactants – soaps and detergents – standing above the rest for their widespread usage. Although soaps, which are derived from natural fats and oil, initially held majority of the market following the second world war, by the 1970s their total market share had decreased to under 20% (Myers, 2020). This is largely attributed to the rise of synthetic detergents driven by various economic and cultural changes.

As of 2012, the global market for laundry detergents is estimated to have reached a value of \$60.9 billion (O. Bianchetti et al., 2015) and although trends may point to the surfactant industry slowing down, there are still various factors that ensure companies must quickly and continuously develop new products in order to capture the largest market share (Fung et al., 2007).

For the successful development of a detergent product, maximising cleaning power is naturally essential for a successful formulation, but with the growing consumer conscience paired with increasingly strict government regulations, it is getting increasingly infeasible to meet every demand. Examples include consumer desires for non-toxic and biodegradable products that make use of renewable feedstocks with less energy demanding processes, whilst restrictions enforced by the government involve reducing VOC emissions in the manufacturing process. (Myers, 2020).

A major aspect for developing detergent products is deciding the chemical structure of the surfactant that should be involved. Developing novel molecule structures via empirical means can be especially laborious as experimenting on a large and diverse set of novel components can be time consuming,

especially considering the various combinations of surfactants and additives in multicomponent mixtures.

Not only this, but maximizing the cleaning power of the surfactants, whilst still considering secondary parameters, such as toxicity and biodegradability, can make finding the best surfactant to meet specific consumer needs especially difficult or even impossible without compromise.

In order to provide an alternative, more robust, approach, this paper employs a computer aided molecular design (CAMD) methodology for the creation of novel structures. Multiple properties that are typically associated with their effectiveness are optimised, with a large emphasis on safety and environmental parameters also being considered.

2 Background

2.1 Surfactant Classification

A surfactant, short for ‘surface-active agent’, is a chemical compound which lowers the interfacial tension at the boundary between two immiscible phases, consequently allowing them to interact with a variety of substances (Kruglyakov, 2000). They do this through the creation of micelles, which are self-assembled molecular clusters, in a solution. Surfactant molecules have an amphiphilic structure meaning they consist of both a hydrophobic alkyl chain, which does not show affinity to water, and a functional hydrophilic group, which conversely does have an affinity for water.

Surfactants can then be classified into either ionic or non-ionic, with the former being further subclassified into anionic, cationic, and amphoteric.

Ionic surfactants have high cleaning efficiency, good foaming properties and the electrostatic interactions with oppositely charged particles allow them to effectively bind to and remove charged soils - an example of this would be anions being able to remove positively charged clay particles.

In aqueous solutions, the hydrophilic group in anionic surfactants dissociates into anions, cations in

cationic surfactants and a mixture of anions and cations in amphoteric surfactants, depending on the pH. In general, anionic surfactants are most widely used due to them being both easy and cheap to manufacture.

On the other hand, non-ionic surfactants do not dissociate into ions in aqueous solutions and are rather subclassified depending on the type of their hydrophilic group. The types of non-ionic surfactants that will be examined in this paper include linear alkyl ethoxylates (LAE), branched alkyl ethoxylates (BAE), ethoxylated amides (EA) and carbohydrate-derivative ethoxylates (CDE). This type of surfactant is frequently used as modifying their physical properties is as simple as adjusting the length of their hydrophilic chain. What is more, non-ionic surfactants are highly versatile and much more stable than their ionic counterparts (Chen and Schechter, 2021).

A binary mixture of non-ionic and anionic surfactants is chosen as the non-ionic component is not affected by hard waters, which is ideal in the context in cleaning products. Unlike the non-ionic, the anionic component is very effective at removing soils but does not cope well in hard water due to the multivalent ions present (Cheng et al., 2020). Combination of the two produce a detergent which has an overall performance greater than those of its individual components. With regards to this paper, binary mixtures of the anionic surfactant of choice, sodium dodecyl sulfate (SDS), and each of the four types of non-ionic surfactants will be analysed in relation to a variety of properties.

2.2 Surfactant Properties

When designing a surfactant molecule, an important initial consideration is the quality factors which are most ideal for the product's use case and their corresponding performance indices. Quality factors can be divided into two different types – primary and secondary (Fung et al., 2007). Primary factors represent the minimum requirement for a detergent product. Examples of these in detergents include fast dissolution, powerful cleaning, and general product stability. High performance in both cold temperatures and hard water and versatility against various soils area also key considerations.

Meanwhile, secondary factors provide additional advantages for the consumer, ultimately setting that product above the rest. In the past, environmental and safety impacts such as biodegradability and toxicity might have been considered secondary quality factors, but with growing consumer awareness leading to stricter governmental regulations these could also be argued to be necessary traits when formulating novel detergents.

The key surfactant properties considered in this detergent formulation include critical micelle concentration (CMC), Hansen solubility parameters (HSP), hydrophilic-lipophilic balance (HLB), cloud point (CP) and toxicity parameters (TP).

2.2.1 Critical Micelle Concentration (CMC)

CMC marks the threshold at which the surfactant molecules begin to aggregate into micelles with hydrophobic tails and hydrophilic heads (Esmaeili et al., 2021). These work to remove soils such as dirt and

grease by binding to the interface between the immiscible phases with their hydrophobic tails. The hydrophilic head of the micelle then allows for the whole structure to be soluble in water and subsequently carried away.

Without being able to form micelles, the surfactant is effectively rendered useless- a low CMC is vital. It is worth noting that the CMC also varies with temperature, pressure and the presence and concentration of other substances such as salts.

2.2.2 Hansen Solubility Parameters (HSP)

The investigation of solubility parameters is typically done in the context of solvents and solutes. The 'like dissolves like' rule of thumb (Barton, 2017) states that solutes with similar intermolecular forces to the solvent have a high likelihood of being dissolved in it, which suggests that if the surfactant and the soil have similar HSP values, then it would make for an effective detergent. (Abbott and Hansen, 2013).

The Hildebrand total solubility parameter is a standard way of representing these intermolecular forces, but its inability to distinguish the solubility behaviour for polar and hydrogen-bonded molecules makes the additional partial solubility parameters developed by Hansen a much better approach (Hansen, 2007). These HSP account for all three types of intermolecular forces- dispersion, polar and hydrogen bonding- and thus have considerably higher predictive capabilities than using just the Hildebrand total solubility parameter.

Although the usage of HSP values in the context of surfactant and detergent product design is surprisingly absent, Abbot and Hansen demonstrate a similar process for choosing the best surfactant based on the specific soils (Abbott and Hansen, 2013).

By considering the three HSP values of a substance as its coordinates in a three-dimensional Euclidian space, theoretically minimising the distance between the surfactant and soil points maximises the cleaning power of the detergent. This also allows for the detergent to be tailored to specific use cases and environments, making this approach especially useful.

2.2.3 Hydrophilic-Lipophilic Balance (HLB)

HLB is a measure of the balance between the hydrophilic and lipophilic groups of a surfactant molecule. This is an effective way of quantifying the effectiveness of a surface structure as an emulsifier (Rosen and Kunjappu, 2012). It also allows prediction of both the solution appearance, as well as its product application, both very important when it comes to meeting consumer demands – the ranges for these are shown in Table 1 (Fung et al., 2007).

Table 1: HLB range for product application and solution appearance

Production Application	HLB Range	Solution Appearance	HLB Range
W/O emulsifiers	3-6	Insoluble	1-4
Wetting agents	7-9	Unstable dispersion	4-7
O/W emulsifiers	8-18	Stable dispersion (opaque)	7-9
Detergents	3-15	Hazy solution	10-13
Solubilizing	15-18	Clear Solution	13+

With regards to the design of a detergent product, the target HLB is commonly reached through the mixing of multiple surfactants. This is because mixtures are generally more effective at stabilising the emulsion in comparison to individual surfactants (Fung et al., 2007).

2.2.4 Cloud Point (CP)

The CP of a non-ionic surfactant represents the boundary at which the surfactant begins to form a separate phase as the temperature increases, causing the solution to appear cloudy. Although in some instances this appearance may be preferred over a clear solution, an insoluble surfactant has a lower cleaning efficiency and therefore must be avoided. This is done by ensuring that the CP is sufficiently above the wash temperature during use (Smulders et al., 2007).

2.2.5 Toxicity Parameters (TP)

LC50 focuses on the lethal concentrations in a medium, with FM standing for flowing medium, such as a stream or river, and DM for a diffusing medium, such as still water or soil (Hukkerikar et al., 2012). Meanwhile, LD50 quantifies the dose of substance required to result in mortality in 50% of the test subjects through either inhalation, ingestion, or dermal exposure after a specified time duration. While this is primarily a measure of acute toxicity of a substance to living organisms, it can be also used to indirectly quantify the potential environmental impact of a substance. For example, values obtained from experiments with aquatic organisms quantifies the harm of the substance on ecosystems.

Development of these models allows for reliable estimates of the environmental impact of the surfactants through allowing the health of various ecosystems, as well as the effects on certain species to be monitored. Ensuring non-toxicity is important- the further away the values are from predetermined hazard classification values, the smaller the detrimental effect on the environment.

2.3 Evaluation of an Existing Approach

The article titled ‘Design and performance optimisation of detergent product containing binary mixture of anionic-non-ionic surfactants’ by Cheng et al. was used as the foundation for this research (Cheng et al., 2020). Cheng focuses on the development of a methodology to design a hospital detergent using CAMD tools. This is done with the main objective being the optimisation of CMC, alongside CP, HLB and MW to make the overall detergent better suited for a hospital.

Four types of non-ionic structures – LAE, BAE, EA and CDE – are considered and later optimised. Additionally, the use of binary mixtures of anionic and non-ionic surfactants is explored to enhance their individual properties, with the anionic surfactant of choice being sodium dodecyl sulfate (SDS).

Cheng utilises a methodological approach to detergent design and successfully quantifies the enhancement a binary system of surfactants has on a purely anionic or non-ionic surfactant system through the integration of computational tools such as group

contribution methods. This has been coupled with a multi-objective optimisation approach to account for the trade-off between different surfactant properties, with CMC being the main property of interest. Not only this, but Cheng also acknowledges the significance of additives such as antimicrobial agents and specific enzymes to strengthen the practical relevance of the research for a hospital setting.

However, there is a limited focus on environmental impact – considering the growing emphasis on sustainability, quantifying, and addressing, this would be valuable. Moreover, modelling the non-linear mixing capabilities of the different properties, and including higher order groups would all have resulted in a more robust final solution. Similarly, as stated by Cheng, the final selection of surfactant mixture and product composition could have been validated through experiments to eliminate the possibility of any undesirable interaction effects. In addition to this, an analysis of cost factors would have been an effective method of ensuring the product provided value for money, as well as whether alternative components were worth exploring.

2.4 Problem Statement

This research paper aims to employ a combination of bi-objective optimisation and CAMD techniques to both address the gaps in Cheng’s paper, as well as provide novel work compared to existing approaches in the field. The environmental impact of the surfactant mixture will be quantified through the calculation of the TP. As there are no predictive models for these properties, their values will be considered based on databases of contributions. Similarly, HSP will be incorporated so as to be able to target the surfactants to specific soils – this will allow us to cater the final surfactant to a wide variety of industries, rather than just Cheng’s hospital setting.

In essence, the proposed objectives of this research are as follows:

- Design feasible binary surfactant mixtures with non-ionic and anionic surfactants to meet current, and future, surfactant demands in a more environmentally sustainable manner
- Satisfy the primary objective of minimising the CMC for each of the four non-ionic structure types to maximise cleaning power
- Fulfil the secondary objective of HSP to better quantify industry specific cleaning power
- Meet the design constraints of the remaining properties

3 Methodology

Bi-objective optimisation was performed in GAMS Studio 40 with the intent of improving the binary mixture of non-ionic and anionic surfactants over CMC, HSP and the other relevant properties.

3.1 Surfactant Design

In Cheng’s work, a total of 4 properties were prioritised, where optimizing CMC was the main objective, while HLB, CP and MW were taken as lower-level problems to be optimised independently using constraints. These

constraints are a certain fraction of the optimal values of the lower-level properties.

This method similarly takes into consideration multiple properties, whilst paying particular attention to the implementation of HSP, a property lacking sufficient research in the area of detergent formulation.

Since additional properties were considered in the formulation of this problem, the HLB and CP constraints were relaxed and fixed to a satisfactory value to prevent the problem from becoming infeasible. The reasoning behind each constraint choice is outlined in the subsequent sections.

3.2 Non-ionic Surfactant Design

3.2.1 Modelling Critical Micelle Concentration

CMC was modelled using the Mattei group contribution method shown in Equation 1 (Mattei et al., 2014).

$$-\log(CMC) = \sum_i N_i C_{i,cmc} + \sum_j M_j D_{j,cmc} + \sum_k O_k E_{k,cmc} \quad [1]$$

Where $C_{i,cmc}$ is the contribution of the first-order group, i , that is present N_i times in that structure. This is similarly the case for the range of second-order groups, j , and third-order groups, k , but only the first-order group contribution was calculated in practice.

When data from this group contribution method wasn't available, the missing values were instead calculated using a Quantitative Structure Property Relationship (QSPR) model (Huibers et al., 1996). QSPR is another property prediction model which relies on signature descriptors that represent the unique structural features of individual molecules.

For the same alkyl chain length, non-ionic surfactants tend to have a lower CMC compared to their anionic counterpart, so the upper boundary for the CMC of the non-ionic surfactant was set to be the CMC value of the anionic surfactant SDS (Tadros, 2013).

3.2.2 Modelling Hansen Solubility Parameters

The HSP values of the surfactant and soil can be interpreted as points in Euclidian space and the closer the points are, the more effective the surfactant is at removing the soil. This distance between the two points is denoted by R_a as shown by Equation 2.

$$R_a^2 = 4(\delta_{d1} - \delta_{d2})^2 + (\delta_{p1} - \delta_{p2})^2 + (\delta_{hb1} - \delta_{hb2})^2 \quad [2]$$

The distance term, R_a , may also be used interchangeably with the HSP distance, or HSP R_a for clarity throughout this paper. δ_d , δ_p and δ_{hb} are the HSP for dispersion, polar and hydrogen bonding, and the 1 and 2 denotes soil or surfactant. The 4 in the equation allows for more convenient representation of the solubility data as a sphere that surrounds a point.

The HSP values are predicted using the group contribution method developed by Stefanis and Panayiotou where the model equation and group contribution data change depending on the size of the parameter. Further details can be found in their work (Stefanis and Panayiotou, 2008).

The soil that was focused on was butyl stearate because of its presence across a variety of industries. In addition to this, its HSP similarity to all the other soils made it a good candidate as the "average" soil. The optimisation algorithm was also tested on other soils to take into account a wide variety of industry use cases—the soils used are shown in Table 2.

Table 2: HSP data for a range of common soils ($MPa^{0.5}$)

Soil	δ_d	δ_p	δ_{hb}
ASTM Fuel "A"	14.3	0.0	0.0
Butyl Stearate	12.6	6.3	6.1
Castor Oil	13.6	6.0	10.5
Ethyl Cinnamate	16.0	10.8	7.5
Linseed Oil	13.5	3.5	3.7
Tricresyl Phosphate	15.9	13.9	13.5

An R_a value of $8 MPa^{0.5}$ or less is considered to have a very good cleaning performance where at least 95% of the soil is removed (Hansen, 2007). This breakpoint indicates that above $8 MPa^{0.5}$, the resultant cleanliness is "unacceptable", and so this is initially used as the maximum boundary value for R_a before being independently optimised as the secondary objective.

3.2.3 Modelling Hydrophilic-Lipophilic Balance

HLB takes into account the ratio between the molecular weight of the hydrophilic portion compared to the total molecular weight of the whole molecule. The HLB system was found to have great utility, leading multiple empirical and theoretical methods to be developed (Rosen and Kunjappu, 2012).

Equation 3 can only be utilised under the condition that the HSP are matched between the surfactant and the soil. When R_a is minimised, HSP values between the surfactant and soil are close, and therefore use of Equation 3 is valid for this research.

$$HLB = \frac{20M_H}{M_L + M_H} = \frac{20M_H}{M_W} \quad [3]$$

In this equation, the M_H is the molecular weight of the hydrophilic group and M_L is the molecular weight of the lipophilic or hydrophobic portion of the molecule. It is assumed that the summation of these is equal to the overall molecular weight, M_W .

The bounds for this property are on the mixture rather than the non-ionic surfactant—these are outlined later in the report.

3.2.4 Modelling Cloud Point

The Mattei group contribution method in Equation 4 was applied to model CP (Mattei et al., 2014).

$$CP^2 = \sum_i N_i C_{i,CP} + \sum_j M_j D_{j,CP} + \sum_k O_k E_{k,CP} \quad [4]$$

The nomenclature of this equation directly mirrors that for CMC in Equation 1, with only the first order group contribution data being taken into account once again.

A lower CP constraint of $55^\circ C$ was employed to enable stable performance under different temperature conditions. This value was chosen as it was the target constraint used in the collation of the database of contributions for non-ionic surfactants for the overall group-contribution method (Mattei et al., 2014).

3.2.5 Modelling Toxicity Parameters

The Marerro and Gani group contribution method (Hukkerikar et al., 2012) was used to model each TP as shown in Equation 5.

$$TP = \sum_i N_i C_{i,TP} + w \sum_j M_j D_{j,TP} + z \sum_k O_k E_{k,TP} \quad [5]$$

The nomenclature of this equation once again mirrors that for CMC in Equation 1, with only the first order group contribution data being taken into account – w and z have values of zero. Simultaneous regression was used to calculate these contributions so all predictors can be considered in the model simultaneously, as opposed to stepwise regression which only incorporates the most significant predictors in the regression model.

Lower bounds of 5 mg L^{-1} on the LC50 values and 5 g kg^{-1} on the LD50 values (Choi and Byeon, 2020) were also set on these parameters, with these values categorised as only being potentially harmful. Meanwhile, values below this range would be classified as toxic, and values even further below this range classified as fatal- the higher the value, the less harmful the substance is.

3.3 Anionic Surfactant Choice

Laundry detergents are typically formulated using a mixture of a non-ionic and anionic surfactant in order to increase product performance. With the focus of this project being on the non-ionic surfactant, selecting an anionic surfactant that would work with a large combination of non-ionic surfactants was the most effective approach.

SDS's widespread usage in both research and industry makes it the surfactant of choice with its properties outlined in Table 3 (Bhattarai et al., 2014).

Table 3: Relevant property data for the anionic surfactant

SDS Property Data	
CMC [M]	8.20E-03
HLB [-]	40.0
CP [°C]	100
MW [g mol ⁻¹]	288

Although SDS is known to be an irritant at high concentrations, the toxicity level on humans is drastically lower in practice due to its diluted nature in detergent formulations. Similarly, the release of SDS into the environment as part of a consumer product is typically non-toxic to aquatic life. Furthermore, biodegradability means that SDS does not persist in the environment (Bondi et al., 2015).

3.4 Surfactant Binary Mixture

Equations for the binary mixture of surfactants were essential for constructing the final detergent product. Although ideal mixing was assumed for the majority of properties, CMC was modelled using Equation 6, developed by Rubingh (Rubingh, 1979).

$$\frac{1}{C_{1,2}} = \frac{\alpha}{f_1 C_1} + \frac{1-\alpha}{f_2 C_2} \quad [6]$$

Where C is the CMC value, f is the activity coefficients of the surfactant and α is the ratio of the non-ionic surfactant- 1 and 2 denote the non-ionic and anionic surfactant respectively. Due to lack of data present, the ideal approximation is assumed where $f_1 = f_2 = 1$.

α was modelled as a continuous variable, constrained from 0.6 to 0.9 to give the mixing ratios more flexibility to meet the product constraints (Azzam, 2001). Since the CMC from mixing non-ionic and anionic surfactant counterparts tend to be lower than their pure counterparts (Myers, 2020), the upper constraint for the CMC of the mixture was also set to the anionic surfactant's CMC, similar to the non-ionic surfactant's CMC constraint.

Although the HLB parameters show evidence of non-linearity when being mixed (Myers, 2020) it was assumed that ideal mixing occurred for simplicity. As shown in Table 1, to achieve detergency, a HLB range of 3-15 for the mixture was required. For the appearance of the product to meet consumer requirements, a clear product was required, which corresponds to HLB values above 13, so the final HLB constraints on the surfactant mixture are 13-15.

An overview of the constraints being enforced on just the non-ionic surfactant, as well as those placed on the final mixture is shown in Table 4.

Table 4: Optimisation constraints on relevant properties

Property	Constrained	Lower	Upper
CMC [M]	Both	0	8.20E-03
Ra [MPa ^{0.5}]	Non-ionic	0	8
HLB [-]	Mixture	13	15
CP [°C]	Non-ionic	55	-
LC50(FM) [mg L ⁻¹]	Non-ionic	5	-
LC50(DM) [mg L ⁻¹]	Non-ionic	5	-
LD50 [g kg ⁻¹]	Non-ionic	5	-

3.5 Bi-Objective Optimisation

It must be recognized that for any bi-objective optimisation problem, the primary objective is influenced by the lower-level objectives. Minimising CMC acts as the primary objective and minimising HSP R_a distance the secondary objective – this must be independently optimised along with the further design constraints.

The secondary objective optimisation approach involved multiplying a weighting, w_{HSP} , against the theoretically ideal HSP distance of $8 \text{ MPa}^{0.5}$.

The optimisation was performed with weightings in the range 0.8125 – 1.125 which accounts for an R_a distance range of 6.5 – 9 $\text{MPa}^{0.5}$.

4 Results & Discussion

Table 5: Initial optimisation results with integer cuts implemented

Structure Type	MW [g mol ⁻¹]	CMC [M]	HSP R_a [MPa ^{0.5}]	HLB [-]	CP [K]	LC50(FM) [mg L ⁻¹]	LC50(DM) [mg L ⁻¹]	LD50 [g kg ⁻¹]
LAE	421	3.04E-07	6.44	13.0	254	1880	860	248
	485	4.01E-08	6.71	13.0	261	2300	1560	531
	576	7.49E-07	6.65	14.4	308	2530	1360	379
	551	5.53E-06	6.51	14.9	314	1690	755	231
	563	2.04E-06	6.56	14.7	311	1850	737	183
BAE	423	6.70E-03	6.43	14.9	329	450	47.4	322
CDE	324	2.00E-04	7.96	13.0	307	1250	905	109
	313	4.00E-04	7.90	13.0	313	1070	688	56.1
	286	2.00E-04	7.95	13.0	288	1280	1040	161
	353	4.00E-04	7.97	13.5	327	1100	809	98.3
	275	4.00E-04	7.87	13.0	293	933	605	47.8

4.1 Initial Optimisation

Table 5 represents the results obtained after an initial optimisation of the mixed-integer programming (MIP) problem with the addition of integer cuts. These were implemented so as to obtain the top 5 surfactants for each structure type, but this was only fully successful for LAE and CDE. This optimisation didn't allow for any EA solutions, but it was later concluded that this was due to the tight constraint of the HSP R_a . Similarly, the LAE structures all display HSP R_a values near the upper constraint value of 8 MPa^{0.5}. This implies the potential existence of surfactants with lower CMC values, and thus higher cleaning power, that have narrowly been missed due to the strict HSP R_a constraint.

In the case of BAE, there was only one feasible structure. After closer analysis of the code, one potential cause could have been the α value. The α value for BAE is its upper constraint of 0.9 implying that a binary mixture even more heavily weighted by the non-ionic surfactant would have been preferable for meeting the main objectives as well as the design constraints.

However, it is still important to note that the mixture results in lower CMC values as opposed to the individual non-ionic or anionic surfactants, proving that they are still more effective in combination than alone. Additionally, the HLB values of the non-ionic surfactants pre-mixing were much lower than the desired 13-15 range where the anionic surfactant was instrumental in bringing the value of this parameter up to the required level. This is representative of the fact that anionic surfactants, such as SDS, are readily employed as solubilising agents in order to increase the solubility of otherwise poorly soluble compounds in aqueous environments (Perinelli et al., 2020).

As part of an external analysis, these results were also compared on a multi-variable plot with the axis range representing the corresponding constraints to allow examination of multiple variables simultaneously. From this, it was clear to see that the optimal MW values were around 400 g mol⁻¹, HLB around 13 and CP around 300K – these values were all consistent to that of Cheng. However, one area where there was a slight difference was the CMC values. While the remaining structures seemed to result in CMC values within at

most two factors of 10 to that of Cheng, the results were on average a thousand times smaller for the LAE structures. A key reason for this is the literature-backed relaxation of Cheng's constraints that was employed before the addition of the further constraints.

The toxicity parameters values are all much higher than their lower bounds of 5 proving that the environmental impacts of all the surfactants is negligible, as is the case with the pre-determined anionic surfactant, with CDE structures being the least harmful overall. These surfactants are in general considered to be more environmentally friendly in comparison to other surfactants as they are derived from renewable feedstock and thus readily biodegrade into non-toxic compounds (Ortiz et al., 2022).

Another essential takeaway is that attempts to change the relative termination tolerance and scale the final results to be in the range of 0-1 in order to remain within the optimality criteria using BARON were largely unsuccessful. Consequently, ANTIGONE was used, where the issue of the solutions only being local, despite using a global solver likely rise due to overly complex nature of certain property models used.

4.2 HSP Weighting

As found from the initial optimisation, the TP values are all so far away from the design constraint that further relaxation to these properties was deemed unnecessary. On the contrary, the effect of the addition of a weighting to the HSP constraint was further explored in order to observe the effect of changing the maximum R_a value from 8 MPa^{0.5}.

The tightening of this constraint would allow for theoretically better cleaning. However, this simultaneously makes finding solutions less likely, or even infeasible in some cases, due to over constraining. Conversely, relaxation of this constraint would mean a drop in cleaning effect but would allow for more flexibility in the design of novel surfactant structures.

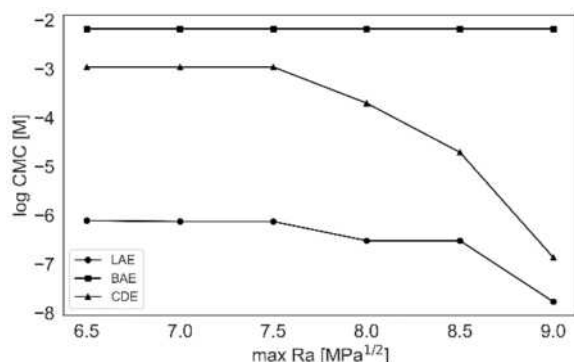


Figure 1: Plot to show effects of changing max Ra value on CMC for each structure type

The graph in Figure 1 was plotted in order to successfully quantify this trade-off. It shows that the

results for BAE remained consistent. This structure being unaffected by ranging the R_a max value between 6.5 and 9 $MPa^{0.5}$ meant that the effects on LAE and CDE were able to be focused on. As seen in the graph, there are much steeper gradients, and thus much larger changes to the overall CMC for R_a max values above 8 $MPa^{0.5}$. This shows that a slight relaxation of this constraint allows for structures with greatly reduced CMC values.

As minimising the CMC is the primary objective, it was decided that only a 10% relaxation of the constraint would be proceeded with to remain close to the theoretical ideal value. This 10% relaxation corresponds to an R_a max value of 8.8 $MPa^{0.5}$ using a weighting (w_{HSP}) of 1.1.

4.3 LAE Bi-Objective Optimisation

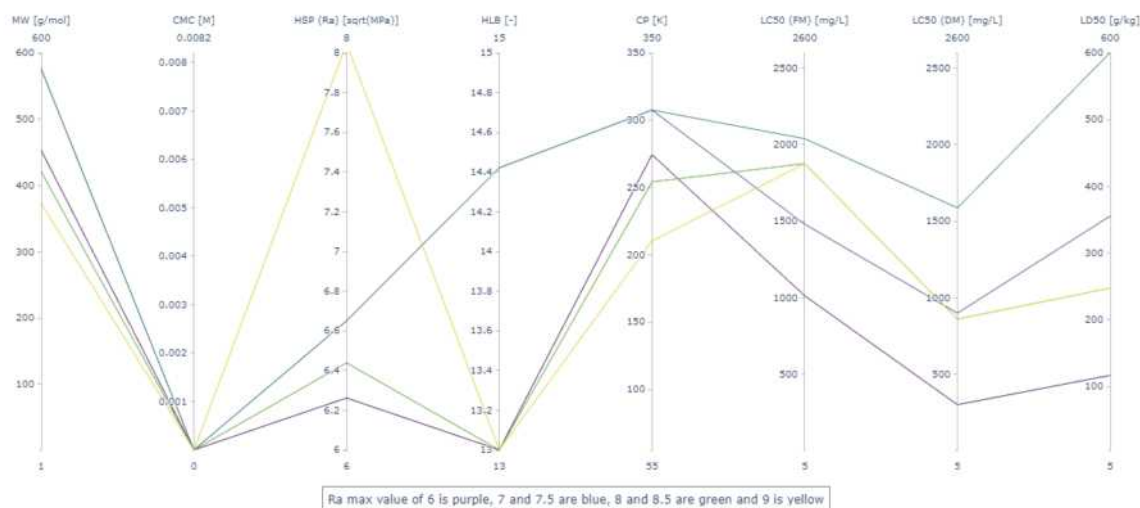


Figure 2: Multi-variable plot of the properties of LAE structures for varying Ra max values

A further analysis was carried out into binary mixtures with LAE structures specifically due to them being the most common non-ionic surfactant used in industry (Evans et al., 1994). Figure 2 is a multi-variable plot which allowed simultaneous examination of multiple variables. The axis range represents the corresponding constraints for all properties except for CMC since the results were far below the upper constraint.

As seen in this figure, the optimal surfactants generated for R_a max values of 7 and 7.5 $MPa^{0.5}$, the blue line, and 8 and 8.5 $MPa^{0.5}$, the green line, were identical – only changes greater than the value of 0.5 $MPa^{0.5}$ from the original 8 $MPa^{0.5}$ had any significant impact. The lines on the graph not aligning suggests different levels of solubility and micelle formation behaviour for each surfactant due to different interactions with their environments.

As minimising CMC is the primary objective, and minimising HSP the secondary, the existence of a potential relationship between the two would be beneficial in accounting for any trade-offs within the optimisation. In theory, a relationship between the two

would be possible as both properties are related to the interactions of the surfactant molecules – with themselves in the case of CMC, and with soils in the case of HSP. It's also important to note that this relationship relies heavily on the intermolecular forces as these are additionally influenced by a variety of other factors including temperature and pH.

Based on the compatibility of surfactants with soils, it would be expected that more optimal CMC solutions could be found for an increased overall solubility parameter radius. In essence, as the surfactant becomes less compatible with the soil, it can form micelles at lower concentrations. One possible cause of this is the stronger hydrophobic effect driving the surfactant molecules to aggregation, rather than soil contact (Kronberg et al., 1995).

In the process of micellisation, there are two opposing forces at work- the hydrophobicity of the tail which favours micelle formation, as well as the repulsion between the hydrophilic head groups. However, the former of these is strong enough to overcome the electrostatic repulsion of the latter,

resulting in the formation of micelles. This demonstrates that the hydrophobic effect plays a significant role in the favouring of surfactant molecule aggregation over direct soil contact.

On the graph, this relationship would be represented by the surfactants with the highest R_a values having the lowest CMC values, which is true for the most part. Swapping of the green and blue lines would mean that all the results were representative of this relationship. The green line having very similar results as the yellow line for the remaining properties further supports that a slight relaxation in the R_a max value does not have too much of an effect on the overall structure, while still allowing for optimal solutions. Subsequently, it is more probable that the error lies within the blue line. A potential reason for a higher-than-expected R_a is that this structure has a considerably higher MW than the others. This would result in a highly different molecular structure, and thus a longer length of hydrophobic tail and greater size and charge of hydrophilic head- these factors significantly influence the surfactants overall performance. Alternatively, this could also be due to the ANTIGONE solver, and further supports the conclusion of local solutions having been obtained.

Following on from that, Figure 3 was plotted after the elimination of confirmed anomalous results. This was done to gain an insight into the actual trade-offs between CMC and HSP and to better understand the boundary of achievable solutions.

The line that would be used to join these points encapsulates the complexity of this relationship, with the optimal CMC solution being obtained from a HSP R_a value of approximately $6.5 \text{ MPa}^{0.5}$. Meanwhile, the

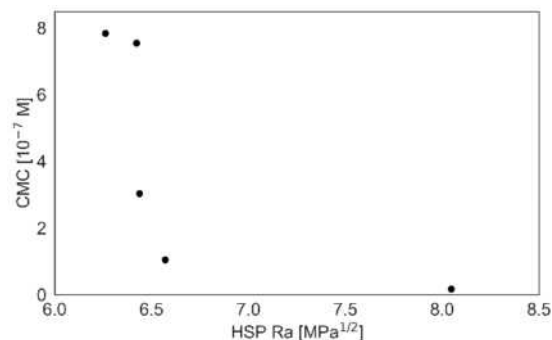


Figure 3: Plot representing the optimal CMC values for ranging R_a max values for LAE surfactants

area on right-hand side of the graph represents the feasible region- the optimised R_a value of $8.8 \text{ MPa}^{0.5}$ lies within this range. As a result, in order to not over-constrain the problem based on only one structure, it was decided to proceed with the previously determined, still feasible, R_a value of $8.8 \text{ MPa}^{0.5}$.

However, since each point represents an individual and optimal solution, interpolation, and extrapolation, through the drawing of a connecting line is not feasible until additional data points are collected. This will also be beneficial in determining whether the solutions obtained are global optimums. In the case they are not, this would further support the fact that ANTIGONE provided local solutions- an initialisation approach of the variables in GAMS would be the first step of checking whether this issue persists with different starting points.

4.4 Final Bi-Objective Optimisation

Table 6: Final optimised solutions for each surfactant type with the implementation of the HSP weighting

Structure Type	Structural Formula	α [-]	MW [g mol^{-1}]	CMC [M]	HSP R_a [$\text{MPa}^{0.5}$]	HLB [-]	CP [K]	LC50 (FM) [mg L^{-1}]	LC50 (DM) [mg L^{-1}]	LD50 [g kg^{-1}]
LAE	$\text{CH}_3 - (\text{CH}_2)_{19} - (\text{OCH}_2\text{CH}_2)_1 - \text{OCH}_2\text{CH}_2\text{OH}$	0.758	373	$1.70\text{E-}08$	8.05	13.0	210	1880	860	248
BAE	$(\text{CH}_2\text{CH}_2\text{CH}_3)_2 - \text{CH} - \text{CH}_2 - \text{O} - (\text{CH}_2\text{CH}_2\text{O})_7 - \text{H}$	0.900	423	$6.70\text{E-}03$	6.43	14.9	329	450	47.4	322
EA	$\text{CH}_3 - (\text{CH}_2)_{17} - \text{CH}_2\text{NH} - \text{CH}_2\text{COO} - (\text{CH}_2\text{CH}_2\text{O})_3 - \text{H}$	0.878	451	$3.97\text{E-}08$	8.25	13.0	227	3220	2150	628
CDE	$\text{CH}_3 - (\text{CH}_2)_{18} - \text{CH}_2\text{COO} - (\text{CH}_2\text{CH}_2\text{O})_2 - \text{CH}_3$	0.814	402	$1.50\text{E-}07$	8.78	13.0	275	3750	3490	718

The final solutions are displayed in Table 6, having been optimised bi-objectively and then subject to further design constraints. Their respective structural formulae are also included, with the red numbers outside of the brackets indicating the change from the general case.

It is essential to note that the culmination of this constrained bi-objective optimisation approach has resulted in achieving solutions for EA- this was not the case after the initial optimisation. This shows that widening the range of effective novel constraints through the relaxation of the HSP R_a constraint was a worthwhile alteration as it resulted in a surfactant with the second-most optimal CMC value of $3.97 \times 10^{-8} \text{ M}$.

Overall, LAE displays the lowest CMC value of $1.7 \times 10^{-8} \text{ M}$ which goes to prove that not only is it the most common, but it is also the best performing with

regards to the primary optimisation. A key contributor to this could be that its linear structure promotes close packing of the molecules in the micelle core, contributing to their overall stability, subsequently resulting in a lower CMC value.

LAE also has the second lowest HSP R_a value of $8.05 \text{ MPa}^{0.5}$, with BAE having the lowest, with a value of $6.43 \text{ MPa}^{0.5}$. However, the BAE structure has a significantly higher CMC value than that of the LAE which is a clear example of the trade-off between these two properties. It is impossible to have the most optimal solution for both of them at once, but the LAE structure results in reasonably optimal values for each property.

4.4.1 Experimental Comparison

The final surfactant solutions are compared to empirical results, with specific attention paid to the analysis of LAE structures due to their availability in literature. According to Cox, physical properties are influenced primarily by ethylene oxide (EO) chain length. Carbon chain length is equally as important for performance- its optimum is shown to depend strongly on surfactant concentration (Cox, 1989).

The optimum LAE solution for this paper is $\text{CH}_3 - (\text{CH}_2)_{19} - (\text{OCH}_2\text{CH}_2)_1 - \text{OCH}_2\text{CH}_2\text{OH}$ with a carbon chain length of 19. This is similar to the carbon chain lengths of the commonly used cetareth-n molecule, which typically range from 16-18 (Śliwa and Śliwa, 2020). The EO chain length is typically very large for these molecules, but the optimum surfactant had a length of only 1. Although not shown in this paper, there were various other, less optimal, solutions generated which better corresponded to empirical surfactant data. While these structures had a lower carbon chain length, they had much larger EO chain lengths- these would be an overall better fit for the cetareth-n molecule.

4.5 Uncertainty Analysis

An analysis of the accuracy for the models is vital for understanding the limits associated with this optimisation process.

If the development of the HSP group contribution data is taken as an example, the potential drawbacks that should be taken into consideration for

any model can be observed. Developing group contribution data requires a set of data in which to base the model off. In Stefanis' case, a both large and diverse set of compounds are considered (Stefanis and Panayiotou, 2008). Utilisation of a broad dataset is great in theory but due to it being inaccessible, it is still unknown as to whether it incorporates the desired types of surfactants.

The average absolute error (AAE) for the first and second groups taken from Stefanis are in Table 7.

Table 7: Absolute error for first and second order groups for HSP

HSP	AAE [$\text{MPa}^{0.5}$]		% Change
	First-Order	Second-Order	
δ_d	0.44	0.41	-6.80
δ_p	1.05	0.86	-18.10
δ_{hb}	0.88	0.80	-9.10

Although the percentage change appears to have quite an effect when using second order groups, the absolute change for a structure is ultimately small, agreeing with the choice to eliminate higher order groups.

Furthermore, the maximum absolute error in R_a as a product of the error of the combined δ parameter is $1.63 \text{ MPa}^{0.5}$. Although the R_a constraint could have been extended to $9.63 \text{ MPa}^{0.5}$ to ensure all feasible structures were considered, the converse could occur where less than satisfactory structures are accepted. Therefore, it is essential that a compromise is and a value of $8.8 \text{ MPa}^{0.5}$ is sufficient until better models are developed.

4.6 Other Industries

Table 8: R_a and CMC values for a variety of industries and their associated soils

Industry	Soil	R_a	CMC [M]
Fuel	ASTM Fuel "A"	5.60	1.84E-08
Foods, textiles, lubricants, paint, ink, perfume agriculture	Butyl Stearate	8.05	1.70E-08
Soaps, lubricants, hydraulics, paints, dyes, coatings, ink, plastics, waxes, polishes, pharmaceuticals, perfumes	Castor Oil	6.96	3.04E-07
Fragrance, flavouring component	Ethyl Cinnamate	7.68	4.18E-08
Painting	Linseed Oil	5.71	5.53E-06
Metalworking, lubricants, plasticizer, detergent, fire-safety	Tricresyl Phosphate	8.42	2.00E-03

One of the key strengths of using HSP is its ability to measure a cleaning agent's effectiveness against specific soils which allows the detergent to be tailored towards specific targets in the design process.

Following the successful optimisation of the surfactant structure against butyl stearate, this approach was tested against each of the other soils in Table 2 using the LAE structure, with the results shown in Table 8. It is interesting to note that despite similarly low values of CMC for some of the soils, the R_a values are considerably smaller than that of butyl stearate. For example, despite ASTM and butyl stearate having almost identical CMC values, ASTM's R_a value is over 25% smaller. This suggests that the optimal LAE structure for ASTM is significantly better than the optimal structure for butyl stearate - further investigation into potential structures would be required to reach the same level of cleaning power.

5 Conclusion

A wide range of properties are implemented into an optimisation framework as an alternative way of developing novel, better suited surfactant molecules for the cleaning industry. To cater to the growing demands of consumers, particular emphasis is placed on cleaning power and its relation to CMC and HSP, whilst simultaneously increasing safety and reducing ecological impact through the implementation of TP.

This paper details the innovative use of HSP to predict a surfactant's affinity towards a specific soil, allowing a non-ionic surfactant to be tailored to specific use cases. This is done before its combination with the anionic surfactant, SDS, as part of a binary mixture. Furthermore, the ratio of the non-ionic surfactant in this binary mixture is also optimised to allow for a higher probability of all of the constraints being met.

After the bi-objective optimisation process was conducted, the theoretical best non-ionic structure for tackling butyl stearate, was an LAE structure. Our process was also easily modified to target other soils which produced more promising structures.

The availability of a more comprehensive and accurate dataset when modelling parameters would improve the validity of results. Further consideration of the non-linearity of mixing relationships for each of the properties would also result in a higher validity.

To comprehensively measure the environmental and cost impacts of the optimised surfactant, a life cycle assessment of the process synthesis could be carried out. This would allow for the identification and analysis of processing paths and designs, which are essential when it comes to the creation of new cleaning products.

The potential for the tailoring of new surfactants to meet demand is a useful tool for the detergent industry as it vastly reduces the large number of surfactant combinations, resulting in experimental validity becoming a much more plausible option.

6 References

- Abbott, S., Hansen, C.M., 2013. Hansen Solubility Parameters in Practice. Hansen-Solubility.
- Azzam, E.M.S., 2001. Effect of alkyl groups in anionic surfactants on solution properties of anionic-nonionic surfactant system. *J. Surfactants Deterg.* 4, 293–296. <https://doi.org/10.1007/s11743-001-0182-4>
- Barton, A.F.M., 2017. CRC Handbook of Solubility Parameters and Other Cohesion Parameters: Second Edition, 2nd ed. Routledge, New York. <https://doi.org/10.1201/9781315140575>
- Bhattacharai, A., Niraula, T., Chatterjee, S., 2014. Sodium dodecyl sulphate: A very useful surfactant for Scientific Investigations. *J. Knowl. Innov.* 2, 111–113.
- Bondi, C.A., Marks, J.L., Wroblewski, L.B., Raatikainen, H.S., Lenox, S.R., Gebhardt, K.E., 2015. Human and Environmental Toxicity of Sodium Lauryl Sulfate (SLS): Evidence for Safe Use in Household Cleaning Products. *Environ. Health Insights* 9, 27–32. <https://doi.org/10.4137/EHI.S31765>
- Chen, W., Schechter, D.S., 2021. Surfactant selection for enhanced oil recovery based on surfactant molecular structure in unconventional liquid reservoirs. *J. Pet. Sci. Eng.* 196, 107702. <https://doi.org/10.1016/j.petrol.2020.107702>
- Cheng, K.C., Khoo, Z.S., Lo, N.W., Tan, W.J., Chemmangattuvalappil, N.G., 2020. Design and performance optimisation of detergent product containing binary mixture of anionic-nonionic surfactants. *Heliyon* 6, e03861. <https://doi.org/10.1016/j.heliyon.2020.e03861>
- Choi, J.-Y., Byeon, S.-H., 2020. HAZOP Methodology Based on the Health, Safety, and Environment Engineering. *Int. J. Environ. Res. Public Health* 17, 3236. <https://doi.org/10.3390/ijerph17093236>
- Cox, M.F., 1989. Effect of alkyl carbon chain length and ethylene oxide content on the performance of linear alcohol ethoxylates. *J. Am. Oil Chem. Soc.* 66, 367–374. <https://doi.org/10.1007/BF02653293>
- Esmaili, H., Mousavi, S.M., Hashemi, S.A., Lai, C.W., Chiang, W.-H., Bahrani, S., 2021. Chapter 7 - Application of biosurfactants in the removal of oil from emulsion, in: Inamuddin, Adetunji, C.O. (Eds.), *Green Sustainable Process for Chemical and Environmental Engineering and Science*. Elsevier, pp. 107–127. <https://doi.org/10.1016/B978-0-12-822696-4.00008-5>
- Evans, K.A., Dubey, S.T., Kravetz, L., Dzidic, I., Gumulka, J., Mueller, R., Stork, J.R., 1994. Quantitative Determination of Linear Primary Alcohol Ethoxylate Surfactants in Environmental Samples by Thermospray LC/MS. *Anal. Chem.* 66, 699–705. <https://doi.org/10.1021/ac00077a019>
- Fung, H., Wibowo, C., Ng, K.M., 2007. Chapter 8 - Product-centered Process Synthesis and Development: Detergents, in: Ng, K.M., Gani, R., Dam-Johansen, K. (Eds.), *Computer Aided Chemical Engineering, Chemical Product Design: Toward a Perspective Through Case Studies*. Elsevier, pp. 239–274. [https://doi.org/10.1016/S1570-7946\(07\)80011-3](https://doi.org/10.1016/S1570-7946(07)80011-3)
- Hansen, C.M., 2007. Hansen Solubility Parameters: A User's Handbook, Second Edition, 2nd ed. CRC Press, Boca Raton. <https://doi.org/10.1201/9781420006834>
- Huibers, P.D.T., Lobanov, V.S., Katritzky, A.R., Shah, D.O., Karelson, M., 1996. Prediction of critical micelle concentration using a quantitative structure-property relationship approach. 1. Nonionic surfactants. *Langmuir* 12, 1462–1470. <https://doi.org/10.1021/la950581j>
- Hukkerikar, A.S., Kalakul, S., Sarup, B., Young, D.M., Sin, G., Gani, R., 2012. Estimation of Environment-Related Properties of Chemicals for Design of Sustainable Processes: Development of Group-Contribution+ (GC+) Property Models and Uncertainty Analysis. *J. Chem. Inf. Model.* 52, 2823–2839. <https://doi.org/10.1021/ci300350r>
- Kronberg, B., Costas, M., Silveston, R., 1995. Thermodynamics of the hydrophobic effect in surfactant solutions: Micellization and adsorption. *Pure Appl. Chem. - PURE APPL CHEM* 67, 897–902. <https://doi.org/10.1351/pac199567060897>
- Kruglyakov, P.M. (Ed.), 2000. Chapter 1 - Physicochemical properties of surfactants used in the definition of hydrophile-lipophile balance, in: *Studies in Interface Science, Hydrophile-Lipophile Balance of Surfactants and Solid Particles*. Elsevier, pp. 4–99. [https://doi.org/10.1016/S1383-7303\(00\)80014-9](https://doi.org/10.1016/S1383-7303(00)80014-9)
- Mattei, M., Kontogeorgis, G.M., Gani, R., 2014. A comprehensive framework for surfactant selection and design for emulsion based chemical product design. *Fluid Phase Equilibria, Special Issue on PPEPPD 2013* 362, 288–299. <https://doi.org/10.1016/j.fluid.2013.10.030>
- Myers, D., 2020. *Surfactant Science and Technology*. John Wiley & Sons.
- O. Bianchetti, G., L. Devlin, C., R. Seddon, K., 2015. Bleaching systems in domestic laundry detergents: a review. *RSC Adv.* 5, 65365–65384. <https://doi.org/10.1039/C5RA05328E>
- Ortiz, M.S., Alvarado, J.G., Zambrano, F., Marquez, R., 2022. Surfactants produced from carbohydrate derivatives: A review of the biobased building blocks used in their synthesis. *J. Surfactants Deterg.* 25, 147–183. <https://doi.org/10.1002/jsde.12581>
- Perinelli, D.R., Cespi, M., Lorusso, N., Palmieri, G.F., Bonacucina, G., Blasi, P., 2020. Surfactant Self-Assembling and Critical Micelle Concentration: One Approach Fits All? *Langmuir* 36, 5745–5753. <https://doi.org/10.1021/acs.langmuir.0c00420>
- Rosen, Milton J., Kunjappu, J.T., 2012. Emulsification by Surfactants, in: *Surfactants and Interfacial Phenomena*. John Wiley & Sons, Ltd, pp. 336–367. <https://doi.org/10.1002/9781118228920.ch8>
- Rubingh, D.N., 1979. *Mixed Micelle Solutions*, in: Mittal, K.L. (Ed.), *Solution Chemistry of Surfactants: Volume 1*. Springer New York, Boston, MA, pp. 337–354. https://doi.org/10.1007/978-1-4615-7880-2_15
- Śliwa, K., Śliwa, P., 2020. The Accumulated Effect of the Number of Ethylene Oxide Units and/or Carbon Chain Length in Surfactants Structure on the Nano-Micellar Extraction of Flavonoids. *J. Funct. Biomater.* 11, 57. <https://doi.org/10.3390/jfb11030057>
- Smulders, E., von Rybinski, W., Sung, E., Rähse, W., Steber, J., Wiebel, F., Nordskog, A., 2007. Laundry Detergents, in: *Ullmann's Encyclopedia of Industrial Chemistry*. John Wiley & Sons, Ltd. https://doi.org/10.1002/14356007.a08_315.pub2
- Stefanis, E., Panayiotou, C., 2008. Prediction of Hansen Solubility Parameters with a New Group-Contribution Method. *Int. J. Thermophys.* 29, 568–585. <https://doi.org/10.1007/s10765-008-0415-z>
- Tadros, T., 2013. Critical Micelle Concentration, in: Tadros, T. (Ed.), *Encyclopedia of Colloid and Interface Science*. Springer, Berlin, Heidelberg, pp. 209–210. https://doi.org/10.1007/978-3-642-20665-8_60

Recovery of 5-(hydroxymethyl)furfural (HMF) from Effluents

Nur Alya Fariesha Affrie Shah¹ & Aashna Jaya Bhugun¹

¹Department of Chemical Engineering, Imperial College London, U.K.

Abstract

Hydroxymethylfurfural (HMF) is a promising chemical for the future, with potential to be used as an intermediate or starting material for functional polymers, pharmaceutical ingredients and biofuels. However, HMF separation from reaction effluents is challenging due to its susceptibility to thermal degradation, and such requires further research. This paper focuses on the separation of HMF via two separation methods: vacuum distillation, and cooling crystallisation. It was found the cooling crystallisation with acetone, was not feasible for temperatures as low as -70 °C. As such recovery of HMF via vacuum distillation with acetone was investigated more thoroughly, with particular focus of the impact of distillation time on recovery. Ultimately, it was found that up to 96 % could be recovered from a reaction effluent, demonstrating that low-boiling point solvent systems are beneficial in optimisation through humin minimisation.

1 Introduction

In recent years, due to the various climate-related initiatives such as the UK's 2050 Net Zero Target, the need for renewable resources to produce energy and various chemicals, has been a topic of increasing interest. One of the potential routes for moving away from the current dependence on fossil fuels is biomass-derived resources, which are available in abundance and at a relatively low cost. As such, these types of materials would allow a more sustainable supply of essential precursors and intermediates for the manufacture of pharmaceuticals, polymers, fuels, and other chemicals [1].

Prime examples of a biomass-derived substance that could shape the future of these sectors, are furans. In fact, furans have been named one of the 'Top 12 Biochemicals from Renewable Resources' by the United States Department of Energy (USDOE) [2], indicating the magnitude of the potential for them in the aforementioned sectors. The USDOE go on to discuss examples of useful substances with furan rings such as furfural and 5-hydroxymethylfurfural (HMF), which was referred to as a 'chemical platform for a lignocellulosic biomass biorefinery' [2].

In addition to a furan ring, HMF also contains two other functional groups (as seen in Figure 1), therefore making it desirable for its versatility to be subjected to a wide range of reaction pathways [3]. To name a few reaction types, HMF can undergo hydrolysis, hydrogenation, and oxidation [4]. Therefore, it can be transformed into numerous useful products, for example, precursors for biofuels and plastics, within the biorefinery sector [1].

HMF is frequently produced via acid-catalysed dehydration of saccharides, such as cellulose, glucose, fructose, or sucrose [5]. Due to its low cost and abundance, glucose is often favoured as the feedstock for said reaction but has typically been shown to produce lower yields, compared to its structural isomer, fructose. This observation has been attributed to the stability of the pyranoside ring glu-

cose, which therefore does not form as much open-chain form glucose within solution [6].

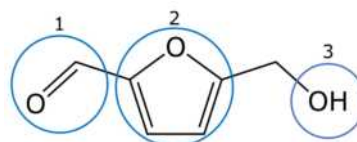


Figure 1: Skeletal Structure of HMF with key functional groups circled; 1: Aldehyde group, 2: Furan ring, 3: Hydroxy group

However, despite these huge benefits, there are some drawbacks to HMF production which pose challenges to its application on a larger scale. HMF is thermodynamically unstable [4], and therefore quite susceptible to side reactions, and in turn by-product formation. Levulinic acid and formic acid are common by-products which are formed via the rehydration of HMF [7]. In addition, the cross-polymerisation of reaction intermediates leads to condensation products – a range of soluble polymers, but also some insoluble humins [6]. The typical reaction pathway to form HMF and its common by-products are illustrated in Figure 2.

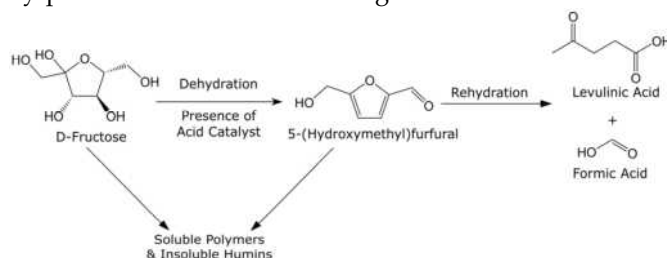


Figure 2: Typical reaction pathway for HMF production from D-Fructose as well as how some of the most common by-products form; Adapted from [5]

2 Background

Previous work has looked extensively into maximising selectivity towards HMF and therefore obtaining higher yields through research experimentation

with various catalysts, solvents and operating conditions. A popular method of generating higher HMF yields is via fructose dehydration reaction using an aprotic solvent such as dimethyl sulfoxide (DMSO) because it can somewhat repress undesired side reactions (such as HMF rehydrations, condensation reactions, and acyclic reaction sequences), thereby generating higher yields of HMF [8]. Alternatively, other papers have discussed effective reactions of fructose using a simple low boiling solvent system, of acetone and water, [9]; this method has the added benefit of lower costs compared to DMSO.

However, many of the hindrances to the widespread industrial application of HMF are not related to the reaction itself. It is in fact the separation of HMF from reaction mixtures which is so complex, for a number of reasons including: the reactivity and thermodynamic instability of HMF and therefore, the formation of by-products. An example of a common undesirable by-product are humins. These are carbonaceous polymers which are thought to be formed from condensation reactions, between a carbohydrate and some intermediates during their conversion to a particular product, such as HMF [10]. Although HMF yields and recoveries are often said to be impeded by humin formation, the exact reaction pathways and changes in morphology and structure, are not fully understood. As a result, the impact of this substance on the separations of HMF from reaction effluents is yet another complexity. Other crucial considerations for HMF isolation, are the reaction and separation conditions. For example, as aforementioned, high boiling points are often favoured for their high yields of HMF (around 90%). In turn, higher temperatures are required for solvent removal. However, this also causes difficulties because HMF itself has a high boiling point and is also susceptible to thermal degradation. Therefore, separation processes such as vacuum distillation, which have been tested for reaction effluents containing DMSO, have often led to sizeable losses of HMF of around 30% [9]. Ultimately, the low recoveries, high energy intensity and high costs associated with high boiling point solvents would not be adequately profitable on a production scale.

Given the challenges with separation even on a small scale, it is evident that a lot more research is required in this field to inform the design of industrial-scale HMF production processes. Furthermore, out of the large number of papers surrounding HMF, it has been reported by Trapasso et al. [4] that “less than 10% address HMF isolation [...] from the reaction mixture”. There are wide range of separation techniques which have been discussed for HMF isolation, including crystallisation, adsorption, vacuum distillation and variations of extraction (liquid-liquid, reactive, in-situ etc.) [3]. However, there has not been much experimental data to address these techniques, thus providing the motivation for this line of research.

Due to the lack of experimental data for the sep-

aration of HMF with low boiling systems, the main aim was to establish if it was possible to separate HMF from such solutions, and if so, which separation techniques would be feasible. Furthermore, the scope of this project covers lab-scale vacuum distillation, as well as cooling crystallisation. The key metric used to assess the success of separations was the recovery of HMF.

Initially, the recovery of HMF was investigated for solutions of pure HMF dissolved in acetone, before moving on to fructose dehydration reaction effluents. The aim of doing so was to assess the impact, if any, of the presence of reaction by-products or unreacted reactants on product separation. Throughout all experiments, the final goal was to work towards gathering data which could eventually be used to inform the process design of HMF production on a larger scale.

3 Method

3.1 Chemicals

All chemicals which were used for this project are as listed in Table 1.

Table 1: Summary of Chemicals Used

Chemical	Purity	Supplier
Hydroxymethylfurfural	>99%	Sigma Aldrich
Dimethylsulfoxide	≥99.5%	Sigma Aldrich
Acetone	≥99.9%	VWR Chemicals
Ethanol	≥99.9%	VWR Chemicals
Deionised (DI) Water	≥99%	N/A ¹
Fructose	≥99%	Sigma Aldrich
Amberlyst-15 ²	N/A	Sigma Aldrich
Sulfuric Acid	99%	Sigma Aldrich
Furfural	>99%	Sigma Aldrich

¹ Onsite purification using Veolia Purelab Chorus

² Hydrogen Form (Dry)

DMSO and Acetone were used as solvents for this research, whilst Sulfuric Acid and Amberlyst-15 were used as acid catalysts. Amberlyst-15 is a strongly acidic cation-exchange resin that is commonly used as a heterogeneous catalyst in acid-catalysed reactions, especially in organic synthesis.

3.2 Experimental Setup

3.2.1 Vacuum Distillation

Solutions of 0.88wt% of HMF in acetone-water (80/20 vol%), a low-boiling solvent and 2.08wt% of HMF in DMSO, a high-boiling solvent, were prepared based on the literature yields of the dehydration reaction of 1wt% fructose in respective solvents [7] [9]. These solvents were selected due to their ability to yield high quantities of HMF hence good potential for production.

The main separation process investigated, vacuum distillation was performed with Buchi Rotavapor R-100 under reduced pressures using vacuum

pump V-100 and heated up to a given operating temperature. These temperatures were determined by trial runs with only pure solvent, ensuring that the entirety of the solvent could be distilled. Starting temperature and pressures were taken from a similar study conducted on the rotavapor [11].

To carry out a fair experiment, several operating conditions on the rotavapor were kept constant such as the flask's rotational speed (level 6 of Buchi Rotavapor R-100) and the cooling water flowrate (maximum). The operating distillation temperature of 220°C for DMSO at reduced pressures was higher than the boiling point of DMSO at atmospheric pressure (189°C), may seem counterintuitive. However, this can be attributed to the loss of heat to the surroundings as proper insulation was infeasible in this setup, suggesting that the temperature of the solution itself may be lower. Due to the limitations of the water bath, a heating plate as seen in Figure 3 was used for DMSO system instead. The water level in the water bath for the acetone-water solvent setup was maintained throughout all separations.

10mL of HMF mixture/reaction effluent were added to a 500mL round-bottomed flask for the separation process. All separations were carried out until the entirety of the bulk solvent was eliminated from the product flask as illustrated in Figure 4. The duration of the distillation process was recorded. Subsequently, the post-rotavap product was redissolved in 10 mL of solvent (either ethanol or acetone) for analysis using analytical equipment mentioned in Section 3.3.



(a) Acetone Solvent System; operating conditions of 85°C and 50mbar, heated with a water bath
(b) DMSO Solvent System; operating conditions of 220°C and 35 mbar, heated with a heating plate

Figure 3: Rotavap Set-Up

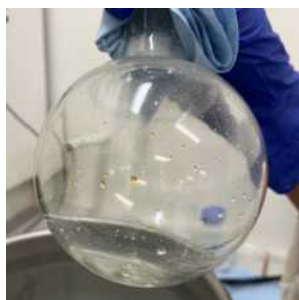


Figure 4: The target end product of distillation, where only small yellow/orange droplets are left and most of the solvent has been removed

3.2.2 Cooling Crystallisation

To investigate cooling crystallisation, initially, 10mL of HMF in acetone-water was used for this process. Further experiments were also conducted in order to understand the effect of varying furan concentrations on the crystallisation process. However, as HMF is very expensive, owing to the high production costs, furfural was used as a substitute. Furfural was considered a good alternative to HMF for this feasibility study because of its structural similarities to HMF; both structures contain an aldehyde functional group, but most importantly a furan pentose ring. 10mL of furfural (0.1, 1 and 10 wt.%) were then prepared in pure acetone. All samples were placed at room temperature, -20°C and -70°C (ThermoScientific Cryogenic Freezer) and examined after 24 hours, and again after 3 days.

3.2.3 Dehydration Reaction Effluents

One of the most common pathways to produce HMF is via the dehydration of reducing sugars in the presence of an acid [5]. Therefore, to reflect a typical industrial manufacturing route, reactions to produce HMF were carried out. The aim was to understand the impact of the presence of reaction by-products on HMF separation, as well as to assess whether there were additional degradation effects as a result.

As described in Section 1, a widely available, and relatively sustainable reactant for this reaction pathway is fructose, which can be produced via the hydrolysis and subsequent isomerisation of cellulose. For the purposes of this project, D-fructose was used directly. The sugar was dissolved into an 80/20 vol% solution of acetone and deionised water to create a fructose solution, with a concentration of 1 wt%.

Subsequently, per 5mL of the 1 wt% fructose solution, 0.01g of catalyst was added; both homogeneous and heterogeneous catalysts were used for different reactions, namely Amberlyst-15 and Sulfuric Acid (1M). These were then reacted in an Anton Paar Microwave Synthesis Reactor, at a temperature of 140°C, for time intervals of 5, 15, 30 and 60 minutes. Prior to separation via vacuum distillation, as described in Section 3.2.1, Amberlyst-15 was filtered out of the reaction effluent. As the sulfuric acid could not be removed as easily, it was possible to assess the impact of the presence of residual acid on the distillation products.

3.2.4 Investigation on Distillation Time

The effects of distillation time towards recovery was also of interest. 45mL of HMF reaction effluent from a 30-minute dehydration reaction catalysed by sulfuric acid was prepared as described in Section 3.2.3. This would be sufficient for three different distillations, ensuring that a fair test would be carried out in assessing the impacts of distillation time on recovery, as initial HMF concentration could also effect the recovery of separation.

From the distillations in previous parts of the experiment, 60-second distillations were found to result in satisfactory recoveries. Thus, recoveries of HMF at half (30s) and double (120s) the time alongside the 60s distillation were chosen to be observed.

3.3 Analytical Techniques

In order to assess the products obtained by: dissolving HMF, dehydration reactions to produce HMF, and separations to isolate HMF, a variety of analytical techniques were used, each with slightly different purposes.

High-Performance Liquid Chromatography with a Variable Wavelength Detector (HPLC-VWD), set to detect $\lambda = 254\text{nm}$, was used to quantify the concentration of HMF within samples before (i.e. dissolved HMF, or reaction effluents) and after separation process (vacuum distillation and crystallisation). A standard error of $\pm 0.005\text{ gL}^{-1}$ for the concentrations of HMF was due to this instrument.

Gas Chromatography with a Flame-Ionisation Detector (GC-FID) was mainly used in comparing the compounds found in the samples of reaction feed, effluent and redissolved post-rotavap product. More specifically, it was used to qualitatively observe compounds which were not UV-active, and therefore went undetected in the HPLC-VWD. It was later also used to quantify the concentration of a well-known by-product of HMF production, Levulinic Acid, within the samples.

The conversion of sugar, fructose, was quantified by utilising a HPLC with Refractive Index Detector (HPLC-RID), alongside a calibration curve. A standard error of $\pm 0.926\text{ gL}^{-1}$ due to this instrument was considered in all fructose concentrations.

Finally, Gas Chromatography Mass Spectrometry (GC-MS) was utilised to identify the unknown compounds detected in post-rotavap products.

Table 2 summarises all of these techniques, as well as details of the equipment model, configurations and operating conditions.

3.4 Quantification of Results

A HMF concentration calibration curve was prepared to quantify the concentration of HMF based on the area of absorption on the HPLC by processing different concentrations of HMF (ranging from 0.01 to 1 wt%) in deionised water.

As the HMF used was purchased around a year ago and stored in a freezer, it was expected that some degradation would occur due to HMF's unstable nature. By producing a calibration curve and comparing it to one created last year, it was confirmed that some degradation had occurred. As a result, throughout the project, any calculated concentrations of HMF were adjusted to account for degradation by considering the difference in gradient of the calibration curves made this year as compared to last year when it was new.

The recovery of HMF from rotavap was calculated from Equation 1 where C_{HMF} is the concentration of HMF post-rotavap in gL^{-1} , C_{HMF_i} corresponds to the initial concentration of HMF solution in gL^{-1} , V_{sol} represents the volume of solvent in L used to redissolve post-rotavap product and V_i the initial volume of HMF-solvent mixture/effluent in L . In Section 4.1.1, the errors of the percentage recoveries were calculated from repeated distillation runs whereas standard errors of $\pm 3.6\%$ were implemented in Section 4.3.2 attributed to the errors from the analytical equipment, different reaction effluents and distillation times.

$$\text{Recovery}_{\text{HMF}}(\%) = \frac{C_{\text{HMF}}V_{\text{sol}}}{C_{\text{HMF}_i}V_i} \quad (1)$$

Besides recovery, another factor that was of interest is the purity of the HMF which was calculated by Equation 2. The mass of product left in the flask after the separation is represented by m_f . The errors in the purity calculations in Section 4.1.1 were obtained from repeated distillation runs.

$$\text{Purity}_{\text{HMF}}(\%) = \frac{C_{\text{HMF}}V_{\text{sol}}}{m_f} \quad (2)$$

To further investigate the carbon balance of the reaction and separation process, a calibration curve for fructose and levulinic acid, were prepared using LC-RID and GC-FID respectively. This allowed for the quantification of said compounds.

4 Results & Discussion

4.1 Pure HMF - Solvent Systems

To investigate the feasibility of separations in a low boiling-solvent system, acetone-water (80/20 vol%) was chosen to be investigated. A high-boiling-point solvent, DMSO was also tested to observe the impact of solvent systems of different boiling points on HMF recovery. All separations were conducted until the bulk solvent was removed from the product flask, as depicted in Figure 4. The time taken to reach this point was measured where the acetone-water system required an average of 170 seconds and the DMSO system 451 seconds.

4.1.1 Recovery and Purity of Product

Due to HMF's thermal instability, it was anticipated that recovery from a high-boiling solvent would be lower as observed in Figure 5. HMF recovery in acetone-water was 90%, while in the DMSO system, it was only 62%. As both HMF (291.5 °C) and DMSO (189 °C) have high boiling points, the separation would imminently be harder than with HMF and acetone-water. The high operating temperature of the DMSO system enhances the degradation of HMF, a thermally unstable compound, leading to significant HMF loss. Nevertheless, a 10% loss in the low-boiling solvent system was also observed,

Table 2: Summary of Analytical Instruments

Instrument Name & Model	Detection Type	Mobile Phase	Flowrate / mL.min ⁻¹	Column Temperature / °C	Column Type	Sample Injection Volume / µL
Agilent 1220 Infinity II - HPLC	Variable Wavelength Detector (VWD) – Ultraviolet (UV)	0.4mL 85wt% H ₃ PO ₄ in 2L of water	0.45	65	MetaCarb	0.5
Agilent 1260 Infinity II - HPLC	Refractive Index Detector (RID) – Ultraviolet (UV)	DI water	0.75	80	HiPlex-Ca	20
Agilent 8890 - GC	Flame Ionisation Detector (FID)	Nitrogen	2.5	Inlet: 230 // Internal Range: 50 - 250	Cp-Wax 52	1
Shimadzu Nexis GC-2030	Mass Spectrometry (MS)	Helium	1.0	Interface: 230 // Internal Range: 50 - 250	SH-Rxi-5ms	1

possibly resulting from the degradation of HMF into larger carbon compounds known as humins or the rehydration process of HMF as shown in Figure 2. Further testing of by-products was conducted in Section 4.1.2 using GC-FID, considering HMF's tendency to react with water to form levulinic acid and formic acid under thermal conditions.

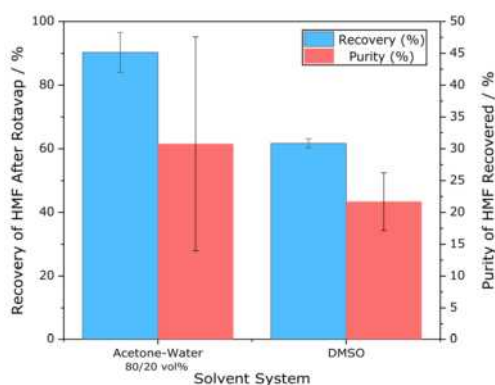


Figure 5: Comparison of recovery and purity of HMF from low-boiling and high-boiling solvents of 0.88wt% (Acetone-Water) and 2.08wt% (DMSO); equations used to calculate these values are in Section 3.4

In addition to understanding the recovery rates of the separation process, attempts were made to assess the purity of the separation product using Equation 2. Purity values of 31% and 22% were obtained for the acetone-water and DMSO systems, suggesting that additional refining process post-distillation would be necessary to obtain purer HMF. However, as depicted in Figure 5, large error bars were observed due to challenges in measuring the mass of the product; this was due to the vapour condensation which occurred once the vacuum pressure from the flask had been released, which therefore added further complexity. Due to these difficulties in obtaining a reliable purity measurement, cooling crystallisation, as described in 3.2.2, was employed. Since it is evident that higher recoveries were achieved in the lower-boiling solvent system, further investigations were conducted using an acetone-water solvent system.

4.1.2 Identifying Degradation Products

GC-FID was utilised in the attempt to detect degradation products from the pure HMF-solvent system as a loss of 10% HMF was observed. A comparison between the GC chromatogram at retention times before and after distillation is shown in Figure 6. The peaks before the 18-minute retention time was not shown because it consisted of only the solvents. The most dominant peak in the chromatogram was at a retention time of 24.8 minutes which is HMF, suggesting that no other detectable compounds were generated in a traceable amount during the distillation process. This implies that the degradation of HMF resulted in the formation of large carbon polymers instead, as levulinic acid and formic acid, if present, would have been identified by the GC-FID.

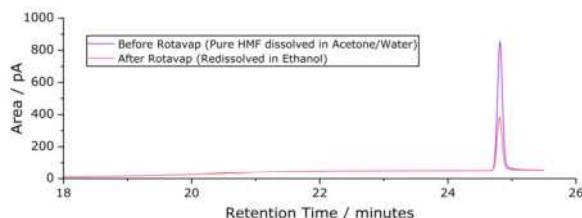


Figure 6: GC-FID chromatogram of pure HMF in acetone/water before and after rotavap

4.2 Cooling Crystallisation

As mentioned in Section 4.1.1, the purity of HMF in vacuum distillation suffered quite largely. Therefore, the HMF solution was placed in a cryogenic freezer for 24 hours, as described in Section 3.2.2. The results of this initial experiment, are pictured in Figure 7.

Though the results seemed promising at first glance, the colour of the liquid indicated that there was a significant amount of HMF in the solution. Furthermore, when the vial was inverted, it became evident that the solid matter was almost completely clear. Within less than 5 minutes at room temperature, whilst the samples were being handled, the solid had melted completely. As a result of these observations, it was eventually concluded that the solid matter was actually just ice, due to a form of

immiscibility between acetone and water at this temperature range. Ultimately, all the HMF appeared to remain dissolved in acetone; therefore, HMF isolation was unsuccessful.



Figure 7: HMF Effluent from Amberlyst-Catalysed Reaction after 24 hours at -70°C . At the top of the vial, a yellow-brown liquid, whilst at the bottom, a solid can be seen covered in the yellow liquid

There are a large range of reasons which could have caused the results aforementioned; to name a few, it may have been due to a much slower rate of crystallisation or perhaps an insufficient initial HMF concentration. As a result, a feasibility study using furfural, as described in 3.2.2, was set up. In order to eliminate the effects of acetone-water separation, pure acetone was used as the solvent.

After 24 hours, none of the samples (at any of the 3 temperatures, or 3 concentrations) showed any formation of crystals. Consequently, the samples were left for a further 2 days, yet even at -70°C , with a 10 wt% solution of furfural in acetone, the solution remained in liquid state. Due to the lack of success in crystallisation using acetone as the solvent, the remainder of the research was centred around the vacuum distillation process with acetone-water.

4.3 Dehydration Reaction Effluent

4.3.1 Reaction Effluent Yield

As aforementioned in Section 3.2.3, all reactions were prepared with 1wt% fructose solution, and either a homogeneous or heterogeneous catalyst, for a few different reaction times. As depicted in Figure 8, a range of reaction yields were observed; these corresponded to reaction effluents of concentrations ranging from 0.4gL^{-1} , to 3.5gL^{-1} .

Generally, for up to 60 minutes, reaction yields seemed to increase with the reaction time. However, the HMF yield with sulfuric acid seemed to decrease again once reaching a 60-minute reaction time. Again looking at the HMF yield, a homogeneous acid (sulfuric acid) seemed to perform very similarly to the heterogeneous one (Amberlyst-15) for a 30-minute reaction, suggesting that the decreased surface area to volume ratio of the solid acid did not lead to significant mass transfer limitations, and therefore had minimal impact to the rate of reaction for this amount of time.

4.3.2 Recovery of Reaction Effluents

The recovery of HMF from reaction effluents was carried out successfully with recoveries as high as 94.3% and as low as 72.6% with a standard error

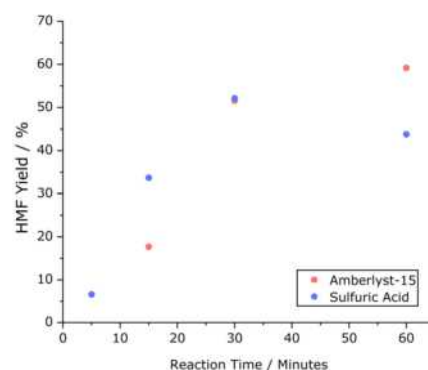


Figure 8: Impact of Reaction Time on HMF Yield at 140°C Reaction

of $\pm 3.6\%$. As the separations were carried out to a certain end-point determined by qualitative observation, different distillation times were recorded for each separation. The recovery of HMF from acid-free reaction effluents (Amberlyst-15) resulted in a similar range of recoveries to the product from pure HMF solvent. Although the distillation time of the pure HMF-acetone system was much larger, a higher recovery would be expected if it were to be distilled at a lower distillation time, similar to the ones of Amberlyst-15 products, as discussed in Section 4.6. Considering that the error bars overlap, the difference in their recoveries is statistically insignificant. Qualitatively similar observations were made on the product from pure HMF-acetone system and from amberlyst-catalysed reaction by comparing Figure 4 and Figure 10a. Both products observed yellow/orange droplets with varying intensities depending on the initial concentration of HMF. This indicates that the presence of unreacted fructose and reaction by-products such as but not limited to levulinic acid and formic acid does not have a significant impact on the recovery of HMF through vacuum distillation.

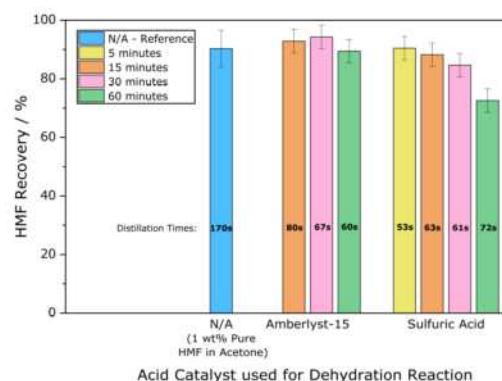


Figure 9: Comparison of HMF Recovery from Pure-HMF Solvent and HMF Effluents from Amberlyst-15 and Sulfuric Acid catalysed Reactions

The acid from sulfuric acid-catalysed reactions was not removed before the separation process to understand the impact of acid on the recovery of HMF. By comparing the recoveries of the 30-minute reaction between amberlyst and sulfuric acid

catalysed reactions, as they have similar distillation times, it can be concluded that the acid too does not have a significant impact towards recovery. While there appears to be a disparity in their recoveries upon initial observation, the overlapping of the error bars implies that the distinction is statistically insignificant. A previous study on the impact of acid-catalyst on the separation process of HMF also saw similar results [11]. Thus, the choice of heterogeneous or homogeneous acid catalyst would not be a factor to be considered in reaching high recoveries.



(a) Amberlyst-15

(b) Sulfuric Acid

Figure 10: Product of Separation from Different Reaction Catalysts

However, the presence of acid poses an additional challenge as black particles were formed during the distillation process as seen in Figure 10b. The black particles formed from the product of sulfuric acid reaction effluent are presumed to be the large carbon polymer, humins as they were insoluble in ethanol. Nevertheless, recovery of HMF is satisfactory despite the formation of black particles. Although no observable humins were detected in the separation product of the acetone/water reaction effluent, the possibility of HMF degradation into humins should not be disregarded. It is conceivable that the formed humins are in a soluble state, potentially serving as intermediates for the subsequent formation of solid humins, as suggested by a study [12].

The presence of acid in reaction effluents does not impact the recovery however, it adds additional complexity to the separation process due to the formation of black carbon particles. This further suggests that an additional separation unit would be required in the scale-up of this process to remove the unwanted black particles from the final product.

4.4 Impact of Solvent Used For Analysis

The formation of ethoxymethylfurfural (EMF) was identified on the GC-MS when the separated product was redissolved in ethanol. This is a result of the etherification reaction between the ethanol and HMF which is common with furans [13]. Further details of how this was found are in the supplementary information. Although the formation of ether did not significantly impede the analysis on the recovery of HMF, as can also be seen in the supplementary information, the production of it should be avoided. Hence a different solvent that would not form reactions with HMF should be used.

Acetone was chosen to redissolve the separation product as a better alternative to ethanol. It was also observed that acetone helps to agglomerate the solid particles in the product, hence a better qualitative observation can be easily seen if needed.

4.5 Understanding Carbon Loss

To further understand the relation between the dehydration reaction to the subsequent downstream process of isolation of the desired HMF product, carbon balance calculations were employed. This enabled a clear depiction of the carbon 'losses' in both the reaction and separation processes but also highlighted the impact of reaction yield on the final product purity.

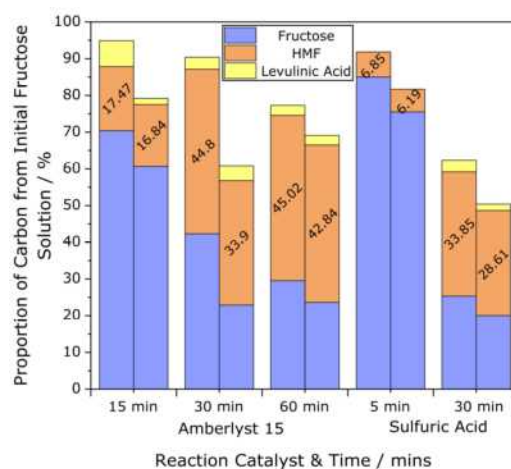


Figure 11: Proportion of Known Carbon in the Reaction Effluents Before Separation (Left) and After Separation (Right)

Assuming that no acid leached out of the solid catalyst, amberlyst-15, it would have been expected that the amount of fructose would have remained constant through the separation process. However, this expectation is contradicted by the actual results. The percentage of fructose post-separation decreased by a constant amount of approximately 10% across both catalysts investigated. Considering an error of $\pm 14\%$ in the percentage of fructose due to the high sensitivity of the LC-RID, the 30-minute reaction with amberlyst would show the same trend. This suggests that the decrease in fructose during separation is attributed to the formation of humins and not a result of further dehydration reactions in the distillation process within the specified conditions and distillation time.

Even at a reaction time as short as 15 minutes, the by-product of the dehydration reaction, levulinic acid, could be detected and quantified using GC-FID. Analysis of Figure 11 indicates that levulinic acid primarily forms during the reaction effluent, with a portion lost after the separation process due to degradation. This observation suggests that rehydration of HMF is not the primary cause of the degradation in the distillation process at this scale.

A key observation that can be made from the two 30-minute reaction in Figure 11, is that for a

homogeneous-catalysed process, the amount of carbon which could not be accounted for was higher. This behaviour is backed up by previous research, which found that the formation of humins is 50% higher on a carbon basis in a homogeneous reaction than a heterogeneous reaction [10]. As humins cannot be detected by any of the analytical methods, there is a possibility that a big portion of the unaccounted carbon is made up of this compound. Despite having carbon losses of around 15% in the distillation process, only a small portion of that loss is from HMF, suggesting that the distillation process is effective in recovering the desired product. This further supports the good recovery of HMF despite the formation of black particles in the distillation process from sulfuric acid-catalysed reactions as seen in Figure 10b.

The percentage of fructose gradually decreases over longer reaction times as seen in Figure 11 indicating that higher fructose conversions were obtained. This corresponds well to the percentage yields of HMF observed in Figure 8 within the investigated reaction times. Thus, achieving a higher fructose conversion, and consequently, a higher yield of HMF, is pivotal in obtaining a final product with a high HMF purity. Since fructose cannot be separated during the distillation process, it would be efficient to decrease the amount of fructose before the separation process through reactions that lead to high fructose conversions and HMF yield. However, further investigations would need to be done in the future to address this.

4.6 Investigating Distillation Time

Distillation time is one of the factors that significantly impact the recovery of HMF in the separation process. From Figure 12, it can be concluded that the recoveries reduce at higher distillation times for both catalysts. Thus, further investigations were carried out to understand why more HMF is lost at longer distillation times.

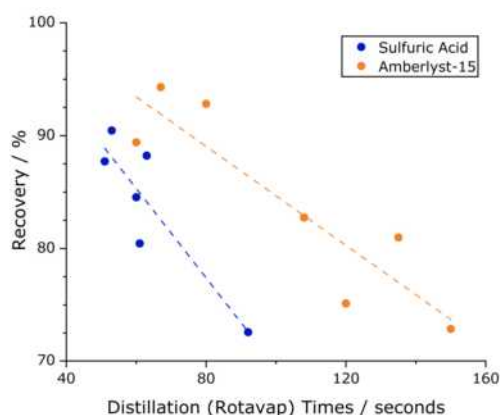


Figure 12: Recovery of HMF from Acid-Catalysed Reactions Across Distillation Times

As black particles were formed in the separation product of the sulfuric acid-catalysed reaction, further investigations on the impact of distillation time

on the formation of humins were also conducted.

4.6.1 Carbon Loss at Varying Distillation Times

For a distillation time of 30 seconds, the evaporation of the solvent was incomplete as there was an observable amount of solvent left in the flask as seen in Figure 14. When redissolved in acetone, the solution formed had a light yellow colour with no particles observed as seen in (a) of Figure 14. The recovery of HMF as expected was high at 90.6% but with low product purity as observed qualitatively. Contrary to this, the 120-second distillation run resulted in dark-coloured droplets alongside insoluble black particles on the walls of the flask. When redissolved in 10mL acetone, most of the particles agglomerated with an observable amount of smaller particles floating independently.

As seen in Figure 13, the recovery of distillation time at 120s went down to as low as 49.3% as compared to the higher recoveries at lower distillation times. Overall, the 30s and 60s distillation times had almost similar recoveries with both achieving over 90% recoveries. However, the 60s distillation run did appear to outperform the 30s one by just over 6%, and as a result, further optimisation should be carried out to investigate more distillation times between the range of 30s and 120s.

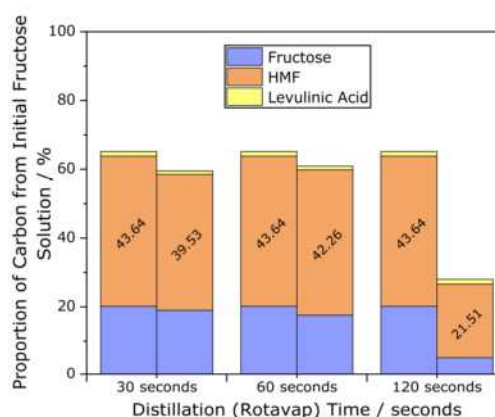


Figure 13: Comparison of the Carbon Balance Before and After Rotavap Assessed at Different Distillation Times

Due to the lack of research on the formation of humins, it is hard to understand the characteristics of these compounds and how they react to different stimuli. Nevertheless, longer exposure to heat was proven to promote greater formation of humins, as observed in this investigation.

As seen in Figure 14, more humins were formed at higher distillation time which corresponds to the high amount of carbon unaccounted for especially shown in the 120s distillation. Not only was there an increase in humin formation, but a substantial portion of HMF was also lost referring to Figure 13, possibly contributing to the humin formation. This is because once most of the solvent has been distilled, higher concentrations of carbon components are left in the flask. Thus, there would be a greater contact area between the carbon-containing components, in-

creasing the probability of the formation of long carbon chains, humin. Despite the 60-second distillation proving optimal for recovery, humins were still present, as depicted in (e) of Figure 14.

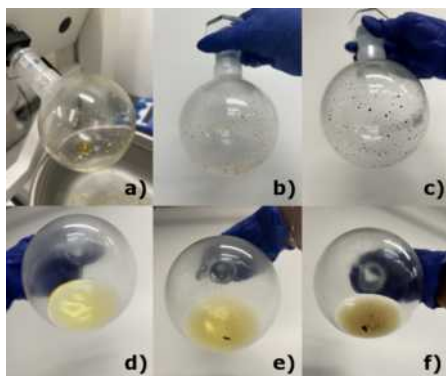


Figure 14: Product of Rotavap. a-c, product left after distillation. d-f, product after dissolving in 10mL acetone; a and c are the products of 30s of distillation, b and e are the products of 60s of distillation, c and f are the products of 120s of distillation

Consequently, addressing the challenge of minimising humin formation while achieving high recovery becomes even more complex when dealing with acid-present reaction effluents. This underscores the importance of finding an optimal point that balances both high recovery and low humin formation.

4.7 Analysis of Solids

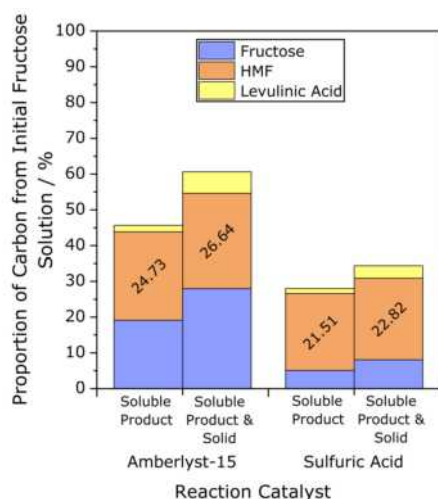


Figure 15: Comparison of the Proportion of Carbon Compounds in the Product that are Soluble in Acetone and Overall Product Including the Water-Soluble Part of Solid

The solids formed in Figure 14 were further analysed by filtering out the solid and redissolving it in water. Although the separation product of amberlyst-catalysed reactions did not produce visible black solid particles, an orange-yellow solid was formed when redissolving in acetone. It was found that the detectable soluble portion of the solid contained HMF, fructose and levulinic acid. This is probably because the carbon components may have formed bonds and agglomerated together with the

solid. For scale-up, where the solids would need to be removed, this would indicate that a portion of HMF may also be removed together with the solids.

As seen in Figure 15, the recovery of HMF including the portion from the solid would be higher. The recovery of HMF improved from 79% to 85%, when including the portion of HMF in the solids for the product of the amberlyst-catalysed reaction. Further research on these solids is suggested to be carried out to increase the efficiencies of HMF separation in obtaining a more economically feasible production route.

5 Conclusions & Outlook

It was found that low-boiling solvent systems particularly acetone-water (80/20 vol%) were a viable alternative to some of the solvents which are currently favoured. The method and conditions used to attempt cooling crystallisation separation proved unsuccessful which could be simply down to insufficient time to crystallisation (i.e. a slow crystallisation process), or the process could be infeasible. Nevertheless, vacuum distillation showed promising results with such solvent systems, achieving values of over 90% recovery although with low purities. To allow the use of HMF in industry, particularly in polymers, plastics and pharmaceuticals, the impurities would need to be removed to prevent adverse effects. As a result, optimisation of the final product purity would require more investigation.

Another important consideration in a scaled-up production would be the catalyst used. In line with previous research, the presence of an acid catalyst within reaction effluents had minimal impact on the recovery of HMF when using a low boiling point solvent. However, holistically the use of a heterogeneous catalyst, such as Amberlyst-15, would be preferable due to the challenge of insoluble humin formation when acid is present, and as such would necessitate an additional separation process unit. Additionally, homogeneous acids are also more difficult to separate from solutions, hence would lead to some losses over time if separation were to be carried out. Therefore, the need to continuously purchase more catalyst, would add yet another cost to the separation process. Generally shorter distillation times were shown to result in higher recoveries as seen in Figure 12. However, further investigations on the same reaction effluent, found that the optimal distillation time concerning recovery occurred for a 60-second run, as opposed to a shorter 30-second run. Thus, further experiments similar to those described in Section 4.6.1 is suggested to be carried out at smaller intervals between 30 and 120 seconds, to establish the point of maximum recovery. The existence of an optimal distillation time on a lab scale is indicative of the behaviour that could be expected when scaled up. Consequently, it would be necessary to test a few different distillation times to inform some of the parameters required to design a distillation column (such as the number of stages or

hold-up times).

Furthermore, by using carbon balance calculations, it can be said that the degradation of HMF mainly leads to the formation of large carbon polymers, as opposed to rehydration products.

It is evident from the conclusions stated, that there is a lot more research to be done before HMF can be used more widely. To address the issue of purity of the end product, a range of factors could be improved. Firstly, improvements to the unsuccessful crystallisation method could be made. An example would be to change the solvent from acetone to an organic one such as Methyl Tertbutyl Ether (MTBE), which based on patent literature, supposedly allows crystallisation to occur within the temperature range tested [14]. Alternatively, seeding is a common practice in crystallisation, particularly for longer crystallisation times or solutions with a low concentration; this is because seeding can promote further nucleation of new crystals. Since this is widely used for semi-batch pharmaceutical applications nowadays, there is certainly some scope for similar industrial applications for HMF.

Potentially, crystallisation could be used as a downstream purification step after another separation technique (such as vacuum distillation). Finally, as much of the previous literature relates to the fact that many of the separation techniques proposed would result in high energy intensity and costs, there is limited information regarding the estimated quantities. Therefore, an investigation into the capital expenditure, operating expenditure, and energy usage should be assessed for separations using the acetone-water low-boiling point solvent system.

6 Acknowledgements

The authors would like to express their gratitude to the Hammond Research Group at Imperial College London for their help when using the lab facilities. Special thanks go to PhD candidate, Laurie Overtoom, for his time, support and guidance throughout the course of this research project.

References

- [1] F.N.D.C. Gomes et al. "Production of 5-Hydroxymethylfurfural (HMF) via Fructose Dehydration: Effect Of Solvent and Salting-Out". In: *Brazilian Journal of Chemical Engineering* 32.1 (Mar. 2015), pp. 119–126. doi: <https://doi.org/10.1590/0104-6632.20150321s00002914>.
- [2] Anuj Kumar Chandel and Fernando Segato. *Production of Top 12 Biochemicals Selected by US-DOE from Renewable Resources*. Elsevier, Oct. 2021. ISBN: 9780128236543.
- [3] Lei Hu et al. "State-of-the-art advances and perspectives in the separation of biomass-derived 5-hydroxymethylfurfural". In: *Journal of Cleaner Production* 276 (Dec. 2020), p. 124219. doi: <https://doi.org/10.1016/j.jclepro.2020.124219>.
- [4] Giacomo Trapasso et al. "Multigram Synthesis of Pure HMF and BHMF". In: *Organic Process Research Development* 26.10 (Sept. 2022), pp. 2830–2838. doi: <https://doi.org/10.1021/acs.oprd.2c00196>.
- [5] Ken-ichi Shimizu, Rie Uozumi, and Atsushi Satsuma. "Enhanced production of hydroxymethylfurfural from fructose with solid acid catalysts by simple water removal methods". In: *Catalysis Communications* 10.14 (Aug. 2009), pp. 1849–1853. doi: <https://doi.org/10.1016/j.catcom.2009.06.012>.
- [6] B. F. M. Kuster. "5-Hydroxymethylfurfural (HMF). A Review Focussing on its Manufacture". In: *Starch - Stärke* 42.8 (1990), pp. 314–321. doi: <https://doi.org/10.1002/star.19900420808>.
- [7] Christof Aellig and Ive Hermans. "Continuous D-Fructose Dehydration to 5-Hydroxymethylfurfural Under Mild Conditions". In: *ChemSusChem* 5.9 (July 2012), pp. 1737–1742. doi: <https://doi.org/10.1002/cssc.201200279>.
- [8] Román-Leshkov Yuriy, Juben N. Chheda, and James A. Dumesic. "Phase Modifiers Promote Efficient Production of Hydroxymethylfurfural from Fructose". In: *Science* 312.5782 (June 2006), pp. 1933–1937. doi: <https://doi.org/10.1126/science.1126337>.
- [9] Ali Hussain Motagamwala et al. "Solvent system for effective near-term production of hydroxymethylfurfural (HMF) with potential for long-term process improvement". In: *Energy Environmental Science* 12.7 (2019), pp. 2212–2222. doi: <https://doi.org/10.1039/c9ee00447e>.
- [10] T. M. C. Hoang et al. "Humin based by-products from biomass processing as a potential carbonaceous source for synthesis gas production". In: *Green Chemistry* 17.2 (2015), pp. 959–972. doi: <https://doi.org/10.1039/c4gc01324g>.
- [11] Erich Muller. *Chemical Engineering Research*. Vol. 5. Authors of paper: Hana Khatib Bethany Lam. Department of Chemical Engineering, Imperial College London, Feb. 2023, pp. 167–176. ISBN: 9781916005044. URL: <https://spiral.imperial.ac.uk/handle/10044/1/102074>.
- [12] Zhanwei Xu et al. "Mechanistic understanding of humin formation in the conversion of glucose and fructose to 5-hydroxymethylfurfural in [bmim]cl ionic liquid". In: *RSC Advances* 10.57 (2020), pp. 34732–34737. doi: [10.1039/d0ra05641c](https://doi.org/10.1039/d0ra05641c).
- [13] L.G. Tonutti et al. "Etherification of hydroxymethylfurfural with ethanol on mesoporous silica catalysts of regulated acidity to obtain ethoxymethylfurfural, a bio-additive for Diesel". In: *Microporous and Mesoporous Materials* 343 (2022), pp. 112–145. doi: [10.1016/j.micromeso.2022.112145](https://doi.org/10.1016/j.micromeso.2022.112145).
- [14] Jacob Jensen et al. *Purification of 5-hydroxymethylfurfural (HMF) by crystallization*. Feb. 2013. URL: <https://patentimages.storage.googleapis.com/12/2a/6b/f8ea8f2f3b5ca8/W02013024162A1.pdf>.

Numerical Modelling and Performance Assessment of Spectral Beam Splitting Based Concentrated Photovoltaic-Thermal Collector

Barry Wang and Tianze Zhang

Department of Chemical Engineering, Imperial College London, U.K.

Abstract: The spectral beam splitting concentrated photovoltaic-thermal (SBS CPVT) collector technologies are attracting more attention for their potential to meet the increasing energy demand. The SBS CPVT collector can effectively decouple the thermal and electrical energy generation by employing the spectral beam splitter, directing specific spectral bandwidths to the PV cells and others to generate heat, mitigating the efficiency loss due to overheating for conventional PV-T collectors. A SBS CPVT collector is designed in this study, which consists of a parabolic trough, a film-based optical filter, a thermal receiver at the bottom, and a triangular top channel where solar cells are attached to the two faces of the channel. Modelling and simulation of this study was conducted in COMSOL physics. A 2D model of the collector was built to investigate geometrical optics, followed by developing a 3D model for collector performance evaluation. An ideal interference filter and a low emissivity glass develop by Saflex was implemented separately as optical filter material in the simulation. Simulation results yield the following maximum efficiencies: 91.86% for ideal filter collector optical efficiency, 17.36% for PV cell electrical efficiency, 33.13% for top channel thermal efficiency, and 41.37% for bottom thermal receiver efficiency. For the Saflex SH filter collector, the corresponding efficiencies are 55.81%, 10.10%, 24.77%, and 20.94%, respectively.

1. Background and Introduction

As a renewable source of power, solar energy plays an important part in mitigating climate change by reducing greenhouse emissions, which is crucial in protecting humans, wildlife, and ecosystems [1]. Solar energy has become one of the most common renewable energy sources. Solar photovoltaic (PV) accounted for 4.5% of total global electricity generation in 2022, and it remains the third largest renewable electricity technology behind hydropower and wind [2]. Silicon cells are by far the most common solar cells studied due to their low cost, large availability, and reasonable efficiency [3]. Research showed that typical commercial panels have electrical efficiencies from 17% to 20%, where the rest of the solar energy absorbed cannot be harnessed by the PV panel and is dissipated as heat instead [4]. For example, a typical silicon cell can utilize a bandgap wavelength of between 550 and 1000 nm [5] while the spectrum out of this range is converted to heat. To utilize the waste heat from the solar cells, the photovoltaic-thermal (PV-T) hybrid collectors were developed.

By integrating a heat exchanger containing heat transfer fluid (HTF) to the back of the PV cells [6,7], the PV-T collectors enabled the waste heat from the solar cells to be effectively converted into thermal energy. Waste heat collection has a notable positive impact on the overall energy efficiency [8] as electricity and valuable thermal energy are produced simultaneously. PV-T collector can cover the energy uses of considerable fields, ranging from domestic uses [6] to large-scale industrial facilities, for instance, sports centres [9]. Despite overall energy efficiency is higher than PV cells alone, conventional hybrid PV-T collectors have limitations. PV-T collectors are thermally coupled, where the thermal and electrical energy outputs are correlated [10]. High solar irradiance leads to a significant temperature increase in solar cells [5].

Although the PV-T collector can achieve a high thermal efficiency, nevertheless, for a 1 K increase in PV temperature, electrical efficiency is reduced by approximately 0.5% [11]. Overheating also leads to a reduced life span of PV cells [12,13], making it capital intensive. Hence, generating high temperature thermal outputs and maintaining a high electrical efficiency simultaneously is the key to greater overall energy output and efficiency.

Spectral Beam Splitting (SBS) design is a strong solution to this challenge. An optical splitting film separates solar radiation based on photon energy: directing photons near the PV cell bandgap energy to the solar cells, while directing photons below and above the bandgap energy to a thermal receiver.[11]. Integration of SBS allows PV cells and thermal receivers to operate in different temperatures since they can be placed spatially separated. Therefore, SBS PV-T can achieve higher thermal efficiency than conventional PV-T [12]. In addition, utilisation of SBS leads to increased electrical efficiency as the filter directs away unabsorbed solar irradiance that could heat the PV element. [14].

Recent research explored the possibility of combining concentrating photovoltaic (CPV) with spectral splitting. The resulting spectral beam splitting concentrating photovoltaic-thermal (SBS CPVT) collectors have several advantages. Firstly, the utilization of CPVT collectors enables a high CR to be achieved [15], hence solar PV cells with smaller surface area and higher efficiency could be used instead of standard ones [15, 16, 17]. In addition, the high CR also enables the SBS CPVT collectors to achieve a high-temperature HTF [18]. High-temperature HTF could be delivered to a steam Rankine Cycle for further application [3].

The design and selection of the concentrator is a vital step for designing a SBS CPVT collector. Parabolic trough is one of the most common concentrators used for SBS CPVT collectors [15,

19]. Parabolic trough SBS CPVT collectors present a good balance of optical efficiency, investment cost, and operational stability [18]. Plenty research was done on parabolic trough. Wingert et al proposed a collector using silicon cell, which achieved a PV cell module electrical efficiency of 12.4% [20]. Zhang et al proposed a collector using plasmonic solar cell that achieved a 21.97% electrical efficiency overall [21]. The collector designed by Wang et al achieved a thermal efficiency of 26.7% [22]. Zhang et al designed a dichroic filter-based collector that achieved a 55% thermal receiver efficiency at an HTF flowrate of 10000 kg/day [23]. Results of aforementioned research are significantly different. Therefore, other than concentrator type, factors including collector structure, component geometry, PV cells, filters etc are also important in SBS CPVT collector design.

Selecting a suitable optical filter is also crucial for the SBS CPVT design. Under the same solar irradiance, integrating CPV without an effective optical filter would lead to a lower electrical efficiency than using a PV-T collector alone [17, 24], because a greater spectrum intensity further increases PV cell temperature. Spectral-splitting technologies can alleviate the excessive heating on PV cells [14, 25]. Many research focus on investigating the performance of interference filters (dichroic filter) and liquid absorptive filters. Film-based filters including interference filters directs part of the spectrum to the PV while reflecting the rest to the thermal receiver [12]. Interference thin film has the advantages of high optical efficiency, a well-defined reflection band, and low polarization dependence [26]. SBS CPVT with an interference filter could achieve high energy and exergy efficiency [11] with a relatively low optical loss of 3% [3]. With the design of multiple thin film layers to cover different bandwidths, interference filter demonstrates greater conversion efficiencies than systems with novel semi-transparent, back-reflecting solar cell beam splitters [27], as well as showing a substantial improvement over a simple bandpass filter [28]. Liquid absorptive filters work similarly to film-based filters, but they absorb part of the spectrum as heat energy instead of reflection [12]. The liquid absorptive filter has the advantages of a low operating temperature for the PV cells and a high energy output [29]. Liquid absorptive filters can act as both HTF and optical filters [30], thus resulting in a high thermal efficiency. In addition to filter type, the structure of the filter is also worth considering. The path of filtered sunlight is dependent on the parabolic curvature of the spectral splitting filter [31]. Therefore, fabricating the dichroic filter on a curved substrate can improve spectral matching to the focused incident light, despite having increased fabrication cost and complexity [26].

Structural design is also a crucial aspect that affects performance of SBS CPVT. Various collector structures are proposed by the previous research. For example, Wang et al developed a collector where PV cells are located on the concentrator [22]. Li et al developed a collector that have a secondary reflector, while the PV cells is located next to the filter [32]. Nevertheless, few studies investigate how the collector behaves if PV cells are placed at focal point of the concentrator. Therefore, an SBS CPVT collector with PV cells placed at the focal point of the parabolic concentrator was designed to fill the research gap. To carry out the study, the SBS CPVT model was built in COMSOL. Then, a simulation of the model using an ideal filter and a real-life filter was conducted. Based on the simulation results, the electrical, thermal, and combined optical efficiencies of the 2 models were calculated and compared afterwards.

2. Methodology

2.1 SBS CPVT Collector Description

The SBS CPVT collector was designed to split the concentrated solar spectrum into 2 bands, one of which is directed to the PV cells for electricity generation and the other to the bottom thermal receiver to generate heat. This achieves thermal decoupling to reduce the overheating of PV cells, which would prevent the reduction in electrical efficiency.

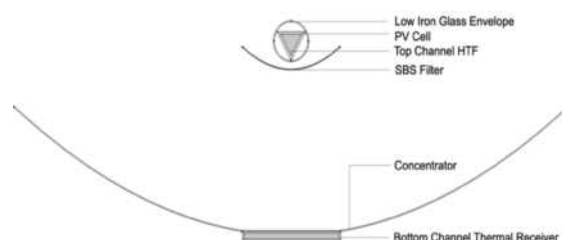


Figure 2.1 SBS CPVT Collector Model

As described in Figure 2.1, a parabolic trough concentrator was designed to concentrate the solar spectrum before being directed to the SBS filter. Part of the spectrum transmitted through the SBS filter enters the low iron glass envelope before it can reach the triangular-shaped PV channel where the silicon PV cells are placed. The triangular shaped PV channel was designed as it can increase the light receiving area. Also, this design is more suitable for parabolic troughs [33]. The silicon cells used in this study operate between the band gap wavelength of 350-1200 nm [34]. A vacuum gap is placed between the PV channel and the outer glass envelope to prevent heat loss. The top PV channel acts as the top channel thermal receiver to utilize the waste heat and prevent the heating up of PV modules. The remaining spectrum reflected by the filter is then directed to the bottom thermal receiver for heat

generation. The bottom thermal receiver was modelled as a rectangular aluminium heat exchanger to receive the thermal energy transmitted to the bottom channel. Water was chosen to be the HTF for both top and bottom channel thermal receivers.

The material of the concentrator, the PV channel and the heat exchanger is aluminum while silicon was selected to be the material to model the low iron glass envelop. The detailed modelling parameters are presented in Table 2.1. Lengths for all components of the CPVT collector were fixed to 3 m.

Table 2.1 Modelling Parameters for the SBS CPVT Collector

	Modelling Parameters	Value
Glass Envelope	Transmittance	0.91
	Thickness	2 mm
Triangular PV Channel	Thickness	5 mm
Bottom Thermal Receiver	Thickness	2 mm
SBS Filter	Thickness	3 mm
Concentrator	Thickness	2 mm
	Reflectivity	0.93

2.1.1 SBS CPVT Collector Geometry

The 2D geometrical model layout is illustrated in Figure 2.2 and the geometrical parameters are presented in Table 2.2. The thermal receiver width (w_t) was kept the same as the filter (w_f) to allow the full reception of the reflected spectrum.

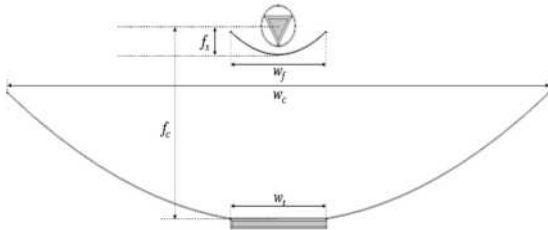


Figure 2.2 Collector Geometry

Table 2.2 Modelling Geometrical Parameters

Parameters	Symbol	Values
Concentrator Focal Length	f_c	340 mm
Filter Focal Length	f_s	60 mm
Filter Aperture	w_f	192 mm
Concentrator Aperture	w_c	1088 mm
Thermal Receiver Width	w_t	192 mm

2.2 Simulation Model in COMSOL

2D simulation of the model was first conducted in COMSOL to test the model against the geometrical optics by observing the ray trajectories. Based on the geometrical optics of the 2D model, a 3D model was

developed to evaluate the collector performance with coupled heat transfer and surface radiation. A physics-controlled finer mesh was implemented for the finite element analysis of the model in COMSOL to generate stable and accurate results with a short computation time. Two separate models were built for top and bottom channel thermal receivers and their thermal efficiencies were evaluated separately. The volume average temperature of the PV channel was approximated as PV cell temperature.

The SBS CPVT model simulation was first conducted with an ideal filter, where the transmittance was assumed to be 1 for the bandwidth between 500 and 1100 nm and 0 for the rest of the spectrum. Since it is not possible to get an ideal filter in real life, the performance of a real-life optical filter collector was also evaluated. Low emissivity glass (LEG) which would provide a similar transmitting range as the ideal filter was selected as real-life example.

Numerical modelling was conducted in MATLAB and the reference electrical efficiency was computed for each LEG base on its transmittance profile and the LEG which has the highest electrical efficiency was selected as the real-life filter. For the selected LEG, average transmittance was calculated within the silicon cell bandgap to estimate the transmittance of the material.

2.3 Numerical Model of SBS CPVT Collector

In the SBS CPVT collector, one should consider both photoelectrical and photothermal processes in the simulation. Numerical models for electrical and thermal efficiency calculations are explained in this section.

2.3.1 Electrical Efficiency

The Electrical Efficiency model was referred to the work of Ni et al [8]. Solar spectrum intensity reaching PV cells (G_{pv}) was calculated:

$$G_{pv}(\lambda) = C \times I_{AM1.5}(\lambda) \times R \times F_{pack} \times TR_g \times TR_f(\lambda) \quad (1)$$

where C is the geometrical concentration ratio, $I_{AM1.5}$ is solar spectrum intensity at air mass (AM) 1.5, F_{pack} is the packing fraction of solar cell, TR_g is transmittance of low iron glass, TR_f is the transmittance of optical filter and R is the reflectance of the concentrator mirror. The geometrical concentration ratio C was calculated by:

$$C = \frac{A_c}{A_{cell}} \quad (2)$$

where A_c is the aperture area and A_{cell} is the total PV cell area.

Solar spectrum intensity and transmittance of optical filter varies with respect to spectrum wavelength, hence they are functions of λ . The dark saturation current (J_0) was calculated as:

$$J_0 = k' T_{pv0}^{\frac{3}{n}} \exp\left(-\frac{E_g}{b k_B T_{pv0}}\right) \quad (3)$$

where k' , b and n are empirical parameters defined by the specific PV cell, T_{pv0} is cell temperature at 298 K, E_g is the bandgap of the cell and k_B is the Boltzmann constant. Then, the short circuit Current (J_{sc}) was calculated by:

$$J_{sc} = A_{cell} \int_{280}^{4000} \frac{G_0}{I_{AM1.5}(\lambda)} \times G_{pv}(\lambda) \times SR(\lambda) d\lambda \quad (4)$$

where $G_0 = 1000 \text{ W/m}^2$ was set as the direct solar irradiance, assuming an ambient condition of AM 1.5. SR is the spectral response of the cell at each wavelength. After that, open circuit voltage of cells (V_{oc}) could be calculated by:

$$V_{oc} = \frac{A' k_B T_{pv}}{e} \ln\left(\frac{J_{sc}}{J_0} + 1\right) \quad (5)$$

where A' is the cell's ideality factor, and e is the charge of an electron. Fill factor (FF) could be therefore calculated by:

$$FF = \frac{v_{oc} - \ln(v_{oc} + 0.72)}{v_{oc} + 1} \quad (6)$$

where $v_{oc} = \frac{V_{oc}}{V_{th}}$ is defined as normalized open circuit voltage with $V_{th} = \frac{A' k_B T_{pv}}{e}$. Then reference electrical efficiency ($\eta_{el,0}$) at 298 K is calculated by:

$$\eta_{el,0} = \frac{2n_{cell} V_{oc} J_{sc} FF}{G A_c} \quad (7)$$

where $2n_{cell}$ represent the total number of cells at both sides of the PV channel, A_c represents aperture area. The reference electrical efficiency is then used to calculate electrical efficiency η_{el} at different operating temperature afterwards.

The electrical efficiency η_{el} at a specific PV cell temperature T_{PV} is given below:

$$\eta_{el} = \eta_{el,0} [1 + \beta(T_{PV} - T_{ref})] \quad (8)$$

where η_{el} is the PV electrical efficiency at the corresponding $T_{PV,i}$, β is the temperature coefficient of power of PV cells. The COMSOL simulation was first performed with the collector reference electrical efficiency ($\eta_{el,0}$), and the simulation was repeated by replacing the $\eta_{el,0}$ with the electrical efficiency calculated from Eqn. 8. The iteration process is conducted due to the initial calculation of the $\eta_{el,0}$, which was performed assuming a constant average cell temperature of 298.15 K. However, in reality, the photovoltaic cell temperature (T_{pv}) varies based

on its inlet conditions. Therefore, iterations were performed until the difference of T_{pv} between each iteration was negligible. Two iterations were performed with COMSOL to get η_{el} in this study as they gave fairly accurate results.

2.3.2 Heat Transfer and Thermal Efficiency

To evaluate the thermal performance and the temperature distribution of the collector, the heat transfer within the collector was studied, which is characterized by convective and radiative heat loss from the surface of thermal receivers and convection heat transfer between HTF and thermal receivers.

Both continuity and Navier-Stokes equation were solved to demonstrate the conservation of mass and momentum of the flow of HTF within the top and bottom channel thermal receivers, where an additional convective term was added to the Navier-Stokes equation to describe the convective acceleration of incompressible fluid of the HTF. ρ_w, μ_w are the density and dynamic viscosity of water and ∇u_f represents the velocity gradient of the flow. I is the identity matrix. Equations used are presented below:

$$\nabla \cdot (\rho_w u_f) = 0 \quad (9)$$

$$\rho_w (u_f \cdot \nabla) u_f = \nabla \cdot [-pI + \kappa] \quad (10)$$

$$\kappa = \mu_w (\nabla u_f + (\nabla u_f)^T) - \frac{2}{3} \mu_w (\nabla \cdot u_f) I \quad (11)$$

where the stress tensor term κ accounts for both shear stresses due to velocity gradients and volumetric change in the HTF and the correlation is given in Eqn. 11. The HTF is assumed to be acting on with no external force and the outflow pressure is assumed to be 0 Pa. Non-slip boundary conditions were assumed at the wall. The convective heat transfer between the HTF and the thermal receiver surface is described as:

$$c_{p,w} u_f \cdot \nabla T_f + \nabla \cdot q_{HTF} = 0 \quad (12)$$

$$q_{HTF} = -k_f \nabla T_f \quad (13)$$

where $c_{p,w}$ is the specific heat capacity of water and k_f is the heat conductivity of water. ∇T_f describes the temperature gradient within the fluid and q_{HTF} is the heat transfer from the thermal receiver to the HTF.

The collector is assumed to be fully thermally insulated and the boundary condition is defined as:

$$-n \cdot q_r = 0 \quad (14)$$

where q_r is the heat radiating across the boundary.

The heat generated at the PV cell q_{gen} due to the radiation from the transmitted spectrum is described as:

$$q_{gen} = (G_0 A_c) \times TR_g \times R \times \alpha_{PV} \times (1 - \eta_{ele}) \quad (15)$$

where α_{PV} is the absorptivity of the PV channel and the term $(G_0 A_c) TR_g R$ describes the part of the radiation reaches the solar cell by accounting for the transmittance TR_g of the glass envelope and the reflectance R of the concentrator. According to Kirchhoff's law:

$$\alpha_{PV}(\lambda, T) = \varepsilon(\lambda, T) \quad (16)$$

where ε is the PV cell emissivity. The PV cell absorptivity is therefore determined to be the same as the PV cell emissivity of 0.88 [35].

The collector does not have an external heating source and the collector is assumed to operate at steady state where the temperature does not change with time.

By conducting the heat balance between the HTF inlet heat flux and specific enthalpy change at the inlet, the thermal boundary condition of the fluid inflow is described as:

$$-n \cdot q_{inlet} = \rho_w \Delta H u_f \cdot n \quad (17)$$

$$\Delta H = \int_{T_{in}}^{T_f} C_p dT \quad (18)$$

where n is the normal vector and q_{inlet} is the inlet flux. ΔH describes the specific enthalpy change at the inlet. T_f and u_f are temperature and velocity of the HTF flow respectively.

For the outlet thermal boundary condition of the HTF, no heat flux is considered across the outlet boundary in the normal direction, and it can be characterized as:

$$-n \cdot q_{outlet} = 0 \quad (19)$$

where q_{outlet} is the outlet outflow heat. The main causes of the heat loss are convective heat loss q_{conv} and radiative heat loss q_{rad} which can be described as:

$$q_{conv} = h(T_{amb} - T_s) \quad (20)$$

$$q_{rad} = \varepsilon_{cell} k_B (T_{sky}^4 - T_s^4) \quad (21)$$

where ε_{cell} denotes the surface emissivity of the PV cell and h is the convective heat transfer coefficient. The correlations for top and bottom convective heat transfer coefficients h_{top} [36], h_{bot} [37] and sky temperature T_{sky} are shown below:

$$h_{bot} = 2.8 + 3v \quad (22)$$

$$h_{top} = v^{0.58} d_{go}^{-0.42} \quad (23)$$

$$T_{sky} = 0.0522 T_{amb}^{1.5} \quad (24)$$

where v is the windspeed from surrounding which was taken as 1 m/s in the simulation, d_{go} is the outer diameter of the glass envelop.

The thermal efficiency of the collector was split into top and bottom channel thermal efficiency, for PV cell and thermal receiver respectively. Temperature measurements were taken from the COMSOL simulation results. The equations for top and bottom channel thermal efficiencies $\eta_{th,top}$ and $\eta_{th,bot}$ calculations are:

$$\eta_{th,top} = \frac{\dot{m}_{top} \times C_{p,w} \times (T_{top,out} - T_{top,in})}{A_c \times I_{AM1.5}} \quad (25)$$

$$\eta_{th,bot} = \frac{\dot{m}_{bot} \times C_{p,w} \times (T_{bot,out} - T_{bot,in})}{A_c \times I_{AM1.5}} \quad (26)$$

Aperture area A_c was obtained from the total projection area of the concentrator, which is 3.264 m² according to the COMSOL simulation model. \dot{m}_{top} and \dot{m}_{bot} are mass flowrates of top and bottom channels respectively.

2.4 Thermal and Electrical Efficiency Plot

Thermal and electrical efficiencies characteristic tests were done for each model by varying the inlet temperatures of the HTF at either the top or the bottom thermal receiver while keeping other parameters constant. The HTF velocities of top and bottom channels were kept at 0.01 m/s and ambient temperature was kept at 298.15 K throughout the simulation. Electrical efficiencies and Thermal efficiencies are plotted against reduced temperature for evaluation, where the reduced temperature T_r was calculated by:

$$T_r = \frac{(T_{out} - T_{amb})}{G_0} \quad (27)$$

where T_{out} is outlet temperature of thermal receiver HTF.

3. Results and Discussion

3.1 SBS CPVT Model Validation

The coupled electrical and thermal model was validated against the work from Bahaidarah et al. [38]. Parameters such as location, time, temperature, tilt angle, wind speed, irradiance, and ambient electrical efficiency at STC conditions from the study were used as input parameters of the model for PV cell temperature estimation. Figure 3.1 shows the comparison of cell temperatures between the PV-T COMSOL simulation and the experimental results of the unglazed photovoltaic compound parabolic concentrator (PV-CPC) system. A relative error below 8% was obtained which indicates a good

match between the simulated thermal and electric model with the experimental results.

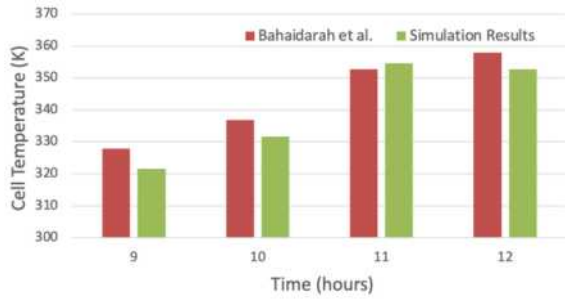


Figure 3.1 Model Validation Against Experimental Results

3.2. Ideal Filter SBS CPVT Collector Efficiency
In this section, the simulation performance of the ideal filter collector is presented and discussed. By varying inlet temperatures from 288.15 K to 328.15 K, outlet temperatures of HTF and the average temperature of the PV channel were obtained from COMSOL and efficiencies for top and bottom channels were calculated. Reduced temperature T_r was calculated using the aforementioned method. Top and bottom channels temperature profiles of the ideal filter collector at an inlet temperature 298.15 K and 323.15 K are presented in Figure 3.2. With the same temperature scale, it can be seen that both channels exhibit a uniform temperature profile and the average HTF temperature is higher with a higher inlet temperature.

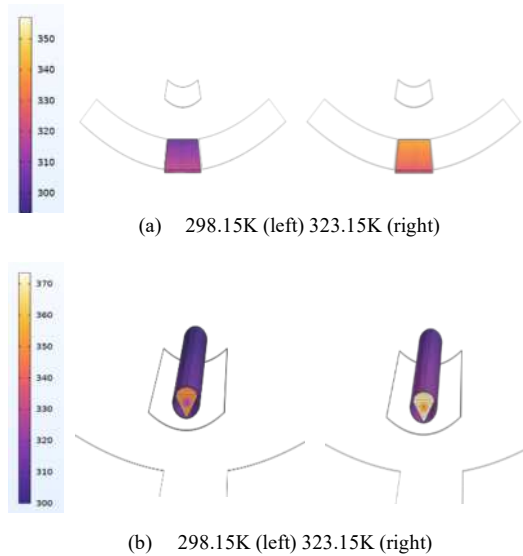


Figure 3.2 Temperature Distribution Profile of SBS CPVT Collector of (a) Top Channel (b) Bottom Channel

Figures 3.3 and 3.4 show the electrical efficiency of PV cells, the thermal efficiency of the top thermal receiver and bottom receiver against the reduced temperature. Electrical efficiency decreased from 17.0% to 13.7% with T_r increases from 0.0044 K/(Wm⁻²) to 0.04736 K/(Wm⁻²). This can be explained by the nature of silicon PV cells, where an increase in PV cell temperature lowers the electrical

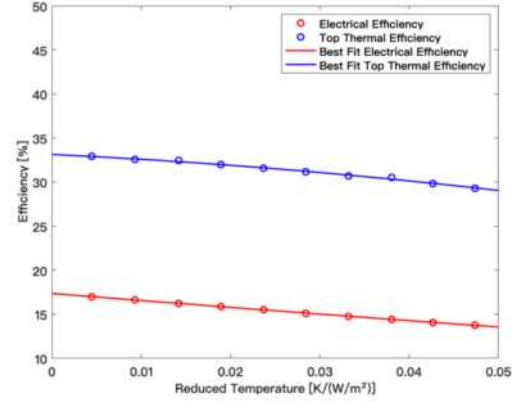


Figure 3.3 Electrical and Thermal Efficiencies against Reduced Temperature for Ideal Filter Collector Top Channel

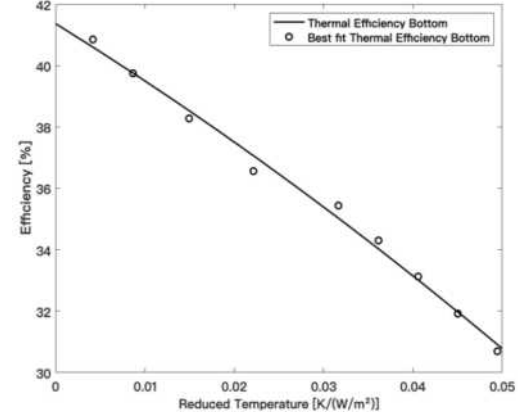


Figure 3.4 Thermal Efficiencies against Reduced Temperature for Ideal Filter Collector Bottom Channel efficiency. The thermal efficiency of the top receiver decreased from 32.9% to 29.3% with T_r increases from 0.0044 K/(Wm⁻²) to 0.04736 K/(Wm⁻²). The thermal efficiency of the bottom thermal receiver at the bottom channel decreases from 40.9% to 30.7% with T_r increases from 0.0042 K/(Wm⁻²) to 0.049 K/(Wm⁻²). The decreases in thermal efficiencies of both receivers at a higher reduced temperature are due to increasing heat loss. However, the heat loss from the bottom receiver is greater than the top part, which is indicated by a larger gradient in thermal efficiency profile. This is because the bottom thermal receiver is not protected by a cover glass, resulting in a more significant convection heat loss than the top thermal receiver.

The PV cells and the bottom thermal receiver are effectively thermally decoupled. Therefore, a higher output temperature can be achieved in the bottom receiver without compromising electrical efficiency. By plotting best fit lines and extrapolate to $T_r=0$ for each efficiency plot, the maximum optical efficiency η_{opt} of the collector could be calculated by

$$\eta_{opt} = \eta_{top} + \eta_{bottom} + \eta_e \quad (28)$$

where η_{top} is efficiency for top thermal receiver, η_{bottom} is the efficiency for bottom thermal receiver, and η_e is the PV cell electrical efficiency. Optimum efficiencies are achieved when $T_r = 0$ due to the negligible heat loss. Table 3.1 presents optimum η_i of collector using ideal filter.

Table 3.1 Optimum Electrical, Top Thermal Receiver, Bottom Thermal Receiver Efficiency of Ideal Filter Collector

η_{top}	η_{bottom}	η_e
33.13%	41.37%	17.36%

For collector using ideal filter, optimum optical efficiency was calculated to be $\eta_{opt}=91.86\%$. η_{top} has a lower value than η_{bottom} at $T_r = 0$, since part of the radiation reflected to the top channel is converted into electricity by the PV cell. Moreover, there is an extra optical loss caused by the glass envelope and the vacuum gap from the top channel.

3.3 Selection of Real-life Optical Filter

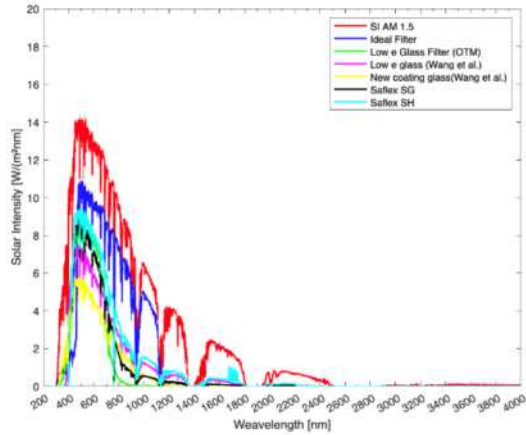


Figure 3.5 Solar Intensity against different wavelengths using different optical filters

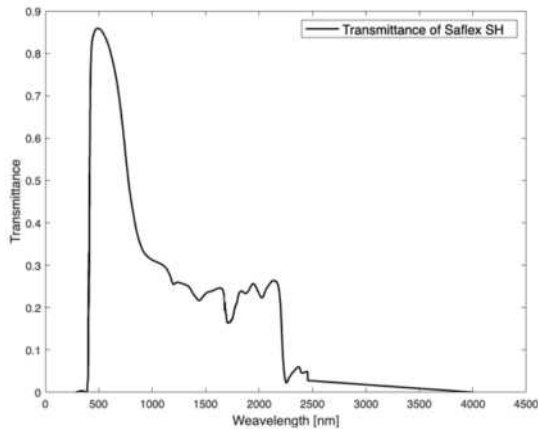


Figure 3.6 Transmittance curve for Saflex SH filter

Table 3.2 Reference Electrical Efficiencies of Different LEGs

Type of LEG	System Electrical Efficiency
OTM Low e Glass	7.1097%
Low e glass (Wang et al)	8.0651%
New Coating Glass (Wang et al)	5.9645%
Saflex SG	7.6042%
Saflex SH	10.4422%

Reference electrical efficiency calculation of SBS CPVT collector using LEGs was carried out using MATLAB. Figure 3.5 describes spectrum reaches the solar cell for ideal dichroic filter and different

types of LEGs. The ideal filter transmittance data was sourced from the research conducted by Wingert et al. [20], and the transmittance data for LEGs was obtained from OTM Solutions Pte Ltd, Wang et al and Saflex [39,40,41,42]. The proportion of the solar spectrum reaching the solar cell is approximately 40%-50% when using LEGs, in comparison to the fraction observed with an ideal dichroic filter. The results of electrical efficiency of PV cells of using test optical filters are presented in Table 3.2. with Saflex SH having the highest efficiency of 10.4422%. Therefore, Saflex SH was selected as the real-life filter for simulation in COMSOL.

Figure 3.6 shows the transmittance curve of Saflex SH. This filter has a high transmittance between 400-1200 nm, which partially aligns with silicon cell's bandgap wavelength 350-1200 nm.

Average transmittance and reflectance in different wavelength ranges are calculated instead of implementing the whole spectrum in COMSOL to save modelling time. The average transmittance between 400-1200 nm is 0.48, while the reflectance for the rest of the spectrum is 0.80.

3.4. Thermal and Electrical Efficiency for Saflex SH Filter SBS CPVT Collector

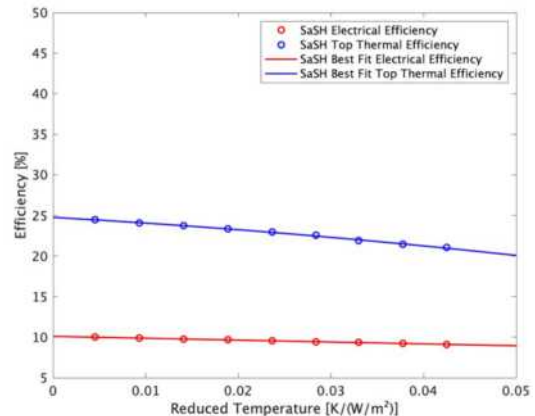


Figure 3.7 Electrical and Thermal Efficiencies against Reduced Temperature for Saflex SH Collector Top Channel

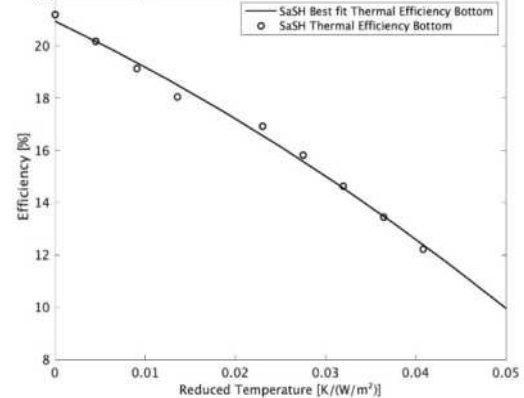


Figure 3.8 Thermal Efficiencies against Reduced Temperature for Saflex SH Collector Bottom Channel
The efficiencies showed a similar trend to ideal filter collector discussed in 3.1. As shown in Figure

3.7 and 3.8, the electrical efficiency and top thermal receiver efficiency decreased from 10.00% to 9.12% and 24.46% to 21.03% respectively with T_r increases from 0.0045 K/(Wm⁻²) to 0.0425 K/(Wm⁻²) for Saflex SH filter. The bottom thermal receiver efficiency decreased from 21.20% to 12.22%, with T_r increases from 0.000002 K/(Wm⁻²) to 0.0408 K/(Wm⁻²). Table.3.3 presents optimum η_i of collector using LEG filter at $T_r = 0$.

For collector using Saflex SH filter, optimum optical efficiency was $\eta_{opt} = 55.81\%$ at $T_r = 0$, which is 36.05% lower than that of the ideal filter collector. The optimum electrical efficiency and top thermal receiver efficiency of Saflex SH filter are 7.26% and 8.36% lower than ideal filter collector's respectively, which is due to the real-life filter having a lower transmittance within the Si-band gap wavelength range.

Table 3.3 Optimum Electrical, Top Thermal Receiver, Bottom Thermal Receiver Efficiency of Saflex SH Collector

η_{top}	η_{bottom}	η_e
24.77%	20.94%	10.1%

On the other hand, the bottom thermal receiver efficiency was 20.94% at $T_r = 0$, which is 20.43% lower than that of the ideal filter collector. This is because Saflex SH has a lower reflectance than ideal optical filter outside the bandgap wavelength of silicon cell, which would be heating up the top channel instead.

5. Conclusions

Numerical modelling of SBS CPVT collector. The optical filter transmits the spectrum within the bandgap of Si-cells (350-1200 nm) while reflects the rest to a thermal receiver. The heat generated by PV cells are absorbed by the top thermal receiver at the back of the solar cell panel. The parabolic trough for this collector had a concentration ratio (CR) of 8.69. By varying HTF inlet temperature, simulations of ideal optical filter collector and Saflex SH filter collector were conducted in COMSOL for top and bottom channels respectively. Both simulations were conducted at an ambient temperature 298.15 K, HTF flowrate 0.01 m/s, wind speed of 1m/s and solar irradiance of 1000 W/m². For ideal filter, spectrum transmittance between 500-1100 nm is 1, and reflectance is 1 at other wavelengths. For Saflex SH filter, average transmittance of filter is 0.48 between 400-1200 nm and average reflectance is 0.80 at other wavelengths.

At reduced temperature $T_r = 0$, the optical efficiency η_{opt} of ideal optical filter collector and Saflex SH collector are 91.68% and 55.81% respectively; top thermal receiver efficiency, bottom receiver efficiency and electrical efficiency (η_{top} , η_{bot} η_e) for ideal filter collector are 33.13%, 41.37% and 17.36% respectively; η_{top} , η_{bot} and η_e for Saflex SH collector are 24.77%, 20.94% and 10.1% respectively. For both collectors,

η_e , η_{top} η_{bot} decreases when T_r increase; η_{top} decreases more rapidly than η_{bot} .

6. Outlook

Some potential improvements are worth considering for a more reliable result, when similar research is conducted in future. A more accurate electrical efficiency at each HTF inlet temperature could be obtained by doing more iterations until each HTF outlet temperature between iterations is less the 1 °C. A more sensible bottom thermal receiver design could be proposed in the future, which could possibly lead to a more realistic result. However, to the structural model complexity may cause a significantly longer simulation time in COMSOL. In addition, implementing the full transmittance profile in COMSOL instead of using the average transmittance for real-life filter could also be explored.

For system performance enhancements, one can explore the following aspects which are not included in this study. In terms of geometry, the curvature and position of the optical filter and parabolic trough could be modified for a compact system design, potentially reducing the space taken and capital cost. Different combinations of filter, thermal receiver, concentrator types and solar cells materials are also worth exploring to achieve a higher optical efficiency. For instance, Zhang et al proposed a Thermal-Electric Generator [34], which could be used as a replacement of the top thermal receiver in this research. This would potentially reduce the operation cost since HTF is not required.

References

- [1] Solar Energy Technologies Office (n.d.). *Solar Energy, Wildlife, and the Environment*. [online] Energy.gov. Available at: <https://www.energy.gov/ee/re/solar/solar-energy-wildlife-and-environment#:~:text=As%20a%20renewable%20source%20of>.
- [2] University of Michigan (2021). *Photovoltaic Energy Factsheet*. [online] Center for Sustainable Systems. Available at: <https://css.umich.edu/publications/factsheets/energy/photovoltaic-energy> factsheet#:~:text=PV%20conversion%20efficiency%20is%20the.
- [3] Widyolar, B., Jiang, L., Abdelhamid, M. and Winston, R. (2018). Design and modeling of a spectrum-splitting hybrid CSP-CPV parabolic trough using two-stage high concentration optics and dual junction InGaP/GaAs solar cells. *Solar Energy*, 165, pp.75–84. doi:<https://doi.org/10.1016/j.solener.2018.03.015>.
- [4] GreenMatch.co.uk. (n.d.). *Solar Power on the Rise: Global Solar Panel Statistics, Facts, and Trends of 2024*. [online] Available at: <https://www.greenmatch.co.uk/solar-energy/solar-pv-statistics>.

- [5] Wang, K., Pantaleo, A.M., Herrando, M., Faccia, M., Pesmazoglou, I., Franchetti, B.M. and Markides, C.N. (2020). Spectral-splitting hybrid PV-thermal (PVT) systems for combined heat and power provision to dairy farms. *Renewable Energy*, 159, pp.1047–1065.
doi: <https://doi.org/10.1016/j.renene.2020.05.120>.
- [6] Kim, J.-H., Park, S.-H., Kang, J.-G. and Kim, J.-T. (2014). Experimental Performance of Heating System with Building-integrated PVT (BIPVT) Collector. *Energy Procedia*, 48, pp.1374–1384.
doi:<https://doi.org/10.1016/j.egypro.2014.02.155>.
- [7] Touafek, K., Khelifa, A. and Adouane, M. (2014). Theoretical and experimental study of sheet and tubes hybrid PVT collector. *Energy Conversion and Management*, 80, pp.71–77.
doi:<https://doi.org/10.1016/j.enconman.2014.01.021>.
- [8] Ni, J., Li, J., An, W. and Zhu, T. (2018). Performance analysis of nanofluid-based spectral splitting PV/T system in combined heating and power application. *Applied Thermal Engineering*, 129, pp.1160–1170.
doi:<https://doi.org/10.1016/j.applthermaleng.2017.10.119>.
- [9] Wang, K., María Herrando, Wang, K. and Markides, C.N. (2019). Technoeconomic assessments of hybrid photovoltaic-thermal vs. conventional solar-energy systems: Case studies in heat and power provision to sports centres. *Applied Energy*, 254, pp.113657–113657.
doi:<https://doi.org/10.1016/j.apenergy.2019.113657>
- [10] Mohit Barthwal and Dibakar Rakshit (2022). Holistic opto-thermo-electrical analysis of a novel spectral beam splitting-based concentrating photovoltaic thermal system. *Journal of Cleaner Production*, 379, pp.134545–134545.
doi:<https://doi.org/10.1016/j.jclepro.2022.134545>.
- [11] Liang, H., Han, H., Wang, F., Cheng, Z., Lin, B., Pan, Y. and Tan, J. (2019). Experimental investigation on spectral splitting of photovoltaic/thermal hybrid system with two-axis sun tracking based on SiO₂/TiO₂ interference thin film. *Energy Conversion and Management*, 188, pp.230–240.
doi:<https://doi.org/10.1016/j.enconman.2019.03.060>.
- [12] Huang, G. and Markides, C.N. (2021). Spectral-splitting hybrid PV-thermal (PV-T) solar collectors employing semi-transparent solar cells as optical filters. *Energy Conversion and Management*, 248, p.114776.
doi:<https://doi.org/10.1016/j.enconman.2021.114776>.
- [13] Ju, X., Xu, C., Liao, Z., Du, X., Wei, G., Wang, Z. and Yang, Y. (2017). A review of concentrated photovoltaic-thermal (CPVT) hybrid solar systems with waste heat recovery (WHR). *Science Bulletin*, 62(20), pp.1388–1426.
doi:<https://doi.org/10.1016/j.scib.2017.10.002>.
- [14] Liew, N.J.Y., Yu, Z. (Jason), Holman, Z. and Lee, H.-J. (2022). Application of spectral beam splitting using Wavelength-Selective filters for Photovoltaic/Concentrated solar power hybrid plants. *Applied Thermal Engineering*, 201, p.117823.
doi:<https://doi.org/10.1016/j.applthermaleng.2021.117823>.
- [15] Mojiri, A., Stanley, C., Taylor, R.A., Kalantar-zadeh, K. and Rosengarten, G. (2015). A spectrally splitting photovoltaic-thermal hybrid receiver utilising direct absorption and wave interference light filtering. *Solar Energy Materials and Solar Cells*, 139, pp.71–80.
doi:<https://doi.org/10.1016/j.solmat.2015.03.011>.
- [16] Adam, S.A., Ju, X., Zhang, Z., Abd El-Samie, M.M. and Xu, C. (2019). Theoretical investigation of different CPVT configurations based on liquid absorption spectral beam filter. *Energy*, 189, p.116259.
doi:<https://doi.org/10.1016/j.energy.2019.116259>.
- [17] O. Elharoun, Tawfik, M., El-Sharkawy, I.I. and Zeidan, E.-S.B. (2023). Experimental investigation of photovoltaic performance with compound parabolic solar concentrator and fluid spectral filter. *Energy*, 278, pp.127848–127848.
doi:<https://doi.org/10.1016/j.energy.2023.127848>.
- [18] Ju, X., Xu, C., Han, X., Du, X., Wei, G. and Yang, Y. (2017a). A review of the concentrated photovoltaic/thermal (CPVT) hybrid solar systems based on the spectral beam splitting technology. *Applied Energy*, 187, pp.534–563.
doi:<https://doi.org/10.1016/j.apenergy.2016.11.087>.
- [19] Daneshazarian, R., Cuce, E., Cuce, P.M. and Sher, F. (2018). Concentrating photovoltaic thermal (CPVT) collectors and systems: Theory, performance assessment and applications. *Renewable and Sustainable Energy Reviews*, 81, pp.473–492.
doi:<https://doi.org/10.1016/j.rser.2017.08.013>.
- [20] Wingert, R., O'Hern, H., Orosz, M., Harikumar, P., Roberts, K. and Otanicar, T. (2020). Spectral beam splitting retrofit for hybrid PV/T using existing parabolic trough power plants for enhanced power output. *Solar Energy*, 202, pp.1–9.
doi:<https://doi.org/10.1016/j.solener.2020.03.066>.
- [21] Zhang, J.J., Qu, Z.G., Wang, Q., Zhang, J.F. and He, Y.L. (2021). Multiscale investigation of the plasmonic solar cell in the spectral splitting concentrating photovoltaic-thermal system. *Energy Conversion and Management*, [online] 250, p.114846.
doi:<https://doi.org/10.1016/j.enconman.2021.114846>.
- [22] Wang, G., Wang, B., Yao, Y., Lin, J., Chen, Z. and Hu, P. (2020a). Parametric study on thermodynamic performance of a novel PV panel and thermal hybrid solar system. *Applied Thermal Engineering*, 180, p.115807.
doi:<https://doi.org/10.1016/j.applthermaleng.2020.115807>.

- [23] Zhang, L. et al. (2024) 'Performance investigation on a concentrating photovoltaic thermal system integrated with spectral splitter and absorption heat pump', *Applied Thermal Engineering*, 237, p.121772. doi:10.1016/j.applthermaleng.2023.121772.
- [24] Ling, Y., Li, W., Jin, J., Yu, Y., Hao, Y. and Jin, H. (2020). A spectral-splitting photovoltaic-thermochemical system for energy storage and solar power generation. *Applied Energy*, 260, p.113631. doi:https://doi.org/10.1016/j.apenergy.2019.113631
- [25] Zhang, G., Wei, J., Xie, H., Wang, Z., Xi, Y. and Khalid, M. (2018). Performance investigation on a novel spectral splitting concentrating photovoltaic/thermal system based on direct absorption collection. *Solar Energy*, 163, pp.552–563. doi:https://doi.org/10.1016/j.solener.2018.02.033.
- [26] Zhang, D., Wu, Y., Russo, J.M., Gordon, M., Vorndran, S. and Kostuk, R.K. (2014). Optical performance of dichroic spectrum-splitting filters. *Journal of Photonics for Energy*, 4(1), pp.043095–043095. doi:https://doi.org/10.1117/1.jpe.4.043095.
- [27] Widyolar, B., Jiang, L. and Winston, R. (2018). Spectral beam splitting in hybrid PV/T parabolic trough systems for power generation. *Applied Energy*, 209, pp.236–250. doi:https://doi.org/10.1016/j.apenergy.2017.10.078.
- [28] Miles, A., Cocilovo, B., Wheelwright, B., Pan, W., Tweet, D. and Norwood, R.A. (2016). Designing spectrum-splitting dichroic filters to optimize current-matched photovoltaics. *Applied Optics*, 55(8), p.1849. doi:https://doi.org/10.1364/ao.55.001849.
- [29] Lucio, C., Behar, O. and Dally, B.B. (2023). Techno-Economic Assessment of CPVT Spectral Splitting Technology: A Case Study on Saudi Arabia. *Energies*, 16(14), pp.5392–5392. doi:https://doi.org/10.3390/en16145392.
- [30] Zhang, C., Shen, C., Zhang, Y. and Pu, J. (2022). Feasibility investigation of spectral splitting photovoltaic /thermal systems for domestic space heating. *Renewable Energy*, 192, pp.231–242. doi:https://doi.org/10.1016/j.renene.2022.04.126.
- [30] Wang, G., Wang, B., Yao, Y., Lin, J., Chen, Z. and Hu, P. (2020a). Parametric study on thermodynamic performance of a novel PV panel and thermal hybrid solar system. *Applied Thermal Engineering*, 180, p.115807. doi:https://doi.org/10.1016/j.applthermaleng.2020.115807.
- [31] Huang, G., Wang, K., Curt, S.R., Franchetti, B., Pasmazoglou, I. and Markides, C.N. (2021). On the performance of concentrating fluid-based spectral-splitting hybrid PV-thermal (PV-T) solar collectors. *Renewable Energy*, 174, pp.590–605. doi:https://doi.org/10.1016/j.renene.2021.04.070.
- [32] Li, J., Yang, Z., Wang, Y., Dong, Q., Qi, S., Huang, C., Wang, X. and Lin, R. (2023). A novel non-confocal two-stage dish concentrating photovoltaic/thermal hybrid system utilizing spectral beam splitting technology: Optical and thermal performance investigations. *Renewable Energy*, 206, pp.609–622. doi:https://doi.org/10.1016/j.renene.2023.02.078.
- [33] Zhang, Y. and Gao, P. (2022). Hybrid Photovoltaic/Thermoelectric Systems for Round-the-Clock Energy Harvesting. *Molecules*, [online] 27(21), p.7590. doi:https://doi.org/10.3390/molecules27217590.
- [34] C60 SOLAR CELL MONO CRYSTALLINE SILICON C60 SOLAR CELL. (n.d.). Available at: http://eshop.terms.eu/data/s_3386/files/137994254_0-sunpower_c60_bin_ghi.pdf.
- [35] Subedi, I., Silverman, T.J., Deceglie, M.G. and Podraza, N.J. (2019). Emissivity of solar cell cover glass calculated from infrared reflectance measurements. *Solar Energy Materials and Solar Cells*, 190, pp.98–102. doi:https://doi.org/10.1016/j.solmat.2018.09.027.
- [36] Chandan, Baig, H., Asif Ali Tahir, Reddy, K.S., Mallick, T.K. and Bala Pesala (2022). Performance improvement of a desiccant based cooling system by mitigation of non-uniform illumination on the coupled low concentrating photovoltaic thermal units. *Energy Conversion and Management*, 257, pp.115438–115438. doi:https://doi.org/10.1016/j.enconman.2022.115438.
- [37] Liu, X., Mehdi Vahabzadeh Bozorg, Xiong, Q. and Li, W. (2022). Numerical analysis of a new design strategy for parabolic trough solar collectors for improved production and consumption of solar energy in the Belt and Road Initiative countries. *Journal of Cleaner Production*, 367, pp.133079–133079. doi:https://doi.org/10.1016/j.jclepro.2022.133079.
- [38] H. M. Bahaidarah, P. Gandhidasan, A. A. B. Baloch et al., "A comparative study on the effect of glazing and cooling for compound parabolic concentrator PV systems – Experimental and analytical investigations," *Energy Convers Manag*, vol. 129, Dec. 2016, doi: 10.1016/j.enconman.2016.10.028.
- [39] OTM Solutions Pte Ltd. (2021). *Low-e coating and glass optical & thermal performances*. [online] Available at: <https://www.otm.sg/glass-low-e-coating> [Accessed 1 Dec. 2023].
- [40] kierantimberlake.com. (2014). *Investigating Low-E Coating Technology for Glass*. [online] Available at: <https://kierantimberlake.com/updates/investigating-low-e-coating-technology>.
- [41] Wang, D., Lu, L. and Zhang, W. (2019). Overall Energy Performance Assessment of a New Heat Blocking Coating. *Journal of Sustainable Development of Energy, Water and Environment Systems*, 7(1), pp.1–12. doi:https://doi.org/10.13044/j.sdewes.d6.0224.
- [42] Saflex. (2017). *Saflex® Solar*. [online] Available at: <https://www.saflex.com/solar> [Accessed 9 Dec. 2023].

The Effect of Hydrodynamics on the Transesterification of Sunflower Oil to Produce FAME

Ajai Gill and Isabel Solomon

Department of Chemical Engineering, Imperial College London, U.K.

Abstract

Hydrodynamics is an important step in the transesterification of sunflower oil for Fatty Acid Methyl Ester (FAME) production. Agitation is important to overcome the mass transfer limitations associated with the immiscible nature of the reactants, sunflower oil and methanol, to maximise the yield of FAME. However, the precise link of the hydrodynamics with the chemical kinetics is poorly understood. Here we show the effect of hydrodynamics on the FAME yield, interfacial surface area and energy consumption. By measuring the formation of FAME over time, we were able to directly link how the interfacial surface area controls the rate and yield of FAME. We found that by varying the impeller type, agitation speed and clearance, a 6-bladed Rushton Turbine, with an agitation speed of 600 RPM, and a clearance of 0.75 C/H, was optimal to overcome the mass transfer limitations compared to a 3-bladed Marine Impeller. This is due to the creation of smaller dispersed methanol droplets, increasing the interfacial surface area. Furthermore, we established how the formation of mono and diglyceride intermediates, directly affects the stabilisation of microdroplets, and how stopping the agitation after the stabilisation, can lead to large energy savings of 83% for FAME production. Our results demonstrated that the type of impeller is the most significant factor affecting FAME production.

1. Introduction

Global biodiesel production has increased significantly over the last two decades. In 2022, total biodiesel production reached over 1.9 million barrels of oil equivalent per day, and the largest producers are the United States, Brazil, and Indonesia [1]. The global biofuel market is expected to reach more than \$200 billion by 2030 [2]. Biodiesel has up to 74% less carbon dioxide emissions than conventional fossil diesel. It is used as a drop-in fuel and blended into diesel to reduce its overall lifecycle emissions [3].

Biofuels such as Fatty Acid Methyl Esters (FAME), are predominantly produced from waste oils and fats [4]. The production of FAME is a relatively simple process involving the transesterification of oils and fats, high in triglycerides, with methanol, to produce FAME and glycerol. Most commercial FAME production is done batch-wise through transesterification in stirred tanks at atmospheric pressure, and a temperature between 40 to 60°C [5]. Mixing has been identified as a crucial processing step, due to the mass transfer limitations caused by the immiscibility between the polar methanol and non-polar oil phase. However, there has been a lack of research into the effects of hydrodynamics, with previous studies only focusing on the effect of agitation speed on FAME production [6].

This paper focuses on the effect of hydrodynamics, including the impeller type, agitation speed, and tank geometry, on FAME production. If a noticeable effect were to be observed at the laboratory scale, this would have far-reaching impacts on the commercial biodiesel industry.

2. Background

2.1 Transesterification

Transesterification is a reaction between triglycerides and methanol, to produce FAME in the presence of a catalyst. It consists of three consecutive stepwise reversible reactions, where mono and diglycerides are produced as intermediates and glycerol as a byproduct.

From stoichiometry, one mole of triglyceride reacts with three moles of methanol to produce three moles of FAME and one mole of glycerol, as seen in *Figure 1*.

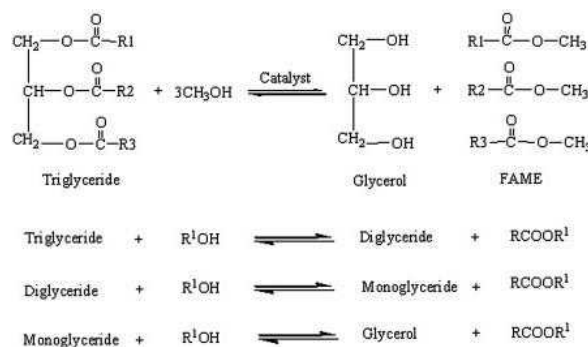
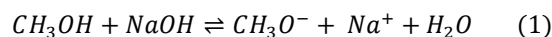


Figure 1. Stepwise transesterification of triglycerides, forming intermediates of mono and diglycerides with methanol in the presence of a catalyst, to produce FAME and glycerol [7]

A base catalyst is used, as opposed to an acid catalyst, due to its high reactivity and its ability to enhance nucleophilic attack. Sodium hydroxide is dissolved in methanol to produce a methoxide ion (CH_3O^-), which is a powerful nucleophile. The methoxide ion (CH_3O^-) is formed from the dissociation of the catalyst, as shown in *Equation 1* [8].



However, the presence of water generated by the disassociation of the catalyst can lead to saponification, which is the production of soap by-products due to the reaction between FAME and water, as shown in *Equation 2* [8].



2.2 Nature of the droplet system over time

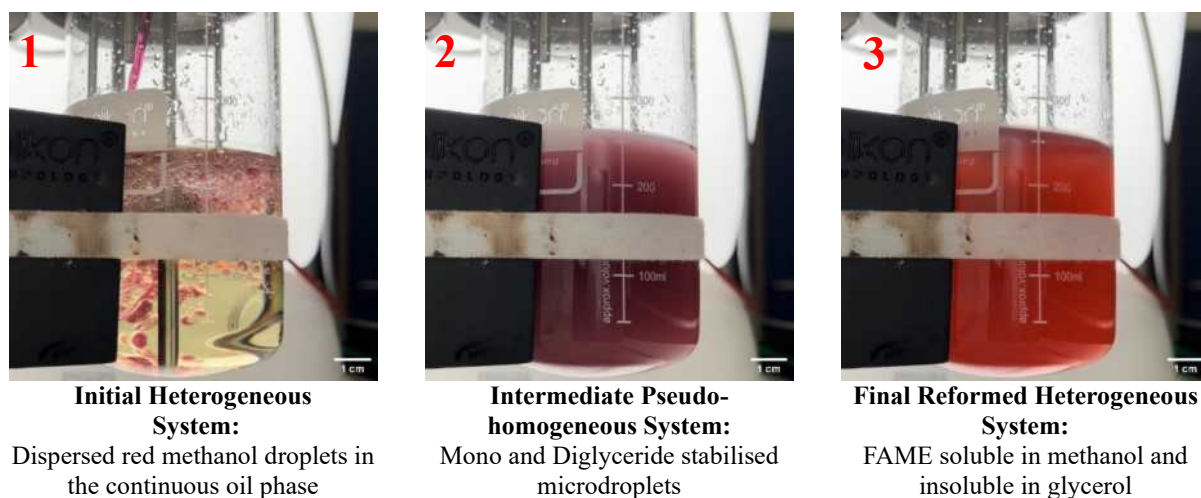


Figure 2. Photographs of the changes in phase nature of the transesterification of sunflower oil over time. A Sudan III dye has been added to methanol and catalyst solution, which turns red in the presence of FAME, to aid the differentiation in the phase of the system

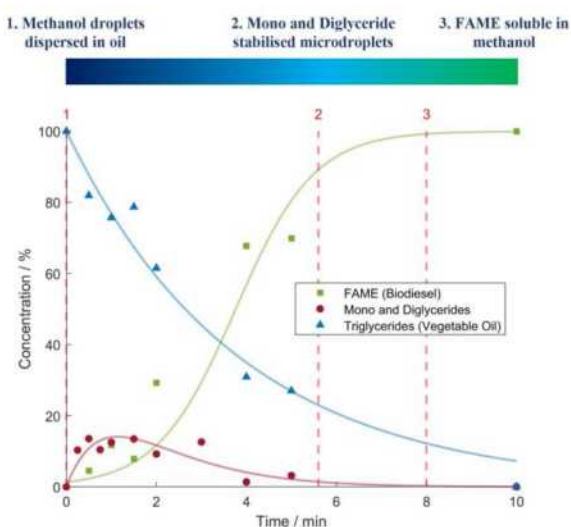


Figure 3. Overall kinetic concentration profile of transesterification reaction for the Marine Impeller at an agitation speed of 600 RPM and a clearance of 0.33 C/H. Positions 1, 2 and 3 directly relate to the system photographs seen in Figure 2

From the kinetic concentration profiles, generated from the experimental data, the change in reactant, intermediates, and product concentration can be observed. The sigmoidal shape of FAME is explained by an initial mass transfer-controlled region followed by a kinetically controlled region [9]. From Figures 2 and 3, initially, the reaction system is heterogeneous consisting of two immiscible phases, a dispersed polar methanol phase and a continuous nonpolar triglyceride phase. The methanol and catalyst solution exist as dispersed droplets due to the formation of reverse micelles. The triglyceride possesses a hydrophilic head that surrounds the methanol solution, and the hydrophobic tails are exposed [10]. Therefore, mass transfer limitations exist due to the immiscibility of methanol and oil, which can be overcome with sufficient agitation. Mixing aids the reaction by creating

dispersion, which increases the interfacial area between the phases, directly controlling the rate of reaction [6].

As the reaction progresses, the mono and diglyceride intermediates are formed. These intermediates are soluble in methanol and act as surfactants to stabilise microdroplets, which begin to remove the phase boundary, forming a pseudo-homogenous system [11]. At the end of the reaction, the byproduct glycerol, is produced, and a heterogeneous system is reformed. Glycerol is a polar molecule and is insoluble in non-polar FAME, and this leads to phase separation.

However, if the agitation is too high, this may lead to the formation of a stable emulsion making it difficult to separate FAME from glycerol, leading to a more energy-intensive separation [12].

3. Methodology

3.1 Feedstocks and reaction conditions

Sunflower oil (KTC) was chosen as the feedstock due to its high concentration of triglycerides, low cost, and widespread availability [13]. Methanol and sodium hydroxide were used as the alcohol and catalyst respectively. A methanol-to-oil molar ratio of 6:1 was employed to ensure there was an excess of methanol to drive the reaction to produce FAME. A 0.75 wt% catalyst loading was chosen to minimise saponification. The reaction occurred at atmospheric pressure and a temperature of 40°C [14].

3.2 Experimental set-up

3.2.1 Reactor configuration

An unbaffled 500ml 'Applikon Minibio Reactor M2' was used and kept at a constant temperature using the integrated heating pad and thermometer. The lid and ports were sealed to minimise methanol loss to the atmosphere due to evaporation. Two ports remained open to introduce the reactants and take samples as the reaction progressed.

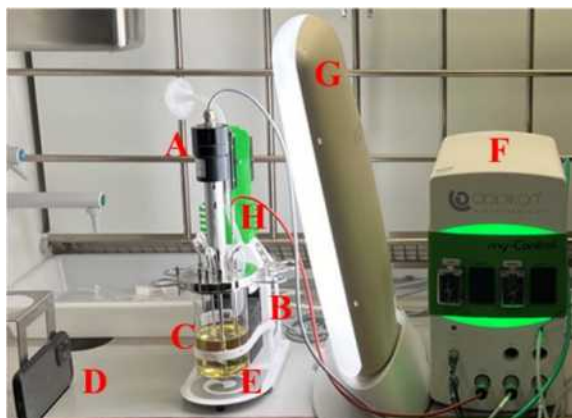


Figure 4. Experimental Set-up. A) Motor, B) Heating Plate, C) 'Applikon MiniBio Reactor M2', D) Camera, E) Impeller, F) 'Applikon my-Control', G) Backlight, H) Thermometer

3.2.2 Standard geometry

The tank geometry must be in standard dimensionless configurations to ensure suitability for scale-up to industry. Standard configurations were applied to the reactor, as seen in *Figure 5*, and this exists when the height of the liquid level is approximately equal to the tank diameter.

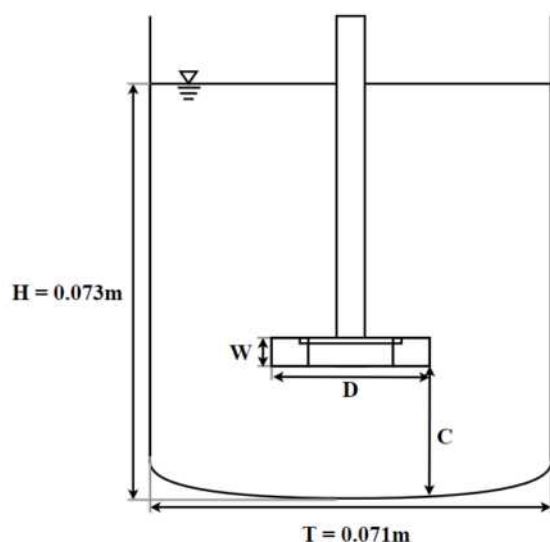


Figure 5. Reactor standard geometry

The clearance is the ratio of the impeller height from the bottom of the tank to the height of the liquid level (C/H). A standard clearance of 0.33 C/H is recommended [15]. However, other sources recommend placing the impeller at the interface for liquid-liquid systems [12]. Three clearances were investigated for each impeller as shown in *Table 1*:

Table 1. Impeller clearance ratios used in experiments.

	Standard	Mid-Point	Interphase
Rushton Turbine	0.33 C/H	0.42 C/H	0.75 C/H
Marine Impeller	0.33 C/H	0.51 C/H	0.62 C/H

3.2.3 Impeller design

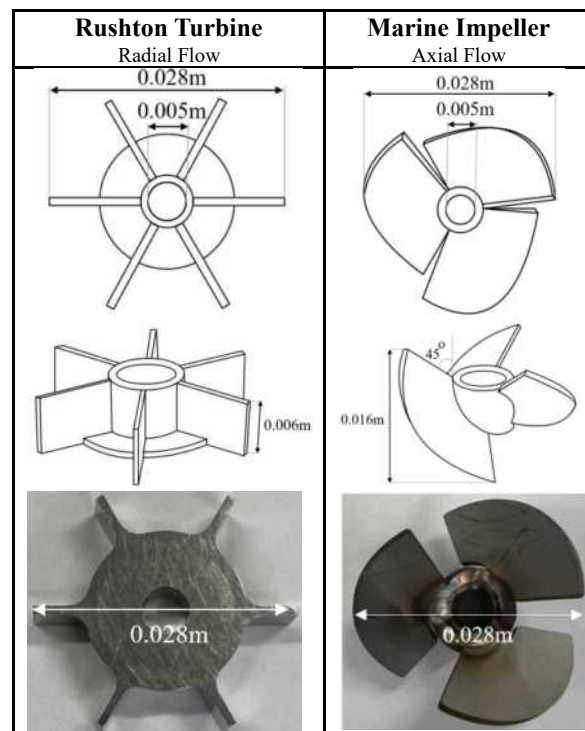


Figure 6. Mechanical drawings and images of the Rushton Turbine and the Marine Impeller with labelled dimensions

The choice of impeller is essential for promoting efficient mixing and distribution of reactants. Common impellers used for liquid-liquid reactions are Disk Turbines and Pitched Impellers [12]. Hence, two different impellers were selected, a 6-bladed Rushton Turbine and a 3-bladed Marine Impeller to satisfy the recommendations respectively, as seen in *Figure 6*.

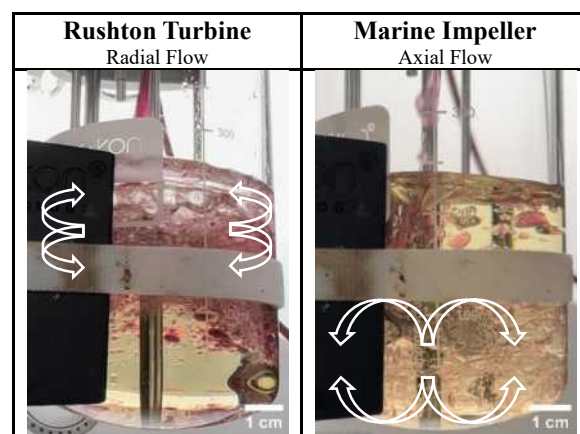


Figure 7. Radial and axial flow patterns created by the Rushton Turbine (600 RPM and 0.75 C/H) and Marine Impeller (600 RPM and 0.33 C/H) respectively. A Sudan III dye has been added to methanol to improve droplet identification

Axial flow is where the fluid flows parallel to the shaft, therefore the fluid rises and falls colliding with the bottom of the vessel. Whereas radial flow is the movement of fluid perpendicular to the shaft, hence the fluid collides with the vessel wall [12]. A Rushton Turbine promotes radial flow whereas a Marine Impeller

causes axial flow, as seen in Figure 7. Therefore, the different mixing types could be explored to evaluate the most effective. It is recommended that the ratio of the diameter of the impeller (D) to the tank diameter (T) is within, $0.2 < D/T < 0.5$ [12]. For both impellers, D/T was equal to 0.4. This provides space for flow patterns to form [15].

3.3 FAME synthesis

3.3.1 Experimental procedure for FAME synthesis

The speed of the agitator was controlled using the 'Applikon my-Control' web UI and varied between 200 to 1000 RPM, for each impeller type and clearance. Sunflower oil (200 g) was added into the bioreactor and heated to 40°C. Sodium hydroxide (1.85 g) was dissolved in methanol (44 g) to produce a 0.75 wt% catalyst solution and preheated to 40°C. The dissolved catalyst mixture was added into the reactor and a timer and camera were started. Ten samples (1 ml) were taken from the reaction mixture over 10 minutes, using syringes. The samples were transferred into Eppendorfs and immediately neutralised with acetic acid (0.25 ml) to quench the reaction. The samples were centrifuged (6000 RPM for 8 minutes) using an 'IKA mini G'. The upper FAME layer was pipetted out and transferred into new Eppendorfs. Deuterated chloroform (0.5 ml) was added, and the samples were then pipetted into NMR tubes.

3.3.2 Analytical method for FAME content using 1H NMR spectroscopy

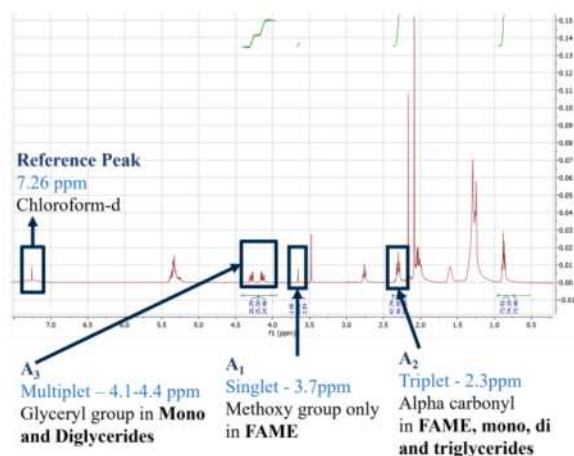


Figure 8. 1H NMR spectrum of upper FAME layer with peak identification

The yield of FAME was calculated using 1H NMR spectroscopy. The spectra, seen in Figure 8, were produced by the 'JeolJNM-ECZS 400-MHz Routine NMR Spectrometer' and analysed using 'MestreNova'. The chloroform peak appeared as a singlet at 7.26 ppm and was used as a reference point when calibrating the spectrum.

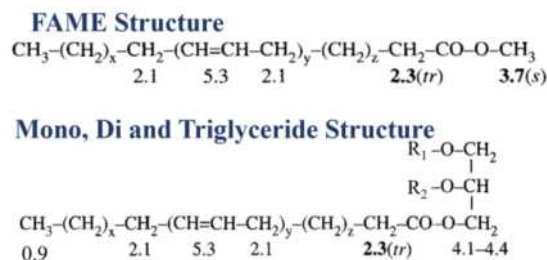


Figure 9. Chemical structure of FAME and mono, di and triglycerides, with the corresponding signals, in ppm, present in the NMR spectrum [16]

An integration was completed across the singlet at 3.7 ppm (A_1), relating to the methoxy group found only in FAME, as seen in Figure 9. An integration was also completed across the triplet found at 2.3 ppm (A_2), which relates to the alpha carbonyl which exists in all the substances present in the upper phase, specifically FAME, mono, di, and triglycerides. The FAME content can be quantified by comparing the areas of the peaks and normalising by the number of hydrogens [16]. It is calculated as shown in Equation 3.

$$FAME \text{ Content (\%)} = \frac{2A_1}{3A_2} \times 100 \quad (3)$$

An estimation of the mono and diglyceride content was calculated following a similar procedure. The multiplet between 4.1 and 4.4 ppm (A_3) corresponds to the glyceryl group and was also normalised by the number of hydrogens associated with the peak. It is calculated as displayed in Equation 4.

$$Mono \text{ and Diglyceride Content (\%)} = \frac{2A_3}{5A_2} \times 100 \quad (4)$$

Manual integration of all areas was repeated three times to calculate an average and standard deviation.

3.4 Droplet size and interfacial surface area

3.4.1 Experimental procedure for droplet size measurement

Two experiments were conducted to observe the effects of hydrodynamics on the dispersed methanol droplets in the reactor. The first experiment, with the sodium hydroxide catalyst, the reactive system, was carried out to observe the effects of the reaction with dispersion. The second experiment, without the presence of a catalyst, the non-reactive system, was performed to only observe the effects of dispersion. The impeller type, speed, and clearance were changed, and a video was taken of the first 10 seconds using the camera from an 'iPhone 13'. A backlight was used to improve the image quality of the droplets and ten random frames were selected.

3.4.2 Analytical method for droplet diameter

An open-source software, 'Bubble Analyser', was used to quantify the droplet diameters and distribution. A background correction image was used to remove artifacts, reducing the error in measurement.

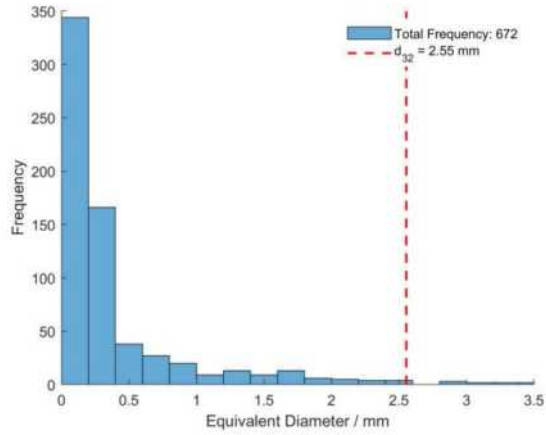


Figure 10. Histogram of droplet diameter generated by ‘Bubble Analyser’ software for the Marine Impeller at 600 RPM and a clearance of 0.33 C/H

A histogram showing a droplet size distribution was generated, seen in *Figure 10*, and from this, the Saunter Mean Droplet Diameter, d_{32} , was calculated, along with the standard deviation. It is a measure of average droplet size and is equivalent to the diameter of a sphere that has the same volume-to-surface area ratio [12]. It is a common measure of droplet size for liquid-liquid dispersion and is calculated using *Equation 5*.

$$d_{32} = \frac{\sum n_i d_i^3}{\sum n_i d_i^2} \quad (5)$$

Where n_i is the number of droplets and d_i is the nominal diameter of the droplets.

3.4.3 Analytical method for the interfacial surface area
The saunter mean droplet diameter was used to calculate the interfacial surface area, a_v [17]. The interfacial surface area governs the rate of reaction and is calculated from *Equation 6*.

$$a_v = \frac{6\phi}{d_{32}} \quad (6)$$

Where ϕ , is the volume fraction of the dispersed methanol phase. This was determined from photographic images.

3.5 Energy Consumption

3.5.1 Experimental procedure for energy consumption
During the FAME synthesis reactions, the total electrical energy consumption of the reactor, E_T , was measured for both impellers at each speed and clearance, using an ‘Energenie Energy Saving Power Meter’, with a resolution of 0.001 kWh.

3.5.2 Analytical method for impeller efficiency

Efficiency is defined as the useful power output over the total electrical energy consumed [18]. The useful power output is the power of the impeller supplied to the fluid. This is determined from the flow regime and the Reynolds number.

The Reynolds number is calculated by applying *Equation 7*.

$$N_{Re} = \frac{\rho N D^2}{\mu} \quad (7)$$

Where the density, ρ , and viscosity, μ , of the reactant mixture were averaged on a mass basis at 40°C, using the initial composition. The Reynolds number for all agitation speeds, from 200 to 1000 RPM, corresponded to a Reynolds number between 87 and 434 respectively. This is entirely within the transitional regime, which occurs at Reynolds numbers in the range of 10 to 10⁴ [12]. The power output of the impeller is calculated using *Equation 8*.

$$P = \rho N_p N^3 D^5 \quad (8)$$

Where N , is the impeller speed, measured in revolutions per second and D , is the diameter of the impeller. The power number, N_p , is determined using correlations from literature for unbaffled reactors in the transitional regime [19]. The power number for the 6-bladed Rushton Turbine is given by *Equation 9*.

$$N_p = 12.2 N_{Re}^{-0.241} \quad (9)$$

The power number for the 3-bladed Marine Impeller is given by *Equation 10*.

$$N_p = 3.77 N_{Re}^{-0.193} \quad (10)$$

Lastly the impeller efficiency is determined from *Equation 11*.

$$\eta = \frac{P t}{E_T} \quad (11)$$

Where t , is the duration of the FAME synthesis production (10 min).

4. Results and Discussion

4.1 FAME production

4.1.1 Final FAME yield

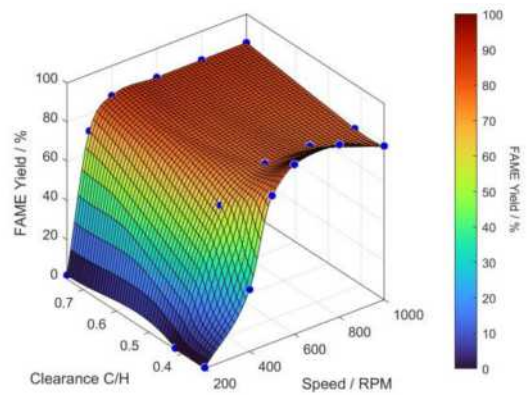


Figure 11. Effect of agitation speed and clearance for the Rushton Turbine on the final FAME yield

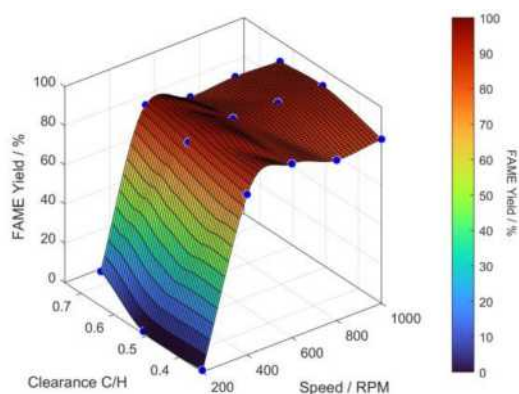


Figure 12. Effect of agitation speed and clearance for the Marine Impeller on the final FAME yield

There is a strong relationship between the final FAME yield and agitation speed for the Rushton Turbine, *Figure 11*, and the Marine Impeller, *Figure 12*. The FAME yield increases significantly with higher agitation speed and plateaus at speeds higher than 600 RPM for both impellers. For the Rushton Turbine, the yield is highest at a clearance of 0.75 C/H, and this is when the turbine is on the interphase. As methanol is less dense than sunflower oil, the methanol droplets rise and coalesce at the top of the reaction mixture. Hence, placing the Rushton Turbine at the interface generated strong radial flow patterns. This prevented the coalescence of droplets, which allowed for efficient breakup and dispersion, as the number of collisions with the impeller blades and reactor walls increased.

However, for the Marine Impeller, the greatest yield occurred at a clearance of 0.33 C/H. When the impeller was at higher clearances, a stagnant laminar region was present at the bottom of the reactor, which is typical of the transitional regime [12]. Consequently, reducing the impeller clearance resulted in a more uniform dispersion of droplets. This phenomenon occurs as the pitch of the Marine Impeller forces the methanol droplets to break up on collision with the bottom of the reactor, which is characteristic of axial flow. Therefore, reducing the impeller clearance increased the frequency of droplet collisions at the bottom of the reactor, causing a greater breakup and dispersion of methanol.

4.1.2 FAME formation profiles

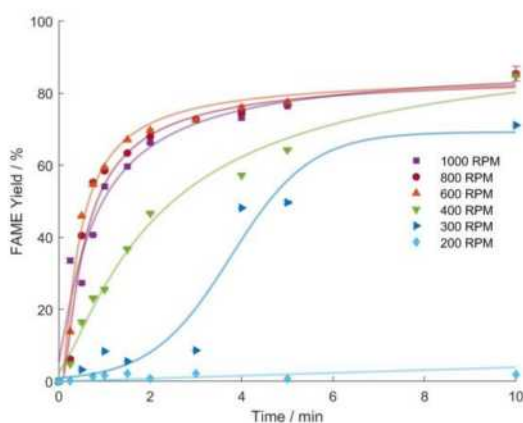


Figure 13. FAME yield against time for varying impeller speeds at an impeller clearance of 0.75 C/H for the Rushton Turbine

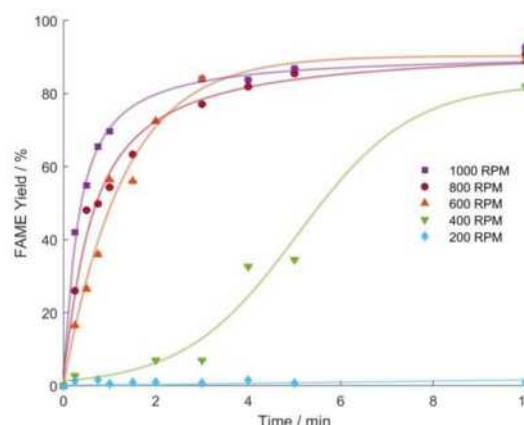


Figure 14. FAME yield against time for varying impeller speeds at an impeller clearance of 0.33 C/H for the Marine Impeller

The FAME yield was monitored throughout the reaction time of 10 minutes and was investigated under varying agitation speeds for the optimum clearance, where the highest yields were produced. For the Rushton Turbine this was at a clearance of 0.75 C/H, *Figure 13*, and at a clearance of 0.33 C/H for the Marine Impeller, *Figure 14*. The FAME formation profiles show that increasing agitation speeds leads to greater yields, since there is a greater dispersion of methanol. There exists a critical agitation speed, between 200 to 300 RPM for the Rushton Turbine and between 200 to 400 RPM for the Marine Impeller. This is the agitation speed required to produce a significant yield of FAME.

A sigmoidal formation profile was seen at 300 and 400 RPM for the Rushton Turbine and Marine Impeller respectively. This is consistent with a slow initial mass transfer-controlled region, followed by a faster kinetically controlled region which plateaus upon reaching an equilibrium, due to a slow reaction rate [9]. At higher agitation speeds, there is a conversion of the final FAME yield since the reactant conditions are constant for all experiments. At higher agitation speeds the kinetic profiles become second order in nature, since there is an elimination of the initial mass transfer-controlled regime, as it has been overcome by the high levels of dispersion [20].

4.1.3 Initial rate of FAME formation

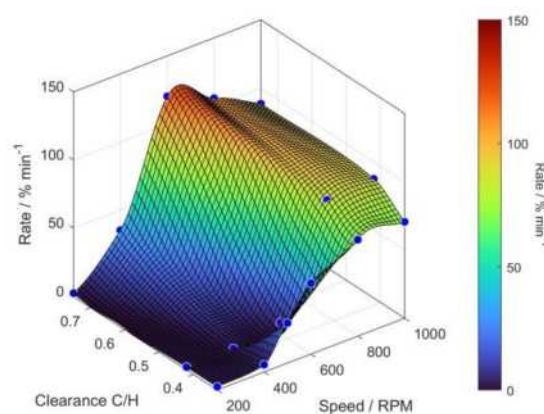


Figure 15. Initial rate of FAME formation against agitation speed and clearance for the Rushton Turbine

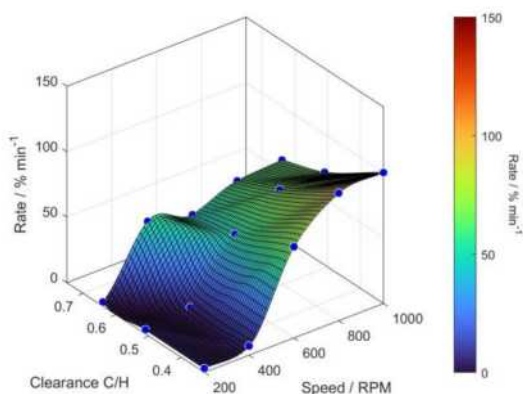


Figure 16. Initial rate of FAME formation against agitation speed and clearance for the Marine Impeller

From the FAME formation profiles, the initial rate of reaction was calculated for varying agitation speeds and clearances for the Rushton Turbine, *Figure 15*, and the Marine Impeller, *Figure 16*. There is a significant rise in the initial rate as the mass transfer region is overcome, and the reaction is dominated by the kinetic terms. The impeller type has the largest effect on the initial rate. The Rushton Turbine is drastically higher than the Marine Impeller, resulting in the faster formation of FAME. For the Rushton Turbine, the maximum rate is seen at an agitation speed of 600 RPM and at a clearance of 0.75 C/H, which aligns with the highest FAME yield seen in *Section 4.1.1*. This is followed by a decrease after 600 RPM which was caused by the formation of a vortex. For the Marine Impeller, the highest rates are seen at a clearance of 0.33 C/H, which also corresponds with the highest FAME yield.

4.2 Droplet size and interfacial surface area

4.2.1 Saunter mean droplet diameter

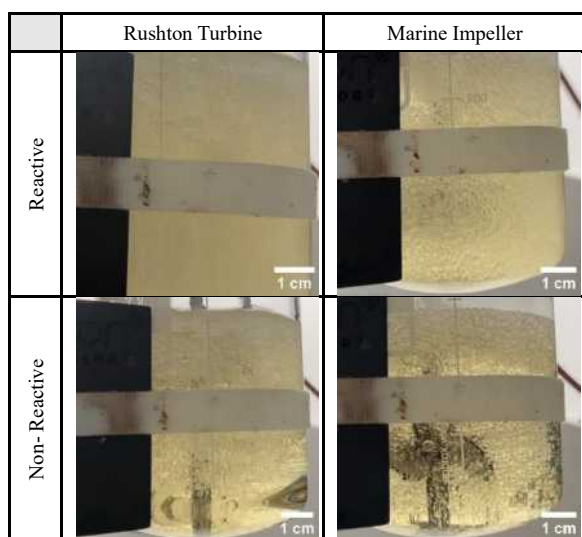


Figure 17. Photographs of the dispersed methanol droplets at 600 RPM at a clearance of 0.75 C/H for the Rushton Turbine, and a clearance of 0.33 C/H for the Marine Impeller, for both the reactive and non-reactive systems

Analysis of the photographic images from the initial 10 seconds show the effect of the droplet sizes with a reaction and dispersion, compared to the system with no

catalyst, to view the effect of dispersion only, seen in *Figure 17*.

The expected droplet size for the reactive system with a catalyst was simulated using the correlation developed by McManamey, *Equation 12* [21].

$$d_{32} = 0.221 \left(\frac{\sigma}{\rho} \right)^{0.6} P_{MI}^{-0.4} \quad (12)$$

Where, σ , is the interfacial surface tension, ρ , is the bulk density and P_{MI} , is the power input per unit mass of the volume swept by the impeller. This was calculated using *Equation 13*.

$$P_{MI} = \left(\frac{4}{\pi} \right) N_p \left(\frac{D}{W} \right) N^3 D^2 \quad (13)$$

The McManamey correlation is a well verified correlation to describe the droplet size for fully dispersed liquid-liquid systems [12].

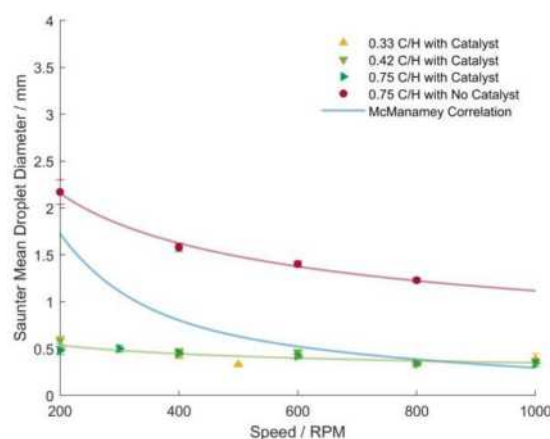


Figure 18. Effect of agitation speed on the saunter mean droplet diameter for reactive systems at varying clearances and a non-reactive system at a clearance of 0.75 C/H for a Rushton Turbine

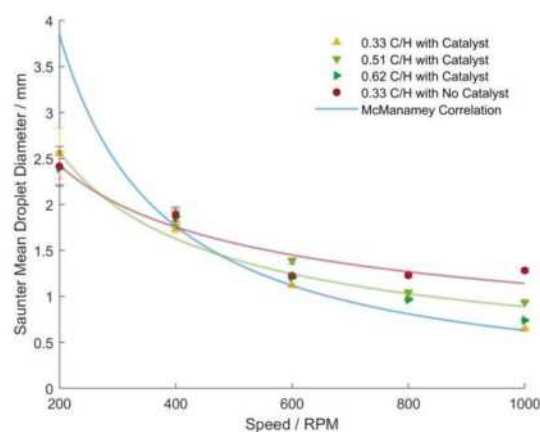


Figure 19. Effect of agitation speed on the saunter mean droplet diameter for reactive systems at varying clearances and a non-reactive system at a clearance of 0.75 C/H for a Marine Impeller

The reactive system is shown for the three clearances, and the non-reactive system at a clearance of 0.75 C/H for the Rushton Turbine, *Figure 18*, and 0.33 C/H for the Marine Impeller, *Figure 19*. There is a trend of decreasing droplet diameter with increasing agitation

speed for both impellers, and this aligns closely with the simulated McManamey droplet sizes when the system is fully dispersed after 600 RPM.

For both the reactive and non-reactive systems the droplet size decreases with agitation speed, as the droplets collide more frequently with the impeller and walls. The droplet size plateaus at high speeds, due to higher surface tension which increases the droplet uniformity and resistance to fragmentation. For the reactive system, the Rushton Turbine has smaller droplet sizes compared to the Marine Impeller for all agitation speeds. The radial flow ensures the droplets collide more frequently with the impeller and walls, preventing coalescence. For the Marine Impeller, initially, there is very poor mixing as demonstrated by the close droplet sizes between the reactive and non-reactive system. However, at higher agitation speeds, clearance has a secondary effect, causing the droplet sizes to be smaller at lower clearances. This is due to the axial flow produced by the Marine Impeller which results in a greater breakup of methanol droplets due to increased collisions with the bottom of the reactor.

4.2.2 Interfacial surface area

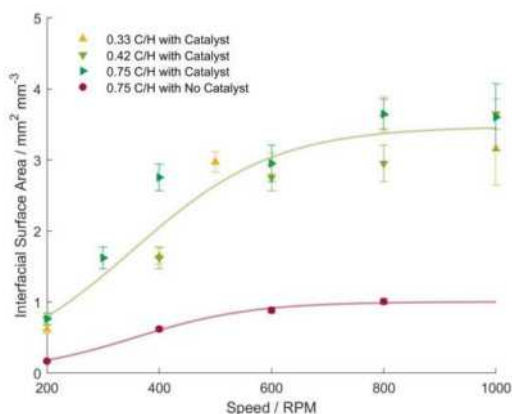


Figure 20. Effect of agitation speed of a Rushton Turbine on the interfacial surface area for reactive systems at varying clearances and a non-reactive system at a clearance of 0.75 C/H

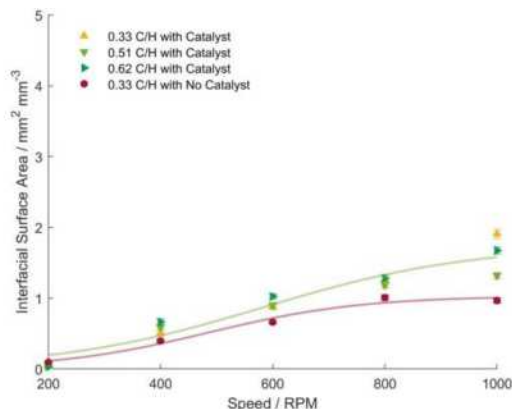


Figure 21. Effect of agitation speed of a Marine Impeller on the interfacial surface area for reactive systems at varying clearances and a non-reactive system at a clearance of 0.33 C/H

As the reaction happens at the interface between triglycerides and methanol, it is important to observe the effect of the hydrodynamics on the interfacial surface

area. A high interfacial surface area is desired since it directly controls the rate of reaction [22].

As the interfacial surface area is inversely proportional to droplet diameter, there is a trend of increasing interfacial surface area with agitation for the reactive and non-reactive systems for the Rushton Turbine, *Figure 20*, and the Marine Impeller, *Figure 21*. The interfacial surface area plateaus at high speeds due to the droplets approaching a constant size. For reactive systems, as the Rushton Turbine produces smaller droplets, there is a significantly higher interfacial surface area than the Marine Impeller. Although both systems follow a similar trend, there is a larger improvement in interfacial surface area, for the reactive system, compared to the non-reactive system, at higher agitation speeds for both impellers. This is due to the higher concentration of mono and diglycerides, acting as surfactants to stabilise microdroplets. Consequently, this increase in interfacial surface area leads to an enhancement in mass transfer and directly explains the higher initial rate of reaction for the Rushton Turbine over the Marine Impeller, as seen in *Section 4.1.3*. Clearance also has a secondary effect when compared to the type of impeller used, which was also seen in *Section 4.2.1*.

The error bars are significantly larger for the reactive system using a Rushton Turbine, due to the smaller droplet sizes and the interfacial surface area being the inverse of the droplet diameter.

4.3 Energy Consumption

4.3.1 Efficiency

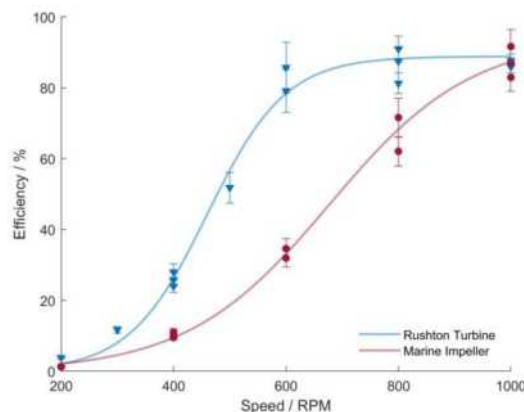


Figure 22. Effect of agitation speed on efficiency for the Rushton Turbine and the Marine Impeller

From *Figure 22*, efficiency increases with agitation speed for both impellers. Additionally, the Rushton Turbine has a higher efficiency at lower agitation speeds. This occurs since the initial rate is higher for the Rushton Turbine, and so there is a faster formation of FAME. Consequently, there is a higher reduction in the overall viscosity of the system since FAME is less viscous than sunflower oil [23]. This reduces the overall resistance faced by the impeller, leading to a lower power consumption for the same agitation speed. The efficiency of the Rushton Turbine plateaus at 600 RPM, signifying that this agitation speed will produce optimal efficiency for the lowest energy consumption.

4.3.2 Effect of agitation time

To minimise energy consumption, agitation can be stopped when mass transfer limitations have been overcome, and this was identified as the time when 90% of the final yield was achieved. Hence, two experiments at the optimal configuration for each impeller were undertaken. The first, when agitation was continuous throughout the 10 minutes, and the second, where agitation was halted when 90% of the final yield was achieved.

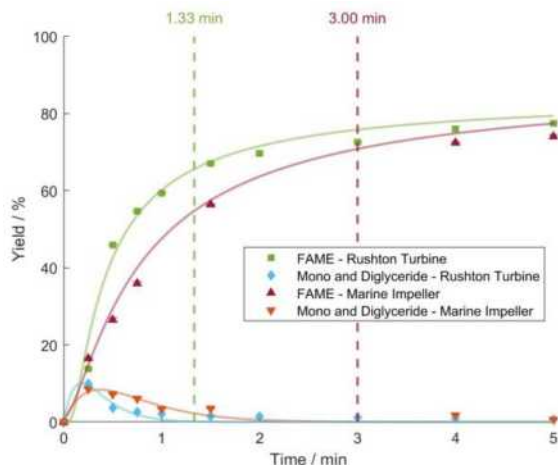


Figure 23. FAME, mono and diglyceride yield over time with identification of the time when agitation was stopped

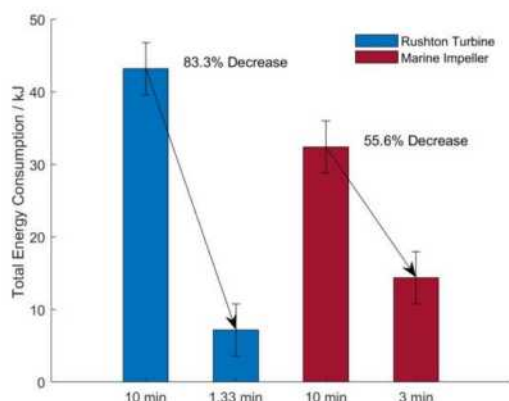


Figure 24. Total energy consumption for the Rushton Turbine and Marine Impeller for the reaction with an agitation time of 10 minutes compared to a reaction with an agitation time corresponding to mass transfer effects being overcome

The optimal configuration for the Rushton Turbine was identified as having an agitation speed of 600 RPM, and a clearance of 0.75 C/H. For the Marine Impeller, the optimal configuration was identified at an identical agitation speed of 600 RPM, but a clearance of 0.33 C/H. For these configurations, the corresponding time to stop the agitation was 80 seconds for the Rushton Turbine and 3 minutes for the Marine Impeller. The reason for this choice is further demonstrated in Figure 23. For the optimal configuration of the Marine Impeller, it has a peak yield of mono and diglyceride intermediates later in time, and they remain higher throughout. This results in the stabilisation of the microdroplets later compared to the Rushton Turbine.

Hence, the Marine Impeller needs a longer agitation time to overcome the mass transfer limitations.

Comparing the total energy consumption, seen in Figure 24, when running the impeller for 10 minutes against the shorter agitation time, the Rushton Turbine had a large reduction in energy consumption of 83.3%, whereas the Marine Impeller had a 55.6% energy reduction. A comparison of the final yields for each of the experiments revealed that both impellers were unaffected by the reduced agitation time. The yields for all four experiments were calculated at 91.4 \pm 0.7 %. Therefore, a reduction in the Rushton Turbine's agitation time produces the greatest decrease in energy consumption, whilst still achieving a very high yield.

5. Conclusion

This research has demonstrated that hydrodynamics has a vital effect on FAME production, specifically the impeller type. By providing a direct link between the hydrodynamics with the chemical kinetics, a molecular foundation was formed. The link between the hydrodynamics with FAME yield, interfacial surface area, and overall energy consumption has been established. The optimal impeller configuration was identified as a 6-bladed Rushton Turbine, at a clearance of 0.75 C/H and an agitation speed of 600 RPM. It is found that the agitation of the impellers could be stopped at 90% of the final yield, as this is when the mono and diglyceride intermediates had stabilised microdroplets. These conditions significantly reduced the overall energy consumption whilst achieving a high yield. This has wide-reaching implications for commercialised biodiesel production and the potential for large energy savings of up to 83%. This investigation has bridged the gap between the impeller type, tank geometry and agitation speed with the chemical kinetics, for the transesterification of sunflower oil to produce FAME.

6. Outlook

There is additional scope for exploration, to optimise the hydrodynamics and minimise total energy consumption. The implementation of baffles, such as beaver tail or finger, would aid in enhancing turbulence by disrupting flow patterns and stopping vortex formation. This could reduce the time for mass transfer limitations to be overcome [12]. Additional impeller types such as Pitched Blade Turbines would also enable the characterisation of mixed flow patterns and the use of hydrofoils could significantly reduce energy consumption due to their lower power numbers [12].

Lastly an investigation into additional reactor configurations could provide a more energy efficient production of FAME. Static inline mixers include obstructions to induce turbulence without agitation, and therefore require no mechanical energy input [12]. An additional agitation mode could also be investigated using an ultrasonic mixer. They produce high and low-pressure waves forming cavitations which produce shockwaves, eliminating the phase boundary. Ultrasonic mixers are more energy efficient as they do not lose energy to frictional heat from rotational equipment [24].

References

- [1] BP. (2023). Leading countries based on biofuel production worldwide in 2022. Statista.
- [2] Precedence Research. (2022). Market value of biofuels worldwide from 2020 to 2022, with a forecast until 2030 (in billion U.S. dollars). Statista.
- [3] Moriarty, K., Milbrandt, A., Lewis, J. and Schwab, A. (2020). 2017 Bioenergy Industry Status Report.
- [4] ETIP Bioenergy. (2023). Fatty Acid Methyl Esters (FAME) Fact Sheet.
- [5] Kianimanesh, H.R., Abbaspour-Aghdam, F. and Derakhshan, M.V. (2017). Biodiesel production from vegetable oil: Process design, evaluation and optimization. *Polish Journal of Chemical Technology*, **19**(3), pp.49–55.
- [6] Stamenkovic, O., Lazic, M., Todorovic, Z., et al. (2007). The effect of agitation intensity on alkali-catalyzed methanolysis of sunflower oil. *Bioresource Technology*, **98**(14), pp.2688–2699.
- [7] Koberg, M. and Gedanken, A. (2013). Using Microwave Radiation and SrO as a Catalyst for the Complete Conversion of Oils, Cooked Oils, and Microalgae to Biodiesel. *New and Future Developments in Catalysis*.
- [8] PennState. (2018). 9.2 The Reaction of Biodiesel: Transesterification | EGEE 439: Alternative Fuels from Biomass Sources.
- [9] Rachmanto. (2014) Monitoring of Biodiesel Transesterification Process Using Impedance Measurement.
- [10] Konda Swathi, Shilpa Kammaradi Sanjeeva, M. Siva, R., et al. (2018). Reverse micelle formation in vegetable oil, 1-butanol and diesel biofuel blends – Elimination of need for transesterification of triglycerides. *Renewable Energy Focus*.
- [11] Alagha, S.M. and Salih, R. (2023). Review the studies of mass transfer and kinetic modelling for production the biodiesel by the transesterification method and the impact of some selected factors. *IOP conference series*, **1232**(1), pp.012014–012014.
- [12] Paul, E.L., Atiemo-Obeng, V.A. and Kresta, S.M. (2004). Handbook of Industrial Mixing. John Wiley & Sons.
- [13] Garcés, R., Martínez-Force, E., Salas, J.J. and Venegas-Calerón, M. (2009). Current advances in sunflower oil and its applications. *Lipid Technology*, **21**(4), pp.79–82.
- [14] A. Singh, B. He, J. Thompson and J. Van Gerpen. (2006). Process optimization of biodiesel production using alkaline catalysts. *Applied Engineering in Agriculture*, **22**(4), pp. 597–600
- [15] EKATO. (2000). Handbook of Mixing Technology.
- [16] Gelbard, G., Brès, O., Vargas, R.M., Vielfaure, F. and Schuchardt, U.F. (1995). ¹H nuclear magnetic resonance determination of the yield of the transesterification of rapeseed oil with methanol. *Journal of the American Oil Chemists' Society*, **72**(10), pp.1239–1241.
- [17] Nagata, S. (1975). Mixing. Halsted Press.
- [18] Ascanio, G., Castro, B. and Galindo, E. (2004). Measurement of Power Consumption in Stirred Vessels—A Review. *Chemical Engineering Research and Design*, **82**(9), pp.1282–1290.
- [19] Zheng Ma. (2014). Impeller Power Draw Across The Full Reynolds Number Spectrum.
- [20] Ezzati, R., Ranjbar, S. and Soltanabadi, A. (2020). Kinetics Models of Transesterification Reaction for Biodiesel Production: a Theoretical Analysis. *Renewable Energy*.
- [21] McManamey, W.J. (1979). Sauter mean and maximum drop diameters of liquid-liquid dispersions in turbulent agitated vessels at low dispersed phase hold-up. *Chemical Engineering Science*, **34**(3), pp.432–434.
- [22] Santacesaria, E., Turco, R., Tortorelli, M., Russo, V., Serio, M.D. and Tesser, R. (2012). Biodiesel process intensification: the role of the liquid-liquid interface area in the achievement of a complete conversion in few seconds. *Green Processing and Synthesis*.
- [23] Borges, M.E., Díaz, L., Gavín, J. and Brito, A. (2011). Estimation of the content of fatty acid methyl esters (FAME) in biodiesel samples from dynamic viscosity measurements. *Fuel Processing Technology*, **92**(3), pp.597–599.
- [24] Veljković, V., et al. (2011). Biodiesel production by ultrasound-assisted transesterification: State of the art and the perspectives. *Renewable and Sustainable Energy Reviews*, **16**(2), pp. 1193-1209

Screening the chemical space: An ML Approach to Predicting the Prices of Chemical Compounds

Cahan O’Driscoll, Denzel Martin Teow

Abstract

Computational techniques in the field of chemical synthesis are constantly improving. As a result, the number of possible molecules is ever-increasing. This makes synthesis of novel compounds highly challenging, as methods are needed to evaluate if production of a molecule is economically desirable compared to others. Current computational methods for determining molecule cost use select features of synthesis routes, but the accuracy of these algorithms could be improved, and they provide limited insight into how these features correlate with cost predictions. In our report, we use a novel approach which utilised a wider range of features as compared to previous literature. These were extracted from a selected retrosynthesis algorithm, AiZynthFinder, and used to train an artificial neural network (ANN) via PyTorch to estimate molecule price. A model was designed and optimised but produced poor prediction accuracy with a mean absolute error of 220 \$/mmol. Results implied that the major contribution to this was an insufficient amount of data points that failed to represent patterns for expensive molecules. Further investigation revealed that the synthesis score input feature was highly redundant and most features were partially correlated, thereby causing lower prediction accuracy. Whereas, the number of precursors was identified as a highly promising model feature as it was shown to have a significant impact on molecule price. These conclusions on feature choice will help guide future research in this field to find the best set of features for accurate price prediction.

Keywords— De-Novo Molecules, Retrosynthesis, Molecular Screening, Machine Learning, Feature Analysis

1 Introduction

In recent decades there has been a stagnation in new drug development in the pharmaceutical industry, with success rates in phase 1 clinical trials remaining at approximately 10% causing drug development investment to increase [1]. In 2004 the estimated cost to bring a new drug to the market was \$800 million and was predicted to double every 5 years[2]. The amount of work that goes into developing novel drugs for most to be discarded is a result of poor resource allocation based on limited insight. For a new drug to be viable for further development, it should meet several feasibility criteria. The criteria of focus for this research is financial feasibility, which projects profitability based on assumed prices for products and cost of production [3]. However, the challenge lies in the absence of reliable methods for accurately predicting the prices of products and by-products [4]. This leads to reduced precision in early-stage molecule screening for eliminating financially unviable options. Identification of new techniques to determine the price of molecules solely based on their structure or synthesis method would therefore prove a highly valuable tool in the early design phases.

Previous research in this field has explored and compared both structural and synthesis route based techniques using a variety of different approaches. Research to identify the most cost-efficient synthesis route has been conducted using the prices of feed stock chemicals and intermediates scaled by reaction yield [5], this inspired a simplified method used by other researchers that only used precursor chemical prices to predict product molecule price. This simplified method was shown to outperform a structural based approach in terms

of prediction accuracy, but could still be improved [6]. Another promising research study found that an artificial neural network was able to learn to predict, with high accuracy, the synthesis scores for compounds based on data about their synthesis routes [7]. The main disadvantage of collecting data on synthesis routes is that it is computationally intensive and slow, the neural network in comparison was shown to be much faster.

These research studies all made use of retrosynthesis algorithms to obtain data about molecule synthesis routes, these are an attractive option for this study as these can be used to make predictions for de-novo molecules. Retrosynthesis analysis is a computational technique where a target molecule is recursively broken down into commercially available precursors, there currently exists retrosynthesis software that utilize various algorithms based on encoded chemistry knowledge to efficiently determine optimal synthesis routes. These algorithms allow the generation of synthesis routes for de-novo molecules based off only their chemical structure and remove the reliance of experienced chemists to create synthesis routes from experience and heuristics while also having the advantage of providing quantifiable justification for choosing an optimal route, such as with a synthesis scoring function[8]. After comparing various retrosynthesis planners for feature extraction, AiZynthFinder was chosen. Other retrosynthesis planners were either inaccessible, such as Synthia, or were too slow, such as Reaxys. In contrast, AiZynthFinder has a relatively fast solving speed, provides in depth documentation [9] and its open-source access allowed for multiprocessing to speed up data generation. This algorithm uses a Monte Carlo search tree to recursively break down molecules into precursors

sors, during synthesis route generation the most promising leaf nodes are expanded and a neural network policy is used to shortlist the possible reaction templates and guide the decomposition of the target molecule [9].

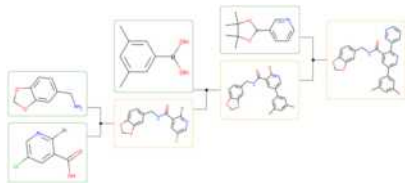


Figure 1: Example of retrosynthesis search tree (output from AiZynthFinder)

The hypothesis investigated within this study is that molecule price can be predicted using correlations with its synthesis route data from retrosynthesis. Price has been predicted before using precursor price, but this research will build upon this literature and aim to achieve higher predictive accuracy using a wider range of synthesis route data, the model will then be analysed to find the correlations between price and each feature. Due to previous reported success of neural network integration with retrosynthesis, this method was incorporated into the research in hopes of generating accurate and fast price predictions.

2 Methodology

2.1 Data Pre-processing

AiZynthFinder software is trained on built-in data sets, such as ZINC the molecule reference database that contains commercially available compounds and is often used for structure-based virtual screening [10]. To increase the efficiency of data generation, Python multiprocessing was introduced to utilize several computer cores simultaneously. Initially, the default AiZynthFinder solving configuration settings were used to generate synthesis routes, but several modifications were made to reduce processing time. The first modification involved analysing each molecule only until the first successful synthesis route was identified, rather than the default approach of finding multiple solved and unsolved routes. The maximum time allocated per molecule was originally 50 seconds as AiZynthFinder can complete searches in under a minute [10], but this was changed to 20 seconds as analysis in the early stages found that 97% of molecules were solved in this time frame. The compiled synthesis routes were then filtered in two stages. The first stage removed unsuccessful solutions based on several criteria: unfeasible synthesis as per scoring function, unrecognized precursors in the ZINC database, no solution found within the allocated time. The second stage discarded any routes that had precursors which could not be cross-referenced with the synthesis cost dataset from Molport [11] as calculating mean, max and min cost would not be possible. To promote efficient training not effected by data distribution, the features were Z-score normalised. To reduce the bias when training and evaluating the neural network, the data was randomly split into training, validation or test using the recommended proportions

7:2:1 [12] respectively. With all data processing complete, data sets for number of precursors, synthesis score, number of reactions, mean precursor price, maximum precursor price and minimum precursor price were ready to be used as model input features.

2.2 Neural Network Training

2.2.1 How a Neural Network Works

This study relied on the application of an artificial neural network (ANN), which are analogous to biological neural networks and contain neurons communicating through a connection network. Python 3.10 and the python library PyTorch were used to create the Neural network models. Figure 2 shows the basic structure of a neural network. To train a neural network, input features must be put into a numerical format, this is represented by the nodes at in the first layer. The number assigned to each node is then multiplied by a weight which is represented by the arrows before being a particular bias is added to it upon reaching a particular node in the next layer to the right. The sum of all these values at each node is passed through an activation function as detailed in Equation 1 below [13] :

$$a(z) = a\left(\sum_{i=1}^n (x_i \times w_i + b)\right) \quad (1)$$

Based on the value of this sum, the activation function will determine if a node will activate. If activated, it will repeat this process and propagate it through all subsequent layers until the output layer. This is known as forward propagation. In the final output layer, a prediction is made. This prediction will be compared against the true value, and a loss function calculates a metric to show how incorrect it is. The level of error is then used to review the model and change the values of the weights and biases. For the purpose of this project, there are known molecule prices as output, therefore this is a supervised machine learning algorithm and the model will learn by minimizing the error between the network’s output and known true results [14].

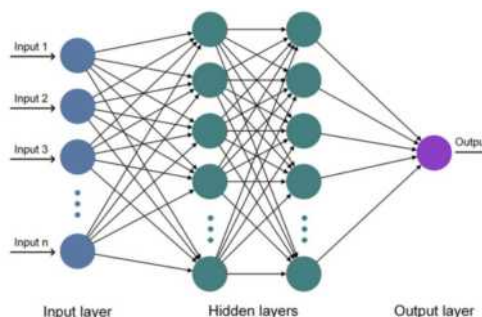


Figure 2: Example of a Neural Network[15]

2.2.2 Base Model Design

The approach taken to design the most suitable neural network for the problem can be separated into two parts. First, a base neural network model is designed to be used as a control

and make preliminary design choices. In conjunction, essential hyperparameters such as batch size, data size and epochs are tuned alongside with regularization technique parameters. In the second part, steps are taken to generate various model architectures and compare their performance to the control model.

Designing the base model involves selecting a model architecture and choosing a loss function and optimiser. The optimal choice of other hyperparameters such as the number of hidden layers, hidden layer size cannot be determined from research as their values depend on the data provided to the neural network, however general design heuristics found in literature were utilized to guide the initial experimentation of network architecture. Heuristics suggest the use of hidden layers with $2n+1$ number of neurons where ‘n’ is number of input features [16], so 13 neurons were first trialed. Theoretical work suggests 1 hidden layer is sufficient to predict nonlinear functions, but research has found using 2 results either causes no improvement or achieves higher efficiency, so for initial trials 2 layers was deemed appropriate [4]. Several layers were not used at the beginning of testing, as too many cause models to overfit [16].

The loss function used for calculating model prediction error during training was mean squared error, seen in equation 2. It is the default choice for PyTorch regression problems, and its quadratic nature incentivizes the removal of large mistakes over small mistakes during training [17]. The optimizing algorithm used to train the algorithm was the stochastic gradient descent method known as the ADAM optimizer because of its high computational efficiency, low memory storage and frequency as a utilized optimiser [18][19]. Heuristics set the learning rate at 0.01 and betas at (0.9, 0.999)[20]. The number of epochs, batch size and dataset size chosen during model training were found through experimenting and based on balancing efficient use of time and minimizing the loss function

$$L = \frac{1}{n} \sum_{i=1}^n (x_i - y_i)^2. \quad (2)$$

As mentioned, this study incorporated various techniques such as drop out, weight decay and early stopping regularization. This is in an effort to prevent overtraining. Drop out causes the random and temporary deactivation of neurons during training, this causes training a different thinned network on each training epoch and causes neurons to be less dependent on their neighbours, resulting in an overall more robust model [21]. The disadvantage this adds to model design is adding another hyperparameter to specify dropout probability. Weight decay reduces the size of the weights and prevents them from becoming too large [22]. Early stopping simply prevents further training once a plateau has been reached in training loss, thereby saving time. As weight decay and dropout effect the model randomly, these were implemented after the best model architecture was found. Further, finding the optimal value for these parameters is a time-intensive process due to the number of possible permutations, so the Hyperopt python Library was used to optimise these. Hyperopt was also used to find the ADAM optimiser parameters of learning rate and betas [23].

2.2.3 Neural Network Architecture Design

Neural network architecture design is essentially varying the number of nodes, hidden layers and the activation function of the layer. Preliminary testing and heuristics indicated the (rectified linear unit) ReLU function as the most viable option compared to functions such as sigmoid, Leaky ReLU, due to its simplicity and efficiency in accelerating the convergence of stochastic gradient descent [24]. To systematically determine the ideal configuration, experiments were conducted by alternately modifying either the number of hidden layers or the number of nodes, while keeping the other parameter constant. For the experiments varying the number of nodes, the use of a single hidden layer was decided to reduce model complexity such that the number of nodes would be the main influencing factor to any changes in performance. The node numbers tested started from 13 and multiplied by 2, up until 13056. The reason for this was that preliminary test runs revealed larger differences in magnitude indicated clearer trends. However, there is no set rule in neural network training that the ideal configuration would consist of the ideal number of layers and the ideal number of nodes within those layers. To stress test this, and the $2n + 1$ and 2 layer heuristic supporting the base model, a further iterative method was designed to generate various model structures based on a maximum number of weight and bias parameters, max number of layers and max number of nodes. The method creates number of node values in powers of 2 up to the max number of nodes, then shuffles those between hidden layers exhaustively to generate the model structures. The optimal model generated is then further optimised by Hyperopt and trained to obtain its average performance data for analysis.

2.3 Model analysis

Neural networks models act like ‘black boxes’ and do not provide much explanation to their reasoning for producing a specific result, so a range of different analysis techniques were incorporated to extract more information on how the models predicted price. The first method used was permutation feature importance, this general analysis aims to identify the relative importance of each feature on model accuracy. This was conducted using the trained optimal model and randomly shuffling each feature separately when making predictions, the average increase in error for each feature indicates how much the model depends on it for its baseline accuracy. A limitation of this method is that it is not a local analysis method that expresses the effect of individual data points, to overcome this barrier a different method was incorporated called SHAP Value analysis. This uses Shapely values from coalition game theory where each value of a feature is ‘player’ for a game where the payout is the predicted output, this breaks down the output predictions into individual contributions from each feature [25] providing more information on how they increase or decrease price predictions. A source of inaccuracy for permutation importance and SHAP value analysis is the in-built assumptions of feature independence, If dependencies do exist, then these analysis techniques will suffer from outlier data points [25]. SHAP values can be used in a clustering function that provide insight into the relationship between input features of a neural network, the clustering algorithms will group features together if they are

below a threshold distance from each other. This distance is scaled from 0 to 1, 0 means the two features are redundant with each other due to being highly correlated and 1 suggests complete independence. This analysis produces a hierarchy of features by their correlation and can give more information of hidden correlations within features and feature redundancy [26].

3 Results and Discussions

Several models were successfully implemented into PyTorch and trained on the feature data, evident by the loss function decreasing per epoch until improvement plateaued. To gauge the accuracy of the trained models, both the mean absolute error (MAE) and MSE were used to intuitively express the average error in predictions of price. The R^2 statistic between true target prices and predicted values was also incorporated, as these ideally produce a straight line, so linear regression metrics can be used. As the value of the R^2 statistic increases, it implies that the model can explain more of the variation in the feature data and therefore predicts prices with higher accuracy. The ideal model would produce a high R^2 but low MSE and MAE [5]. Section 3.1 details the results of neural network experiments, having designed the model section 3.2 displays the analysis of the model features.

3.1 Neural Network Training

3.1.1 Base-Model Results

During the process of collating retrosynthesis results, the effect of data set size on the performance of neural networks could be appreciated. Retrosynthesis is a relatively slow process even with the augmentations made to increase efficiency, so analysis with the beginning heuristic model was conducted as the data set grew with time, as indicated in Figure 3. Training with 5000 data points produced the results R^2 value of 0.390 and a test MSE loss of 129181, in comparison to the improved performance at 50,000 data points where test loss decreased to 112627 and R^2 increased to 0.476. This is because more data points reinforce patterns within the data recognized by the neural network, increasing prediction accuracy.

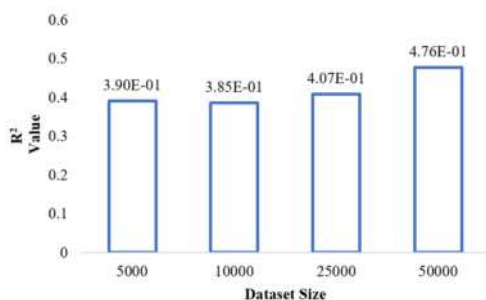


Figure 3: Effect of Dataset Size on R^2

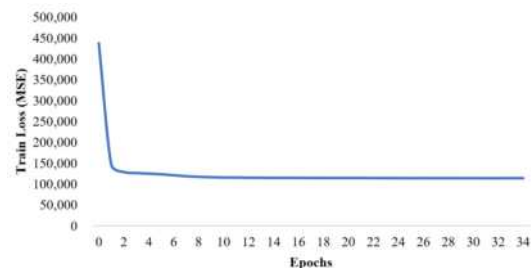


Figure 4: Train Loss (MSE) of Baseline Model

The initial value for number of epochs for this study was set arbitrarily to 100. Figure 4 shows that the training loss plateaus quickly for the basic model and by the 6th epoch shows very marginal improvements, if any. This prompted to test early stopping regularisation, and the patience level was set at 20.0, meaning if no improvement is made within 20 epochs after the last best training loss, then training is stopped. This often activated in the range of 20 to 30 epochs, providing an ample buffer for training the model, and thus 100 epochs was used for following experiments and was never reached for any of the other training runs. Batch sizes $\in \{8, 16, 32, 64, 128, 256\}$ were tested on the basic model to determine which provided the minimal time to train the model while not sacrificing accuracy, the results in Figure 5 display no apparent correlations. After averaging multiple runs, batch size 8 and 256 took the least time at approximately 32 and 34 seconds respectively and had equivalent R^2 values of 0.469. The other batch sizes on average took considerably longer, up to 72 seconds per run, and did not provide better R^2 results. To ensure experiments were more time efficient the batch size was set to 256, 8 was not chosen as it may make the model more susceptible to noise in the data depending on the train-test split. Combining all the results of design specification, the 6-13-13-1 model with batch size 256 and 100 epochs produced an average R^2 value of 0.434 and test MSE Loss of 122846.

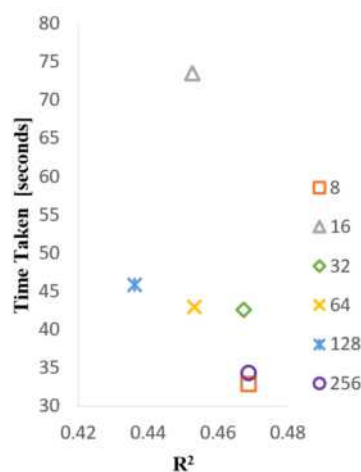


Figure 5: Effect of batch size on R^2 and time taken to train basic model

3.1.2 Network Architecture Optimization

With training hyperparameters completed and the best parameters chosen, exploration of more complex network architecture was conducted in the aim of increasing prediction accuracy. Varying the number of nodes for a single layer model as shown in Figure 6, it can be observed that increasing the number of nodes exhibits a trend of improving performance with the peak point at 6528 nodes in the first layer. It is possible that increasing the number of nodes by doubling it further might improve performance and should be investigated further. The average results for the varying the number of 13-node hidden layers on the basic model is depicted in Figure 7. It shows that having 3 hidden layers produces the best R^2 value of 0.466, the second best is 6 hidden layers. However, the model with 6 13-node hidden layers does take 94 seconds to train, as compared to the 77 seconds it took to train the basic model of 3 hidden layers. The data does suggest that increasing the number of hidden layers past 3 might have a detrimental effect on model performance, although more data points are needed to verify this. As the 3 hidden layer model produced the best performance, this was used as an input for the maximum number of layers into the model generator, which produced models like those seen in Table 1.

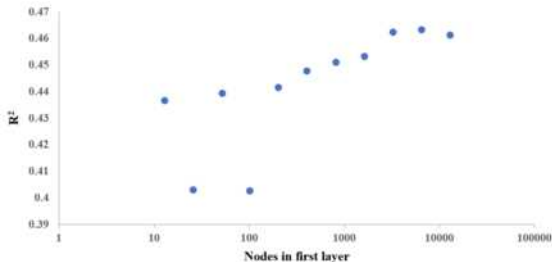


Figure 6: Effect of nodes on R^2 value

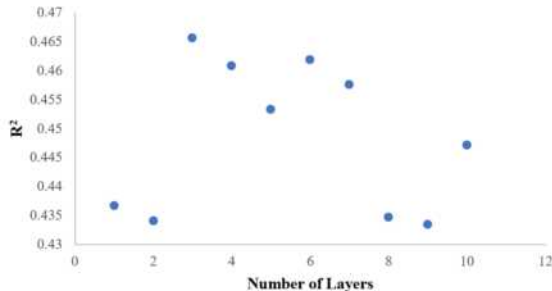


Figure 7: Effect of number of layers on R^2 value

Figure 8 depicts the performance of 77 models with difference structures that have a test MSE loss range of 103491 to 125088 and R^2 range from 0.396 to 0.495. The tables 1 and 2 contains the statistics of the 5 best and worst network architectures. The optimal design in this generation utilized the hidden layer sizes 64-32-128, had a test MSE loss of 103491 and R^2 of 0.495. After further tuning the parameters for implementing regularization via dropout and weight decay on this optimal model and obtaining an average, the result is

test MSE loss of 109560 and R^2 of 0.481. Comparing this to the average performance of the base model performance of MSE Loss of 122846 and R^2 value of 0.434, improvement can be observed. However, the R^2 value for both models is still low. It indicates the existence of a correlation between true prices and predictions, but not a strong one. The MAE of the tuned and optimised model shows that predictions are on average 220\$/mmol above or below the true price. The large discrepancy between MAE and MSE is indicative of either outliers or specific instances where the model's predictions are significantly off. This could be due to limitations in the model's architecture, insufficiently representative training data, or the inherent complexity of the prediction task.

Table 1: Table of best performing models by R^2

Model	R^2	MSE loss	MAE loss
64_32_128	4.95E-01	1.03E+05	2.19E+02
64_32_32	4.89E-01	1.09E+05	2.23E+02
128_16_32	4.87E-01	1.05E+05	2.18E+02
128_16_128	4.87E-01	1.08E+05	2.19E+02
64_16_128	4.85E-01	1.09E+05	2.19E+02

Table 2: Table of worst performing models by R^2

Model	R^2	MSE loss	MAE loss
64_32	3.96E-01	1.23E+05	2.41E+02
32_64	3.98E-01	1.23E+05	2.43E+02
16_16_16	4.10E-01	1.25E+05	2.42E+02
64_16	4.14E-01	1.22E+05	2.43E+02
16_32	4.17E-01	1.20E+05	2.45E+02

Even though heuristics recommend single and double hidden layer structures, the experiments for Tables 2 show that the best models instead all have 3 hidden layers, which could be due to more layers being able to detect the complex patterns being present within the data. This correlates well with the previous experiment, where solely the number of hidden layers are varied. An interesting observation is that the majority of the most accurate models have 3 hidden layers with a smaller central layer instead of an even distribution. The dwarfed middle hidden layer may be acting as a bottleneck and restricting the information the model can learn from the training data, causing the neural network to prioritise only the most important patterns and mitigating overfitting.

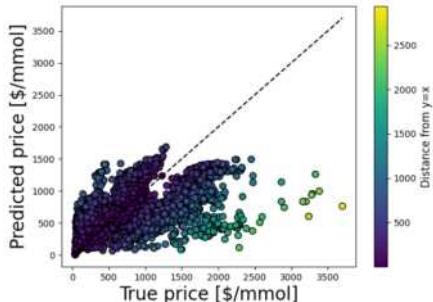


Figure 9: Plot of true prices against predicted prices

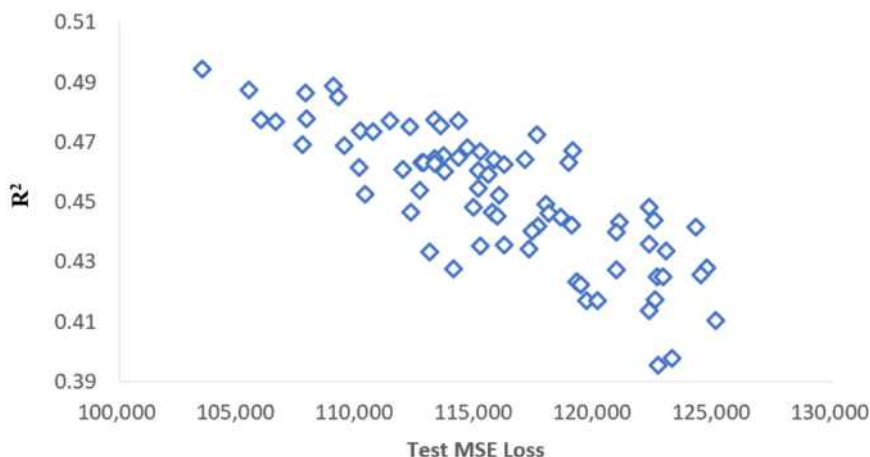


Figure 8: Performance of the 77 generated model architectures

Figure 9 is the parity plot of true prices and predicted prices, along with the straight line the data should lie upon. This plot highlights a major issue with this study, that is the lack of representation for higher prices. For molecules priced below \$1000/mmol the predictions gather close around the expected straight line correlation, but as you go past this point the data starts to veer to the right and not increase in predicted price above approximately \$2000/mmol. This highlights that the data obtained for this study is too small to express patterns for more expensive molecules, as data in this region becomes scarce, causing the average accuracy to be very low. However, even when cherry-picking the results and using data below \$1000/mmol using the best model structure, the average R^2 is only 0.632. This is still not a convincing level of accuracy, so there may be limitations with this approach, which the following section explores in more depth.

3.2 Feature Analysis

Permutation feature analysis was simulated using several random shuffles for each feature to find the average increases in error. The main findings of this analysis were that number of precursors was the most important feature for price prediction whereas minimum price of precursors was least, with relatively little variance in error for the other 4. It would be expected that number of reaction steps would be of similar importance to number of precursors, as more reaction steps would generally imply having more chemicals to react. But the large difference in importance implies that these are decoupled for patterns for predicting price. Another interesting decoupling of expected relationships is that mean precursor price is ranked third most important, but the minimum and maximum price are least influential. One would expect the average price of precursors to be more important, as spending more money on feed precursors should be offset by selling at a higher price, but this model argues it is less influential than theorised. The minimum and maximum feed stock price would not have been predicted the least influential, as one sets the minimum expenditure on feed materials and one

measures the relative magnitude of capital expenditure, similar to mean price you would expect more spending on feed material needs to be offset with higher price.

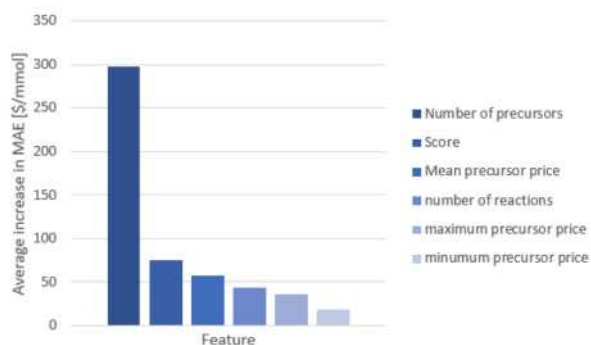


Figure 11: Permutation Importance of Features

The results of SHAP value analysis are seen in Figure 10, this summary plot is based on the optimal model where each point is a data point from our data set. The x-axis value is the effect on the predicted price, the colour shows the relative magnitude of the feature value. Before investigating feature correlations, cross-reference with permutation importance can be conducted as the top to bottom ordering of features in the plot is the SHAP estimation of feature importance. The overall ranking for feature importance is generally the same over both techniques, the only difference being the rankings of score and mean precursor price are swapped, which is likely due to differences in fundamental theory behind the techniques, nevertheless most placements are identical and confirms that random results were not being produced by the permutation method. The SHAP plot displays that the number of precursors increasing causes an increase in molecule price, possible reasoning for this is that more precursors implies a more complex product structure, regardless of number of reactions as mentioned above, and that there is some price premium that comes with complexity. A chemical

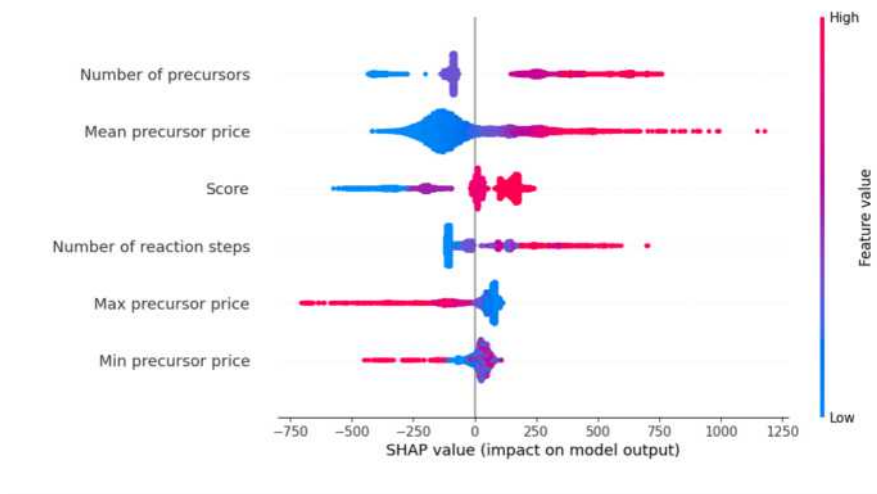


Figure 10: SHAP value analysis - Summary plot

engineering approach to this correlation is that more precursors requires a more intricate separation system for high purity, which raises capital and operation costs which need to be offset with a higher product price. Surprising correlations are obtained for maximum and minimum precursor price, as discussed for permutation importance, one would expect these to increase in price as they increase due to more operational costs. However, the opposite is displayed in the results, this may be due to the lack of patterns in the data for price prediction as surmised by its low importance, so the theory is that these trends are the product of data noise. Number of reactions increasing appears to cause an increase in price, this may be explained by more reaction vessels and equipment being required to synthesise the product which, similar to needing more separation equipment, increases molecule price to offset capital/operation costs. Synthesis score provided a surprising result of having a positive result with price, implying a molecule more likely to be synthesised is more likely to be expensive. However, it should be noted that the correlations for score and number of reactions are less likely to be as accurate as the other correlations for reasons discussed below.

The charts in Figure 12 are clustering hierarchies using different threshold clustering distances. The main inference from this analysis is that synthesis score and number of reactions are highly correlated, being clustered at distances as low 0.1 and implying high partial redundancy. The redundancy between them means the model receives the same patterns from both and fundamentally removes the amount of input features by 1. High correlation also effects SHAP summary plot analysis for score and number of reactions, as their relationships to price are more likely to be effected by data set outliers [26]. This may be the result of using the same scoring function used by the AiZynthFinder as it finds synthesis routes, causing some unknown bias towards reaction steps when results are generated. The second limitation of this approach is apparent from the 0.5 clustering cutoff graph, it shows that most features exhibit 50% partial redundancy with at least one other feature. This will cause the model to be less sensitive to data patterns, as each feature is partially corre-

lated and providing less unique information than if they were independent.

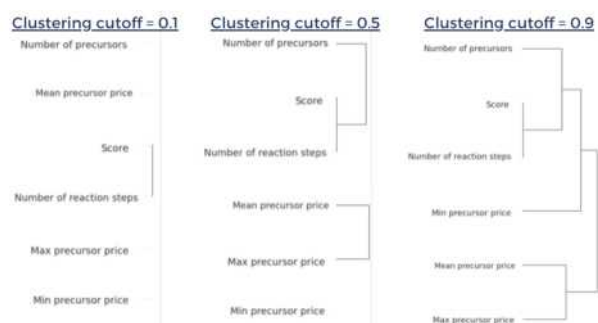


Figure 12: SHAP cluster analysis

4 Conclusion

A retrosynthesis algorithm, AiZynthFinder, was utilized to generate synthesis routes for a range of commercially available chemicals, which were converted to quantifiable statistics such as number of reaction steps, number of precursors and mean precursor price among others. These were used as input features for training an artificial neural network to predict molecule price. A range of neural network architectures were systematically tested and models were refined using hyperparameter tuning methods. The optimal model was found to provide low accuracy predictions, with mean average errors of 220 \$/mmol. It is implied that the data set size used in this research was too small to encapsulate patterns across a wide range of molecule prices. Even when analysing molecules with prices below 1000 \$/mmol, which had better predictions, the accuracy was still inadequate, suggesting limitations in the novel approach attempted in this research. To identify factors limiting model performance, a range of neural network analysis techniques were utilized to investigate feature importance,

how features affected price predictions and feature independence. Permutation importance and SHAP value analysis indicated that the number of precursors was the most important feature for making price predictions, and the other features did not contribute significantly to the models' performance. Correlations between each feature and price predictions were analysed and used to connect possible real world explanations for these patterns emerging. Feature redundancy found that 2 features, score and number of reactions were incredibly correlated and that most features had significant partial redundancy with one another, which could have contributed to the limited model accuracy and highlights potential issues with the feature selection.

5 Research recommendations

Having concluded the findings of this research, a variety of possible recommendations for future work in this field were identified. Firstly, the search space for possible chemical structures is very large so more data is required to find the necessary patterns to predict the price of a wider range of molecules, this involves increasing computation time and power to retrieve results. A large limitation of data generation was the filtering process, 115,000 molecules were analysed but only 51,383 were eligible for use in the neural network. This can be improved by increasing AiZynthFinder's success rate by configuring its code to use deeper search trees and training the algorithm on larger data sets, the other major improvement would be to increase the size of the pricing index, so more precursors can be identified. The results from the neural network experiments imply that the use of several hidden layers with a small central layer may improve model accuracy when creating network architectures for this type of research. Future research could focus on training more complex model architectures with a much larger number of nodes and hidden layers to better establish trends in performance of the models. Feature analysis highlighted that the number of precursors exhibited potential for effectively predicting molecule price from retrosynthesis features, and its use in future research is recommended. Synthesis score and number of reactions were highly correlated so their use together in future models is not advised, this may be mitigated using a scoring function independent of the retrosynthesis algorithm to generate synthesis routes, but this requires more research to prove. The overall significant partial redundancy implies that different features that are independent of the ones researched in this study may improve prediction accuracy as more patterns in the data can be identified, preliminary literature identifies some machine learning algorithms that predict reaction yields and temperatures that could be promising choices.

References

- [1] Tohru Takebe, Ryoka Imai, and Shunsuke Ono. The current status of drug discovery and development as originated in united states academia: the influence of industrial and academic collaboration on drug discovery and development. *Clinical and translational science*, 11(6): 597–606, 2018.
- [2] Sandra Kraljevic, Peter J Stambrook, and Kresimir Pavelic. Accelerating drug discovery: Although the evolution of ‘-omics’ methodologies is still in its infancy, both the pharmaceutical industry and patients could benefit from their implementation in the drug development process. *EMBO reports*, 5(9):837–842, 2004.
- [3] PS Ananthanarayanan. Project technology and management. In *Treatise on Process Metallurgy*, pages 1145–1191. Elsevier, 2014.
- [4] Guoqiang Zhang, B Eddy Patuwo, and Michael Y Hu. Forecasting with artificial neural networks:: The state of the art. *International journal of forecasting*, 14(1):35–62, 1998.
- [5] Tomasz Badowski, Karol Molga, and Bartosz A Grzybowski. Selection of cost-effective yet chemically diverse pathways from the networks of computer-generated retrosynthetic plans. *Chemical science*, 10(17):4640–4651, 2019.
- [6] Ruben Sanchez-Garcia, Dávid Havasi, Gergely Takács, Matthew C Robinson, Frank von Delft, Charlotte M Deane, et al. Coprinet: graph neural networks provide accurate and rapid compound price prediction for molecule prioritisation. *Digital Discovery*, 2(1):103–111, 2023.
- [7] Cheng-Hao Liu, Maksym Korablyov, Stanislaw Jastrzebski, Paweł Włodarczyk-Pruszyński, Yoshua Bengio, and Marwin HS Segler. Retrognn: Approximating retrosynthesis by graph neural networks for de novo drug design. *arXiv preprint arXiv:2011.13042*, 2020.
- [8] Yinjie Jiang, Yemin Yu, Ming Kong, Yu Mei, Lutotian Yuan, Zhengxing Huang, Kun Kuang, Zhihua Wang, Huaxiu Yao, James Zou, Connor W. Coley, and Ying Wei. Artificial intelligence for retrosynthesis prediction. *Engineering*, 25:32–50, 2023. ISSN 2095-8099. doi: <https://doi.org/10.1016/j.eng.2022.04.021>. URL <https://www.sciencedirect.com/science/article/pii/S2095809922005665>.
- [9] Samuel Genheden, Amol Thakkar, Veronika Chadimová, Jean-Louis Reymond, Ola Engkvist, and Esben Bjerrum. Aizynthfinder: a fast, robust and flexible open-source software for retrosynthetic planning. *Journal of cheminformatics*, 12(1):70, 2020.
- [10] John J Irwin and Brian K Shoichet. Zinc- a free database of commercially available compounds for virtual screening. *Journal of chemical information and modeling*, 45(1):177–182, 2005.
- [11] molport. Website title, 2023. URL <https://www.molport.com/>.
- [12] Daniel Bourke. 01. pytorch workflow fundamentals. URL https://www.learnpytorch.io/01_pytorch_workflow/.
- [13] Sagar Sharma, Simone Sharma, and Anidhya Athaiya. Activation functions in neural networks. *Towards Data Sci*, 6(12):310–316, 2017.
- [14] Jinming Zou, Yi Han, and Sung-Sau So. Overview of artificial neural networks. *Artificial neural networks: methods and applications*, pages 14–22, 2009.
- [15] Amir Sahraei, Alejandro Chamorro, Philipp Kraft, and Lutz Breuer. Application of machine learning models to predict maximum event water fractions in streamflow. *Frontiers in Water*, 3:652100, 06 2021. doi: 10.3389/frwa.2021.652100.

- [16] Steven Walczak and Narciso Cerpa. Heuristic principles for the design of artificial neural networks. *Information and software technology*, 41(2):107–117, 1999.
- [17] Alfrick Opidi. Pytorch loss functions: The ultimate guide, Jul 2022. URL <https://neptune.ai/blog/pytorch-loss-function>.
- [18] Chieh-Huang Chen, Jung-Pin Lai, Yu-Ming Chang, Chi-Ju Lai, and Ping-Feng Pai. A study of optimization in deep neural networks for regression. *Electronics*, 12(14):3071, 2023.
- [19] Marco Taboga. Loss function. URL <https://www.statlect.com/glossary/loss-function>.
- [20] Ange Tato and Roger Nkambou. Improving adam optimizer. 2018.
- [21] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.
- [22] Anders Krogh and John Hertz. A simple weight decay can improve generalization. *Advances in neural information processing systems*, 4, 1991.
- [23] James Bergstra, Dan Yamins, David D Cox, et al. Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. In *Proceedings of the 12th Python in science conference*, volume 13, page 20. Citeseer, 2013.
- [24] Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 41–50, 2019.
- [25] Christoph Molnar. *Interpretable machine learning*. Lulu.com, 2020.
- [26] Documentation by example for shap.plots.bar. URL https://shap-lrjball.readthedocs.io/en/latest/example_notebooks/plots/bar.html.

Utilizing the SAFT- γ Mie GC Equation of State to Assess the pH-Dependent Solubility of Active Pharmaceutical Ingredients

Babafemi Mustapha and Harrison Fraser

Department of Chemical Engineering, Imperial College London, U.K.

ABSTRACT

The potency of active pharmaceutical ingredients (APIs) is directly related to their bioavailability, that is, how easily the drug reaches the biological site of interest. Currently, experimentally determined pH-Solubility profiles are used to access this metric, however a significant number of APIs have extremely low solubilities in water, making experiments a difficult process. In this study, we explore the use of computational thermodynamics, in particular the SAFT- γ Mie group-contribution equation of state combined with quantum mechanical calculations to generate pH-solubility profiles. Ibuprofen and mefenamic acid were explored in this study, both showing very good accuracy on pKa prediction and the qualitative shape of their pH-solubility profiles. Reasonable quantitative accuracy was observed on solubility values against experimental data.

1. Introduction

The efficacy of substances and active pharmaceutical ingredients (APIs) is undoubtedly affected their properties, with solubility being one of most importance as it determines the extent to which the drug is completely accessible to the desired biological destination (Price & Patel, 2023). Ambient conditions as a result also affect drugs, since temperature, pressure but also pH will influence solubility. The pH in the human body ranges greatly, going from a value of roughly 2 in the stomach, to 7 in the blood and around 8 in the colon (Surat, 2022). It is therefore of great necessity to accurately determine the relationship between pH and solubility to help determine bioavailability across various regions of the body.

There are a few experimental methods to create pH-solubility profiles. These include: the saturation shake-flask (SF) method, where a flask containing the API and buffer solution are shaken together and let to reach saturation. pH is then typically measured with a pH electrode, and the concentration is determined by performing a liquid chromatography on a sample of the supernatant (liquid left behind after precipitation), and UV spectrophotometry (Raja & Barron, 2023)– Drug solutions at various pH levels are created, and the absorbance of the solution is measured, meaning the concentration of the API can be evaluated using Beer-Lambert's law (Shoghi, et al., 2013).

While these techniques are all relatively straightforward, the determination of pH-dependent solubility is still a time-intensive and expensive process. Most APIs have extremely low solubilities, especially in water, complicating matters further as results between independent sources usually differ

considerably. Experimental data on the salts of APIs and other valuable data, such as temperatures of melting, and heats of fusion, which are important parameters in general, are also very sparse. This motivates the work towards creating theoretical methods that can efficiently and accurately predict the behaviour of APIs and other substances of interest. The development of advanced theories such as density functional theory and solvation models have made computational prediction of important thermodynamic parameters possible in recent years (Gui, et al., 2023), and in this work, these quantum mechanical calculations were used to estimate pKa, which directly relates to solubility, alongside the state-of-the-art Statistical Associating Fluid Theory (SAFT) equation of state (EoS) to attempt to predict the pH-dependent solubility of APIs.

SAFT is an EoS based on Wertheim's perturbation theory with the ability to model large and complex molecules, including species with strong intermolecular interactions such as hydrogen bonding (Febra, et al., 2021). It is important to note that there are a multitude of SAFT models, such as: perturbed chain (PC)-SAFT, which uses spherical particles in the context of hard chains as a reference fluid for the dispersion term (Gross & Sadowski, 2001), soft-SAFT, which uses the Lennard–Jones intermolecular potential for the reference fluid in the equation, with dispersive and repulsive forces explicitly considered into the same term, instead of the perturbation scheme based on a hard-sphere reference fluid plus dispersive contributions to it (Belkadi, et al., 2010), or SAFT-VR (variable range) which models chains as homonuclear, and allows attractive potentials with variable widths, hence the name (Lafitte, et al., 2013). Each of these versions of SAFT have distinct properties and

advantages, however the model used in this study is the SAFT- γ Mie group contribution approach, chosen as it has been shown to be able to predict, to a high degree of accuracy, a wide-range of thermodynamic properties, including phase equilibria and excess properties of mixing for a plethora of mixtures (Wehbe, et al., 2022).

The pH-dependent solubility of two non-steroidal anti-inflammatory drugs (NSAIDs) are explored in this study: ibuprofen, and mefenamic acid, primarily used for pain relief. Ibuprofen is the weaker agent out of the two and is usually the first form of pain relief taken by individuals since it can be purchased over the counter. Mefenamic acid is much stronger and requires prescription, however it is still very commonly used especially in comparison to more intense analgesics such as opioids.

2. Theory and Methods

2.1 pKa Estimation

A substance's solubility is defined as the amount of substance that will form a saturated solution in a specified amount of solvent, at a given temperature and pressure (Soult, 2023). For an API, solubility will be related to its inherent pKa value, a measure of how strongly the acid dissociates into its respective ionic elements, and the pH of the environment. pKa was quantified computationally in this study, using the thermodynamic cycle method reported by Ho & Ertem, where the Gibbs free energy of deprotonation in the aqueous phase, directly related to pKa, can be calculated through the free energies of solvation of the reacting species and of deprotonation in the gas phase. These free energies can in turn be calculated from the solute electronic energies (Raamat, et al., 2012) computed on geometries primarily optimised in Gaussian 16, a general-purpose computational chemistry software.

In this work, the geometries were initially optimised through a GMMX conformational search within Gaussview 6, using a MMFF64 molecular mechanics force field, in which each heavy atom of the molecule is moved a small distance in each Cartesian dimension, allowing for an exploration of the conformational space (Gaussian, n.d.). The conformational search then provides a set of candidate conformations, with the most stable confirmation being continued with. Two further optimisations were then performed using Gaussian 16, first using a low-level of theory, specifically Hartree-Fock with a limited basis set (3-21g), using SMD as the solvation model. This provided a good starting guess for a more rigorous optimisation using SMD as the solvation model, in conjunction with the M062X level of theory and 6-31+G(D) basis set. As mentioned, electronic energies calculated by Gaussian were then used to find pKa

following the method outlined by Ho & Ertem, a more detailed explanation of which now follows.

$$pKa = \frac{\Delta G_{aq}^*}{RT \ln 10} \quad (1),$$

where ΔG_{aq}^* is the (Gibbs) free energy change of the reaction. This is illustrated in the diagram below.

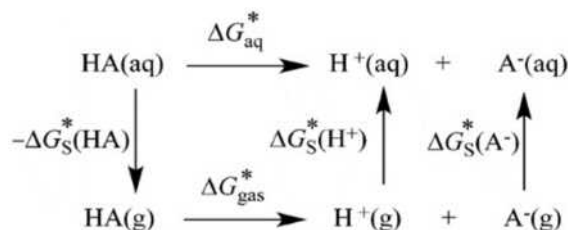


Figure 1. Generic thermodynamic cycle reproduced from: (Ho & Ertem, 2016)

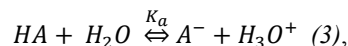
As the cycle shows,

$$\Delta G_{aq}^* = -\Delta G_s^*(HA)^{[L]} + \Delta G_{gas}^{*[H]} + \Delta G_s^*(H^+) + \Delta G_s^*(A^-)^{[L]} \quad (2),$$

where $\Delta G_s^*(HA)$ is the free energy change of solvation of the acid, ΔG_{gas}^* is the free energy change of ionisation of the acid, and $\Delta G_s^*(H^+)$, $\Delta G_s^*(A^-)$ are the free energy changes of solvation of hydrogen, taken as equal to -265.9 kcal/mol (Kelly, et al., 2007), and the conjugate base respectively. The superscripts [L] and [H] indicate the level of theory used to calculate each term, referring to low and high levels of theory respectively.

2.2 pH-Solubility Profile Prediction

With the computationally determined value of pKa, pH-dependent solubility could be modelled. To best capture bodily conditions, predictions were performed at 298.15K and atmospheric pressure. An explanation of the system of equations used for modelling follows. We begin by considering the acid dissociation reaction:



note that the API is depicted as a weak protic acid HA, which forms a conjugate base A^- .

$$K_{a,Ha} = \frac{a_{A^-} a_{H_3O^+}}{a_{HA} a_{H_2O}} \quad (4),$$

$K_{a,Ha}$ being the acid dissociation constant. The term a_i refers to the activity of a species i , which is essentially the effective concentration of the species.

$$\mu_i(T, P, x) = \mu_i^\emptyset + RT \ln a_i(T, P, x) \quad (5),$$

where μ_i is the chemical potential (a thermodynamical property) of a species i , and μ_i^\emptyset is the chemical potential at a reference state.

To be consistent with thermodynamical expressions, molal units were used instead of molar units:

$$K_{a,HA}^m = K_{a,HA} a_{H_2O} \quad (6),$$

Where $K_{a,HA}^m$ is the acid dissociation constant in molal units, which was calculated from pKa as shown in equation 7 below:

$$pKa = -\log_{10} (K_{a,HA}^m) \quad (7).$$

In this study, compounds which are chemically and structurally similar to the API of interest (referred to subsequently as moieties) were used as a means of calibrating the pKa value obtained from Gaussian to experimental data. This was done through regression plots.

Above, the solubility of the API has been considered; however, at low pH, the API will not dissociate meaning the intrinsic solubility, corresponding to the solid-liquid equilibrium of the API in solution, must also be considered (Wehbe, et al., 2022). This is calculated via the following equation:

$$\ln x_{HA}^L(T, P) = \frac{\Delta h_{HA}^{fus}(T_{HA}^{fus}, P)}{R} \left(\frac{1}{T_{HA}^{fus}} - \frac{1}{T} \right) + \frac{1}{RT} \int_T^{T_{HA}^{fus}} \Delta C_{p,HA}(T', P) dT' - \frac{1}{R} \int_T^{T_{HA}^{fus}} \frac{\Delta C_{p,HA}(T', P)}{T'} dT' - \ln \gamma_{HA}(T, P, x^L) \quad (8),$$

with x^L being the composition vector of the liquid phase containing the solvent, buffer ions, and API (in molecular and dissociated form). Δh_{HA}^{fus} is the molar enthalpy of fusion of the API at the melting temperature, T_{HA}^{fus} . R is the universal gas constant, $\Delta C_{p,HA}$ is the difference between the isobaric heat capacity of the API in the liquid and solid phases, and γ_{HA} is the symmetric activity coefficient of HA in the liquid phase at the given T and P .

The heat capacity terms have been left in equation 8 for completeness, however their contribution was found to be negligible and increased the likelihood of computational errors. These terms were removed from calculations. This simplification is commonly done in literature, as supported in the paper by (Febra, et al., 2021). The activity coefficient γ_{HA} is also related to the fugacity coefficient $\hat{\phi}_{HA}$, (a quantity calculated by SAFT) through the relationship:

$$\gamma_{HA} = \frac{\hat{\phi}_{HA}(T, P, x^L)}{\hat{\phi}_{HA}^*(T, P)} \quad (9),$$

with $\hat{\phi}_{HA}^*$ being the fugacity coefficient of pure HA at the specified T and P . In this solvation model, we take a reference state as an infinitely dilute solution in water meaning:

$$a_i = x_i \tilde{\gamma}_i = x_i \frac{\gamma_i}{\gamma_i^\infty} \quad (10),$$

where a_i is the activity as above, $\tilde{\gamma}_i$ is the asymmetric activity coefficient, γ_i^∞ the symmetric activity coefficient at infinite dilution and x_i the mole fraction, of species i .

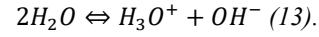
As aforementioned, molal units are used for quantities instead of molar units but these can be approximated as equal for very dilute aqueous systems (where the density of the solution is virtually equal to that of pure water), allowing for the following equation to be derived:

$$K_{a,HA}^m = \left(\frac{m_{A^-} m_{H_3O^+}}{m_{HA}} \right) \left(\frac{\tilde{\gamma}_{m,A^-} \tilde{\gamma}_{m,H_3O^+}}{\tilde{\gamma}_{m,HA}} \right) \quad (11),$$

where m_i is the molality of species i and $\tilde{\gamma}_{m,i}$ is the molal asymmetric coefficient calculated as:

$$\tilde{\gamma}_{m,i} = x_{H_2O} \tilde{\gamma}_i \quad (12).$$

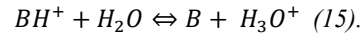
Water also undergoes a small reaction with itself regardless of the other ions and species present:



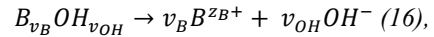
The degree of dissociation of water (in the reaction shown above) was quantified through the ionic product, K_w , found using an analogous method to K_a , as is shown below:

$$K_w = (m_{H_3O^+} m_{OH^-}) (\tilde{\gamma}_{m,H_3O^+} \tilde{\gamma}_{m,OH^-}) \quad (14).$$

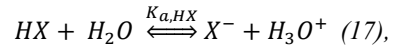
From this point, we can consider the action of the base, which is added to increase the pH of the solution (which moves the system from intrinsic solubility to ‘solubility’). For a weak base B , the dissociation reaction is as follows:



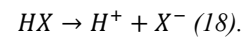
For a strong base, we can assume complete dissociation:



where B^{z_B+} is the anion of the strong base added, z_i , v_i is the charge and stoichiometric coefficient of species i respectively. It is important to highlight, that lowering the pH below the pKa requires the action of an acidic buffer. It is also needed if the API was a base, since:

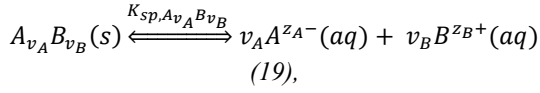


which is virtually identical to the equation for the API above. If a strong acid is used, complete dissociation can be assumed once again, where:



Eventually, the pH of the solution reaches a value where the API-salt ($A_{v_A} B_{v_B}$) precipitates out. This occurs when the solubility product $K_{sp,A_{v_A} B_{v_B}}$ is reached, the pH at which is occurs was described by

Wehbe as pH^{\max} (Wehbe, et al., 2022) and the same nomenclature will be used in this work. The precipitation reaction and equation used to calculate the solubility product are shown in equations 19 and 20:



$$K_{sp, A_{v_A} B_{v_B}} = (m_{A^{z_A-}} m_{B^{z_B+}}) (\tilde{\gamma}_{m, A^{z_A-}} \tilde{\gamma}_{m, B^{z_B+}}) \quad (20),$$

the solubility product $K_{sp, A_{v_A} B_{v_B}}$. Note that this equation is different to the one referenced in Wehbe's paper. This reformulation was chosen due to a lack of information on the solubility limit, one of the terms in Wehbe's formulation.

All the necessary equations, quantities and values for modelling are expressed above; with the addition of the condition of charge electroneutrality:

$$\sum_{i=1}^N q_i m_i = 0 \quad (21),$$

where q_i is the charge of species i and N being the total number of species in the model, allows for a profile of pH-dependent solubility to be found, pH and solubility were calculated as:

$$\text{pH} = -\log_{10}(a_{\text{H}_3\text{O}^+}) \quad (22),$$

$$S_{API} = \rho M_w (m_{\text{HA}} + m_{\text{A}^-}) \quad (23),$$

$$S_{API, \text{lim}} = \rho M_w m_{\text{A}^-, \text{lim}} \quad (24),$$

where S_{API} , $S_{API, \text{lim}}$ is the solubility of the API and solubility of API at the solubility limit (in moles per volume) respectively, ρ is the molar density of the solution and M_w is the molecular mass of water can be produced.

Experimental values from literature were used to obtain the values of $K_{a, \text{HA}}^m$, K_w , and $K_{sp, A_{v_A} B_{v_B}}$ used in the model. SAFT- γ Mie GC calculated the activity coefficients and the density of solution, and in this study, gPROMS was the software used to simultaneously solve these equations and generate the profile.

2.3 SAFT- γ Mie GC model and theory

SAFT- γ Mie GC (group contribution) is an EoS that models chemical species as segmented, round and fused heteronuclear chains, with each of these segments being depictions of the functional groups existing in the species. Interactions between these segments are then modelled through Mie potentials of variable range. Close range directional forces are considered by association sites found on the segments. The Mie potential is a pair potential, representing the potential energy relationship between two particles. A diagram of the Mie potential is shown in Figure 2 below:

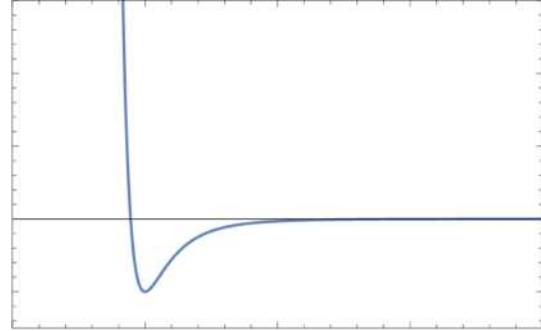


Figure 2. Example of the Mie potential used in the SAFT- γ Mie GC EoS. Reproduced from (University of Liverpool, 2023). The x-axis is the intersegment distance and the y-axis the intermolecular potential energy.

It is a generalised version of the more commonly known Lennard-Jones potential, where the repulsive and attractive exponents are adapted depending on the behaviour of the segments within each compound (Wehbe, et al., 2022). The potential is a function of the intersegment distance r_{kl} , it is of the form:

$$\Phi_{kl}^{Mie}(r_{kl}) = C_{kl} \epsilon_{kl} \left[\left(\frac{\sigma_{kl}}{r_{kl}} \right)^{\lambda_{kl}^r} - \left(\frac{\sigma_{kl}}{r_{kl}} \right)^{\lambda_{kl}^a} \right] \quad (25),$$

where σ_{kl} is the segment diameter, ϵ_{kl} is the depth of the potential well and λ_{kl}^a , λ_{kl}^r are the attractive and repulsive exponents respectively. The coefficient C_{kl} is present to mathematically satisfy the minimum energy being $-\epsilon_{kl}$. It is a function of the exponents:

$$C_{kl} = \frac{\lambda_{kl}^r}{\lambda_{kl}^r - \lambda_{kl}^a} \left(\frac{\lambda_{kl}^r}{\lambda_{kl}^a} \right)^{\frac{\lambda_{kl}^a}{\lambda_{kl}^r - \lambda_{kl}^a}} \quad (26).$$

2.4 Thermodynamic Relationships

The total Helmholtz free energy of these molecules is expressed as the summation of six independent terms, five of which capture the deviation from ideality the molecule will intrinsically exhibit:

$$A = A^{ideal} + A^{mono} + A^{chain} + A^{assoc.} + A^{ion} + A^{born} \quad (27).$$

A^{ideal} represents the contribution to the Helmholtz free energy assuming the individual segments can be taken as ideal gas molecules, with no interparticle interaction. A^{mono} considers the interaction between these segments via a Mie potential. A^{chain} handles the effect of the formation of chains and $A^{assoc.}$ quantifies the energy due to the association of molecules via bonding sites. A^{ion} embodies the Coulombic ion-ion interactions using the Mean Spherical Approximation, and A^{born} examines the ion-solvent electrostatic interactions through the Born model (Wehbe, et al., 2022).

Following from this, the pressure, P , residual chemical potential, μ_i^{res} , and fugacity coefficient, $\hat{\phi}_i$, were all obtained through the following thermodynamic identities:

$$P = - \left. \frac{\partial A(T, V, \mathbf{x})}{\partial V} \right|_{T, N} \quad (28),$$

$$\mu_i^{res}(T, P, \mathbf{x}) = \left. \frac{\partial A^{res}(T, V, \mathbf{x})}{\partial N_i} \right|_{T, V, N_{j \neq i}} - RT \ln Z(T, P, \mathbf{x}) \quad (29),$$

$$\ln \hat{\phi}_i(T, P, \mathbf{x}) = \frac{\mu_i^{res}(T, P, \mathbf{x})}{RT} \quad (30),$$

where $A^{res} = A - A^{ideal}$ is the residual free energy, $Z = \frac{Pv_p}{RT}$ is the compressibility factor, with $v_p = \frac{V_p}{N_{mol}}$ as the molar volume at the specified pressure, N is the vector of moles, N_{mol} is the total number of moles and $\mathbf{x} = \frac{N}{N_{mol}}$ is the vector of the mole fraction.

2.5 Group Contributions

As SAFT- γ Mie GC is a group contribution approach, the groups within each compound, and the parameters governing their interactions with one another, needed to be characterised. The groups used to model each API, as well as the solvent and buffers, are detailed in Figures 3 and 4, with the interaction parameters between each group being taken from prior work in which the SAFT- γ Mie GC was used to successfully model thermodynamic properties of ibuprofen and mefenamic acid in water (Febra, et al., 2021) (Wehbe, et al., 2022).

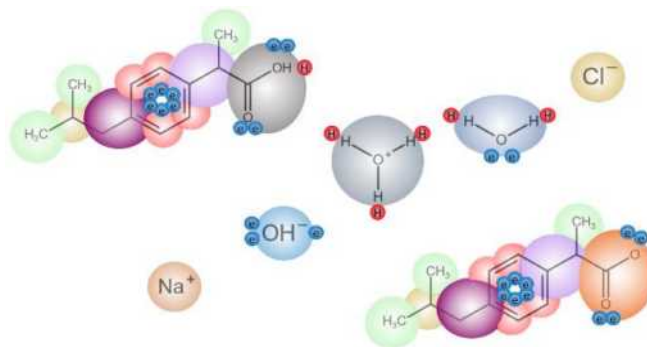


Figure 3. (Going left to right, bottom to top) SAFT representation of ibuprofen, sodium cation, hydroxide ion, hydronium ion, water, chloride anion and ibuprofen's anion.

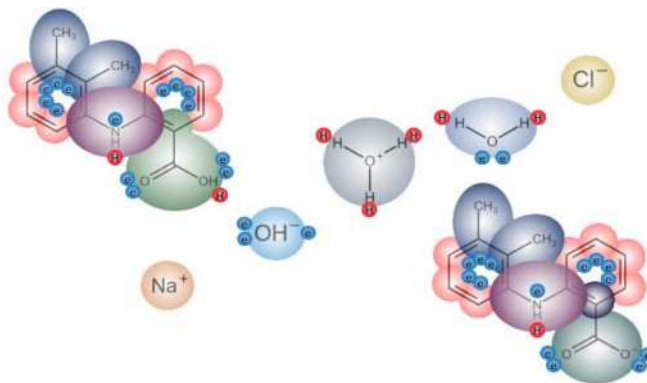


Figure 4. (Going left to right, bottom to top) SAFT representation of mefenamic acid, sodium cation, hydroxide ion, hydronium ion, water, chloride anion and mefenamic acid's anion.

Ibuprofen, in its neutral form, was broken down into the following groups: 3 x CH₃, 4 x aCH, 1 x aCCH, 1 x aCCH₂, 1 x CH, and 1 x COOH. In its deprotonated form, the COOH group was replaced by a COO⁻ group. The groups for mefenamic acid were: 1 x aCCOOH, 1 x aCNHaC, 7 x aCH, and 2 x aCCH₃, with its anion being modelled with: 1 x aCNHaC, 7 x aCH, 2 x aCCH₃, 1 x C and 1 x COO⁻ instead. Using an aCCOO⁻ group to represent the aromatic carbon attached to the speciated carboxylic acid is likely more accurate, but due to a lack of experimentally determined thermodynamic data of the required quality on the behaviour of mefenamic acid or other benzoic acid derivatives, this group is yet to be parametrised within the SAFT- γ Mie GC framework. Furthermore, combining rules currently must be used to model the interaction parameters between the aCNHaC group of mefenamic acid and water, again due to the lack of experimental data on compounds containing this group. In addition, to model the pH-dependent solubility of compounds of interest, one needs to also characterize the interactions between the solute, solvent, and buffer. The solvent, water, was modelled using a H₂O group, with the acid dissociation reaction then producing a hydronium and hydroxide ion, modelled by H₃O⁺ and OH⁻ groups respectively. Below the value of pK_a for each API, the variation of pH was modelled by the addition of a HCl buffer, represented by the H₃O⁺ and Cl⁻ groups. Above the value of pK_a, pH variation was

achieved by modelling the addition of NaOH as a buffer, itself represented by the OH⁻ and Na⁺ groups. Figures 3 and 4 also indicate the association sites present for each group, marked by e and H, which allowed for representation of association interactions by the model.

Note that, to analyse the quantitative difference between the theoretically determined relationship of pH against solubility versus experimental results, the percent average absolute deviation (%AAD) of a property, x was used:

$$\%AAD\{x\} = \frac{100}{N_x^D} \sum_{i=1}^{N_x^D} \left| \frac{R_{x,i}^{exp} - R_{x,i}^{calc}}{R_{x,i}^{exp}} \right| \quad (31),$$

with N_x^D being the number of data points of the property of interest x , $R_{x,i}^{exp}$ being the i th measured value of property x , and $R_{x,i}^{calc}$ being the respective value calculated by SAFT- γ Mie GC.

3. Results & Discussion

3.1. Ibuprofen

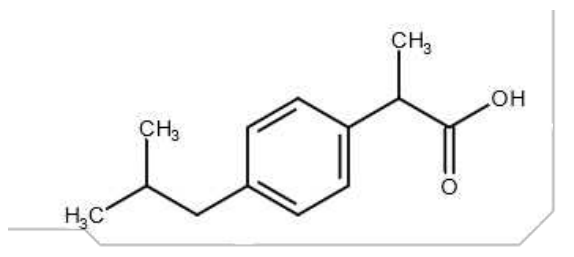


Figure 5. Chemical structure of ibuprofen

3.1.1. Initial pKa Prediction

Utilising Gaussian 16 and the thermodynamic cycle method, the value of pKa for ibuprofen was initially found to be 7.03, far from any reported value of pKa at 298.15K in literature (Wehbe, et al., 2022) (Raamat, et al., 2012) (Sangster, 1994). This is an expected difference however, as Ho & Ertem reported that the thermodynamic cycle method, when used in conjunction with the SMD-M062X solvation model, has a mean absolute error of 2.0 pH points from the experimental value (Ho & Ertem, 2016).

3.1.2. Correction of pKa Using Experimental Data

Due to the large error in the initially calculated value of ibuprofen's pKa, a regression was performed to correct the value of pKa closer to experimental values. Moieties of ibuprofen, that is, smaller, simpler compounds similar in structure to the parent molecule,

are listed below. The goal of this is to remove any error present in Gaussian's optimisations of geometry due to the groups specifically within ibuprofen, whilst also accounting for any additional systematic error.

The selected compounds for the regression were: acetic acid, propanoic acid, phenylacetic acid, 2-phenylpropanoic acid, 4-methylbenzoic acid, and 4-tert-butylbenzoic acid.

Each of these molecules' pKas was calculated by the same method as performed for ibuprofen, using a thermodynamic cycle, with the SMD-M062X solvation model again being used in this work to compute molecular geometries and free energy changes. The value of pKa calculated from Gaussian for each compound is reported in Table 1 below, along with the experimentally determined values of pKa for each compound from literature, used in the regression (Haynes, 2015).

Compound	Acetic acid	Propanoic acid	Phenylacetic acid
Calculated pKa	6.93	6.98	6.77
Experimental pKa	4.76	4.87	4.31
Compound	2-phenylpropanoic acid	4-methylbenzoic acid	4-tert-butylbenzoic acid
Calculated pKa	7.21	5.79	6.05
Experimental pKa	4.66	4.37	4.38

Table 1. Calculated versus experimental pKa (sourced from: (Haynes, 2015)) for ibuprofen moieties.

Notably, both phenylacetic and 2-phenylpropanoic acid were outliers: with phenylacetic acid having the lowest experimental value of pKa, but a middling value for the dataset of pKa calculated from Gaussian, and 2-phenylpropanoic acid having the largest calculated value of pKa, but only the third largest experimental pKa of the compounds reported. As such, to ensure a strong correlation between the experimental and calculated values, these two compounds were removed from the set used in the regression. A polynomial regression was also performed as opposed to a linear one to further improve the R² value of the regression, which is shown in Figure 6.

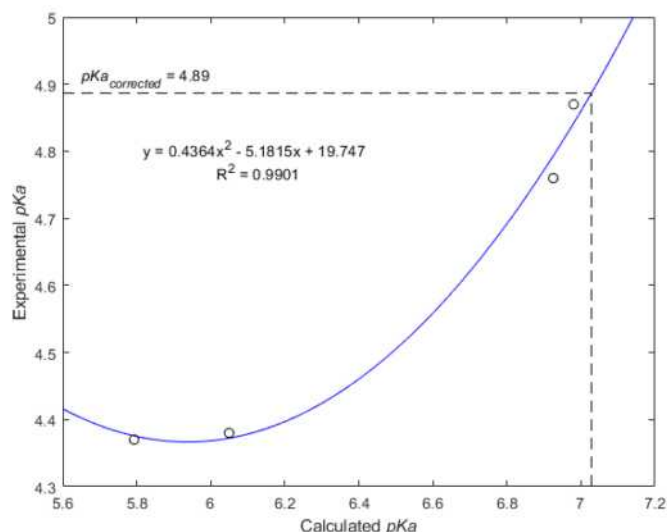


Figure 6. Regression plot for ibuprofen. The moiety data is denoted by the empty black circles, and the regression line is in blue. The dotted line displays how the corrected pKa was found.

As Figure 6 shows, the regression model for ibuprofen shows a great correlation between the experimental and calculated values of pKa, with an R^2 of 0.9901, appearing to remove much of the systematic error present in the calculations. This is further supported by the corrected value of ibuprofen's pKa, which is used throughout the rest of this section, falling well within the range of reported values from literature, lying only 0.05 pH points from the mean of the experimental pKas reported in Table 2. However, it must be noted that, due to none of the calculated pKas exceeding that of ibuprofen, to obtain the corrected value of pKa, an extrapolation of the regression model outside of the set of moieties was performed. Despite this, the adjusted value of pKa for ibuprofen is likely still valid as this extrapolation was only by 0.05 pH points.

Experimental Values of Ibuprofen's pKa			Corrected Calculated Value of Ibuprofen's pKa
4.45	4.91	5.17	4.89

Table 2. Comparison of ibuprofen pKas, experimental data was sourced from (from left to right) (Raamat, et al., 2012) (Sangster, 1994) (Wehbe, et al., 2022)

3.1.3. pH-dependent Solubility Profile

Utilising the corrected calculated value of pKa for ibuprofen alongside the model presented earlier in the report, the pH-dependent solubility profile for ibuprofen can be calculated, shown in Figure 7 below, compared to experimental data (Wehbe, et al., 2022), with pH being plotted against the logarithm of solubility to improve the legibility of the figure.

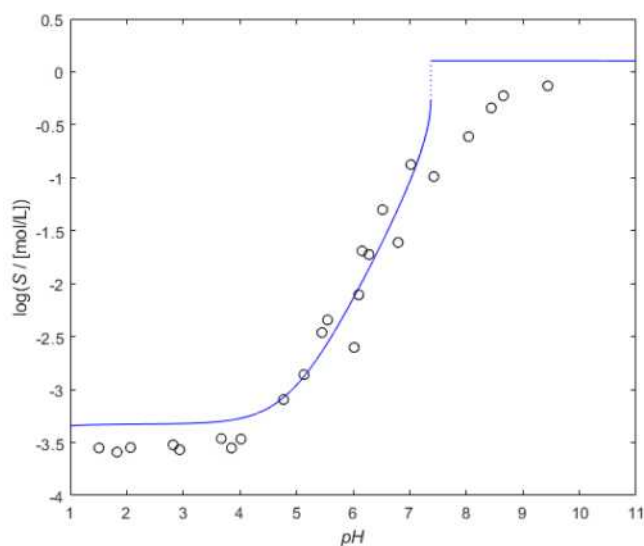


Figure 7. Regression plot for ibuprofen. The moiety data is denoted by the empty black circles, and the regression line is in blue. The dotted line displays how the corrected pKa was found.

As highlighted by Figure 7 above, an entirely computational approach to modelling solubility as a function of pH produced a decent qualitative result for ibuprofen, with the profile, especially for pH values between $4 < \text{pH} < 7$, fitting well to experimental data. In this region, the %AAD in calculated solubility from the reported literature data is 53.7%, compared to 125.6% across the entire pH range, although this value for the full region is inflated by the poor representation of solubility above pH^{max} by the model. pH^{max} was calculated by the simultaneous solution of the regular model equations in addition to the equation for solubility product, producing a value of 7.16, falling short of the measured value of 8.55 (Wehbe, et al., 2022), explaining why the profile fails to accurately predict solubility at these higher pHs. There was also a discontinuity produced by the model, shown by the dotted line on Figure 7, between the regions above and below pH^{max} , where the model fails to represent the "tailing-off" nature of the profile as pH approaches pH^{max} , as is seen in the experimental data. Furthermore, the model over-predicted solubility at low pH, likely due to the calculated value of pKa, equal to 4.89, being below the value reported alongside the experimental data shown in the figure, equal to 5.17 (Avdeef, 2007). If the experimental value of pKa is used instead when modelling, the profile's fit improves significantly, reducing the AAD for the solubility prediction from 0.18 mol/L to 0.05 mol/L, although still marginally overpredicting solubility generally.

The solubility of ibuprofen can also be modelled with the inclusion of the $\Delta C_{p,HA}$ terms of the solubility equation, but their addition, which is frequently neglected in literature (Febra, et al., 2021), makes no apparent difference to the predicted value of solubility. However, using these terms does surprisingly lead to numerical convergence issues

within gPROMS, causing the model to fail at values of pH close to the value of pKa, in the acidic region. It then may be of great interest to future work to directly quantify the importance of these terms to the modelling of thermodynamic properties, as if they provide little to no benefit, it is likely that removing them results in increased computational efficiency and may eliminate non-convergent regions when modelling.

3.2 Mefenamic Acid

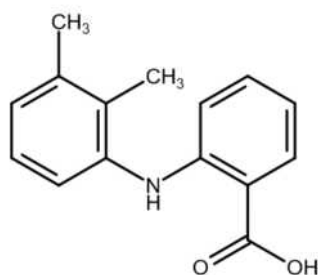


Figure 8. Chemical structure of mefenamic acid

3.2.1. Initial pKa Prediction

As with ibuprofen, the thermodynamic cycle method from Ho & Ertem was used to predict the pKa of mefenamic acid, utilising geometries optimised within Gaussian 16. From this, the initially calculated pKa for mefenamic acid was 5.68, outside of the range of reported experimental pKas from literature (Haynes, 2015) (Avdeef, 2007).

3.2.2. Correction of pKa Using Experimental Data

Due to the significant error present between the initially calculated pKa value and experimental data, a regression to correct the error generated by the thermodynamic cycle method was also performed for mefenamic acid. Due to its particularly unique structure however, only 3 moieties for mefenamic acid, with reliable literature-reported pKas, were considered in this study. These being: benzoic acid, 2-aminobenzoic acid, and fenamic acid. Values of experimental pKa (Haynes, 2015) (Kortum, et al., 1960) (Zapata, et al., 2014) and calculated pKa, found using same thermodynamic cycle method as for mefenamic acid, are reported in Table 3.

Compound	Benzoic acid	2-Aminobenzoic acid	Fenamic acid
Experimental pKa	4.20	4.95	2.86
Calculated pKa	5.77	6.30	5.16

Table 3. Experimental versus calculated pKas for mefenamic acid moieties

Due to only 3 moieties being considered, only a linear regression could be performed for mefenamic acid, which is shown in Figure 9 below:

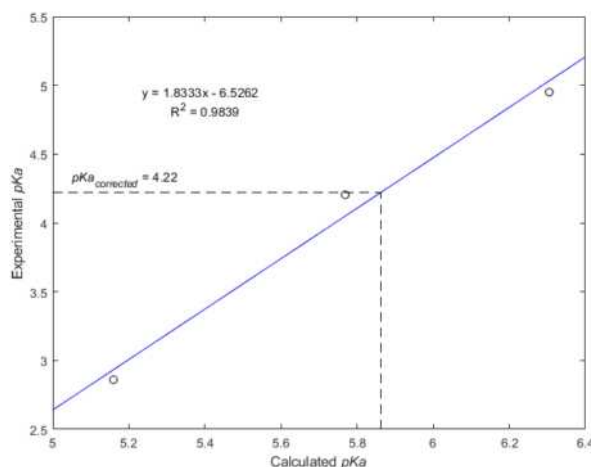


Figure 9. Regression plot for mefenamic acid. The empty black circles represent the moiety data with the regression line plotted in blue. The dotted line shows how mefenamic acid's corrected pKa was found.

The correction of pKa to experimental values for mefenamic acid, as with ibuprofen, was very successful, exhibiting a good fit, with an R^2 value of 0.9839. The corrected pKa value of mefenamic acid, used in the rest of the paper, was 4.22, 0.01 pKa points away from the average of the available values of pKa from literature, reported in Table 4. The improvement seen on Ho & Ertem's thermodynamic cycle method here was very encouraging, with the AAD from the average experimental pKa of the two compounds being 0.03 pKa points, a marked improvement from using the thermodynamic cycle method alone, which was reported to produce results with an AAD of 2.00 pKa points in literature.

Experimental Values of Mefenamic Acid's pKa			Corrected Calculated Value of Mefenamic Acid's pKa
3.88	4.20	4.54	4.22

Table 4. Comparison of mefenamic acid pKas, with experimental values from: (Domanska, et al., 2010), (Haynes, 2015), and (Avdeef, et al., 2007) respectively

3.1.3 pH-dependent Solubility Profile

With mefenamic acid's corrected calculated value of pKa obtained, the prediction of pH-dependent solubility for the compound could now be performed. Notably, due to a lack of experimental data on the solubility product of mefenamic acid, likely due to its extremely low solubility (Avdeef, et al., 2007) (meaning its dissolution into an ionic salt could not previously be analysed), there is no region of $\text{pH} > \text{pH}^{\text{max}}$ modelled in this work. However, plots of

solubility against pH using experimental data from literature also do not exhibit this region (Avdeef, 2007), meaning mefenamic acid's precipitation into a salt may not occur physically because of the limited amount of the API available in solution, causing the idea of mefenamic acid having a value for pH^{max} to likely be incorrect. As with ibuprofen, the $\Delta C_{p,HA}$ terms of the solubility equation were neglected when modelling mefenamic acid, and although, as previously stated, removing these terms may lead to easier model convergence, this was done out of necessity due to a lack of reported data on the value of $\Delta C_{p,HA}$ for the compound (Febra, et al., 2021). The pH-dependent solubility profile for mefenamic acid, compared to reported experimental values of solubility at various pHs (Avdeef, et al., 2007), is shown in Figure 10.

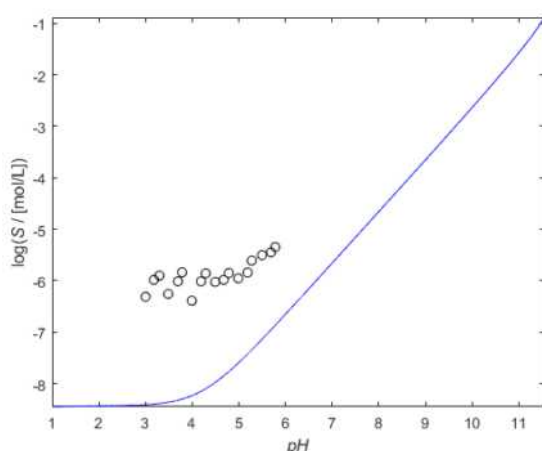


Figure 10. pH-dependent solubility profile for mefenamic acid. Experimental data is displayed as the empty black circles, and the calculated plot is in blue.

Similar discussion around the prediction of solubility as a function of pH for ibuprofen also applied to mefenamic acid. The entirely computational method produced a decent fit to values of experimentally determined solubility, and the shape of the profile corroborated well qualitatively to pH-solubility profiles for mefenamic acid produced using the Henderson-Hasselbalch equation in literature (Avdeef, et al., 2007). Although the apparent underestimation of solubility by the model may seem significant from the profile alone, the reader is reminded that the plot is against the logarithm of solubility, meaning the distance between the experimental points and the curve is enlarged, and that both the experimental and calculated values for mefenamic acid's solubility are extremely low. This means that, despite the %AAD of the profile from the experimental data being 98.57%, the AAD in solubility was only 1.53×10^{-6} mol/L, meaning, in terms of absolute values, the calculated solubility was close to that of the experimental data.

The model used to quantify the solubility of mefenamic acid used was also suboptimal, with the interactions between the aCNH₂ group and water being found using combining rules, and, for the anion,

the deprotonated carboxylic acid group and its adjacent aromatic carbon being modelled with a C and COO⁻ group, as opposed to the more accurate depiction using one aCCOO⁻ group. We attribute much of the error from experimental data to these shortcomings of the model, and as such, the accuracy of the model to experimental data despite these issues is highly promising for the efficacy of the use of the thermodynamic cycle method and SAFT- γ Mie GC in combination to predict pH-dependent solubility computationally.

4. Conclusion

As can be seen from the data presented, an entirely computational approach to predicting pH-dependent solubility can provide a good qualitative estimate of solubility for active pharmaceutical ingredients, even for nearly insoluble compounds such as mefenamic acid. However, the method outlined in this paper likely does not achieve the required precision in solubility prediction by the pharmaceutical industry and traditional experimental techniques such as the saturation shake-flask method or UV spectrophotometry should still be used to ensure an entirely accurate assessment of solubility as a function of pH. The method presented can however act as an initial screening step in the process of drug development in which candidate compounds too insoluble to be bioavailable could be identified without the need for the expensive and time-consuming methods outlined above, and with computational methods and SAFT constantly improving, extensions of the method detailed in this work will only continue to increase in accuracy.

The improvements on the prediction of pKa from the thermodynamic cycle method first presented by Ho & Ertem through an additional regression step are significant, with the deviation in calculated pKa from experimental values falling from 2.0 pH points for carboxylic acids (Ho & Ertem, 2016) to 0.05 and 0.01 pH points for ibuprofen and mefenamic acid respectively. Notably, the pKa and pH-dependent solubility prediction outlined in this work was only performed at 298.15K and atmospheric pressure in water and as such it may be of interest to future work to assess the viability of the methods presented at different temperatures and pressures as well as with different solvents. A major contributor to the error seen in the solubility prediction for mefenamic acid was attributed to the use of combining rules and improper modelling of the groups within the anion, and as such, more work must be done to expand the library of groups, and the parameters of their interactions with one another, to improve the predictive capacity of the SAFT- γ Mie GC EoS. The introduction of new parameters and groups to the framework does, however, still require more experimental assessments

of the thermodynamic properties of compounds to be performed. Furthermore, the impact of including the $\Delta C_{p,HA}$ terms in the solubility equation should be assessed more thoroughly as, for ibuprofen, no benefits were seen in this work by the introduction of these terms on the predictive ability of SAFT.

5. References

- Avdeef, A., 2007. Solubility of sparingly-soluble ionizable drugs. *Advanced Drug Delivery Reviews*, Volume 59, pp. 568-590.
- Avdeef, A. et al., 2007. Solubility-Excipient Classification Gradient Maps. *Pharmaceutical Research*, Volume 24, pp. 530-545.
- Belkadi, A. et al., 2010. Soft-SAFT modeling of vapor-liquid equilibria of nitriles and their mixtures. *Fluid Phase Equilibria*, Volume 289, pp. 191-200.
- Domanska, U. et al., 2010. Solubility and pKa of select pharmaceuticals in water, ethanol, and 1-octanol. *The Journal of Chemical Thermodynamics*, Volume 42, pp. 1465-1472.
- Febra, S. A. et al., 2021. Extending the SAFT- γ Mie approach to model benzoic acid, diphenylamine, and mefenamic acid: Solubility prediction and experimental measurement. *Fluid Phase Equilibria*, Volume 540.
- Gaussian, n.d. *GMMX Conformer Search*. [Online] Available at: <https://gaussian.com/gv6gmmx/>
- Gross, J. & Sadowski, G., 2001. Perturbed-Chain SAFT: An Equation of State Based on a Perturbation Theory for Chain Molecules. *Industrial & Engineering Chemistry Research*, Volume 40, pp. 1244-1260.
- Gui, L. et al., 2023. Integrating model-based design of experiments and computer-aided solvent. *Computers and Chemical Engineering*, Volume 177.
- Haynes, W., 2015. *CRC Handbook of Chemistry and Physics, 95th Edition*. Boca Raton (FL): CRC Press LLC.
- Ho, J. & Ertem, M. Z., 2016. Calculating Free Energy Changes in Continuum Solvation Models. *The Journal of Physical Chemistry*, Volume 120, pp. 1319-1329.
- Kelly, C. P., Cramer, C. J. & Truhlar, D. P., 2007. Single-Ion Solvation Free Energies and the Normal Hydrogen Electrode Potential in Methanol, Acetonitrile, and Dimethyl Sulfoxide. *The Journal of Physical Chemistry B*, 111(2), pp. 408-442.
- Kortum, G., Vogel, W. & Andrussow, K., 1960. Dissociation Constants of Organic Acids in Aqueous Solution. *Pure and Applied Chemistry*.
- Lafitte, T. et al., 2013. Accurate statistical associating fluid theory for chain molecules formed from Mie segments. *The Journal of Chemical Physics*, Volume 139.
- Papaioannou, V. et al., 2014. Group contribution methodology based on the statistical associating fluid theory for heteronuclear molecules formed from Mie segments. *The Journal of Chemical Physics*, Volume 140.
- Price, G. & Patel, D. A., 2023. *Drug Bioavailability*. Treasure Island (FL): StatPearls Publishing.
- Raamat, E. et al., 2012. Acidities of strong neutral Brønsted acids in different media. *Journal of Physical Organic Chemistry*, Volume 26, pp. 162-170.
- Raja, P. M. V. & Barron, A. R., 2023. *UV-Visible Spectroscopy*. [Online] Available at: [https://chem.libretexts.org/Bookshelves/Analytical_Chemistry/Physical_Methods_in_Chemistry_and_Nano_Science_\(Barron\)/04%3A_Chemical_Speciation/4.04%3A_A_UV-Visible_Spectroscopy](https://chem.libretexts.org/Bookshelves/Analytical_Chemistry/Physical_Methods_in_Chemistry_and_Nano_Science_(Barron)/04%3A_Chemical_Speciation/4.04%3A_A_UV-Visible_Spectroscopy)
- Sangster, J., 1994. *LOGKOW Databank*. Montreal, Quebec: Sangster Res. Lab..
- Shoghi, E., Fuguet, E., Bosch, E. & Ràfols, C., 2013. Solubility-pH profiles of some acidic, basic and amphoteric drugs. *European Journal of Pharmaceutical Sciences*, Volume 48(Issues 1–2).
- Soult, A., 2023. *Solubility*. [Online] Available at: [https://chem.libretexts.org/Courses/University_of_Kentucky/UK%3A_CHE_103_-_Chemistry_for_Allied_Health_\(Soult\)/Chapters/Chapter_7%3A_Solids_Liquids_and_Gases/7.7%3A_Solubility](https://chem.libretexts.org/Courses/University_of_Kentucky/UK%3A_CHE_103_-_Chemistry_for_Allied_Health_(Soult)/Chapters/Chapter_7%3A_Solids_Liquids_and_Gases/7.7%3A_Solubility)
- Surat, P., 2022. *pH in the Human Body*. [Online] Available at: <https://www.news-medical.net/health/pH-in-the-Human-Body.aspx>
- University of Liverpool, 2023. [Online] Available at: https://hep.ph.liv.ac.uk/~burdin/phys132/lecture_2.pdf
- Wehbe, M., Haslam, A. J., Jackson, G. & Galindo, A., 2022. Phase behaviour and pH-solubility profile prediction of aqueous buffered solutions of ibuprofen and ketoprofen. *Fluid Phase Equilibria*, Volume 560.
- Zapata, L., Woznicka, E. & Kalemekiewicz, J., 2014. Tautomeric and Microscopic Protonation Equilibria of Anthranilic Acid and Its Derivatives. *Journal of Solution Chemistry*, Volume 43, pp. 1167-1183.

On the Solubility and Recovery of Gypsum with Ionic Liquids

Sinclair Mabon and Zhongqi Zhuang

Department of Chemical Engineering, Imperial College London, U.K.

Abstract

The solubility of calcium sulphate dihydrate, commonly known as gypsum, was investigated in two ionic liquids: triethylammonium hydrogen sulphate [TEA][HSO₄] and monoethanolamine citrate [MEA][Cit.], with water as a co-solvent. A design of experiments was employed to systematically evaluate the influence of varying water content, temperature, acid-base ratio, and solid loading on gypsum solubility. Results indicated limited solubility in [TEA][HSO₄] ionic liquid, while [MEA][Cit.] exhibited, based on visual observations, remarkable gypsum solubilization, reaching preliminary estimates of up to 81 g_{CaSO₄} kg_{IL}⁻¹. The study also explored antisolvents and additives for solvent recovery, including acetone, ethanol, methanol, ammonium hydroxide, and sulfuric acid. Unfortunately, none of the tested antisolvents allowed for simultaneous recovery of solute and solvent. While [MEA][Cit.] IL presents a promising avenue for gypsum removal in industrial applications, further research is essential to devise efficient methods for solvent recycling. These findings underscore the potential of ionic liquids in addressing challenges associated with gypsum solubility, while highlighting the need for continued investigation into solvent recovery strategies.

Introduction

Calcium sulphate dihydrate [CaSO₄•2H₂O], better known as gypsum, is encountered in a multitude of industries; its uses include construction material in drywall, mineral removal in hydrometallurgical processes, and water treatment^[1-3]. However, gypsum's relatively low solubility in aqueous solutions and propensity to precipitate poses operational issues. The inorganic salt is known to form blockage-inducing crud as well as hold principal responsibility for scaling in reverse osmosis and desalination processes, and in unit equipment such as heat exchangers. The presence of gypsum in these systems thusly decreases efficiency and exacerbates maintenance costs^[4-7]. Estimates place the cost fouling in major industries at USD 4.4 billion annually^[8]. As such, gypsum's widespread commercial usage and challenges justifies further investigation into its solubility.

In recent years, there has been much excitement on the use of ionic liquids (ILs) as solvents. ILs are organic salts which remain in liquid form below an arbitrary temperature limit, usually $T_m < 100^\circ\text{C}$. This comparatively low melting temperature is possible due the constituent ions' bulkiness and asymmetry resulting in weakened ionic interactions and lattice energy, see Figure 1. ILs may be regarded as designer solvents in which the selection and properties of the constituent ions may be altered to formulate an optimised solvent. For example, the acid-base ratios may be manipulated to produce non-stoichiometric ILs^[9-10]. This has led to research into the effect of using IL additives in brine solutions to dissolve gypsum^[11-12]. Yet, no investigation has been undertaken to test the solubility of gypsum in ILs doped with H₂O. To bridge this gap, this paper will study the solubility of calcium sulphate dihydrate in triethylammonium hydrogen sulphate IL [TEA][HSO₄] and monoethanolamine citrate IL [MEA][Cit.] with varying water contents, acid-base ratios, solid loadings, and temperatures as well as explore appropriate antisolvents for solvent recovery.

Background

A considerable portfolio of research has been accumulated over the years on the solubility of calcium sulphate. Taherdangkoo *et al.* (2022) compiled 42 of these papers into a detailed literature review on the dissolution of gypsum and anhydrous calcium sulphate in aqueous brine solutions^[13].

Given the composition of ionic liquids, works relating to acidic solutions were of particular significance; few investigations have developed

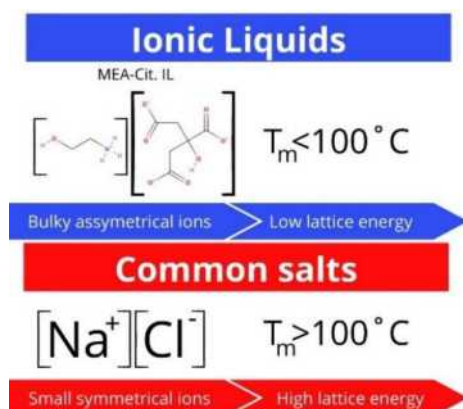


Figure 1. Comparison of ionic liquids and common salts

alkaline solvents. In studies where H₂SO₄ solutions were used to solubilize gypsum, it was found that solubility increases with increasing acidity at constant temperature up to ~0.6M. Above this threshold, increases in acidity are instead inversely proportional to solubility. This is believed to be caused by the common ion effect of deprotonated H₂SO₄ forming SO₄²⁻ in tandem with a salting out effect. A similar trend was observed with temperature at constant acidity: increasing solubility up to 80°C after which further raising the temperature will reduce the dissolved gypsum content^[14-19]. Muhammed & Zhang (1989) detailed that at 80°C gypsum converts into its even more insoluble anhydrous form, explaining the witnessed result^[20].

Most studies, however, have focused on using salt solutions to dissolve gypsum. Solutions of NaCl specifically are most prominently featured within literature. Zhang *et al.* (2013), Kumar *et al.* (2005), Li & Demopoulos (2005), and Nakayama (1971) all recognized the relationship of increasing gypsum solubility with increasing NaCl in solution^[21-24]. It was within this context of NaCl solutions that the effect of IL addition was first investigated. Shukla *et al.* (2018) and Shukla *et al.* (2019) studied the dissolution of gypsum in NaCl aqueous solutions upon the introduction of up to 15wt% IL. Shukla *et al.* (2018) employed ethylammonium lactate IL and imidazolium-based hydrogen sulphate ILs while Shukla *et al.* (2019) used hydroxyalkyl ammonium acetate ILs. It was found that imidazolium-based hydrogen sulphate IL additives reduced the solubility of gypsum by ~60% attributed to the common ion effect of SO₄²⁻. Oppositely, the ILs utilizing carboxylic acids as anions instead experienced a significant increase in solubility. Moreover, the effect of acetate was less pronounced than that of lactate^[11-12].

A possible explanation lies in the chelating properties of lactate. Chelation arises when an organic complexing agent, known as a chelating agent, forms multiple encircling bonds with a single metal ion. A chelating agent is a complex with several bonding sites; denticity is the measure of bonding sites in a given complex. Bidentate describes a complex with two possible bonding sites, for three or more sites, tridentate and polydentate can be used. Carboxylic and hydroxyl groups are both known as possible bonding sites contributing to chelation^[25-26]. Bidentate lactate has been shown to chelate to Cu²⁺, Ni²⁺, Co²⁺, and Zn²⁺ hence supporting the hypothesis of Ca²⁺ simultaneously coordinating to the deprotonated oxygens of the carboxylic and hydroxyl groups^[27]. Hence, chelating complexes can form the basis for

sequestering agents of metals. For example, Murtaza *et al.* (2022) used a strong chelating agent, hexadentate EDTA, at high temperatures to solubilize gypsum with superb results^[28]. Of course, chelating agents are not without complications. These compounds may form gels or gelatinous precipitates when introduced to metallic cations. The gelation mechanism is explained by the metal ions acting as crosslinking agents for chelates which promotes hydrolysis, condensation, and finally polymerisation. This gelation reaction is favoured by high temperatures. Indeed, EDTA and tridentate citrate, due to their chelating properties, can form such polymers, and Ca²⁺ has been shown as a crosslinking agent^[29-31], for example in inducing pectate gelation^[32].

Lastly, good solvents should be recyclable which entails describing routes for the removal of the dissolved solutes in solution. One such route is the employment of antisolvents to facilitate precipitation. The basic principle involves a reduction in solute solubility caused by the addition of the antisolvent which, through a salting out effect, results in the precipitation of the dissolved compound of interest. This antisolvent should then be easily separated from the solution to allow for the repetition of the process. A major disadvantage of antisolvent solute removal is the high dependency on good mixing; otherwise, a heterogeneous mixture may form with remaining solute in solution^[33-34].

Methods

For the synthesis of triethylammonium hydrogen sulphate [TEA][HSO₄], a round bottom flask containing triethylamine (TEA) was placed in an ice bath for temperature control and concentrated sulfuric acid (97%) was introduced drop wise using an addition funnel. The reaction mixture was left to stir overnight to ensure a complete reaction takes place.

In the case of monoethanolamine citrate [MEA][Cit.] synthesis, a similar procedure was followed with the addition of monoethanolamine (MEA) dropwise, via an addition funnel, to the round bottom flask containing solid citric acid. The round bottom flask was once more immersed in an ice bath to control temperature and prevent the decomposition of amines. The reaction mixture was also stirred overnight.

To determine water content of the synthesized ILs, volumetric titration was employed, using a Mettler Toledo Karl Fischer volumetric titrator. Acid-base ratio of [TEA][HSO₄] was quantified using a Mettler Toledo KF benchtop titrator and NMR was employed to determine [MEA][Cit.]’s acid-base ratio.

An initial 2 level, 4 factor experimental design (DoE) was configured for [TEA][HSO₄], where four selected variables were employed to ascertain the predominant factor influencing solubility. The 4 variables under investigation include acid-base ratio, water content, reaction temperature, and solid loading. A 2 level, 4 factor DoE requires 16 experimental runs. Table 1 displays the 16 necessary experimental runs along with their

respective configurations. Table 2 outlines the positive, negative and centre levels for each variable in the DoE. Replicates were carried out for each experiment to increase validity. Triplicate centre level experiments were carried out to further reduce errors. The total weight of all samples was 10g.

Table 1: [TEA][HSO₄] DoE

2 level, 4 factor (2 ⁴ = 16)	Acid – base ratio	Water content (%)	Experiment Temperature [C]	Solids loading (%)
1	+	+	+	+
2	+	+	-	+
3	+	-	+	+
4	-	+	+	+
5	-	+	-	+
6	-	-	+	+
7	+	-	-	+
8	-	-	-	+
9	+	+	+	-
10	+	+	-	-
11	+	-	+	-
12	-	+	+	-
13	-	+	-	-
14	-	-	+	-
15	+	-	-	-
16	-	-	-	-

Table 2: Levels of DoE

Variable	Positive level	Negative level	Centre level
Acid - base ratio	1.50	1.00	1.25
Water content (%)	80	20	50
Temp. [C]	100	20	60
Solid loading (%)	20	10	15

Preliminary testing with the citrate ionic liquid showed gelation at elevated temperatures. Thus, a 2 level, 3 factor DoE was devised for [MEA][Cit.] IL, excluding temperature as a variable. Table 3 details the necessary experimental runs and Table 4 specifies the positive, negative and centre levels. Note that for the first DoE, acid-base ratio was the first variable, whilst for the second DoE the variable becomes MEA-Citrate ion ratio, more akin to base-acid ratio. This was due to the poor results acidic conditions gave and thus a change in approach was needed. All sample weights were again 10g.



Figure 2: Experimental setup

Figure 2 shows the experimental setup. To quantify solubility of CaSO₄ in the ionic liquids, inductively coupled plasma mass spectrometry (ICP-MS) was employed to measure concentrations of Ca²⁺ ions in solution. Serial dilutions at a factor of 100,000 were carried out to ensure sufficiently low concentrations of IL in the solutions, thereby securing optimal results from the ICP analysis.

Antisolvents and additives, including, ethanol, acetone, methanol, ammonium hydroxide (NH₄OH, 37%), and sulfuric acid (72%), were

investigated to assess their ability to precipitate dissolved CaSO_4 . Methanol, NH_4OH and sulfuric acid were miscible, whilst ethanol and acetone were immiscible. For immiscible systems, thorough

mixing was employed with a vortex machine to enhance the dispersion and emulsification of the two phases.

Table 3: $[\text{MEA}]/[\text{Cit.}]$ DoE

2 level, 3 factor ($2^3 = 8$)	MEA – citrate ion ratio	Water content (%)	Solids loading (%)
1	+	+	+
2	+	-	+
3	+	+	-
4	+	-	-
5	-	+	+
6	-	-	+
7	-	+	-
8	-	-	-

Table 4: Levels of DoE

Variable	Positive level	Negative level	Centre level
MEA – citrate ion ratio	3	1	2
Water content (%)	80	40	60
Solid loading (%)	7.5	2.5	5.0

Results

Figures 3, 4, 5 and 6 show samples post-experiment. As seen in Figures 3 and 4, $[\text{TEA}][\text{HSO}_4]$ exhibited minimal dissolution of CaSO_4 . Figures 5 and 6 however, shows $[\text{MEA}][\text{Cit.}]$'s significantly improved results. Although the intention was to employ ICP-MS for precise analysis, due to technical failures experienced by the detector at the time of writing, the analysis of samples was not possible. In light of this, an estimation of the maximum concentration of dissolved CaSO_4 in $[\text{MEA}][\text{Cit.}]$ was derived based on the known information that 9.25g of $[\text{MEA}][\text{Cit.}]$ fully dissolved 0.75g of CaSO_4 in sample 2. This equates to 81.10 $\text{gCaSO}_4/\text{kg}_{\text{solvent}}$. Gelation was apparent in sample 1, and samples containing excess MEA displayed superior solubility of gypsum. A dataset containing solubilities of CaSO_4 in many different solvents from 42 studies complied by Taherdangkoo *et al.* (2022)^[11] was analysed and represented graphically in Figure 11. Only studies that investigated gypsum ($\text{CaSO}_4 \cdot 2\text{H}_2\text{O}$) were included in the analysis.



Figure 3: $[\text{TEA}][\text{HSO}_4]$ post-experiment (1-8)

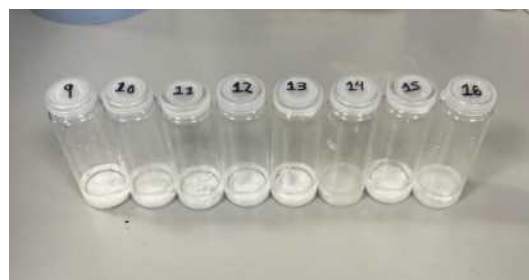


Figure 4: $[\text{TEA}][\text{HSO}_4]$ post-experiment (9-16)



Figure 5: $[\text{MEA}][\text{Cit.}]$ post-experiment

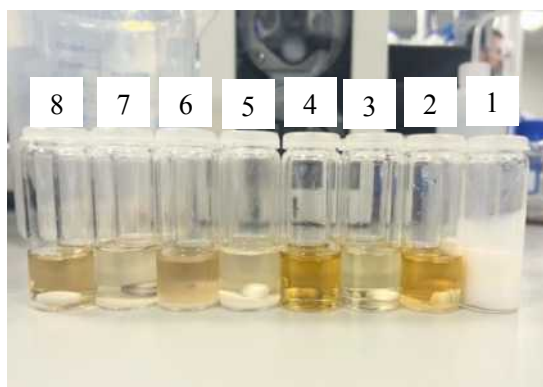


Figure 6: [MEA][Cit.] post-experiment

Out of the immiscible systems, only acetone resulted in the precipitation of CaSO_4 . A triphasic system emerged after vortex mixing, with acetone constituting the uppermost layer, the lower layer comprising the ionic liquid, and CaSO_4 interposed in the intermediary layer, as seen in Figure 7. Different amounts of acetone were investigated to examine potential effects on precipitation and no discernible effect was observed.

Ethanol, when subjected to vortex mixing, also formed a triphasic system, as shown in Figure 9; however, following a short settling period, the system reverted back to a biphasic state.

Ammonium hydroxide (37%) resulted in the gelation of samples after vortex mixing, as seen in Figure 8.

Sulfuric acid and methanol showed the most promising results. Concentrated sulfuric acid (97%) resulted in the immediate precipitation of CaSO_4 , as seen in Figure 10.

Methanol caused precipitation of solute after vortex mixing, and the resulting system can be observed in Figure 8.

Table 5 provides a summary of the antisolvents employed and their corresponding effects.

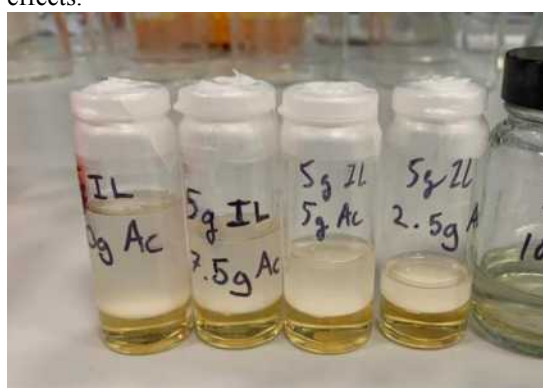


Figure 7: Antisolvent - acetone



Figure 8: Antisolvent - NH_4OH (37%) (Left) and methanol (Right)



Figure 9: Antisolvent – ethanol



Figure 10: Antisolvent: H_2SO_4 (97%)

Table 5: Solvent summary

Solvent	Number of phases	Precipitation
Methanol	1	Yes
Sulfuric acid (97%)	1	Yes
NH ₄ OH (37%)	1	No – gelation
Ethanol	2	No
Acetone	3	Yes

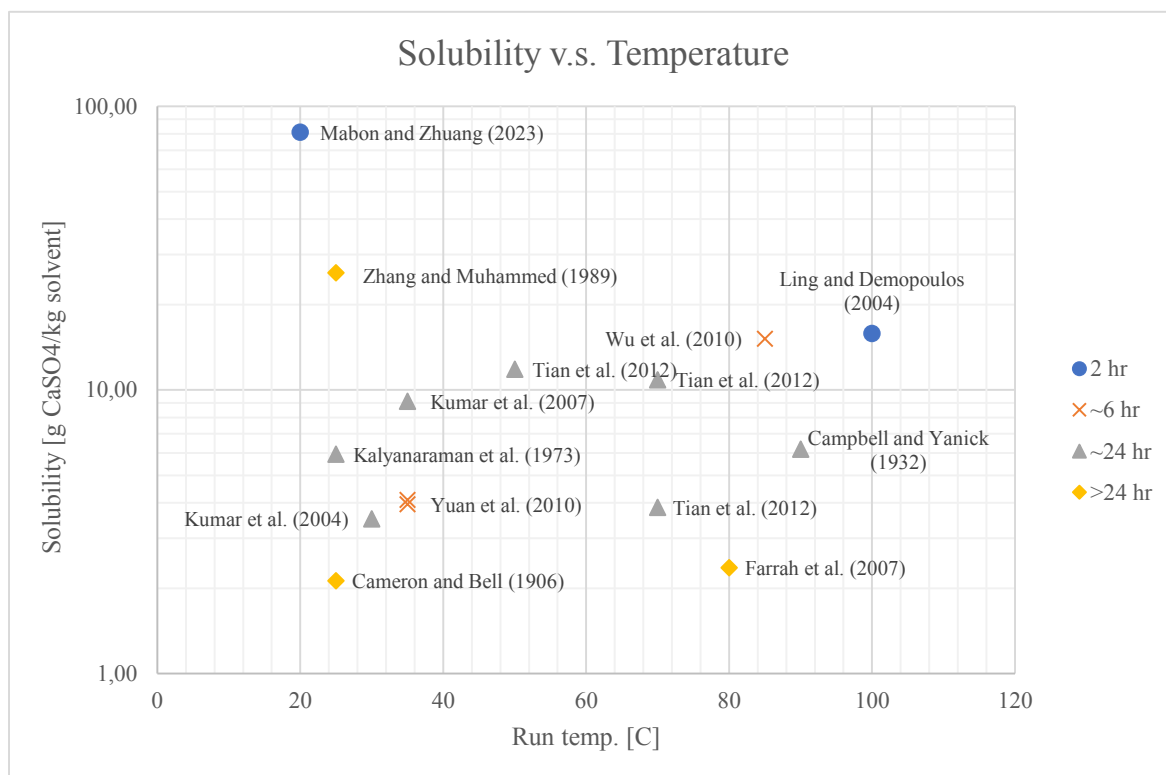


Figure 11: Semi-logarithmic comparative plot of solubilities of gypsum in literature^[13, 17, 19, 35-43]

Discussion

As seen in Figures 3 and 4 there is a large quantity of solid gypsum in solution implying poor solubility in [TEA][HSO₄] IL. This result agrees with that determined in Shukla *et al.* (2018) which utilized imidazolium-based hydrogen sulphate IL added to brine solution leading to a decrease in solubility of gypsum^[11]. This low solubility may be explained through a combination of the common ion effect of SO₄²⁻ formed from deprotonated hydrogen sulphate shifting the equilibrium to solid gypsum and a salting out effect caused by the large concentration of ions in solution (see Figure 12). This theory concurs with the solubility of gypsum witnessed in solutions of high molarity H₂SO₄^[14-20].

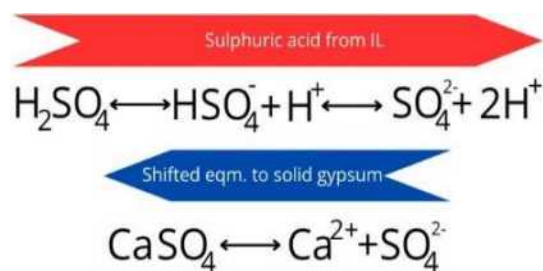


Figure 12: Equilibrium of H₂SO₄ & CaSO₄

Figures 5 and 6 demonstrate the high affinity of gypsum to solubilize in the newly synthesised IL [MEA][Cit.]. As seen in Figure 11, [MEA][Cit.] IL could potentially hold up to 81 g_{CaSO₄} kg_{IL}⁻¹. Shukla *et al.* (2018) and Shukla *et al.* (2019), which also employed carboxylic acids as anions, likewise noted an increase in gypsum dissolution^[11-12]. Promising results were similarly detailed by Murtaza *et al.* (2022) using EDTA at high temperatures^[28]. This may be described by the chelating properties of the above compounds, also found in citrate. Crucially, citrate composes of three carboxylic groups and one hydroxyl group, all of which may deprotonate to form bonding sites in the form of negatively charged oxygen atoms. These bonding sites then allow the simultaneous encircling coordination of the cationic divalent calcium in solution to form a chelate complex thus solubilizing gypsum^[25-26], an adapted schematic of which is seen in Figure 13.

Another observation made evident by Figures 5 & 6 and Table 3 is the dependency of base ratio on solubility: with excess base, the solubility increases. This relationship could be explained through the accompanying increase in denticity by introducing proton receptors. The samples with excess alkaline MEA likely experienced a heightened deprotonation of the carboxyl and hydroxyl groups present in citrate hence increasing the possible number of bonding sites to chelate to Ca²⁺. As such, these samples seemingly displayed

superior solubility of gypsum as compared to those using stoichiometrically proportionate ILs^[25-26]. Citrate's deprotonation dependency on pH is visualized in Figure 14.

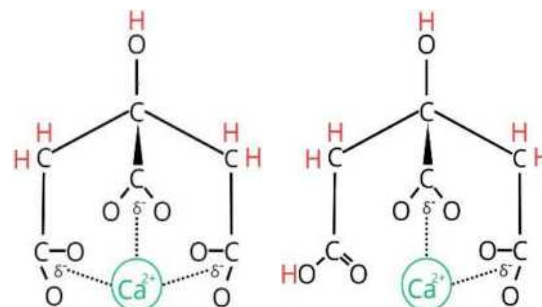


Figure 13: Proposed Cit-Ca²⁺ chelate complexes^[44]

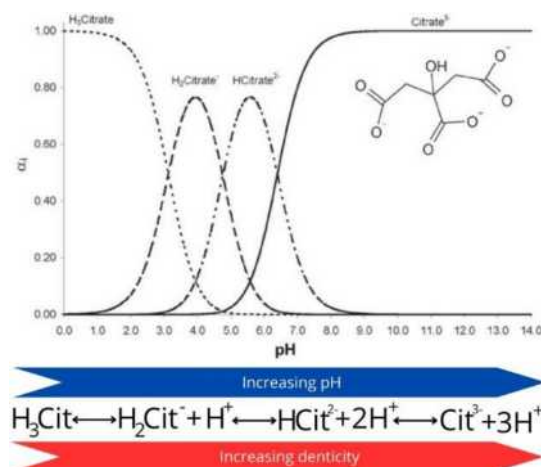


Figure 14: Cit's deprotonation dependency on pH^[45]

The gelation observed in high temperature preliminary tests of [MEA][Cit.] IL may be explained by the reaction being favoured at elevated temperatures. Furthermore, the gelation witnessed in Figures 5 and 6 in sample 1, representing high water content and solid loading with excess base (see Table 3) may be interpreted through a combination of denticity, crosslinking agents, and steric hindrance. Firstly, an excess of base leads to an increase in citrate denticity, as detailed previously, which forms more chelate complexes^[25-26]. Secondly, metallic ions, such as Ca²⁺^[32], can act as crosslinking agents between chelate complexes hence causing a polymerisation reaction which ultimately constructs a gel or gelatinous precipitate. The higher solid loading used in the sample thusly ensures there is a sufficient amount of Ca²⁺ to form chelates and to function as crosslinking agents in this polymerisation^[29-31]. Although both acid-base ratio and solid loading were identical in sample 2, no gelation was observed. It is hypothesized that the effect on gelation of increased water content lies in a reduction of steric hindrance. In low water content

conditions, it is believed that the bulky constituent ions of the IL introduce steric hindrance to the formation of the 3D structures necessary for the polymerisation to progress. Studies have shown that gelation may be suppressed by steric hindrance of bulky molecules, bolstering this theory^[46].

Regarding the performance of the tested antisolvents, it was found that acetone, ethanol, and methanol were all ineffective at precipitating gypsum to allow for solvent recycling, see Figures 7, 8, and 9. Antisolvents function by invoking a salting out effect of the solute; however, this effect can be inhibited by poor mixing^[33-34]. This is the believed reasoning of the dissatisfactory performance of the above antisolvents. The immiscibility of the antisolvents in the high ionic strength IL solution is believed to be due to their relatively low dielectric constants of 21.1, 24.5, and 33 respectively as compared to H₂O at 78 which is fully miscible^[47]. Therefore, these antisolvents induce constant demixing promoting a heterogeneous mixture. The comparatively higher dielectric constant, leading to superior mixing, of methanol may describe its improved performance. Couple this with the high stability of the chelate complex, due to the simultaneous coordination of several ligands to the Ca²⁺ ^[28], and a sparingly precipitable metal complex is formed. Alkaline ammonium hydroxide NH₄OH was similarly inefficient in precipitating gypsum; instead forming a gel, see Figure 8. Briefly explained, this is hypothesized to be due to the basic salt causing deprotonation of the citrate, thus increasing the denticity and number of chelates to polymerize^[25-26, 29-31]. The delivery in 37% aqueous solution is postulated to also limit steric hindrance via the introduction of H₂O^[46]. These factors function in tandem to form a gel. Lastly, using sulphuric acid as an additive was recognized as precipitating considerable amounts of gypsum, see Figure 10. It is posited that this is caused by a trinity of complimentary effects. Most prominently, the high acidic strength of H₂SO₄ is believed to protonate the citrate hence reducing its denticity and ability to chelate^[25-26]. Moreover, the addition of sulphate ions in solution, as seen in many other solubility studies, leads to the precipitation of gypsum due to the combined common ion effect and salting out effect^[14-24]. Although sulphuric acid allows for excellent solute recovery, the presence of SO₄²⁻ in solution prevents full solvent recycling.

Conclusion & Outlook

To address the potential challenges posed by gypsum in industry, this study has delved into the solubility of calcium sulphate dihydrate in IL

solvents and explores potential antisolvents. The findings reveal that gypsum exhibits limited solubility in [TEA][HSO₄] IL, while [MEA][Cit.] IL demonstrates significant promise, reaching preliminary estimates of up to 81 g_{CaSO₄} kg_{IL}⁻¹ at mild conditions. Despite these encouraging results, the study identifies a lack of suitable antisolvents or additives for solvent recovery, hindering the potential industrial applicability of [MEA][Cit.]. Thus, an emphasis is placed on the need for further exploration into developing effective antisolvent or electrochemical separation techniques.

This paper suggests a potential future avenue: the alternating introduction of H₂SO₄ for solute recovery and Ca(OH)₂ for solvent recovery and gypsum formation. Additionally, this study proposes experimenting with lower-order polydentate chelating agents like lactate in IL solvents to weaken the chelate complex, facilitating more readily precipitable solutions. Furthermore, optimization of reaction conditions is highlighted as a crucial aspect to maximize solubility while minimizing the risk of gelation. These insights underscore the potential of ILs in addressing gypsum solubility challenges, prompting further research to refine and expand the practical applications of this novel approach.

Acknowledgements

We would like to thank Dr. Pedro “Pedrito” Nakasu for his expertise, advice, and friendship, without which this research could never have been realized. And to Dr. Amir Dezashibi, our serial dilution guru and advisor.

Lastly, we would like to extend our gratitude to M.Res Jiaying “Suri” Chen whose kind and humorous personality made laboratory work a joy.

References

1. Bumanis, G. et al. (2022) ‘Processing of gypsum construction and demolition waste and properties of secondary gypsum binder’, *Recycling*, 7(3). doi:10.3390/recycling7030030.
2. Zhang, D. et al. (2015) ‘Incorporation of arsenic into gypsum: Relevant to arsenic removal and immobilization process in hydrometallurgical industry’, *Journal of Hazardous Materials*, 300, pp. 272–280. doi:10.1016/j.jhazmat.2015.07.015.
3. Igwegbe, C. et al. (2019) ‘Utilization of calcined gypsum in water and wastewater treatment: Removal of phenol’, *Journal of Ecological Engineering*, 20(7), pp. 1–10. doi:10.12911/22998993/108694.
4. Kristian Hirsi, T., Vaarno, J. and Salonen, P. (2013) ‘Outotec® cooling towers provide cooling efficiency and low emissions in gypsum removal in SX plants’, *Base Metals Conference 2013*

- [Preprint]. Available at: <http://saimm.org.za/Conferences/BM2013/065-Hirsi.pdf>.
5. Rolf, J. et al. (2022) 'Inorganic scaling in membrane desalination: Models, mechanisms, and characterization methods', *Environmental Science ; Technology*, 56(12), pp. 7484–7511. doi:10.1021/acs.est.2c01858.
 6. Liu, Q., Xu, G.-R. and Das, R. (2019) 'Inorganic scaling in reverse osmosis (Ro) desalination: Mechanisms, monitoring, and inhibition strategies', *Desalination*, 468. doi:10.1016/j.desal.2019.07.005.
 7. Chagwedera, T.M., Chivavava, J. and Lewis, A.E. (2022) 'Gypsum seeding to prevent scaling', *Crystals*, 12(3). doi:10.3390/cryst12030342.
 8. Ben-Mansour, R. et al. (2023) 'Experimental/Numerical Investigation and prediction of fouling in Multiphase Flow Heat Exchangers: A Review', *Energies*, 16(6), p. 2812. doi:10.3390/en16062812.
 9. Krossing, I. et al. (2006) 'Why are ionic liquids liquid? A simple explanation based on lattice and solvation energies', *Journal of the American Chemical Society*, 128(41), pp. 13427–13434. doi:10.1021/ja0619612.
 10. Kaur, G., Kumar, H. and Singla, M. (2022) 'Diverse applications of ionic liquids: A comprehensive review', *Journal of Molecular Liquids*, 351. doi:10.1016/j.molliq.2022.118556.
 11. Shukla, J., Mehta, M.J. and Kumar, A. (2018) 'Effect of ionic liquid additives on the solubility behavior and morphology of calcium sulfate dihydrate (gypsum) in the aqueous sodium chloride system and physicochemical solution properties at 30 °C', *Journal of Chemical ; Engineering Data*, 63(8), pp. 2743–2752. doi:10.1021/acs.jced.8b00093.
 12. Shukla, J., Mehta, M.J. and Kumar, A. (2019) 'Solubility behavior of calcium sulfate dihydrate (gypsum) in an aqueous sodium chloride system in the presence of hydroxyalkyl ammonium acetate ionic liquids additives: Morphology changes and physicochemical solution properties at 35 °C', *Journal of Chemical Engineering Data*, 64(12), pp. 5132–5141. doi:10.1021/acs.jced.9b00354.
 13. Taherdangkoo, R. et al. (2022) 'Experimental data on solubility of the two calcium sulfates gypsum and anhydrite in Aqueous Solutions', *Data*, 7(10). doi:10.3390/data7100140.
 14. Wang, W. et al. (2013) 'Experimental determination and modeling of gypsum and insoluble anhydrite solubility in the system $\text{CaSO}_4\text{--H}_2\text{SO}_4\text{--H}_2\text{O}$ ', *Chemical Engineering Science*, 101, pp. 120–129. doi:10.1016/j.ces.2013.06.023.
 15. Azimi, G. and Papangelakis, V.G. (2010) 'Thermodynamic modeling and experimental measurement of calcium sulfate in complex aqueous solutions', *Fluid Phase Equilibria*, 290(1–2), pp. 88–94. doi:10.1016/j.fluid.2009.09.023.
 16. Dutrizac, J.E. (2002) 'Calcium sulphate solubilities in simulated zinc processing solutions', *Hydrometallurgy*, 65(2–3), pp. 109–135. doi:10.1016/s0304-386x(02)00082-8.
 17. Farrah, H.E., Lawrance, G.A. and Wanless, E.J. (2007) 'Solubility of calcium sulfate salts in acidic manganese sulfate solutions from 30 to 105 °C', *Hydrometallurgy*, 86(1–2), pp. 13–21. doi:10.1016/j.hydromet.2006.10.003.
 18. Mutalala, B.K., Umetsu, Y. and Tozawa, K. (1989) 'Solubility of calcium sulfate in acidic copper sulfate solutions over the temperature range of 298 to 333 K', *Materials Transactions, JIM*, 30(6), pp. 394–402. doi:10.2320/matertrans1989.30.394.
 19. Calmanovici, C.E., Gabas, N. and Laguerie, C. (1993) 'Solubility measurements for calcium sulfate dihydrate in acid solutions at 20, 50, and 70 °C', *Journal of Chemical Engineering Data*, 38(4), pp. 534–536. doi:10.1021/je00012a013.
 20. Zhang, Y. and Muhammed, M. (1989) 'Solubility of calcium sulfate dihydrate in nitric acid solutions containing calcium nitrate and phosphoric acid', *Journal of Chemical Engineering Data*, 34(1), pp. 121–124. doi:10.1021/je00055a032.
 21. Li, Z. and Demopoulos, G.P. (2005) 'Effect of NaCl, MgCl₂, FeCl₂, FeCl₃, and AlCl₃ on solubility of CaSO₄ phases in aqueous HCl or HCl + CaCl₂ solutions at 298 to 353 K', *Journal of Chemical Engineering Data*, 51(2), pp. 569–576. doi:10.1021/je0504055.
 22. Zhang, Y. et al. (2013) 'Effect of chloride salts and bicarbonate on solubility of CaSO₄ in aqueous solutions at 37 °C', *Procedia Environmental Sciences*, 18, pp. 84–91. doi:10.1016/j.proenv.2013.04.012.
 23. Kumar, A. et al. (2005) 'Ionic interactions of calcium sulfate dihydrate in aqueous sodium chloride solutions: Solubilities, densities, viscosities, electrical conductivities, and surface tensions at 35 °C', *Journal of Solution Chemistry*, 34(3), pp. 333–342. doi:10.1007/s10953-005-3053-0.
 24. Nakayama, F.S. (1971) 'Calcium complexing and the enhanced solubility of gypsum in concentrated sodium-salt solutions', *Soil Science Society of America Journal*, 35(6), pp. 881–883. doi:10.2136/sssaj1971.03615995003500060013x.
 25. Pilgrim, C.D. (2018) 'Chelation', *Encyclopedia of Earth Sciences Series*, pp. 233–234. doi:10.1007/978-3-319-39312-4_46.
 26. Crisponi, G. and Nurchi, V.M. (2016) 'Chelating agents as therapeutic compounds—basic principles', *Chelation Therapy in the Treatment of Metal Intoxication*, pp. 35–61. doi:10.1016/b978-0-12-803072-1.00002-x.

27. Cariati, F. et al. (1977) 'Chelating properties of lactate anion. Perturbing effect of additional ligands on bis(dl-lactato)-metal(ii) complexes', *Inorganica Chimica Acta*, 21, pp. 133–140. doi:10.1016/s0020-1693(00)86249-0.
28. Murtaza, M. et al. (2022) 'Single step calcium sulfate scale removal at high temperature using tetrapotassium ethylenediaminetetraacetate with potassium carbonate', *Scientific Reports*, 12(1). doi:10.1038/s41598-022-14385-6.
29. Bai, H. et al. (2011) 'On the gelation of graphene oxide', *The Journal of Physical Chemistry C*, 115(13), pp. 5545–5551. doi:10.1021/jp1120299.
30. Magami, S.M. and Williams, R.L. (2019) 'Gelation via cationic chelation/crosslinking of acrylic-acid-based polymers', *Polymer International*, 68(12), pp. 1980–1991. doi:10.1002/pi.5910.
31. Kakihana, M. (1996) 'Invited review "sol-gel" preparation of high temperature superconducting oxides', *Journal of Sol-Gel Science and Technology*, 6(1), pp. 7–55. doi:10.1007/bf00402588.
32. Donati, I., Benegas, J. and Paoletti, S. (2021) 'On the molecular mechanism of the calcium-induced gelation of pectate. different steps in the binding of calcium ions by pectate', *Biomacromolecules*, 22(12), pp. 5000–5019. doi:10.1021/acs.biomac.1c00958.
33. Mostafa Nowee, S., Abbas, A. and Romagnoli, J.A. (2008) 'Antisolvent crystallization: Model identification, experimental validation and dynamic simulation', *Chemical Engineering Science*, 63(22), pp. 5457–5467. doi:10.1016/j.ces.2008.08.003.
34. Dighe, A.V. et al. (2022) 'Three-step mechanism of antisolvent crystallization', *Crystal Growth & Design*, 22(5), pp. 3119–3127. doi:10.1021/acs.cgd.2c00014.
35. Cameron, F.K. and Bell, J.M. (1906) 'The system lime, Gypsum, water, at 25°', *Journal of the American Chemical Society*, 28(9), pp. 1220–1222. doi:10.1021/ja01975a015.
36. Campbell, A.N. and Yanick, N.S. (1932) 'The system NiSO₄—CaSO₄—H₂O', *Trans. Faraday Soc.*, 28(0), pp. 657–661. doi:10.1039/tf9322800657.
37. Kalyanaraman, R., Yeatts, L.B. and Marshall, W.L. (1973) 'High-temperature Debye-Huckel correlated solubilities of calcium sulfate in aqueous sodium perchlorate solutions', *The Journal of Chemical Thermodynamics*, 5(6), pp. 891–898. doi:10.1016/s0021-9614(73)80051-5.
38. Ling, Y. and Demopoulos, G.P. (2004) 'Solubility of calcium sulfate hydrates in (0 to 3.5) mol·kg⁻¹ sulfuric acid solutions at 100 °C', *Journal of Chemical Engineering Data*, 49(5), pp. 1263–1268. doi:10.1021/je034238p.
39. Kumar, A. et al. (2004) 'Ionic interactions of calcium sulfate dihydrate in aqueous calcium chloride solutions: Solubilities, densities, viscosities, and electrical conductivities at 30°C', *Journal of Solution Chemistry*, 33(8), pp. 995–1003. doi:10.1023/b:josl.0000048049.62958.f9.
40. Kumar, A., Sanghavi, R. and Mohandas, V.P. (2007) 'Solubility pattern of CaSO₄·2H₂O in the system NaCl + CaCl₂ + H₂O and solution densities at 35 °C: non-ideality and ion pairing', *Journal of Chemical Engineering Data*, 52(3), pp. 902–905. doi:10.1021/je0604941.
41. Yuan, T., Wang, J. and Li, Z. (2010) 'Measurement and modelling of solubility for calcium sulfate dihydrate and calcium hydroxide in NaOH/KOH Solutions', *Fluid Phase Equilibria*, 297(1), pp. 129–137. doi:10.1016/j.fluid.2010.06.012.
42. Wu, X. et al. (2010) 'Solubility of calcium sulfate dihydrate in Ca—Mg—K chloride salt solution in the range of (348.15 to 371.15) K', *Journal of Chemical Engineering Data*, 55(6), pp. 2100–2107. doi:10.1021/je900708d.
43. Tian, P. et al. (2012) 'Determination and Modeling of Solubility for CaSO₄·2H₂O—NH₄⁺—Cl—SO₄²⁻—NO₃⁻—H₂O System', *Journal of Chemical Engineering Data*, 57(12), pp. 3664–3671. doi:10.1021/je300871p.
44. LLussi, A. and Jaeggi, T. (2006) 'Chemical factors', *Monographs in Oral Science*, pp. 77–87. doi:10.1159/000093353.
45. Gervais, C. et al. (2010) 'Cleaning marble with ammonium citrate', *Studies in Conservation*, 55(3), pp. 164–176. doi:10.1179/sic.2010.55.3.164.
46. Matsumoto, A. et al. (1991) 'Steric control of gelation in monovinyl-multivinyl polymerization leading to preparation of self-crosslinkable polymer having pendant vinyl groups', *European Polymer Journal*, 27(12), pp. 1417–1420. doi:10.1016/0014-3057(91)90245-j.
47. Barwick, V.J. (1997) 'Strategies for solvent selection — A literature review', *TrAC Trends in Analytical Chemistry*, 16(6), pp. 293–309. doi:10.1016/s0165-9936(97)00039-3.

Lead-free halide Ternary Perovskite Composites for CO₂ Photocatalytic Reduction to Solar Fuels

Joyce Cheung and Yiheng Shao

Department of Chemical Engineering, Imperial College London, U.K.

Abstract

Lead-free halide perovskites have received significant attention in photocatalytic reduction of CO₂ to CO due to their good optoelectronic properties. Despite having suitable properties, challenges associated with low photocatalytic performance are posed such as instability in H₂O and high charge recombination. To boost the photocatalytic performance, better charge separation and effective suppression of charge recombination are necessary to enhance the overall efficiency of photocatalyst. Heterojunctions proposed a propitious solution to enhance spatial charge separation. This report demonstrates the synthesis of a heterojunction of Cs₃Bi₂Br₉ and multilayer Ti₃C₂T_x at different ratios, achieved by anti-solvent crystallisation process. Characterisation techniques such as scanning electron microscopy, x-ray photoelectron spectroscopy and x-ray diffraction were utilised to demonstrate the successful synthesis of the photocatalytic materials. The optimal Cs₃Bi₂Br₉ composite of 11.31 at % Ti showed an enhanced CO₂ reduction performance of $(5.94 \pm 0.18) \mu\text{mol CO g}^{-1}\text{h}^{-1}$ over pure Cs₃Bi₂Br₉ of $(1.33 \pm 0.27) \mu\text{mol CO g}^{-1}\text{h}^{-1}$ in the presence of water as a proton donor. Photoelectrochemical measurements have also displayed enhanced photocurrent density in composites, indicating improvement in charge separation.

Keywords

Lead-free halide perovskites, MXene, metal-semiconductor junction, photocatalysis, CO₂ reduction

1 Introduction

In the pursuit of addressing challenges posed by global warming, finding a sustainable energy source is one of the biggest challenges faced by mankind at present.[1] Extensive efforts have been devoted to the CO₂ emission reduction, carbon capture and storage (CCS), and the conversion of CO₂ into valuable carbon-based products such as methanol, formic acid, and formaldehyde. As a renewable, inexhaustible energy source, solar energy shows great potential in energy conversion. Photocatalysis is a green technology that converts solar energy into chemical energy. Developing new photocatalysts is a central focus in advancing photocatalytic technology. To enable CO₂ reduction, the selected photocatalyst must possess a band structure that encompasses with redox potentials of the target reactions. To achieve a high photocatalytic performance, the photocatalyst should have strong visible light absorption, good generation of electron-hole pairs, good charge separation, and a low recombination rate.

In the recent years, halide perovskites have recently emerged as a promising candidate in photocatalysis because of their exceptional optoelectronics properties, great light-harvesting capabilities, efficient charge generation, extended carrier diffusion lengths, a perfectly aligned redox potential of CO₂. [2][3] As one of the rising stars, caesium bismuth bromide (Cs₃Bi₂Br₉) has received great interest, due to their low cost, low toxicity, and relative stability against air, light, heat [4]. To further enhance the charge separation and reduce charge recombination in lead-free perovskites, a solution has been proposed to couple perovskites with other materials such as oxides, semiconductors, and metal nanoparticles to form heterojunctions [4]. The heterojunction forms a

built-in electric field at the interface which enhances the separation and transportation of charge carriers. [5]

Ti₃C₂T_x, a two-dimensional (2D) nanomaterial also known as MXene, has also attracted notable attention due to their outstanding properties including superior metal-like conductivity, outstanding chemical and mechanical stability, and excellent hydrophobicity [4][6]. Moreover, with precise control over surface termination groups, the work function of Ti₃C₂T_x is tuneable at the surface contacts in the effort of improving charge separation [7] and has theoretically shown that it can be potentially tuned from less than 2 eV to above 6 eV.

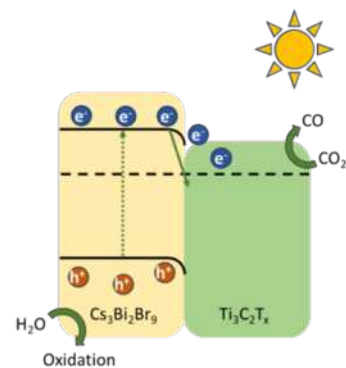
2 Experimental Details

2.1 Materials

CsBr (99%, Sigma-Aldrich), BiBr₃ (99%, Alfa-Aesar), anhydrous dimethyl sulfoxide ($\geq 99.9\%$, Sigma-Aldrich), anhydrous 2-propanol (99.9%, Sigma-Aldrich), LiF (300 mesh, Sigma-Aldrich), HCl (37 wt% in H₂O, Fisher Chemical), Ti₃AlC₂ (400 mesh, Xinx Technology Co., Ltd), acetone ($\geq 99\%$, VWR Chemicals), Nafion® perfluorinated resin solution (5 wt% in lower aliphatic alcohols in water, Sigma-Aldrich), anhydrous acetonitrile (98%, Sigma-Aldrich), and tetrabutylammonium hexafluorophosphate, TBAPF₆, (98%, Sigma-Aldrich) were used without any further purification. CsBr and BiBr₃ were stored in a nitrogen regulated glovebox (< 0.5 ppm H₂O, < 0.5 ppm O₂). LiF and TBAPF₆ were stored in a desiccator.

2.2 Synthesis of Ti₃C₂T_x nanocrystal (Multilayer)

Multilayer titanium carbide (M-Ti₃C₂T_x) was synthesised by acid etching. 2 g of LiF was added into



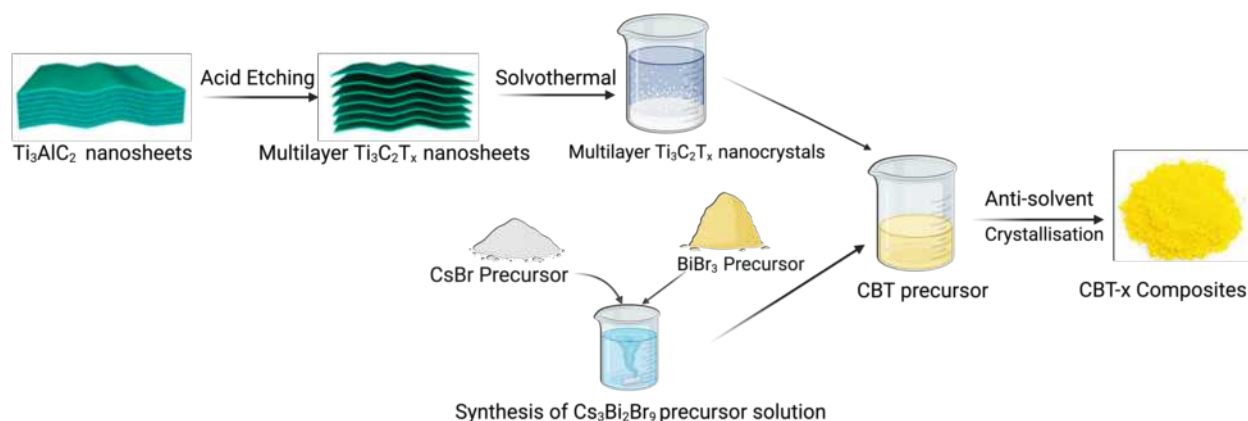


Figure 1: Schematic representation of the synthesis process of CBT composites

a 100 mL Teflon vessel containing 50 mL of 9M of HCl. The mixture was stirred at 1500 rpm for 30 min. 2 g of titanium aluminium carbide Ti_3AlC_2 was slowly introduced under stirring. The acid etching was maintained at 35°C for 36 h to completely remove the aluminium layer. The resulting precipitate underwent two rounds of centrifugation with 160 mL of 1M HCl each time to eliminate fluoride anions. The precipitate was then washed with deionised water (DI water) six to eight times until the pH exceeded 6. The final precipitate was vacuum dried at 60°C for 12 h, yielding M- $\text{Ti}_3\text{C}_2\text{T}_x$. $\text{Ti}_3\text{C}_2\text{T}_x$ nanocrystals (TNC) were produced through solvothermal treatment. 2 g M- $\text{Ti}_3\text{C}_2\text{T}_x$ is dissolved in 50 mL of anhydrous dimethyl sulfoxide (DMSO) by stirring for 1 h at room temperature. The mixture is transferred into an autoclave and placed in an oven at 120°C for 6 h. After the mixture has cooled down, the transparent supernatant is obtained through centrifugation at 10100 rpm for 2 min and repeated twice for thorough collection. The resulting supernatant was filtered with a $0.22\ \mu\text{m}$ syringe filter to ensure the removal of any remaining particulate matter. To adjust the concentration of TNC, 1 mL of TNC in DMSO underwent drying in a vacuum oven, and the mass difference before and after the drying process was calculated. The TNC concentration was then adjusted to $1\ \text{mg mL}^{-1}$ for further use.

2.3 Synthesis of $\text{Cs}_3\text{Bi}_2\text{Br}_9$ (CBB) nanocrystals

CBB was synthesised using an anti-solvent crystallisation process. The CBB precursor was prepared using $1.2\ \mu\text{mol}$ of CsBr and $0.8\ \mu\text{mol}$ of BiBr_3 , and the solids were introduced into 20 mL of DMSO. The resulting mixture was sonicated for 30 min until a homogenous solution was achieved. 4 mL of the solution was swiftly injected into 100 mL of 2-propanol (IPA), yielding a yellow-coloured solution. The precipitate was separated through centrifugation at 4500 rpm for 2 min. The supernatant was disposed of, and fresh IPA was added. This centrifugation process was repeated twice. The final precipitate was then placed in a vacuum oven at 40°C and left overnight to facilitate the evaporation of any remaining IPA.

2.4 Synthesis of CBB/TNC composites

To prepare CBB/TNC composites, 0.1 mL of TNC solution was added to 10 mL of CBB precursor to form a CBT precursor. The newly formed precursor was sonicated for 10 min. Following this, 4 mL of the CBT precursor is injected into 100 mL of IPA. The CBT sample was obtained through a process of centrifugation and subsequent drying. The specific experimental conditions remained consistent with those used in the synthesis of CBB.

In this work, 4 CBB/TNC composites were synthesised with CBB precursor and $\text{Ti}_3\text{C}_2\text{T}_x$, in which the volume of TNC added into the making CBT precursor differ. 0.1, 0.25, 0.5 and 1 mL of TNC added to the CBB precursor were abbreviated to CBT-1, CBT-2.5, CBT-5, and CBT-10 respectively.

2.5 Characterisation

To understand about the morphology of the pure CBB, $\text{Ti}_3\text{C}_2\text{T}_x$, and composites, scanning electron microscopy (SEM) was performed with ZEISS AURIGA Cross Beam at 5 kV with 15 nm of gold coating. The powders were deposited onto a carbon conductive tape as support, and the liquid was drop casted onto a silica glass before placing onto the support. X-ray diffraction (XRD) was conducted with Malvern PANalytical Aeris at 40 kV and 15 mA using Cu $\text{K}\alpha$ radiation ($\lambda = 1.54\text{\AA}$) in the 2θ range 5° to 70° . X-ray photoelectron spectroscopy (XPS) was conducted to understand the elemental composition of the samples by using Thermo Fisher K-Alpha+ with a monochromatic Al $\text{K}\alpha$ X-ray source. Valence band XPS measurement for CBB was also taken with the same machine and situate the valence band edge from the fermi level. All data processing and peak deconvolution were performed with Avantage and all binding energies were carbon calibrated to 286.8 eV. Ultraviolet-visible spectroscopy (UV-Vis) was performed, with barium sulfate (BaSO_4) as a reference, using the SHIMADZU UV-2600 UV-Vis spectrophotometer with an integrated sphere to understand the reflectance of the materials and the material band gap (E_g) of CBB and the synthesised composites.

2.6 Photocatalytic performance

Photocatalytic CO₂ reduction was conducted in a 20 mL gas-tight stainless steel photoreactor with a quartz window. The samples were tested by using 32 mm Cytiva Whatman™ quartz filter. 1 mg mL⁻¹ of the desired sample was prepared using the powder sample and anhydrous IPA. The filter paper was weighed, and the solution was drop casted evenly onto a filter paper at 70 °C by utilising a hot plate. The filter paper with the deposited catalyst is placed in a vacuum oven at 40 °C overnight for drying. The dried filter paper is weighed again to acquire the deposited catalyst mass. The photoreactor is cleaned with and the test filter paper is placed carefully in the centre of the photoreactor. 40 µL of distilled H₂O is added at the side of the photoreactor as hole scavenger (proton source). With the quartz window firmly secured in place and valve to the gas chromatography (GC) is closed, the photoreactor is evacuated twice by using a vacuum pump and refill of CO₂. Once the pressure is at 0 bar, the valve to the GC is opened and the system and GC were purged further at 20 mL min⁻¹ for 15 min to equilibrium and to ensure no remaining unwanted gases were left in the GC. This was also to fully saturate the CO₂ with H₂O. The adjacent valves to the photoreactor were closed and an initial GC measurement was taken. Finally, a 300 W Xe source from LOT-Quantum Design equipped with AM 1.5 G filter was shone onto the sample for 1 h, such that the sample was subjected to an intensity of 100 mW cm⁻² (equivalent to 1 sun). After 1 h of light exposure, both valves before and after were opened, and 5 mL min⁻¹ of CO₂ was pumped for 2.5 min. An after-reaction GC measurement was taken immediately after 2.5 min. A batch 5 h recyclability test was also conducted to determine the stability of CBT-2.5.

2.7 Photocurrent measurement

The suspension of the desired sample was prepared mixing 50 mg of the powder sample, 50 µL of Nafion, 1 mL of anhydrous IPA. The suspension is sonicated until a homogeneous liquid is formed. The fluorine-doped tin oxide (FTO) coated glass was cut into 2.5 × 2.5 cm² pieces and were cleaned with three consecutive stages of sonication by using DI water, acetone, and IPA for 10 min each. The cleaned FTO glass is left to dry at 80 °C. The 1 mL of solution is uniformly drop casted at 80 °C onto the conductive side of the FTO glass, with one edge untouched for later electrical connection. The catalyst deposited FTO glass is left on the hot plate for further evaporation of IPA solvent. Photoelectrochemical (PEC) measurements were conducted using a half cell configuration. The electrolyte was a solution of 0.1 M of TBAPF₆ in anhydrous ACN. The non-aqueous Ag/Ag⁺ reference electrode and a Pt wire counter electrode were used. The prepared FTO glass was used as a working electrode with an area of 0.28 cm². Sunlight was simulated with a Lot Quantum Design Xe Lamp equipped with a AM 1.5 G filter and a potentiostat from the Ivium Technologies was used to control the applied potential of the working electrode.

3 Results and Discussion

3.1 Materials characterisation

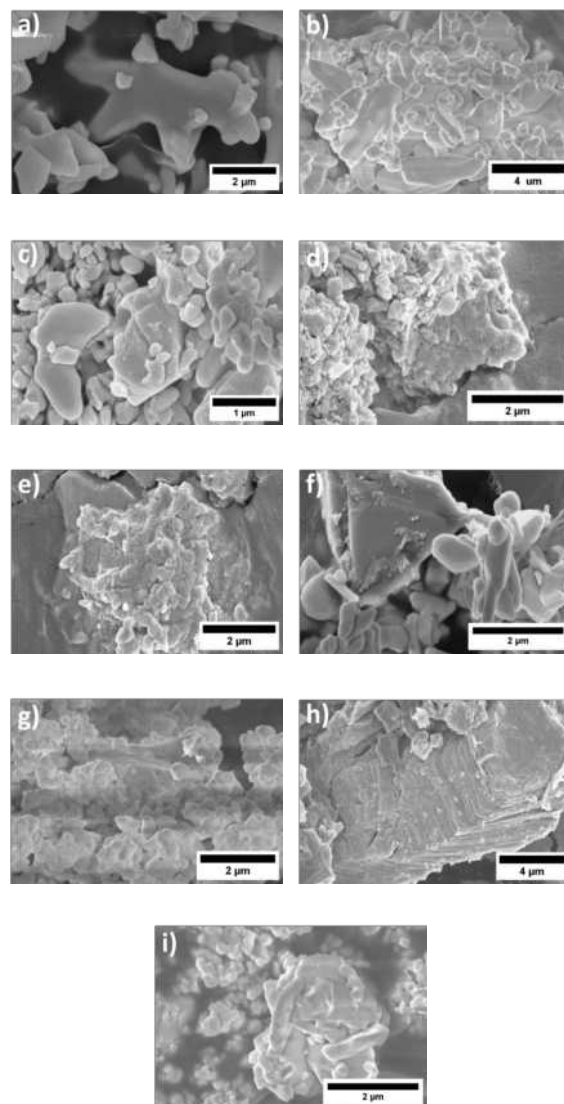


Figure 2: SEM micrographs of (a-b) CBB, (c) CBT-1, (d-e) CBT-2.5, (f) CBT-5, (g) CBT-10, (h) Ti₃C₂T_x, and (i) CBT-2.5 after 5 h of recyclability test.

The morphology of the above samples were characterised by SEM (Figure 2). For pure CBB, the surface morphology had a big deviation. Different sizes of plates and clusters of CBB were seen, this suggested the sample was not homogenous. This happens in anti-solvent process as the stirring can cause inhomogeneities. In CBT-1 (Figure 2c), Ti₃C₂T_x is surrounded by CBB bulk while in CBT-5 (Figure 2f) and CBT-10 (Figure 2g), a large bulk of Ti₃C₂T_x is attached to the CBB were spotted. In CBT-2.5 (Figures 2d & 2e), appropriate amount of Ti₃C₂T_x attached to CBB is recognised. In Figure 2h, it can be seen clearly that M-Ti₃C₂T_x was successfully synthesised. However, the synthesised sheets were not nanocrystals, with a diameter of 5.66 µm, indicating that the syringe filtration was insufficient to

achieve nanocrystals. The spacing between the layers were quite small, this could potentially suggest the acid-etching wasn't as successful. After the recyclability test, the CBT-2.5 showed no major change in morphology, despite having a lower photocatalytic performance over repeated cycles. Charging of samples were seen in some SEM images despite coating the sample with gold. This was a piece of evidence to demonstrate that the synthesised material had a bandgap.

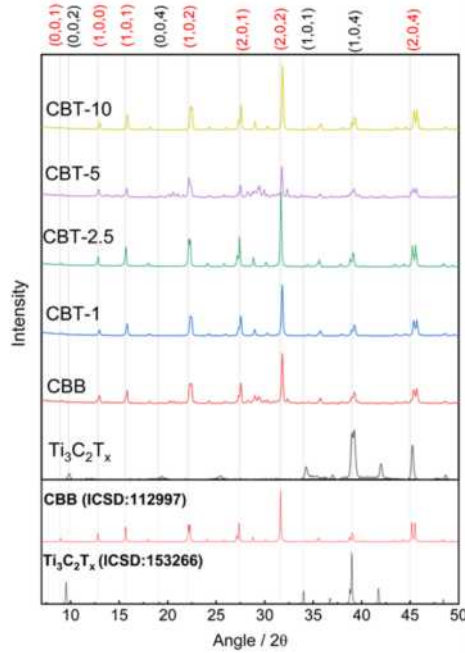


Figure 3: XRD spectra of $\text{Ti}_3\text{C}_2\text{T}_x$ [8] and $\text{Cs}_3\text{Bi}_2\text{Br}_9$ [9] in comparison to the synthesised $\text{Ti}_3\text{C}_2\text{T}_x$, CBB and their composites.

XRD technique was utilised to determine the crystal structure of photocatalyst materials synthesised and enhance the successful synthesis of target catalysts. The reference peaks of $\text{Cs}_3\text{Bi}_2\text{Br}_9$ and $\text{Ti}_3\text{C}_2\text{T}_x$ are shown at the bottom of Figure 3 to represent the position of intrinsic crystal planes. The experimental data of pure $\text{Cs}_3\text{Bi}_2\text{Br}_9$ and $\text{Ti}_3\text{C}_2\text{T}_x$ were obtained and portrayed a consistent trend with their reference peaks respectively. The peaks for CBT photocatalysts highly align with both experimental and reference data for $\text{Cs}_3\text{Bi}_2\text{Br}_9$, which emphasised that there is a significant amount of CBB in CBT catalysts and CBB acts as the main component of CBT catalysts. The main diffraction peaks were centered at 15.6° , 22.1° , 27.4° , 31.6° , 39.0° and 45.0° (2θ), corresponding to the planes (1,0,1), (1,0,2), (2,0,1), (2,0,2), (1,0,4), and (2,0,4), respectively. The highest-intensity peaks for all CBT composites were present at 31.6° corresponding to the (2,0,2) crystal plane which belongs to $\text{Cs}_3\text{Bi}_2\text{Br}_9$, which further proved the clear presence of CBB in CBT photocatalysts. The fact that CBT-2.5 showed the largest intensity for (2,0,2) plane compared to other CBT samples indicated that it contained the highest amount of CBB. The deviation between experimental peaks and reference crystal planes

peaks can be explained by the difference in sample holder height during XRD reflection-mode operation for CBT powders [10] as the change in angle of 2θ leads to the shifting in XRD pattern.

XPS was conducted to further understand the surface composition of $\text{Ti}_3\text{C}_2\text{T}_x$, CBB, and the composites. Wide scan was conducted to gain insight in the surface composition (Figure 4a). The wide spectra have detected the elements Cs, Bi, Br, C and O in CBB and in all composites. Ti was also detected in the composites. In addition, F and Cl were both detected in the $\text{Ti}_3\text{C}_2\text{T}_x$ survey, this was likely residue left from the washing step in the $\text{Ti}_3\text{C}_2\text{T}_x$ synthesis. The elemental XPS spectra were illustrated in Figure 4. The C 1s scan showed four types of bonding. In $\text{Ti}_3\text{C}_2\text{T}_x$, a binding energy (BE) of 282.08 eV can be corresponded to Ti-C (purple). The pure CBB and composites have shown an additional C species, around 286.74 eV, which was attributed to the C-O (yellow) bonding. A BE of 284.80 and 288.60 eV were attributed to the C-C (red) and O-C=O (green) bonding respectively. In the O 1s spectrum, two different peaks were observed. Both peaks are represented the Ti-O bonding, with the difference in BE caused by oxygen bounded to titanium at different lattice sites [11]. The Ti 2p scan revealed 3 possible bonding, which were Ti-O (red), Ti-O (yellow), Ti-C (green), and Ti-F (purple). A doublet peak of 724.44 and 738.40 eV was observed in the Cs 3d scan, attributed to the Cs $3d_{5/2}$ and Cs $3d_{3/2}$ respectively. The presence of the Bi^{3+} species was demonstrated in the Bi 4f scan, with a doublet at 159.16 and 164.46 eV 159.16 and 164.46 eV, which was attributed to Bi $4f_{7/2}$ and Bi $4f_{5/2}$ respectively. Finally, the Br 3d scan demonstrated Br $3d_{5/2}$ (red) and Br $3d_{3/2}$ (green) respectively. For CBT-10, higher shifts in BE were identified in the Cs 3d and Bi 4f scans. The higher shift in BE implied a lower electron density in those atoms. This highlighted the electrons flowed from the CBB to $\text{Ti}_3\text{C}_2\text{T}_x$, suggesting that there were more electrons on the metal side to participate in the CO_2 reduction. The elemental composition of the semiconductors were also acquired from XPS and compared to the theoretical compositions (Table 1). XPS showed that the experimental compositions do not reassemble with the theoretical compositions. However, as expected, the Ti at % in the composites increased with the addition of $\text{Ti}_3\text{C}_2\text{T}_x$.

Semiconductor band gaps can be categorised into two types, direct and indirect. Direct band gap is when the maximum energy of valence band and minimum energy of conduction band occur at the same value of momentum (k). If the maximum and minimum lie at different values of k , then it is indirect. Direct band gap is always larger than indirect band gap.

The Tauc plot of $(F(R)h\nu)^{(1/\gamma)}$ vs $h\nu$, where h , ν , $F(R)$ and γ represent the Planck's constant, frequency, Kubulka-Munk function, and a constant which is equivalent to 0.5 for direct and 2 for indirect E_g calculations.

The Kubulka-Munk function is known as

$$F(R) = \frac{K}{S} = \frac{(1 - R)^2}{2R}$$

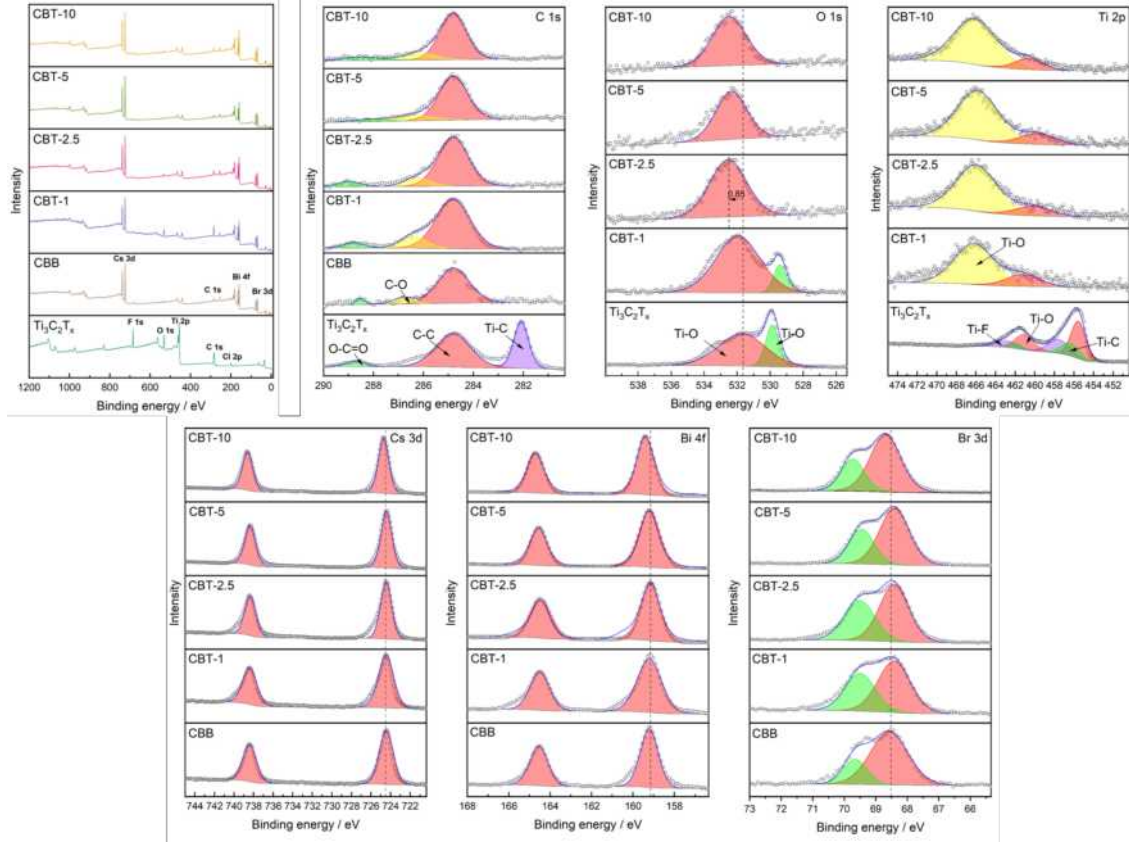


Figure 4: XPS of (a) survey scans, (b) C 1s, (c) O 1s, (d) Ti 2p, (e) Cs 3d, (f) Bi 4f, and (g) Br (4d) of the pure CBB semiconductor and their composites with $\text{Ti}_3\text{C}_2\text{T}_x$.

Table 1: Theoretical composition of CBB and their composites compared to XPS experimental composition.

Sample	Composition [at %]							
	Theoretical				Experimental			
	Cs	Bi	Br	Ti	Cs	Bi	Br	Ti
CBB	21.43	14.29	64.29	0	15.41	6.34	43.89	0
CBT-1	21.42	14.28	64.26	0.04	5.65	2.86	18.09	6.67
CBT-2.5	21.41	14.27	64.23	0.09	11.79	5.15	33.55	11.31
CBT-5	21.39	14.26	64.17	0.19	19.14	6.68	41.74	11.55
CBT-10	221.35	14.23	64.05	0.37	14.92	6.51	39.34	14.95

where K , S , and R correspond to molar absorption coefficient, scattering coefficient, and diffused reflectance of the material respectively.

Reflectance of all powder samples are plotted and presented in Figure 5a and the visible light region from 380 to 700 nm were also outlined. $\text{Ti}_3\text{C}_2\text{T}_x$ had a relatively high reflectance when compared to CBB and their composites. CBB and all composites had a good absorption performance in the lower wavelength, up to 450 nm. Moreover, all showed similar reflectance results throughout the different wavelengths. Thus, no trend for the CBB and composites were observed. Through the extrapolation of a tangent to the onset linear region in the Tauc plot (Figure 5b & 5c), the material E_g were obtained. For the E_g calculations, CBB has a direct and indirect E_g of 2.70 and 2.63 eV respectively. The indirect calculation of CBB was better fitted with a tangent than

the direct calculation, therefore the material E_g was determined to be indirect.

3.2 Photocatalytic performance

The testing of different photocatalysts with various TNC loading on CBB has been conducted to optimise the CO production rate Figure (6a). CBT-1, CBT-5, CBT-10 and CBB were tested at the first stage which demonstrated that CBT-1 had a relatively higher CO production rate, which led to the further manipulation of CBT-2.5. After optimisation, CBT-2.5 showed the best performance for CO_2 reduction at $5.94 \mu\text{mol CO g}^{-1}\text{h}^{-1}$ which was more than three times larger compared to pure CBB samples. The increase in photocatalytic performance can be attributed to the appropriate TNC loading and regular morphology. As presented previously in SEM micrographs, CBT-1 was shown to have $\text{Ti}_3\text{C}_2\text{T}_x$

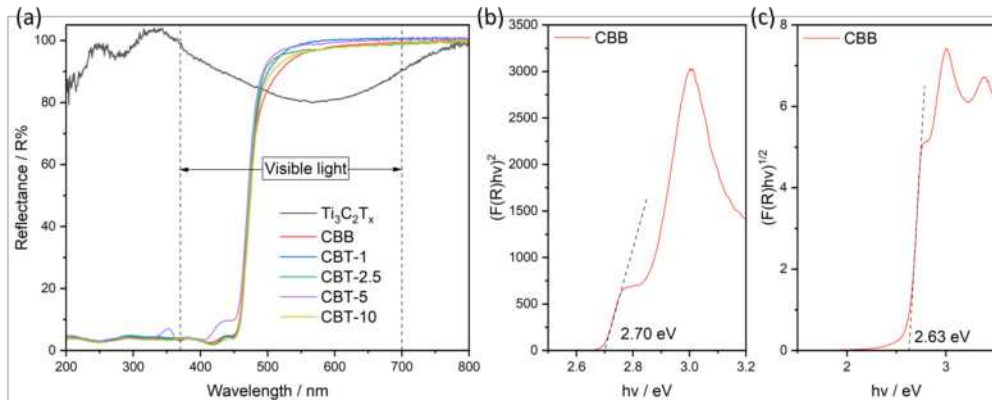


Figure 5: (a) Reflectance of $\text{Ti}_3\text{C}_2\text{T}_x$, CBB, and composites. Tauc plot of CBB assuming (b) direct and (c) indirect behaviour.

was surrounded by CBB bulk. As the contact of the metal and semiconductor formed active sites for CO_2 reduction, too much CBB attached to $\text{Ti}_3\text{C}_2\text{T}_x$ will lead to insufficient active sites for the reduction, resulting to a lower photocatalytic performance. For CBT-5 and CBT-10, the lower photocatalytic performance was explained by excessive $\text{Ti}_3\text{C}_2\text{T}_x$ attached to CBB, leading to potential blockage of light absorption. The error bars, which indicated the stability of photocatalytic performance of each sample, were calculated by repeating the experiments for over three times under the same conditions. Selectivity towards CO , H_2 , and CH_4 were also calculated and presented in Table 2. The optimal sample, CBT-2.5, also yielded the highest selectivity of CO with 97.43%.

To investigate the stability of the optimal sample, a recyclability test was conducted over five consecutive 1 h cycles with cleaning and adding fresh water drops into the reactor after each run (Figure 6b). The overall decrease in CO production rate was tested to be approximately 69% after five cycles with most of the loss taking place after the second hour. This may be due to the degradation of CBB or the filter was damaged by the water vapour.

Table 2: CBB and their composites selectivity towards CO , CH_4 , and H_2 .

Sample	Selectivity [%]		
	CO	H_2	CH_4
CBB	95.14	2.82	2.03
CBT-1	97.18	0.41	17.57
CBT-2.5	97.43	0.44	2.13
CBT-5	50.24	49.40	0.37
CBT-10	95.45	6.57	6.22

PEC tests were operated to determine the performances in terms of current generation and charge separation ability of photocatalysts. The photocurrent density-voltage curves (Figure 7a) demonstrated the higher photocurrent of CBT composites compared to CBB under the same chopped-light irradiation and applied potential range. The intensity of the photocurrent was calculated by subtracting the dark current from

the photocurrent. Although the CBT composites did not show an obvious trend in photocurrent density, all composites showed a more obvious 'jump', illustrating the CBT photocatalysts contain more photogenerated electrons compare to CBB, which can be attributed to the better charge separation. The higher negative photocurrent under a large negative applied potential could be contributed by the potential reduction of thin-film samples and the change in the composition of the cell electrolyte.

Figure 7b illustrated the significant improvement of photocurrent density difference from the dark and light conditions. There was a clear enhancement in the photocurrent density for CBT composites compared to CBB.

Valence-band XPS measurement was taken to estimate the energy between the fermi level (E_f) and the valence band edge. The energy between the two levels was attained at the x-value of the two tangents intersection (Figure 8a). The work function describes the minimum energy to remove an electron from the material to vacuum level. This was used to locate the E_f of the material. Coupling with the E_g value acquired from the Tauc plot (Figure 5c), the conduction band (E_c) and valence band (E_v) can be situated (Figure 8b). As the work function (ϕ) of the CBB is deeper than the $\text{Ti}_3\text{C}_2\text{T}_x$, an accumulation layer of electrons in the space charged region will be formed when the metal and semiconductor are in contact. This contact is a metal-semiconductor junction. In the space charged region, the semiconductor energy band edges are shifted due to the electric field formed by the charge transfer between the $\text{Ti}_3\text{C}_2\text{T}_x$ and CBB. Consequently, CBB is subjected to a downward band bending.[12] Since $\phi_m < \phi_s$ and CBB is n-type type semiconductor, the contact will be ohmic and no Schottky barrier will be formed can also be concluded.

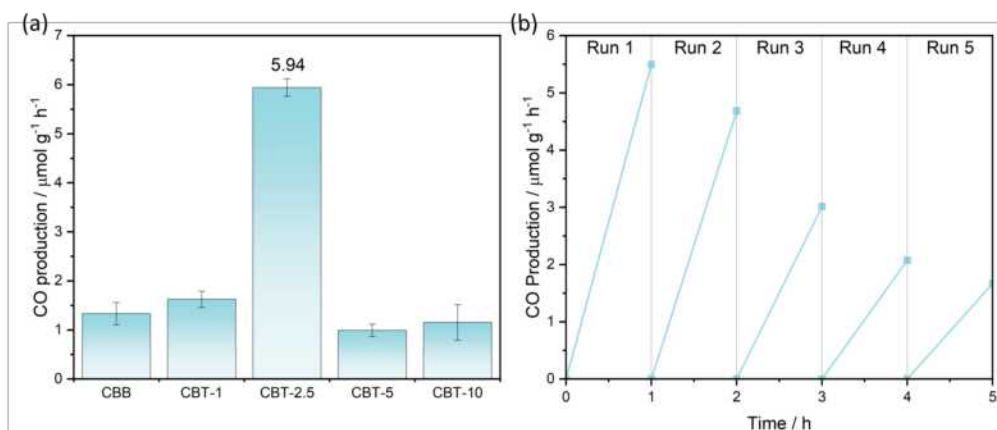


Figure 6: (a) Photocatalytic CO_2 reduction rate to CO of CBB and its composites and (b) Recyclability test of CBT-2.5 over 5 consecutive cycles.

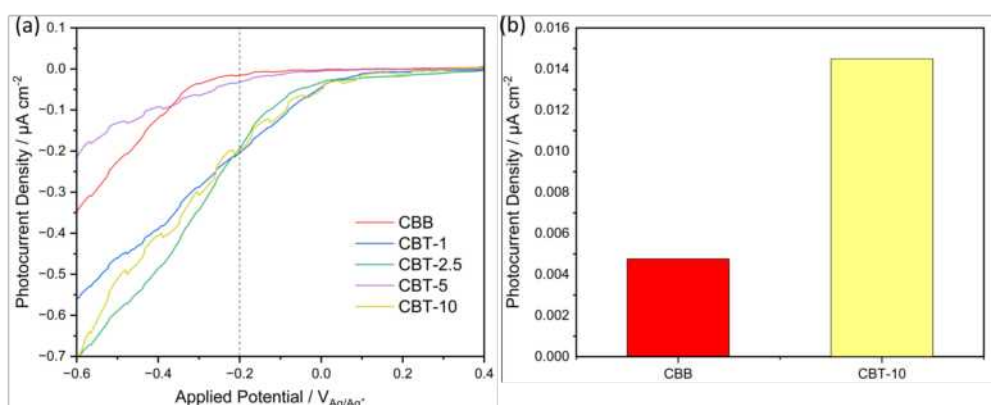


Figure 7: (a) Photocurrent density-voltage curves of pure CBB and the composites obtained under chopped simulated-sunlight illumination of 100 mW cm^{-2} in a solution of 0.1 M TBAPF_6 in anhydrous acetonitrile. (b) Measurements of photocurrent density achieved at -0.2 V .

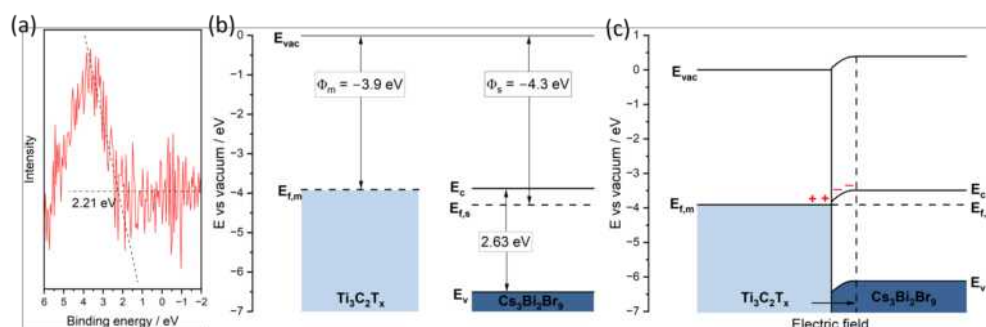


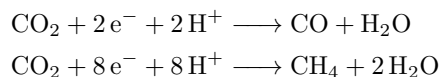
Figure 8: (a) Valence band XPS measurement of CBB. (b) Energy band diagram of CBB and $\text{Ti}_3\text{C}_2\text{T}_x$ where ϕ_m is obtained from SECO [11], ϕ_s is obtained from XPS [2], and (c) energy band diagram of CBB with contact with $\text{Ti}_3\text{C}_2\text{T}_x$.

4 Conclusion

In conclusion, we have successfully constructed an all-inorganic lead-free perovskite heterostructure with multilayer $\text{Ti}_3\text{C}_2\text{T}_x$. The results have demonstrated that the heterostructure had a higher photocatalytic performance than pure CBB. Out of all the composites, CBT-2.5 has yielded the greatest performance in CO_2

reduction with a CO production rate of $(5.94 \pm 0.18) \mu\text{mol CO g}^{-1} \text{h}^{-1}$ and a selectivity of 97.44 % to CO. Comprehensive characterisations and experiments have highlighted that the optimal sample was CBT-2.5 and the photocatalytic performance was enhanced owing to the morphologies and a better charge separation in the composites of CBB and $\text{Ti}_3\text{C}_2\text{T}_x$. The mechanism was

established with multiple techniques. The energy band diagram was partially constructed with the UV-Vis and valence-band XPS and it was confirmed as downward band bending. The following reduction reactions were proposed to happened on the $\text{Ti}_3\text{C}_2\text{T}_x$ side during photocatalytic CO_2 reduction reactions:



These findings demonstrate the potential for heterojunction of CBB and $\text{Ti}_3\text{C}_2\text{T}_x$ and will open new avenues to explore more efficient and stable lead-free halide perovskite in CO_2 conversion.

5 Outlook

To further improve this study, control tests can be performed to evaluate the effects of experimental conditions on CO production rate and selectivity, such as using a different light and proton source such as hydrogen gas. As highlighted in SEM, the synthesised material morphologies are not homogeneous and the filtered $\text{M-Ti}_3\text{C}_2\text{T}_x$ are not nanocrystals. Addition of ligands can produce homogeneous CBB crystals and dialysis can be used further for TNC solution as it is a superior separation technique over syringe filtration. Transmission electron microscopy (TEM) can be utilised to gain a deeper understanding of the materials morphology, especially with $\text{Ti}_3\text{C}_2\text{T}_x$. Coupling with energy dispersive x-ray (EDX) spectroscopy with TEM, the distribution of elements can also be identified in the material to further comprehend the dispersion of $\text{Ti}_3\text{C}_2\text{T}_x$ in CBB. Further characterisations on surface area can be conducted using the BET analysis to determine the relationship between surface area and photocatalytic performance. Simulation tools such as density functional theory (DFT) and molecular theory (MD) can be employed to optimise the performance of photocatalysts by stimulating the electronic structure of materials. With further characterisations, a more comprehensive study can be achieved.

6 Acknowledgement

The authors would like to acknowledge the support from Chen, Lu and Baghdadi, Y, and Eslava Group.

References

- [1] Kim J, Kwon EE. Photoconversion of carbon dioxide into fuels using semiconductors. *Journal of CO2 Utilization*. 2019 Oct;33:72–82. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S2212982019303865>.
- [2] Baghdadi Y, Temerov F, Cui J, Daboczi M, Rattner E, Sena MS, et al. $\text{Cs}_3\text{Bi}_2\text{Br}_9$ /g-C₃N₄ Direct Z-Scheme Heterojunction for Enhanced Photocatalytic Reduction of CO_2 to CO.

Chemistry of Materials. 2023 Oct;35(20):8607–8620. Available from: <https://pubs.acs.org/doi/10.1021/acs.chemmater.3c01635>.

- [3] Getachew G, Wibrianto A, Rasal AS, Kizhepat S, Dirersa WB, Gurav V, et al. Lead-free metal halide perovskites as the rising star in photocatalysis: The past, present, and prospective. *Progress in Materials Science*. 2023 Dec;140:101192. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S007964252300124X>.
- [4] Li Q, Song T, Zhang Y, Wang Q, Yang Y. Boosting Photocatalytic Activity and Stability of Lead-Free $\text{Cs}_3\text{Bi}_2\text{Br}_9$ Perovskite Nanocrystals via In Situ Growth on Monolayer 2D $\text{Ti}_3\text{C}_2\text{T}_x$ MXene for C–H Bond Oxidation. *ACS Applied Materials & Interfaces*. 2021 Jun;13(23):27323–27333. Available from: <https://pubs.acs.org/doi/10.1021/acsami.1c06367>.
- [5] Chi Q, Zhu G, Jia D, Ye W, Wang Y, Wang J, et al. Built-in electric field for photocatalytic overall water splitting through a $\text{TiO}_2/\text{BiOBr}$ P–N heterojunction. *Nanoscale*. 2021;13(8):4496–4504. Available from: <http://xlink.rsc.org/?DOI=D0NR08928A>.
- [6] Patil SA, Marichev KO, Patil SA, Bugarin A. Advances in the synthesis and applications of 2D MXene-metal nanomaterials. *Surfaces and Interfaces*. 2023 Jun;38:102873. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S2468023023002432>.
- [7] Aydin E, El-Demellawi JK, Yarali E, Aljamaan F, Sansoni S, Rehman AU, et al. Scaled Deposition of $\text{Ti}_3\text{C}_2\text{T}_x$ MXene on Complex Surfaces: Application Assessment as Rear Electrodes for Silicon Heterojunction Solar Cells. *ACS Nano*. 2022 Feb;16(2):2419–2428. Available from: <https://pubs.acs.org/doi/10.1021/acsnano.1c08871>.
- [8] Wu E, Wang J, Zhang H, Zhou Y, Sun K, Xue Y. Neutron diffraction studies of $\text{Ti}_3\text{Si}_0.9\text{Al}_0.1\text{C}_2$ compound. *Materials Letters*. 2005 Sep;59(21):32715–2719. Available from: <https://www.sciencedirect.com/science/article/pii/S0167577X05004015>.
- [9] Samanta D, Saha P, Ghosh B, Chaudhary SP, Bhattacharyya S, Chatterjee S, et al. Pressure-Induced Emergence of Visible Luminescence in Lead Free Halide Perovskite $\text{Cs}_3\text{Bi}_2\text{Br}_9$: Effect of Structural Distortion. *The Journal of the Physical Chemistry*. 2021 Feb;125(6):3432–3440. Available from: <https://doi.org/10.1021/acs.jpcc.0c10624>.
- [10] Chen X, Bates S, Morris KR. Quantifying amorphous content of lactose using parallel beam X-ray powder diffraction and whole pattern fitting. *Journal of Pharmaceutical and Biomedical Analysis*. 2001 Aug;26(1):63–72. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0731708501003466>.

- [11] Schultz T, Frey NC, Hantanasirisakul K, Park S, May SJ, Shenoy VB, et al. Surface Termination Dependent Work Function and Electronic Properties of $\text{Ti}_3\text{C}_2\text{T}_x$ MXene. *Chemistry of Materials*. 2019 Sep;31(17):6590–6597. Available from: <https://pubs.acs.org/doi/10.1021/acs.chemmater.9b00414>.
- [12] Zhang Z, Yates JT. Band Bending in Semiconductors: Chemical and Physical Consequences at Surfaces and Interfaces. *Chemical Reviews*. 2012 Oct;112(10):5520–5551. Available from: <https://pubs.acs.org/doi/10.1021/cr3000626>.

Environmental Impact Assessment of Sustainable Aviation Fuels Against Planetary Boundaries

Nicholas Gerard and Elton Lam

Department of Chemical Engineering, Imperial College London, U.K.

Abstract

Anthropogenic activity has been rapidly pressuring Earth systems beyond their safe operating space. The aviation sector has seen substantial growth in the past decades, with a projected CAGR of 3.3% for the next two decades. Despite merely contributing approximately 3% of global emissions, it has been classed as a difficult sector to decarbonise. Sustainable aviation fuel (SAF) poses itself as one of the most promising interim fixes for aviation. To assess the absolute environmental impacts of SAF, a methodology comprising of life cycle assessment (LCA) and planetary boundary framework (PBF) are used in conjunction to delineate its effect on Earth systems. Aspen HYSYS and OpenLCA were utilised to obtain the inventories of the system. Characterisation factors were used to translate LCA inventories to PBF in conjunction with a grandfathering allocation principle used for assigning the share of safe operating space (SOS) to the aviation sector to ultimately indicate the levels of transgression (LT). The resulting environmental impacts and LT were obtained for SAF, where five of the nine control variables assessed outperformed conventional aviation fuels. Despite the better performance, SAF still exceeds the SOS for two control variables, namely radiative forcing and biogeochemical flow of nitrogen, ultimately still threatening the Earth. The aviation sector sees SAF as an interim fix, and its future is to be met with greener propulsion systems based on hydrogen and batteries.

Keywords: sustainable aviation fuel, life-cycle assessment, planetary boundaries framework, safe operating space

1. Introduction

Humans have been reshaping the Earth for at least 3,000 years [1], but these alterations have snowballed into permanent pressures since 1970's [2]. Humanity's impact on the globe is often assessed through the narrow lens of climate change and many overlook the broader spectrum of environmental influences. Effects on the Earth system extend beyond climate change, encompassing aquatic, atmospheric and terrestrial systems, along with resource degradation and ecological crises. This calls for a standardised and accurate assessment of the impacts such that humanity can monitor and mitigate the deterioration of the environment.

The aviation sector poses an appreciable contribution to the world in all facets. Global air traffic has been growing, on average, at a rate of 5.0% annually since the mid-1990s; in that same period, the world economy has been growing at 2.8% [3]. Building upon the robust growth rate, the aviation sector is poised for continued expansion

for the next two decades, with industry experts forecasting an average annual growth rate of 3.3% [4, p. 13]. The aviation sector contributed 3.6% to the world's GDP in 2018 [5, p. 4] and accounted for 2.0% of global energy-related CO₂ emissions in 2022 [5]. These figures seem rather trivial, but aviation has been identified as one of the most challenging sectors to decarbonise [5].

To minimise the sector's impact on the environment; green propulsion systems are contenders, but they are years away from commercial viability [6]; this is where sustainable aviation fuels (SAFs) take the spotlight for the best alternative. SAF is produced from a sustainable feedstock which includes waste oil and solid waste; it is also produced synthetically via direct air carbon capture (DAC) [7]. Current usage of SAF is estimated to be less than 1%; it is not surprising as the production costs of SAFs are threefold those of conventional aviation fuels (CAFs) [6]. In the context of environmental impact, how can one ensure the process of SAF outperforms CAF?

A conventional method to monitor the impacts of a process (in this study, the production and combustion of SAF) is through a life-cycle assessment (LCA). LCA is a process of “evaluating the effects that a product has on the environment over the entire period of its life thereby increasing resource-use efficiency and decreasing liabilities” [8]; it is a powerful tool which supports engineering decisions due to its robustness. A major limitation of LCA is its sole assessment on a relative scale; the analysis of the processes does not provide insight into the impact on the Earth system in its entirety.

To assess the absolute impact of a process on the Earth system, Rockström *et al.* proposed a framework based on planetary boundaries which define the “safe operating space for humanity concerning the Earth system” [9, p. 1]. Many subsystems of Earth are sensitive around the safe operating threshold. Transgressing the threshold could cause important subsystems to shift into a new state which would cause, at best, deleterious consequences for humanity [9, p. 1]. The planetary boundaries framework (PBF) is the first framework proposed that assesses the absolute environmental impact of anthropogenic activities.

With the rudiments of the aviation sector laid out and its potential effect on Earth systems highlighted, this study aims to assess the absolute environmental performances of SAF production and combustion with planetary boundaries framework in conjunction with LCA methodology.

2. Background

This paper explores the absolute environmental performances of SAF by bridging the gap between LCA and PBF through the lens of this fuel. Previous studies have conducted life cycle, economic and environmental assessments on aviation fuels, and defined general methods of translating LCA to PBF. This study hopes to combine the attributes of different studies and classify a method of translating LCA to PBF for SAF.

2.1. Quantifying environmental impacts of aviation fuel

Previous life-cycle assessments on aviation fuels

suggested SAF outperformed CAF in terms of resource and global warming impacts [10, p. 19] [11, p. 19]. These LCA investigations quantified resource depletion and emissions by considering different providers and sources. Despite the conclusions these assessments have provided, its shortcomings lie in its inability to gauge the absolute global impact of processes.

Research has significantly focused on the ramifications of global warming, often overlooking alternative environmental impacts [12, p. 2]. While evaluating the status of global warming is essential, it is imperative to recognise that this concern constitutes just one facet of the broader spectrum of environmental impact.

2.2. Planetary boundaries framework

Based on the shortcomings of LCA, the PBF is one of the first methods to quantify the absolute impact of processes on Earth systems. The framework was first proposed in 2009 by Rockström *et al.* [13, p. 2] and has been revised continuously such that all planetary boundaries are mapped out [14, p. 4]. This framework paints an overarching picture of the conditions of Earth systems but does not delineate the specific contributions from specific sectors and processes.

2.3. LCA to PBF: bridging the gap

The optimal assessment of specific processes on Earth systems in a comprehensive and absolute manner can be achieved by leveraging the benefits of LCA and PBF. Thus far, no conventional standard has been established to link the two systems, but substantial work in this field has been published. The challenges lie in translating inventories of the LCA into absolute environmental impact and justifying the share of safe operating space (SOS) for anthropogenic activity relative to global SOS.

Numerous works on translating LCA inventories to absolute environmental impact predominantly explore the utilisation of characterisation factors (CFs). The use of CFs allows impacts of inventories of LCA to “be expressed in the same metric as the control variables of the Earth system processes” and reflects “the proportional change in environmental impact per change in quantity of

environmental interventions” [15, p. 4]. Ryberg *et al.* have developed CFs for LCA inventories to PBF translations [16, pp. 3,8-9]. This translation is not limited to the use of characterisation factors; an alternative investigation considered the “weighted average of substance-specific contributions” [17, p. 5]. It is crucial to highlight the application of this weighted average approach was restricted to greenhouse gas emissions and only explored “climate change” – one of the nine planetary boundaries.

Justifying contributions of the aviation sector is based on the allocation principle. The allocation principle is a framework which assigns a fair share of the global SOS to an anthropogenic process [18, p. 4]. There is extensive study on different methods of allocation based upon distribution justice theory [19]; different theories have then spawned subsequent allocation principles.

Table 1 - Previous relevant studies which have explored environmental impacts and at least one of the following topics: SAF, LCA, and PBF

Study	Year	SAF focus	Use of LCA	Use of PBF	Environmental impacts explored
Rockström <i>et al.</i> [9]	Sep-09	No	No	Yes	Climate change and other indicators
van der Giesen <i>et al.</i> [20]	May-14	Yes	Yes	No	Climate change only indicators
Ryberg <i>et al.</i> [16]	Dec-17	No	Yes	Yes	Climate change and other indicators
Magone <i>et al.</i> [10]	Apr-20	Yes	Yes	No	Climate change only indicators
Bjørn <i>et al.</i> [18]	Jul-20	No	Yes	Yes	Climate change and other indicators
Ryberg <i>et al.</i> [19]	Jul-20	No	Yes	Yes	Climate change and other indicators
Sherwin <i>et al.</i> [21]	May-21	Yes	No	No	Climate change only indicators
D'Angelo <i>et al.</i> [22]	Jul-21	No	Yes	Yes	Climate change and other indicators
Ordóñez <i>et al.</i> [12]	Aug-22	Yes	Yes	No	Climate change and other indicators
Petersen <i>et al.</i> [17]	Aug-22	No	Yes	Yes	Climate change only indicators
Sacchi <i>et al.</i> [23]	Jun-23	Yes	Yes	No	Climate change only indicators
Rojas-Michaga <i>et al.</i> [11]	Jul-23	Yes	Yes	No	Climate change only indicators
Richardson <i>et al.</i> [14]	Sep-23	No	No	Yes	Climate change and other indicators

2.4. Rationale for this study

There is an opportunity within the research domain to focus on the evaluation of SAF using LCA methodology along with PBF. While existing studies have delved into the environmental impact of aviation fuels, there is a notable absence of insight into the absolute global impact. Other papers have explored translating LCA to PBF through CF and have extensively justified different allocation principles, but they did not address the SAF process. This presents an ideal prospect to bridge the gap between the SAF process and the translation from LCA to PBF, enabling a holistic examination of the environmental impacts associated with this anthropogenic process. None of the studies in **Table 1** consider the absolute global environmental impacts of SAF production and combustion

to Earth systems through LCA methodologies; this study aims to bridge the gap between SAF, LCA, and PBF.

3. Methods

The comprehensive assessment of the SAF process, from cradle to gate with combustion, in absolute terms involves a methodological unification of LCA inventories and planetary boundaries framework. The work presented here was repeated for CAFs, namely kerosene; this provides a basis to compare the performance of SAF.

3.1. Defining the scope

A novel process for SAF production was developed by OXCCU Tech Ltd. by utilising CO₂ and renewable energy to produce aviation fuels with a lower environmental impact (EI). The process converts CO₂ obtained via DAC

[24] and H₂ from water electrolysis [25] into a blend of hydrocarbons to be used as aviation fuels.

A simulated plant using Aspen HYSYS was constructed to assess the hourly calorific production of SAF, established as the functional unit for this LCA, according to the ISO 14040 standards [26]. The model incorporates raw material inputs (CO₂ and H₂), yielding the corresponding output of SAF. The study adopted a cradle-to-gate scope employing an attributional approach. Given the significant influence of SAF combustion on the LCA, their impacts were incorporated into the analysis. This was estimated by performing a mass balance on the carbon in the fuel produced.

To produce results representative of a global scale, the SAF produced in the simulated plant was linearly scaled-up to meet the annual global demand for aviation fuels, which was estimated to be 300 million tonnes in 2019 [27].

3.2. Quantifying the SAF process

To quantify the process of SAF production and combustion, a life-cycle inventory (LCI) analysis was conducted. Foreground data (mass and energy flows of the process) and models were obtained from the simulated plant. Background data on the extraction of the raw materials were obtained from Ecoinvent (v3.6), a LCI database. OpenLCA (v1.1), a LCA software, was employed to integrate these models and datasets for the computation of the LCIs.

3.3. Evaluating environmental impact

The results generated from OpenLCA were analysed to assess the EIs of the SAF process. Characterisation factors (CFs), developed by Ryberg *et al.* [16, p. 3], were applied to quantify the environmental impacts on the control variables (CVs) of the planetary boundaries introduced by Rockström *et al.* [13, pp. 8-9]. In this study, two planetary boundaries, namely “biosphere integrity” and “introduction of novel entities” were excluded from consideration. Their boundaries were only defined in 2023 [14, p. 4], and the characterisation methods are therefore deemed immature.

The following approach developed by Ryberg *et al.* [16, p. 5] was employed to quantify the environmental impacts of the SAF process; this is shown in Eq(1).

$$EI_j^{SAF} = \sum_{i \in I} LCI_i^{SAF} \cdot CF_{j,i} \cdot CP^{SAF}, \quad \forall j \in CV \quad (1)$$

EI_j^{SAF} refers to the environmental impact of SAF in each CV j . LCI_i^{SAF} is the inventory i associated with the functional unit of SAF; these are obtained from OpenLCA. $CF_{j,i}$ is the characterisation factor corresponding to CV j for inventory i . CP^{SAF} is the annual calorific production of SAF by the process scaled-up to global production.

3.4. Determining the level of transgression

After the EIs have been evaluated, the subsequent step involves calculating the level of transgression (LT). This calculation considers the proportion of the Safe Operating Space (SOS) delimited by the PBF and is downscaled to SAF production. Allocating a segment of the SOS to SAF production was achieved through an allocation principle called grandfathering (GF), wherein the allocation is proportional to the environmental impact of the anthropogenic activity [18, p. 13]. This allocation is shown in Eq(2).

$$SOS_j^{SAF} = SOS_j^{GLO} \cdot \frac{EM^{AVI}}{EM^{GLO}}, \quad \forall j \in CV \quad (2)$$

SOS_j^{SAF} refers to the allocated SOS of the SAF process for each CV in j . SOS_j^{GLO} refers to the SOS delimited by the PBF for each CV in j . EM^{AVI} refers to the annual emissions of the aviation sector, and EM^{GLO} refers to the annual global anthropogenic emissions. The ratio of the EMs is estimated to be 3.5% [28].

LT is calculated with Eq(3) after the SOS for the aviation sector has been determined; it delineates the environmental impact of SAF production and combustion against its portion of the SOS.

$$LT_j^{SAF} = \frac{EI_j^{SAF}}{SOS_j^{SAF}}, \quad \forall j \in CV \quad (3)$$

LT_j^{SAF} refers to the LT of SAF for each CV in j .

4. Results

This section highlights the processed data obtained from

Aspen and OpenLCA. **Table 2** presents the findings of EI of SAF and CAF, along with the share of SOS allocated to the aviation sector. The units for each CV are presented, and the last two columns present the LT for each of the

fuels; all LT greater than 100% are highlighted in red. **Figure 1** is the planetary boundaries mapped out for each CV for SAF and CAF; the red dotted circle shows the allocated share of SOS for aviation.

Table 2 – Summary of the EIs and LTs of SAF and CAF, and the corresponding share of SOS of the aviation sector

CV	EI of SAF	EI of CAF	Share of SOS	Unit	SAF LT (%)	CAF LT (%)
Climate change: radiative forcing	9.45×10^{-2}	0.385	3.50×10^{-2}	[W·m ⁻²]	270	1100
Climate change: atmospheric CO ₂ conc.	7.26	29.5	12.3	[ppm]	59	240
Stratospheric ozone depletion	3.76×10^{-4}	1.88×10^{-7}	9.63	[DU]	3.90×10^{-3}	1.95×10^{-2}
Ocean acidification	2.21×10^{-2}	8.97×10^{-2}	0.10	[Ω _{arag}]	23	93.1
Biogeochemical flows: nitrogen cycle	2.32	0.315	2.17	[Tg·N·yr ⁻¹]	107	14.5
Biogeochemical flows: phosphorus cycle	0.641	1.504	0.917	[Tg·P·yr ⁻¹]	69.9	164
Land-system change	1.59×10^{-9}	9.65×10^{-11}	2.63	[%]	6.03×10^{-8}	3.67×10^{-9}
Freshwater use	0.157	1.79×10^{-2}	140	[km ³ ·yr ⁻¹]	0.11	1.28×10^{-2}
Atmospheric aerosol loading	2.44×10^{-7}	1.04×10^{-7}	8.75×10^{-3}	[-]	2.79×10^{-3}	1.19×10^{-3}

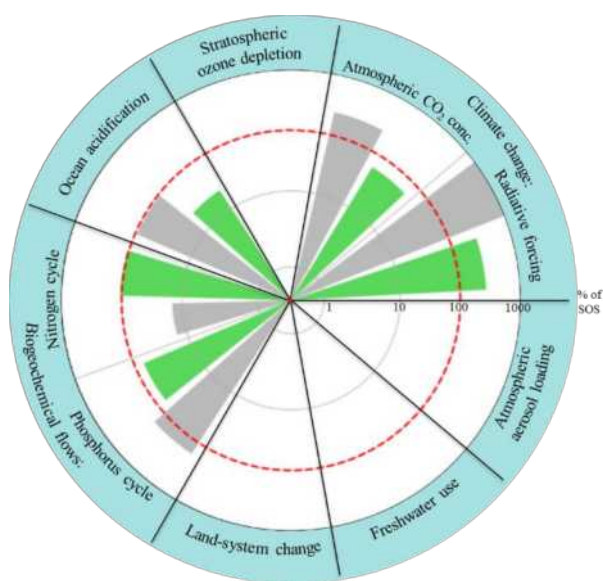


Figure 1 – Planetary boundaries mapped out for SAF and CAF

5. Discussion

5.1. Significance of results and implications

5.1.1. Implications of environmental impact

The EIs of each CV quantify the potential stress a particular activity exerts on Earth systems. Knowledge of EIs is fundamental for promoting sustainable practices by highlighting the hotspot inducing factors in current practices. **Figure 1** suggests that the hotspots associated with SAFs are the CVs: radiative forcing, atmospheric CO₂ conc., ocean acidification, nitrogen and phosphorus cycle;

what are the factors behind these large EIs? **Figure 2** reveals the contribution of each elementary flow on the CV.

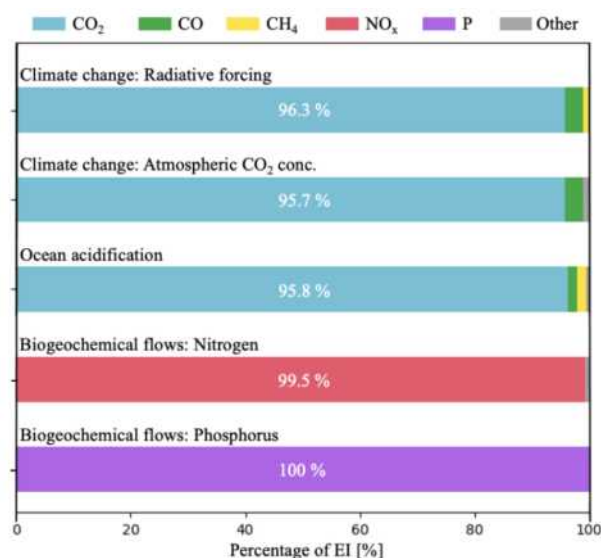


Figure 2 – Elementary flow contributions towards control variables

Climate change and ocean acidification are dominated by the emissions of CO₂, which is coherent with other studies conducted [29]. Considering this, and that radiative forcing of the aviation sector currently surpasses the ecological budget, it suggests the need to phase out the use of carbon-based aviation fuels. Ammonia is recognised as a viable alternative to carbon-based fuels [30]. However, the main culprit for biogeochemical flow of nitrogen is

NO_x, which is a by-product of ammonia combustion. Caution is advised for the utilisation of ammonia-based aviation fuel as the established boundary for nitrogen flows in **Table 2** has already been exceeded.

The transition from CAF to SAF process leads to the nitrogen cycle surpassing the share of SOS, which is extremely undesirable for a process. A hypothesis for this transgression is attributed to the complete combustion of the lights, a subprocess of the SAF production process. NO_x is the main culprit for nitrogen cycle transgression, and its emission primarily stems from the combustion of atmospheric oxygen and nitrogen in flames [31]. The energy generated from the combustion process surpasses the requirements of the simulated plant. Opting to burn only the necessary mass of lights may result in reduced NO_x emissions and consequently minimise the LT of nitrogen cycle.

Aside from reducing the major contributors to EI, prudence is required when handling elementary flows characterised by large CF values as they possess a greater influence on the computation of EI. It is imperative to reduce the emissions of these compounds to minimise environmental impact.

EIs are helpful in evaluating the combined effects of individual elementary flows, however, its significance is contingent upon understanding the tolerance of Earth's environmental systems. Therefore, a comparison between EIs and the SOS is necessary to assess meaningful impact of human activities. While EIs identify contributing factors, discerning the specific LT of each boundary is crucial for effective change implementation. Without this element, the analysis remains a standard life cycle assessment.

5.1.2. Implications of the level of transgression

The LTs are rather sensitive to the allocation principle, and a contender for GF is economic value added (EVA); this allocation principle assigns a share proportional to economic value added [18, p. 13]. GF exclusively addresses environmental impact, whilst EVA exclusively focuses on economic value. In the current socioeconomic context, these concepts are engaged in a zero-sum game.

The focus of this study is environmental impact; therefore, the allocation of the share of SOS should be based upon the anthropogenic activity's emissions. An allocation principle reasoned with economic value does not shed light on the activity's environmental impact, thus GF aligns better with the research objectives than EVA.

It is crucial to note the allocation principle of GF is based upon relative emissions not environmental impact; it is therefore most representative for the CVs of atmospheric CO₂ conc. and radiative forcing, whereas other CVs, such as freshwater use, may not be represented accurately. This could detail why the LT for freshwater use is so low; freshwater use is not classed as a type of emission and with other sectors (e.g. agriculture) dominating its use, the EI of this CV is dwarfed. It could be reasoned that different allocation principles can be applied for each CV, but for the sake of uniformity and generality, GF is adopted for all boundaries.

With all the LTs mapped out, what are the implications on the environment within the aviation context? EIs improve when SAF is used; the LT for atmospheric CO₂ conc. and the phosphorus cycle will stay within the assigned SOS, implying these CVs will see massive improvements. Despite improvements of SAF, the mass adoption of it will not save the earth. Whilst seven of the nine CVs are within the SOS, two of them, namely radiative forcing and the nitrogen cycle exceed the assigned SOS. The transgression of this space could cause detrimental effects to Earth systems. The reduction of the LT for radiative forcing from 1100% to 270% (**Table 2**) from the transition would still impose damage to the Earth, albeit representing a substantial improvement.

5.1.3. Next steps

Figure 1 clearly illustrates that Earth systems remain vulnerable even with the full adoption of SAF, but what implications does this hold for the aviation sector? Aviation has considerable impacts on our world, both environmentally and economically. The debate over whether policy initiatives should address the environmental impact of the aviation sector often hinges

on its economic contributions. With the aviation sector contributing 3.6% to the global GDP but accounting for only 3.5% of global emissions, policymakers might be hesitant to impose stringent restrictions, considering that aviation contributes to economic prosperity more rapidly than it poses environmental challenges. This leaves few solutions and may start from the individual; practising conscious consumerism could greatly reduce environmental impacts.

From the inventory analysis phase of the LCA, the total CO₂ emissions to the air amount to 23,600 kg·hr⁻¹. Out of this, 13,900 kg·hr⁻¹ of CO₂ is emitted during the SAF production process from burning the lights. This represents a maximum capture of CO₂ if a carbon capture and storage (CCS) facility were to be implemented within the process. In order for the LT of radiative forcing to align with its allocated share of the SOS, it is necessary to reduce CO₂ emissions to at least 2.88 times lower than the current level, which translates to a target of no more than 8,190 kg·hr⁻¹ of CO₂. With the implementation of a CCS process, the reduction in CO₂ emissions would be (23,600 - 13,900) kg·hr⁻¹, resulting in a residual emission of 9,700 kg·hr⁻¹ in the best-case scenario, which unfortunately, still exceeds the target of 8,190 kg·hr⁻¹ required to prevent transgression of the radiative forcing limit. This analysis indicates that while CCS can substantially reduce emissions, its implementation in the SAF production process will at best contribute to Earth's degradation at a diminished rate.

5.2. Limitations

5.2.1. Uncertainties

While LCA is a valuable tool, the results inherently involve uncertainties. These uncertainties mainly stem from background databases and the methodology employed to convert these data into the LTs.

The uncertainties for LTs stem from EI, and from Eq(1), they can be traced back to LCIs and CFs. The background uncertainties linked to each of the LCIs were curated from the Ecoinvent database. However, the associated uncertainties of CFs are unknown. In [16], the EIs calculated by Ryberg's methodology and the ILCD

2011 midpoint yielded an appreciable Spearman's correlation coefficient of 0.85 for most CVs. Leveraging this, a rough estimate for the CF uncertainties is obtained.

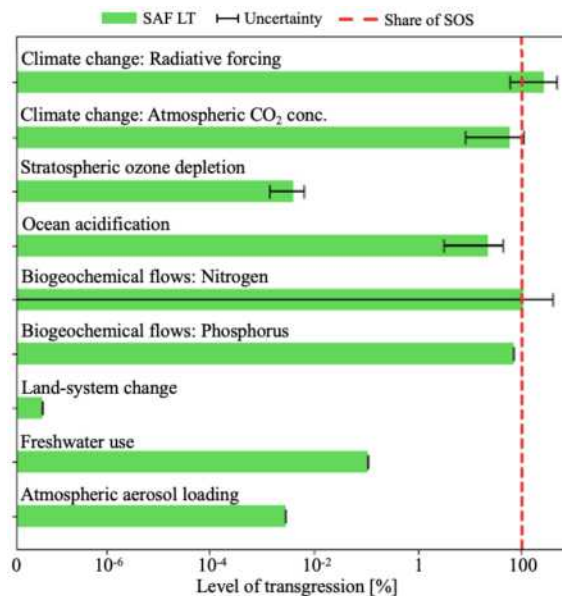


Figure 3 – Estimated uncertainties for each CV

Following this, the combined uncertainty of the LCIs and CFs used in this study were enumerated by performing a Monte Carlo simulation on OpenLCA, with ILCD 2011 method, yielding the uncertainties of the EIs, and consequently the LTs, depicted in **Figure 3**. Uncertainty analysis of the LTs provides crucial insights into the reliability of the calculated values. The error bars illustrated in **Figure 3** span six standard deviations, indicating an almost 100% probability that the LTs fall within these bounds.

A notable concern arises in the case of radiative forcing for the climate change boundary, where a 99% probability of transgression suggests an urgent need for corrective action. Atmospheric CO₂ conc., which was initially perceived to be safely within the share of SOS, now has a 1% probability of transgressing. This presents a negative outlook on the climate change boundary.

The significant uncertainty associated with the biogeochemical flow of nitrogen is believed to be attributed towards the lack of a directly comparable impact category in the ILCD 2011 method. Hence the uncertainties of the nitrogen cycle are deemed unreliable as the method may account for irrelevant elementary flows.

On a positive note, there is still confidence that the remaining CVs will not violate their share of the SOS.

5.2.2. Blackbox: OpenLCA and Aspen HYSYS

Both Aspen HYSYS and OpenLCA were treated as black boxes in this study as the internal mechanisms of these applications were not directly accessible. This limitation introduced challenges in fully understanding the intricacies of the calculations produced and assumptions made by these tools, which emphasised the need for mindful interpretation of the outputs obtained.

The fundamental understanding of the LT of phosphorus cycle and freshwater use is currently not well understood; in-depth analysis into these black boxes could shed light on these values. OpenLCA states the main source of phosphorus stems from the treatment of sewage water, but the nature of the treatment is not well understood. On that note, source of water is primarily sewage water.

5.2.3. Planetary boundaries framework

The PBF plays an instrumental role in guiding the understanding of SAF's position in the global environmental context. While this framework has been influential, one of its key shortcomings is the treatment of each boundary as an isolated system [32], failing to account for interactions between CVs. For example, the ocean acidification CV is heavily influenced by the atmospheric CO₂ conc. CV. Changes in one boundary can also alter the limit of another; for example, increased climate change may increase the SOS of freshwater.

Furthermore, there are uncertainties associated with the SOS values defined by Rockström *et al.* [13], which has been attributed to the lack of scientific data and the intrinsic complexity of natural systems. The SOS values used in this study correspond to the lowest end of the uncertainty to ensure a risk-averse approach was adopted.

5.3. Sensitivity analysis

Sensitivity analysis is a key technique for evaluating the robustness of a process by examining the outcome when changes to different variables are made.

In the context of SAF, it is essential to understand how changes in the manufacturing process affect EIs,

particularly in relation to the share of SOS. Can the production process be improved by making the process greener? By changing the inputs of the DAC process such that the electricity is sourced from renewable offshore wind and the natural gas furnace is replaced with an electric furnace, the observed impacts on the LTs are minimal, shown by the green dots of **Figure 4**, indicating that the EIs are desensitised to the DAC process.

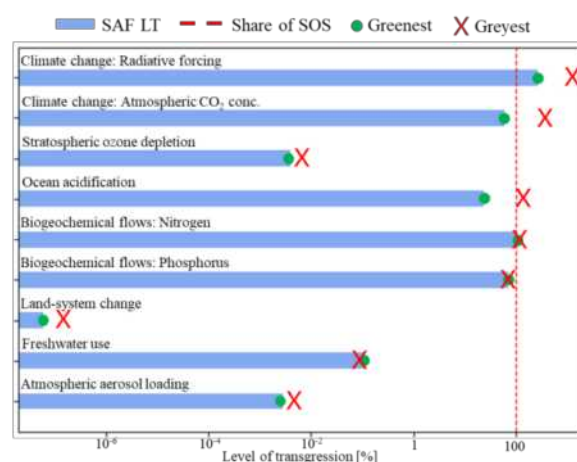


Figure 4 – LT of greenest and greyest power source for feedstock

Another central aspect of the SAF process is the generation of green hydrogen, which is currently achieved through the electrolysis of water using renewable electricity powered by offshore wind. Changing this to electricity generated from fossil fuels significantly exacerbates the LTs for the climate change boundary, rendering them almost five times larger. This is observed in the red crosses in **Figure 4**. This emphasises the dominance of green hydrogen generation on the LTs and the need for renewable electricity in this process. The energy-intensive nature of green hydrogen production [11, p. 14] is well known and is attributed to technology's immaturity [25].

Besides this, the LTs are heavily dependent on the demand for aviation fuel. Projections indicate a substantial increase in the global middle-class population, expected to reach 5.3 billion by 2030, showcasing a CAGR of 3.4% from 2018 [33]. Anticipating that over 60% of the population in 2030 will be part of the middle class, there is an expected surge in the demand for normal goods. The shift of aviation to a normal good, coupled with the growth

of the middle-class demographic, is poised to contribute to a rising demand for air travel.

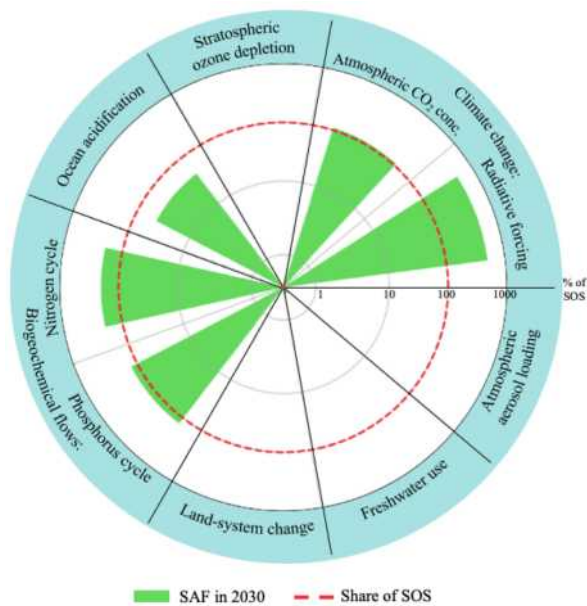


Figure 5 – Predicted LTs for SAF in 2030 based on the PBF

The anticipated CAGR for flight-passenger demand from 2019 to 2030 is projected at 3.3%, culminating in an estimated 5.64 billion passengers by 2030 [4, p. 13]. This increase in demand will lead to escalated production levels which could have a profound impact on the EIs associated with the production and use of SAF. The resulting LTs are illustrated in **Figure 5**, showing multiple limits transgressed, reinforcing the notion that SAFs are just an interim solution. Greener propulsion systems or changes in consumer behaviour are required to ensure aviation stays within its SOS in the future.

6. Conclusion

6.1. Key takeaways

This study concludes SAF outperforms CAF, with five of the nine EIs for each CV lower than that of kerosene. Only two of nine CVs have their SOS transgressed for SAF, whilst three are transgressed for CAF.

Despite SAF presenting to be a better alternative to CAF, it is detrimental to Earth systems. As two CVs have transgressed the assigned SOS, it shows the transition to SAF is insufficient for the preservation of Earth systems. SAF is not the be the be all end all solution for aviation.

SAF only demonstrates itself as an interim fix. Alternatives such as hydrogen and battery-electric aircrafts have proven to be viable solutions for the future of aviation. ZeroAvia [34] and Eviation [35] are at the frontier of hydrogen and battery-electric development respectively. Green propulsion systems will only be poised for continued growth in the coming decades.

6.2. Outlook and further research

The study only explores the environmental impact of SAF on the aviation sector, but this assessment can extend beyond a unidimensional assessment.

As SAF has been identified as a temporary solution, an obvious next step is to perform the environmental assessment for more permanent solutions, namely hydrogen and batteries. An absolute environmental impact comparison of these solutions to SAF would provide a direction for aviation development.

Shifting away from theoretical models and assumptions, utilising real-life data from existing SAF production plants can provide more accurate insights and offer a more realistic and reliable view of the environmental performance of SAF.

Additionally, understanding and consequently reducing the uncertainty associated with the study is crucial to uphold the reliability of the results obtained. Further research could look to validate uncertainties of the CF's used in this study [16] by directly quantifying them.

The attributional scope adopted in this LCA evaluates EIs at a particular point in time and allocates impacts based on pre-established parameters. As such, it fails to capture the interactions and changes of the dynamic system. Alternatively, the consequential LCA tracks the interplay across different supply chains and responds to system changes, which could provide a nuanced understanding of the SAF process. While the consequential approach offers a more accurate view, it is difficult to implement due to the added complexity of intertwined models and data scarcity.

Despite the stellar environmental performance of SAFs, a critical dimension absent from this analysis is the

economic feasibility of its implementation, which is greatly influenced by variables such as manufacturing costs and market dynamics. Disregarding economics constrains a comprehensive understanding of the overall sustainability of SAF, as it is essential to weigh both environmental benefits and economics for an informed decision-making process. Future research initiatives should include economic assessment into the study, ensuring a thorough evaluation of SAF's suitability.

Acknowledgements

We would like to express utmost appreciation for Dr. Andrea Bernardi and Mr. Ben Lyons for their expertise and unceasing assistance. Their unwavering support since day one has been crucial for the development of this study.

References

- [1] D. Biello, "3,000 Years of Abusing Earth on a Global Scale," 30 April 2013. [Online]. Available: www.scientificamerican.com/article/humans-had-global-impacts-thousands-of-years-ago/.
- [2] T. Begum, "Humans are causing life on Earth to vanish," 12 December 2019. [Online]. Available: www.nhm.ac.uk/discover/news/2019/december/humans-are-causing-life-on-earth-to-vanish.html.
- [3] ICAO, "World Aviation and the World Economy," 22 November 2023. [Online]. Available: www.icao.int/sustainability/pages/facts-figures_world-economydata.aspx.
- [4] IATA, "Global Outlook for Air Transport," IATA, Montréal, 2022.
- [5] IEA, "Aviation," November 2023. [Online]. Available: www.iea.org/energy-system/transport/aviation.
- [6] B. Prentice and A. DiNota, "Aviation Is Poised For A Decade Of Growth, But There Are Headwinds Besides Covid-19," 24 February 2022. [Online]. Available: www.forbes.com/sites/oliverwyman/2022/02/24/why-aviation-will-grow-for-a-decade-but-there-are-headwinds-besides-covid-19/.
- [7] IATA, "Developing Sustainable Aviation Fuel," November 2023. [Online]. Available: www.iata.org/en/programs/environment/sustainable-aviation-fuels/.
- [8] EEA, "Life cycle assessment," 15 March 2023. [Online]. Available: www.eea.europa.eu/help/glossary/eea-glossary/life-cycle-assessment.
- [9] J. Rockström, W. Steffen, K. Noone and Å. Persson, "A safe operating space for humanity," *Nature*, London, 2009.
- [10] L. G. Magone, L. Peltz and A. Barker, "A Life Cycle Assessment of Producing Synthetic Fuel via the Fischer-Tropsch Power to Liquid Process," *American Institute of Aeronautics and Astronauts*, Reston, 2020.
- [11] M. F. Rojas-Michaga, S. Michailos, E. Cardozo, M. Akram, K. J. Hughes, D. Ingham and M. Pourkashanian, "Sustainable aviation fuel (SAF) production through power-to-liquid (PtL): A combined techno-economic and life cycle assessment," Elsevier, Sheffield, 2023.
- [12] D. F. Ordóñez, T. Halfdanarson, C. Ganzer, N. Shah, N. M. Dowell and G. Guillén-Gosálbez, "Evaluation of the potential use of e-fuels in the European aviation sector: a comprehensive economic and environmental assessment including externalities," *Royal Society of Chemistry*, London, 2022.
- [13] J. Rockström, W. Steffen, K. Noone and Å. Persson, "Planetary Boundaries: Exploring the Safe Operating Space for Humanity," *Ecology and Society*, Stockholm, 2009.
- [14] K. Richardson, W. Steffen, W. Lucht and J. Bendtsen, "Earth beyond six of nine planetary boundaries," *Science Advances*, Copenhagen, 2023.
- [15] M. W. Ryberg, M. Owsianiak, K. Richardson and M. Z. Hauschild, "Challenges in implementing a Planetary Boundaries based Life-Cycle Impact Assessment methodology," Elsevier, Copenhagen, 2016.
- [16] M. W. Ryberg, M. Owsianiak, K. Richardson and M. Z. Hauschild, "Development of a life-cycle impact assessment methodology linked to the Planetary Boundaries framework," Elsevier, Copenhagen, 2017.
- [17] S. Petersen, M. W. Ryberg and M. Birkved, "The safe operating space for greenhouse gas emissions," Aarhus University, Aarhus, 2022.
- [18] A. Bjørn, C. Chandrakumar, A.-M. Boulay and G. Doka, "Review of life-cycle based methods for absolute environmental sustainability assessment and their applications," *Environmental Research Letters*, Montréal, 2020.
- [19] M. W. Ryberg, M. M. Andersen, M. Owsianiak and M. Z. Hauschild, "Downscaling the planetary boundaries in absolute environmental sustainability assessments," Elsevier, Copenhagen, 2020.
- [20] C. v. d. Giesen, R. Kleijn and G. J. Kramer, "Energy and Climate Impacts of Producing Synthetic Hydrocarbon Fuels from CO₂," ACS, 2014.
- [21] E. D. Sherwin, "Electrofuel Synthesis from Variable Renewable Electricity: An Optimization-Based Techno-Economic Analysis," ACS, 2021.
- [22] S. C. D'Angelo, S. Cobo, V. Tulus, A. Nabera, A. J. Martín, J. Pérez-Ramírez and G. Guillén-Gosálbez, "Planetary Boundaries Analysis of Low-Carbon Ammonia Production," ACS, 2021.
- [23] R. Sacchi, V. Becattini, P. Gabrielli, B. Cox, A. Dirnaichner, C. Bauer and M. Mazzotti, "How to make climate-neutral aviation fly," *Nature*, 2023.
- [24] F. Sabatino, A. Grimm, F. Gallucci, M. v. S. Annaland, G. J. Kramer and M. Gazzani, "A comparative energy and costs assessment and optimization for direct air capture technologies," *Joule*, 2021.
- [25] G. Brändle, M. Schönfisch and S. Schulte, "Estimating long-term global supply costs for low-carbon hydrogen," Elsevier, 2021.
- [26] ISO, *Environmental management - Life cycle assessment - Principles and framework*, First ed., ISO, 1997.
- [27] Statista, "Total fuel consumption of commercial airlines worldwide between 2005 and 2021, with a forecast until 2023," 13 November 2023. [Online]. Available: www.statista.com/statistics/655057/fuel-consumption-of-airlines-worldwide/.
- [28] J. Overton, "The Growth in Greenhouse Gas Emissions from Commercial Aviation," 9 June 2022. [Online]. Available: www.eesi.org/papers/view/fact-sheet-the-growth-in-greenhouse-gas-emissions-from-commercial-aviation.
- [29] British Geological Survey, "The greenhouse effect," 2023. [Online]. Available: [www.bgs.ac.uk/discovering-geology/climate-change/how-does-the-greenhouse-effect-work/#:~:text=In%20descending%20order%2C%20the%20gase,s,nitrous%20oxide\(N2O\)](http://www.bgs.ac.uk/discovering-geology/climate-change/how-does-the-greenhouse-effect-work/#:~:text=In%20descending%20order%2C%20the%20gase,s,nitrous%20oxide(N2O)).
- [30] M. Otto, L. Vesley, J. Kapat and M. Stoia, "Ammonia as an Aircraft Fuel: A Critical Assessment From Airport to Wake," ASME, 2023.
- [31] APIS, "Nitrogen Oxides (NO_x)," 20 July 2022. [Online]. Available: www.apis.ac.uk/overview/pollutants/overview_nox.htm.
- [32] S. J. Lade, W. Steffen, W. d. Vries, S. R. Carpenter, J. F. Donges, D. Gerten, H. Hoff, T. Newbold, K. Richardson and J. Rockström, "Human impacts on planetary boundaries amplified by Earth system interactions," *Nature*, 2019.
- [33] H. Kharas and K. Hamel, "A global tipping point: Half the world is now middle class or wealthier," 27 September 2018. [Online]. Available: www.brookings.edu/articles/a-global-tipping-point-half-the-world-is-now-middle-class-or-wealthier/.
- [34] ZeroAvia, "ZeroAvia Makes Aviation History, Flying World's Largest Aircraft Powered with a Hydrogen-Electric Engine," 19 January 2023. [Online]. Available: zeroavia.com/do228-first-flight/.
- [35] Eviation, "Eviation's Alice Achieves Milestone with First Flight of All-Electric Aircraft," 27 Sep 2022. [Online]. Available: www.eviation.com/Press%20Release/eviations-alice-achieves-milestone-with-first-flight-of-all-electric-aircraft/.
- [36] ATAG, "Aviation Benefits Beyond Borders," ATAG, Geneva, 2018.

Superhydrophobic Cotton: Fabrication and Application in Oil/Water Separation

Afiqah Alias and Ke Jing Toh

Department of Chemical Engineering, Imperial College London

Abstract

In the context of ongoing rapid industrial evolution, the inevitability of oily wastewater production poses challenges to conventional treatment methods due to their inefficiency and high costs. This study addresses this issue by exploring the use of superhydrophobic cotton for efficient oil/water separation. The superhydrophobic cotton, synthesised through co-hydrolysis and co-condensation of tetraethoxysilane (TEOS) and octadecylsilane (ODTMS), exhibits a water contact angle (WCA) of 154.7° and minimal water sorption capacity (0.06g/g) that is much lower than raw cotton (43.3g/g). With an oil/water separation efficiency exceeding 97.5% across various conditions, the modified cotton demonstrates high durability, maintaining a WCA above 150°. Moreover, the superhydrophobic cotton shows its reusability with a separation efficiency above 98.6% even after 20 separation cycles, proving its potential as a sustainable solution for oily wastewater treatment. In summary, the prepared superhydrophobic cotton presents a promising and eco-friendly alternative for the efficient separation of the oil/water mixtures, offering excellent chemical durability, high-efficiency performance and recyclability.

Keywords: Superhydrophobic, Oil/water separation, Durability, Recyclability

1. Introduction and Background

According to United Nations Sustainable Development Goal 6, access to clean water, sanitation and hygiene stands as the fundamental human necessity for health and well-being [1]. However, with the rapid industrialisation especially in the oil and gas, petrochemical and food sectors nowadays, the generation of oily wastewater is unavoidable, whether as a result of accidents or operational processes [2]. For example, the Deepwater Horizon oil spill as one of the largest oil spills in history, was deemed to have significant impacts on marine lives, human health and socioeconomic [3]. To address these challenges, numerous methods for oil/water separation have been introduced. Mechanical tools such as booms and skimmers are used in the industry to clean up oil spills but energy or high pressure is required to operate [4][5]. Besides, traditional methods such as coagulation-flocculation, skimming, centrifugation, flotation and sedimentation although are conventionally used, they generally have low separation efficiency, bulky equipment and high cost [6]. In addition, in situ burning of oil can potentially cause air pollution despite the fact that it can remove up to 98% of an oil spill [5]. Given the inherent flaws in the mentioned methods, there is a proclaimed need for the development of highly efficient and cost-effective materials for oil/water separation.

The idea of using superhydrophobic and superoleophilic (mesh-based) membranes for separating oil

and water was first proposed in 2004 drawing inspiration from the characteristics of the lotus leaf [7]. Since then, many studies have been done on producing novel superhydrophobic materials because of their ability to showcase high oil/water separation efficiency by selectively only absorbing oil while repelling water [8][9][10]. The superhydrophobicity of a material is defined as the phenomenon with the contact angle between water and the corresponding material surface greater than 150° [11]. For instance, Spathi et al. reported that superhydrophobic powder with a WCA of ~153° can be produced by dry milling paper sludge ash (PSA) in the presence of a fatty acid surface functionalising agent [12]. Shah et al. looked into the modification of eggshells with stearic acid to produce superhydrophobic powder with WCA of 169° and Liu et al. fabricated a superhydrophobic polyurethane sponge that showed outstanding oil/water separation efficiency (greater than 97%) [13][14]. Other than that, it was proven by Xu et al. that the fabrication of superhydrophobic stainless steel mesh resulted in a high oil/water efficiency of 97% [15]. Cotton, which is an easily accessible crop, is also undoubtedly suitable for the fabrication of superhydrophobic surfaces for oil/water separation. Superhydrophobic cotton fibres were deemed to be ideal in absorbing oil due to their lightweight, and loose internal structure with a large liquid adsorption capacity [16]. However, current techniques for the

fabrication of superhydrophobic cotton wool, which includes the spraying method, layer-by-layer assembly method, and ultrasound-assisted in situ growth method are complex, involve multiple steps or are time-consuming [17][18].

The main aim of this research was to synthesise superhydrophobic cotton that is durable, cheap, and environmentally friendly via simple chemical modification for oil/water separation. The other objectives of this research were: (1) to investigate the effects of different pre-treatment methods on the superhydrophobicity of the cotton, (2) characterisation of the superhydrophobic cotton surface, (3) to evaluate the water sorption capacity of raw and modified cotton, (4) to analyse the oil/water separation efficiency at different conditions (temperature and mixing speed) and (5) to investigate the chemical and laundering durability of the superhydrophobic cotton as well as its recyclability.

2. Materials and Methodology

2.1. Materials

Cotton wool, TEOS ($\geq 99.0\%$), sodium chloride (NaCl) ($\geq 99.0\%$), oil red O, methylene blue solution, sodium dodecyl sulfate (SDS) surfactant, pyridine ($\geq 99.8\%$) ammonia solution (28%), anhydrous heptane ($\geq 99.0\%$), paraffin oil and anhydrous hexane ($\geq 99.0\%$) were sourced from Sigma Aldrich (Gillingham, UK). Ethanol ($\geq 99.8\%$), acetone ($\geq 99.8\%$), sodium hydroxide (NaOH) pellets, cyclohexane ($\geq 99.8\%$), acetonitrile ($\geq 99.0\%$), hydrochloric acid (HCl) ($\geq 37\%$), toluene ($\geq 99.5\%$) and ODTMS (C18) (90%) were purchased from VWR (Lutterworth, UK). De-ionised (DI) water was provided by PURELAB Chorus 1 (ELGA LabWater) water purification system. Cooking oils (rapeseed oil, sunflower oil and olive oil) were sourced from a local supermarket.

2.2. Pre-treatment and synthesis of superhydrophobic cotton

Three different methods for cotton pre-treatment were used in this research, namely pre-treating with (a) ethanol and acetone, (b) NaOH and (c) cyclohexane. The purpose of pre-treatment was to remove surface impurities and wax on the cotton.

- (a) For cotton pre-treatment with ethanol and acetone, cotton wool was washed twice in ethanol followed by acetone once. The cotton was squeezed to remove excess acetone before drying in a vacuum oven at 70°C and 500 mb for 1h.

- (b) Cotton was washed with acetone once and dried at 80°C and 200 mb. The dried cotton was then mixed with 1M NaOH solution at 60°C for 10 min. The cotton was washed with DI water five times before drying it in the oven overnight at 60°C and 300 mb.
- (c) The same procedure as (b) was used except for mixing the cotton with cyclohexane for 2h instead of 10 min.

The method to synthesise superhydrophobic cotton was the same for all three pre-treatment methods. A mixture of DI water (0.6 mol), ethanol (1.7 mol), TEOS (0.03 mol) and ODTMS (6 mmol) was prepared and mixed with a magnetic stirrer for 5 min. 1000 mg of pre-treated cotton was added and soaked in the solution for 5 min followed by adding 0.08 mol of ammonia dropwise into it. The solution was left to react overnight at room temperature with a stirring speed of 500 rpm. Subsequently, the modified cotton was recovered and washed with ethanol twice and acetone once. The fabricated cotton was dried in the oven at 80°C and 800 mb for 1h. Compressed air was then used to blow the silica nanoparticles off the dried and modified cotton.

2.3. Characterisation

WCA was measured to assess the cotton's superhydrophobicity by using 590 goniometer/ tensiometer (Ramé-hart Instrument Co., USA). This was done by using the sessile drop method and by placing water droplets (7 μL) on at least three different positions on a pressed cotton surface to obtain average measurements. The surface morphology was observed using Gemini 1525 scanning electron microscope (Carl Zeiss AG, Germany) at 5kV with the cotton samples coated with gold. The surface chemistry was analysed by X-ray photoelectron spectroscopy (XPS) (K-Alpha+, Thermo-Fisher Scientific Inc., USA).

2.4. Oil/water separation experiments

For batch experiments with cooking oils, 2 ml of oil was added to 20 ml of DI water and ~200 mg of cotton was placed into the mixture for 5 min. 5 ml of hexane (solvent) was added to the mixture to extract the remaining oil which was quantified using a UV-Vis spectrometer (nanodrop 2000c, Thermo-Fisher Scientific Inc., USA). The separation efficiency (η) was calculated from:

$$\eta(\%) = \frac{(V_I - V_R)}{V_I} \times 100\% \quad (1)$$

where V_I is the initial oil volume (2 ml) and V_R is the residual oil volume. For batch experiments with UV inactive oils, the initial mass of the glass vial with its lid and around 20 ml of DI water was first measured using a weighing balance. ~2 ml of oil was added to the glass vial and the mass of the glass vial containing oil with its lid was then measured. The difference between the two measurements is the initial mass of oil, m_I . ~200 mg of cotton was left soaked in the mixture for 5 min and the mass of the glass vial was measured again. To find out the final mass of oil, the final mass of the glass vial is subtracted from its initial mass. To ensure minimal vapourisation of oil to the surroundings, the lid must be closed tight immediately all the time. The separation efficiency was quantified using Equation [2]:

$$\eta(\%) = \frac{(m_I - m_R)}{m_I} \times 100\% \quad (2)$$

where m_I and m_R represent the initial and final mass of oil respectively.

For oil/water separation in a flow system, a laboratory-scale filtration system was set up, comprising a 5 ml syringe packed with approximately 200 mg of modified cotton, connected to a reservoir containing oil/water mixture. For this experiment, olive oil was selected due to its distinctive colour compared to other cooking oils. The reservoir and syringe were linked through a pump, enabling control over the flow rate of the mixture. A beaker was positioned under the syringe to collect the effluent. The objective of this setup is to examine the correlation between filtrate volume and oil concentration.

2.5. Chemical and laundering durability

To assess the chemical durability of the cotton, the WCA and water sorption capacity were measured at different conditions: immersing ~10 mg of cotton in different solvents (toluene, acetonitrile, pyridine and NaCl solution) and extreme pHs (HCl solution (pH1.3) and ammonia solution (pH13.5)) for 24h. For a laundry test, the cotton was mixed with 0.15 wt% SDS surfactant at 200 rpm for 24h. The cotton samples were then dried and soaked in 10 ml of water. The water sorption capacity was calculated from:

$$S_w(g/g) = \frac{m_2 - m_1}{m_1} \quad (3)$$

where m_1 and m_2 are the initial and final mass of cotton. The chemical and laundering durability was also examined by measuring the WCA.

2.6. Recyclability

The modified cotton was placed in a mixture of 2 ml of oil and 20 ml of water. After each separation cycle, the modified cotton was washed with hexane to remove the oil absorbed and was dried in preparation for another separation cycle. A total of 20 cycles was carried out in this research and measurement was taken for every 5 cycles. Measurement was taken by quantifying the amount of remaining oil using UV-Vis. The oil-water separation efficiency was then calculated using Equation [1].

3. Results and Discussion

3.1. Opting for NaOH as the pre-treatment material

The necessity for pre-treatment arises from the imperative to expose a greater number of hydroxyl groups on the cotton surface, facilitating a more facile attachment of silica nanoparticles [19]. As illustrated in Figure 1, the WCAs resulting from three distinctive pre-treatment methods exhibited a range from 155° to 148°. The WCAs were 154.7°, 147.0° and 133.8° for NaOH, ethanol with acetone and cyclohexane pre-treated cotton respectively. This finding was consistent with the research conducted by Nguyen-Tri et al. in 2019, where employing a chemical pre-treatment involving NaOH resulted in contact angle values higher (147°– 173 °) than those obtained through pre-treatment with water and ethanol alone (91°) [20].

Further investigation into the effectiveness of pre-treatment methods involved an analysis of their respective water sorption capacities. NaOH pre-treated cotton demonstrated the lowest water sorption capacity at 0.06 g/g, followed by acetone, with ethanol (0.11 g/g) and cyclohexane (0.15 g/g) pre-treatment method. Although cotton pre-treated with cyclohexane exhibited water sorption more than twice that of the NaOH-treated cotton, it still performed better than the raw cotton that had water sorption of 43.3±0.3 g/g. Consequently, NaOH as the pre-treatment material was utilised for subsequent experiments in this research.

3.2. Surface Morphology of raw cotton and modified cotton

3.2.1. Cotton Modification

Pure cotton has hydrophilic and superoleophilic properties, consisting of 88%-97% cellulose, waxes, protein, and pectin [21]. The cotton fibre can undergo chemical modification using TEOS and ODTMS. The pre-treatment involved removing the plant wax to

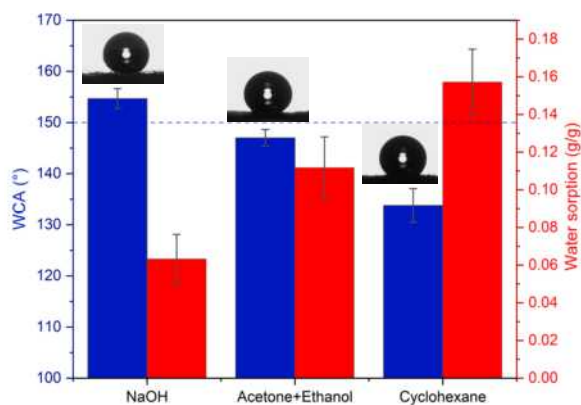


Figure 1: Water contact angle and water sorption (g/g) of cotton pre-treated with different materials: NaOH, ethanol+acetone and cyclohexane

facilitate easy attachment of silica nanoparticles as shown in Figure 2. SEM was utilized to observe the visual differences between raw cotton and modified cotton. There was no significant distinction between the two in their raw and pre-treated states. Initially, the surface of the cotton was smooth; however, after modification, a substantial amount of silica adhered and covered the entire surface of the cotton. This modification provided hydrophobic chains which contributed to the water-repellent property to the cotton fibre.

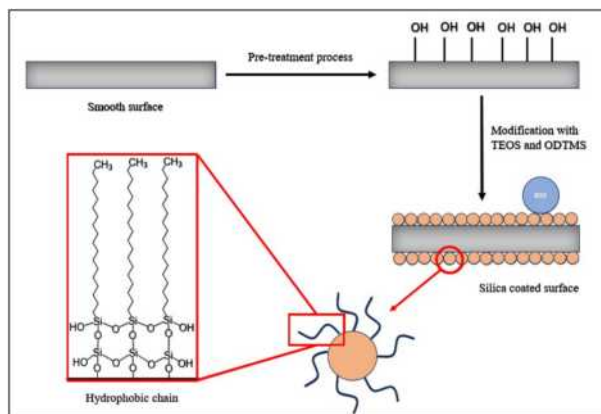


Figure 2: Schematic illustration of superhydrophobic cotton modification.

3.2.2. XPS Analysis

This section focused on analysing the cotton surface chemical properties by utilising XPS and the result was shown in Figure 3. The XPS data analysis for pure cotton showed the presence of two distinct peaks with binding energies of 533.6 eV and 287.1 eV, which were attributed to O1s and C1s respectively. These peaks proved the presence of oxygen and carbon molecules in the raw cotton. Upon the modification of cotton using ODTMS and TEOS, a shift in the peaks was observed. Specifically, the oxygen peak was now po-

sitioned at 532.6 eV which could be due to contributions from the C-Si bond. The carbon peak underwent a significant change, shifting to 384.6 eV, indicating the presence of Si-O-Si bonds. Additionally, the introduction of silicon (Si) was evident, presented by two peaks at 103.1 eV and 154.1 eV. The XPS also suggested that there was coverage of Si at 11.08%. The shifting in peaks in the XPS spectra signified the successful attachment of silica into the cotton structure which provided the hydrophobic property to the cotton. The presence of coated silica nanoparticles on the cotton fibre surfaces was also shown through the observed mass gain of cotton, measuring 143 mg for every 1027 mg of cotton, after removing the loosely bound silica nanoparticles.

3.2.3. Wettability

Surface wettability is defined as the attraction of a liquid phase to a solid phase and it is usually characterised by the water contact angle (WCA) [22]. A surface is considered hydrophilic when the WCA is less than 90° and hydrophobic when the WCA is more than 90° but less than 150° [22]. Regarding a superhydrophobic surface, it requires the WCA to surpass 150° [23]. The WCA for untreated cotton was zero as it absorbed the water droplet upon contact. As mentioned in Section 3.1, NaOH pre-treated cotton that had a WCA of 154.7° equipped the requirement as a superhydrophobic surface. However, both cotton pre-treated with cyclohexane and ethanol with acetone were deemed to be just hydrophobic as they had WCA less than 150°. The differences in the cotton WCA between three different pre-treatment methods were visualised in Figure 1.

3.3. Oil/water separation in a batch system

In this section, the oil/water separation efficiency in a batch system was investigated from several aspects. The purpose of carrying out the experiment in a static condition was to simulate an oil spill incident where cotton could be used to absorb the oil. From Figure 4, the cotton showed an impressive capability in absorbing oil which was evident with the absorption of red dyed oil and no visible residue was left. Additionally, the modified cotton also had a high affinity to oil since it did not absorb water during the process. Following this, an investigation was done to determine the efficiency of fabricated superhydrophobic cotton in absorbing various oil-based systems. Other than that, further study was conducted which involved the impact of different operational parameters, including temperatures and mixing speeds, on the separation

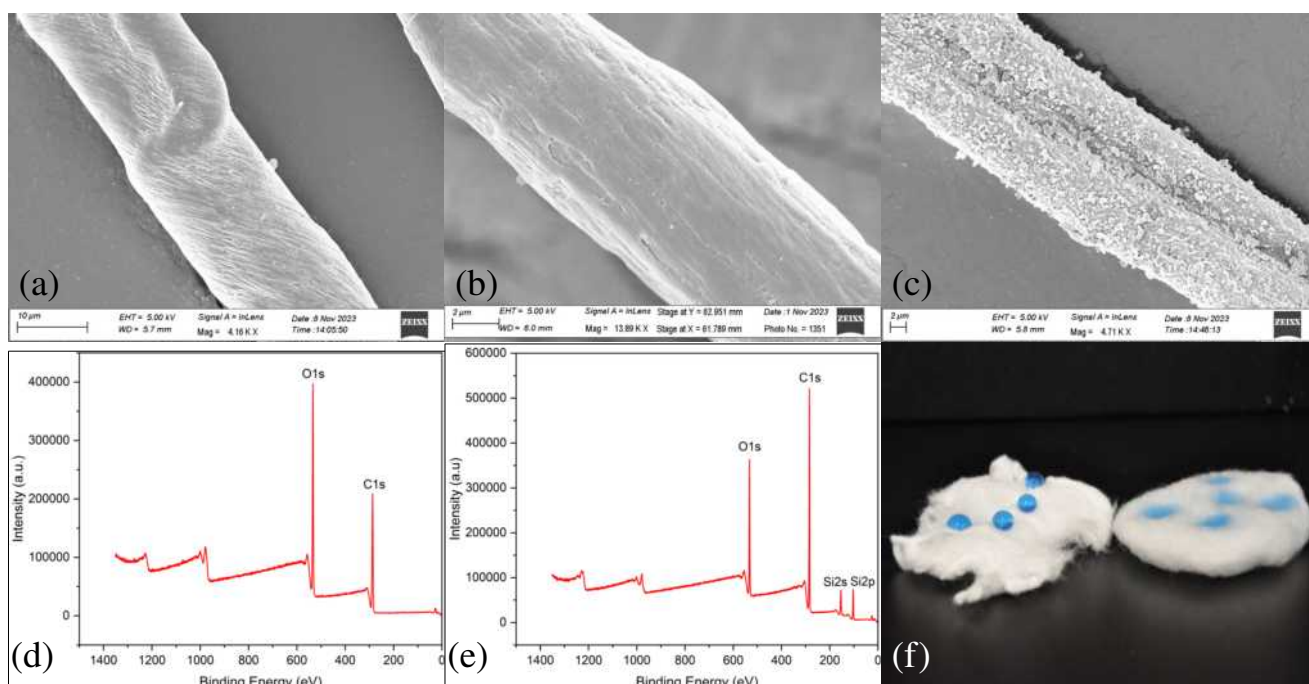


Figure 3: (a) SEM image of the surface of raw cotton fibre, (b) SEM image of the pre-treated cotton fibre surface, (c) SEM image of the modified cotton fibre surface coated densely with silica nanoparticles, (d) XPS spectra of raw cotton fibre surface with only C1s and O1s, (e) XPS spectra of modified cotton fibre surface with the presence of Si2s and Si2p at 103.1 and 154.1 eV and (f) the cotton showed superhydrophobicity by repelling water dyed in blue (left) and the raw cotton absorbed water (right).

efficiency of the superhydrophobic cotton.

their respective oil/water mixtures was investigated.

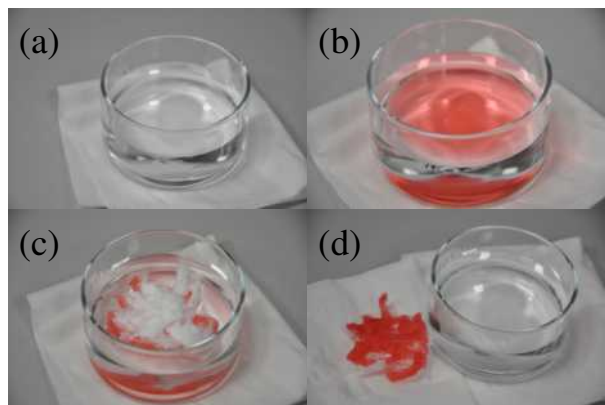


Figure 4: Photographs for oil/water separation by superhydrophobic cotton in a batch system. (a) clear water, (b) oil dyed in red floating on the water surface, (c) cotton submerged in the mixture and (d) oil was absorbed by the modified cotton.

3.3.1. Efficiency in oil/water separation for different oil-based systems

Other than petroleum, it was observed that there was a substantial volume of waste cooking oil that caused water pollution and there was inefficient cooking oil waste management taking place [24]. Therefore, in this research, the efficiency of superhydrophobic cotton in absorbing various oils – sunflower oil, rapeseed oil, olive oil, toluene, heptane, and paraffin oil from

For three cooking oils, the separation efficiency was evaluated using Equation 1 whereas the separation efficiency for toluene, heptane, and paraffin oil involved the use of Equation 2. The separation efficiencies of superhydrophobic cotton for sunflower oil/water, rapeseed oil/water, olive oil/water, toluene/water, heptane/water and paraffin oil/water were 98.9%, 98.6%, 98.7%, 98.9%, 99.1% and 98.8% respectively. Despite two methods: gravimetry and UV-Vis were used to quantify the oil absorbed by the modified cotton, the results were deemed reliable as the error bars were minimal with the largest value being $\pm 0.32\%$. In comparison to other reported materials such as stainless-steel mesh coated with cellulose nanofiber (CNF), although almost the same toluene/water separation efficiency (97.6%) was reported, the preparation process was more complicated than fabricating a superhydrophobic cotton [25]. Apart from that, Pan et al. conducted a study indicating that the fabrication of cotton through an ultrasound-assisted in situ growth method achieved a separation efficiency of more than 96%, but this approach was complex and time-consuming [18]. As the efficiencies of oil-water separation were nearly identical, further experiments utilised olive oil because of its darker colour, which facilitated easier observation.

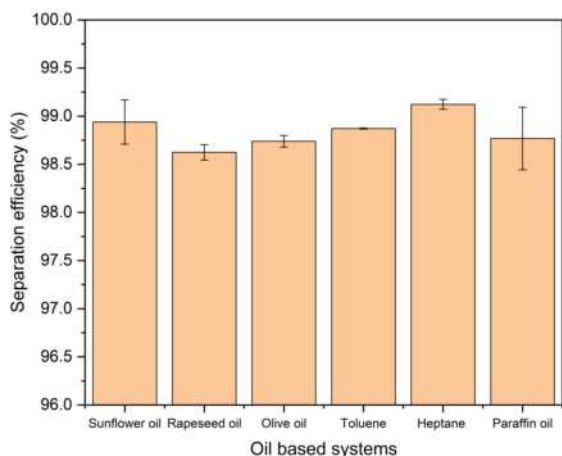


Figure 5: Oil/water separation efficiency (%) of superhydrophobic cotton in different oil-based systems.

3.3.2. Influence of temperature on the oil/water separation efficiency

Temperature has always been one of the most fundamental aspects to investigate when it comes to separation processes. In this research, it was essential to investigate the impact of temperature on the efficiency of oil/water separation, given the variations in temperature across different countries. For instance, the average water temperature of the North Sea is around 12°C over the year whereas the seawater temperature can reach around 26.5°C in the South China Sea [26]. Furthermore, once the relationship between temperature and oil/water separation is understood, an optimal operating temperature can be determined for effective oily wastewater treatment.

Experiments were conducted by placing ~200mg cotton into 20 ml of water with 2 ml of olive oil at 10°C, room temperature (23°C) and 60°C. As depicted in Figure 6, elevating the temperature of the oil and water mixture had been observed to enhance separation efficiency. There was a slight increase in the oil/water separation efficiency of around 0.7% when transitioning from 10°C to 60°C. This finding could be explained by the fact that an increase in temperature caused a reduction in oil viscosity. As the mobility ratio of oil to water rose, it led to an improved flow capability of the oil phase, thereby enhancing the mass transfer rate of oil into the cotton [27]. The observed outcome also could be ascribed to the expansion of pores and the formation of new active sites on the adsorbent's surface, resulting from distortion due to the temperature increment [28]. Based on these results, the cotton was able to maintain approximately the same separation efficiency in this temperature range which indicates its capability to perform well in a similar environment. While it was believed that raising

the temperature could improve separation efficiency, the actual enhancement was marginal and there is a potential for increased costs at higher temperatures.

3.3.3. Effect of mixing speed on the oil/water separation efficiency

To model the weather and wave movements in the ocean, it was crucial to take into account the impact of agitation speed on the oil/water separation. The mixing rate was a key element that influenced both the development of the outer boundary layer and the dispersion of oil within the bulk solution. Therefore, an experiment was done to explore the impact of mixing speed on the separation efficiency.

The oil and water were mixed for 5 min before soaking the cotton in. When the mixing speed increased from 200 rpm to 400 rpm, the separation efficiency decreased. A reduction in the separation efficiency with the increase of mixing speed could be due to the formation of oil-water emulsion which weakened the attraction forces between the oil molecules and the adsorption sites on the surface of the cotton. It was observed that the oil-water emulsion was unstable when the mixing rate was 400 rpm where macroemulsion was formed. Since macroemulsion droplets will be separated into two-phase mixtures again in minutes after the oil extraction by modified cotton, the UV-Vis picked up the residual oil, making the oil/water separation efficiency lower [29]. However, an increase in the mixing speed to 600 rpm resulted in a significant efficiency improvement, possibly due to the formation of a stable oil-water emulsion at this speed, leading to a poor extraction of oil from the sample.

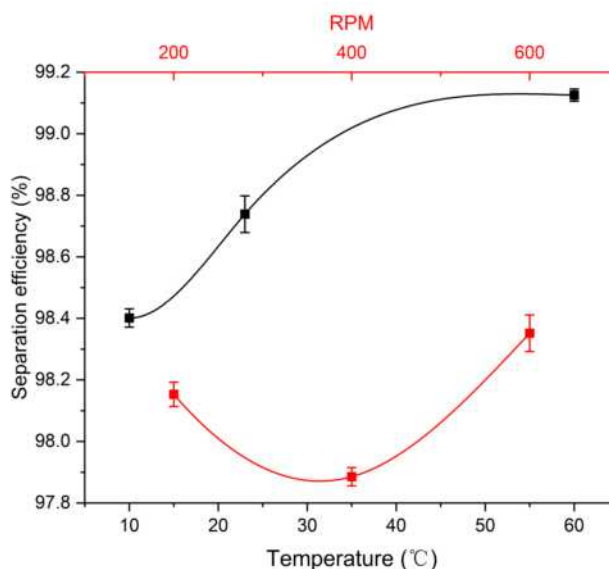


Figure 6: Oil/water separation efficiency (%) of modified cotton in different RPMs and temperatures.

3.4. Oil/water separation in flow system

Fixed bed columns are commonly employed for practical purposes because of their convenient and continuous operational characteristics, for example in H₂S adsorption separations [30]. Hence, this section centred on the oil/water separation process in a continuous flow system by using a syringe as a column packed with the modified cotton.

3.4.1. Breakthrough curve

To simulate the implementation of the superhydrophobic cotton in an industrial condition, an oil concentration of 8000 ppm was used to construct a breakthrough curve under a consistent flow rate of 0.62 ml/s at room temperature. The $t=0$ point was marked at the moment when a droplet of the mixture emerged from the syringe. Samples of 4 ml each were extracted at 60-second intervals to measure the oil volume left in the filtrate. A plot of C_e/C_0 against time was constructed where C_e represents the outlet concentration of oil, and C_0 denotes the initial oil concentration in the reservoir. As depicted in Figure 7, the oil concentration remained at 0 initially, gradually increasing around 420 seconds. There was a distinct jump after 420 seconds, where the value of C_e/C_0 almost reached unity. Following this jump, the concentration remained constant. Analysis of the breakthrough curve indicated that at an oil concentration of 8000 ppm and a steady flow rate of 0.62 ml/s, the cotton became saturated with oil by 600 seconds.

The flow system exhibited non-homogeneity as it evolved into a two-phase mixture along the pipeline. During the experiment, it was observed that oil adhered to both the inner pipe and syringe walls. This led to the value of C_e/C_0 not precisely equal to 1, indicating that the outlet concentration differs from the inlet concentration after reaching the breakthrough point. The point where C_e/C_0 reached its peak before decreasing to a constant could be attributed to the non-uniform distribution of oil and water within the system which led to variations in concentration along the flow path.

3.4.2. Effect of oil concentration on the filtrate volume

Typically, the concentration of oil in industrial oily wastewater systems ranges from 10 ppm to 200000 ppm [31]. Different regulatory thresholds for the maximum oil concentration in the discharge of oily wastewater were established by many countries and the limits normally fell within the range of 5 to 100 ppm [2]. Therefore, it is important to treat the oily

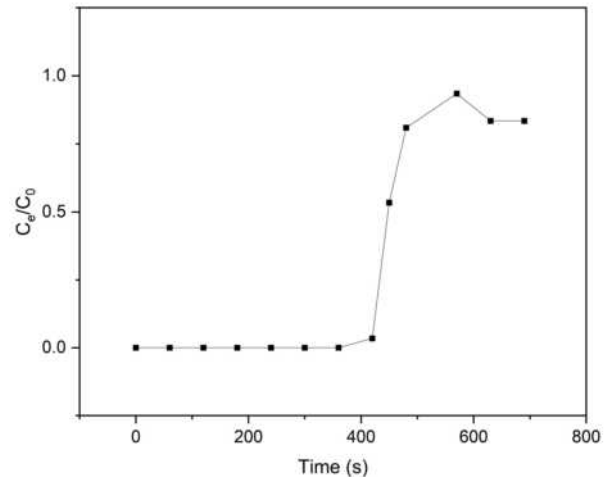


Figure 7: Breakthrough curve for oil (8000ppm) sorption using superhydrophobic cotton.

wastewater before discharging into the ocean. Since it was reported by Rasouli that there will be no significant impact on the oil/water separation efficiency with the changes in oil concentrations, it was decided to explore the relationship between oil concentration and the filtrate volume in this research [32].

The same experimental setup for the breakthrough curve was applied but with varying oil concentrations – 4000 ppm, 8000 ppm, 12000 ppm, 16000 ppm and 20000 ppm. As indicated in Figure 8, when the oil concentration increased, less filtrate was collected prior to the breakthrough of the filtration system. This was due to the fact that at a consistent flow rate of 0.62 ml/s, increasing oil concentrations will cause more oil particles to be absorbed into the cotton, resulting in an increased oil absorption rate, thereby yielding a smaller volume of filtrate. This aligned with the findings of Huang and Lim, who clarified that the decrease in filtrate volume with increasing oil concentration resulted from a reduction in the hydraulic conductivity of the filtration system [33]. It was also noticed that the time taken for the column to be saturated became shorter with the increase in oil concentration. From a theoretical aspect, according to Darcy's law, with a reduction in both the hydraulic conductivity and time taken and provided that all other parameters are kept constant, the filtrate volume will decrease as well [34].

$$\frac{1}{A} \frac{dV}{dt} = \frac{K(-\Delta P)}{l} \quad (4)$$

where A is the cross-sectional area of the bed, V is the filtrate volume, K is the hydraulic conductivity, ΔP is the pressure drop across the bed and l is the bed thickness.

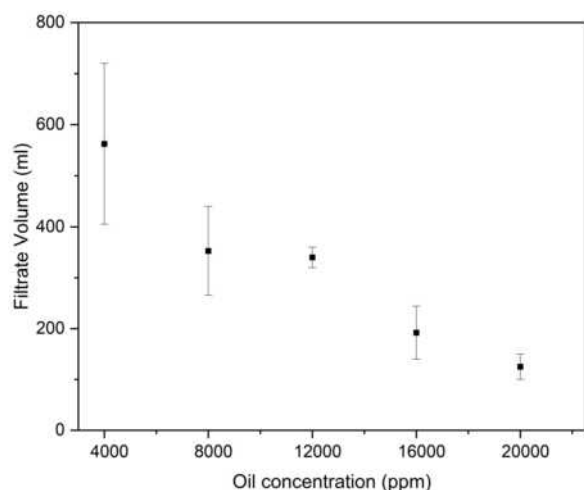


Figure 8: The filtrate volume associated with different oil concentrations.

3.5. Chemical and laundering durability

The limited durability of superhydrophobic cotton wool poses a challenge to their practical use. Therefore, the chemical and laundering durability of superhydrophobic cotton is an important aspect to ensure that the cotton will be long-lasting and can be reused. The chemical durability of modified cotton was determined by submerging it in six distinct solutions including solvents such as acetonitrile, NaCl, pyridine and toluene and solutions with extreme pH values (pH 1.3 and pH 13.5). The laundering durability of the modified cotton was examined by a laundry test with the modified cotton being mixed in a 0.15 wt% SDS surfactant at 200 rpm for 24h. Post-testing, it was observed that the modified cotton surfaces were still under good condition with no noticeable difference in the WCAs of the cotton before and after the exposure to these solutions; they consistently maintained an angle of around 150°. Compared to the pristine modified cotton with a WCA of 154.7°, there was only a slight decrease of around 5° in the cotton WCA after experiencing the harsh conditions.

The chemical and physical stability can also be validated by quantifying the water sorption of the modified cotton. As illustrated in Figure 9, the cotton water sorption remained under 0.5 g/g after the durability tests. The results for cotton immersing in acetonitrile and pyridine were comparable to the water sorption of the clean modified cotton wool (0.06 g/g). For the laundry test (32 cycles with 1 laundry cycle = 45 minutes), although the water sorption was around 8 times higher than the clean modified cotton, it still exhibited a lot less water sorption compared to the raw cotton that had a water sorption capacity of 43.3 g/g. These results demonstrated that the superhydrophobic cotton

had outstanding resistance towards rigorous chemical environments which displayed its ability to be applied in the oily wastewater treatment.

It was expected for the water sorption capacity to have nearly identical values across all conditions as the WCAs were almost the same. However, contrary outcomes were observed and there were substantial margins of uncertainty in the measurements, which may result from the different degrees of shedding of cotton fibres while submerged, causing the actual cotton mass to be less than the initial cotton mass measured.

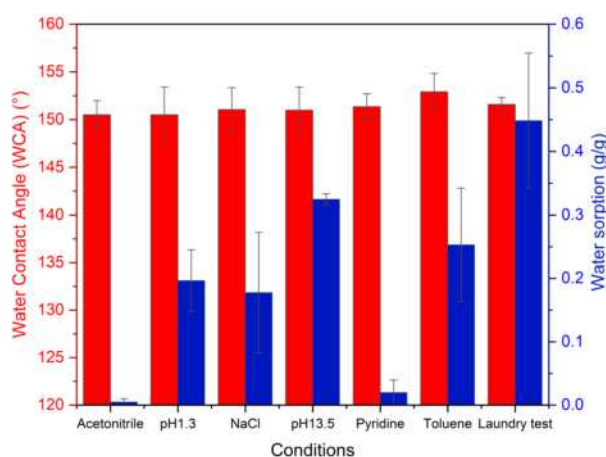


Figure 9: WCAs and water sorption capacities of the modified cotton after being immersed in acetonitrile, NaCl solution, pyridine, toluene, HCl (pH 1.3), ammonia solution (pH13.5) and SDS surfactant for 24 hours.

3.6. Recyclability

The ability to recycle and reuse is a crucial factor that determines the reusability and effectiveness of the sorbent, in this case, the superhydrophobic cotton [35]. Several studies have been done on investigating the recyclability of superhydrophobic cotton wool, but they mainly focussed on the changes in the cotton WCA [36]. As the main purpose of this research was to fabricate durable superhydrophobic cotton for oily wastewater treatment, the oil/water separation efficiency was the key measure in this section instead.

Figure 10 outlined the relationship between the oil/water separation efficiency and the number of separation cycles. The cotton was washed with hexane after each separation cycle to remove oil from the cotton for sorbent regeneration so that it could be reused for the next oil/water separation cycle. Following 20 times of sorption/desorption, the oil/water separation efficiency of the superhydrophobic cotton showed an approximate decrease of 1.4%. The decrease in the oil sorption could be due to the destruction of the hydrophobic layer on the surface of cotton fibres owing to repeated uses and multiple washes [37]. This can

be validated by noting that the cotton absorbed more water as the number of separation cycle increased. Another factor of the reduction in oil/water separation efficiency was the residual oil inside the cotton wool [38]. The residual oil resulted in a reduction of active sites on the cotton surfaces, leading to a decreased capacity for oil sorption. Regardless, the separation efficiencies remained above 98.6% for 20 separation cycles, highlighting excellent recyclability for its application in oil/water separation.

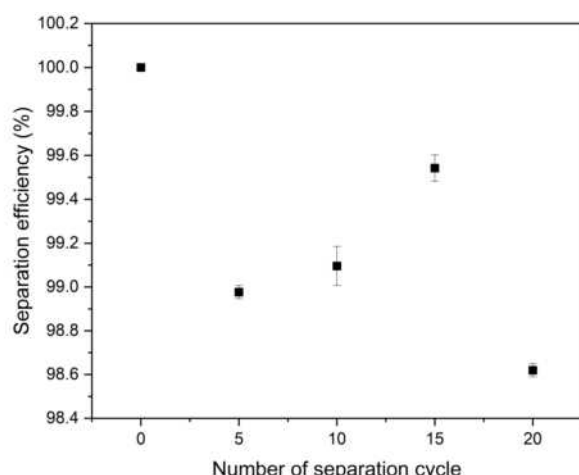


Figure 10: Oil/water separation efficiencies of modified cotton after a different number of separation cycles.

4. Conclusion and Outlook

The pre-treatment process of cotton using NaOH, followed by chemical modification using TEOS and ODTMS had proven to be successful in introducing a superhydrophobic property, as evidenced by a contact angle of 154.7° . Upon the cotton modification, SEM analysis revealed complete coverage of cotton fibres by silica nanoparticles. The hydrophobic chains in the silica contributed to the water-repellent property of the cotton. This superhydrophobic cotton showed significant applicability in oil/water separation systems, proven by a separation of nearly 99% when using different oil-based systems. Durability and laundering tests confirmed that the modified cotton exhibited great durability when immersed in various solutions, with only a decrease in WCA of around 5° . Additionally, the water sorption capacity of the cotton maintained below 0.5 g/g, proving its endurance in harsh environments. Furthermore, the modified cotton was proved to be sustainable and environmentally friendly by being reusable for up to 20 cycles. In a separate experiment, it was observed that the superhydrophobic cotton could treat less water when the oil/water mixture had higher oil content. Improvements in oil/water separation within a continuous system can be achieved by increasing the temperature and maintaining a low

mixing speed. These collective findings have positioned the superhydrophobic cotton as a promising solution with great potential in oil/water separation.

Looking ahead, the upcoming research will centre on the synthesis of cotton, focusing on understanding how TEOS and ODTMS concentrations influence superhydrophobicity to ensure the optimal amount of TEOS and ODTMS are used to minimise wastage. The investigation will extend to various operational parameters in the continuous flow system such as flow rate. This also involves experimenting with parameters like void fraction, diameter, or thickness of the packing. The research scope also involves testing the modified cotton's effectiveness in handling heavier oils, particularly petroleum-based substances. Moreover, it is important to address the inaccuracies in data resulting from the oil's affinity to wall surfaces to enhance the findings and ensure the reliability of the results. By delving into these aspects, the aim is to enhance the robustness and applicability of superhydrophobic cotton in oil/water separation, hence broadening its potential in diverse environments.

Acknowledgments

The authors would like to thank the PhD student, Tsaone Gosiamemang and the research fellow, Vivek Verma for their guidance throughout the research.

References

- [1] United Nations. *The Sustainable Development Goals Report*. United Nations Publications, 2020.
- [2] Khaled Abuhasel, Mohamed Kchaou, Mohammed Alquraish, Yamuna Munusamy, and Yong Tzyy Jeng. Oily wastewater treatment: Overview of conventional and modern methods, challenges, and future opportunities. *Water*, 13(7):980, 2021.
- [3] Alesia Sandifer, Paul A. and Ferguson, , Melissa Finucane, Melissa L. and Partyka, Ann Haywalk Solo-Gabriele, Helena M. and Walker, Kateryna Wowk, Rex Caffey, and David Yoskowitz. Human health and socioeconomic effects of the deepwater horizon oil spill in the gulf of mexico. *Oceanography*, 34(1):174–191, 2021.
- [4] Zunaira Asif, Zhi Chen, Chunjiang An, and Jinxin Dong. Environmental impacts and challenges associated with oil spills in shorelines. *Journal of Marine Science and Engineering*, 10(6):762, 2022.
- [5] Chadetrik Rout and Arabinda Sharma. Oil spill in marine environment: Fate and effects. 2013.
- [6] X. Qin and S. Subianto. *Electrospun nanofibers for filtration applications*, chapter Electrospun Nanofibers. Woodhead Publishing, 2017.
- [7] Lin Feng, Zhongyi Zhang, Zhenhong Mai, Yongmei Ma, Biqian Liu, Lei Jiang, and Daoben Zhu. A superhydrophobic and super-oleophilic coating mesh film for the separation of oil and water. *Angewandte Chemie International Edition*, 43(15):2012–2014, 2004.
- [8] Dula Daksa Ejeta, Chih-Feng Wang, Shiao-Wei Kuo, Jem-Kun Chen, Hsieh-Chih Tsai, Wei-Song Hung, Chien-Chieh

- Hu, and Juin-Yih Lai. Preparation of superhydrophobic and superoleophilic cotton-based material for extremely high flux water-in-oil emulsion separation. *Chemical Engineering Journal*, 402, 2020.
- [9] Qingzhen He, Zhibiao Guo, Shiyu Ma, and Zhiwei He. Recent advances in superhydrophobic papers for oil/water separation: A mini-review. *ACS omega*, 7, 2022.
 - [10] Yu-Qing Zhang, Yu-Hui Jiang, Ya-Nan Qin, Qing-Da An, Ling-Ping Xiao, Zhan-Hua Wang, Zuo-Yi Xiao, and Shang-Ru Zhai. Cooperative construction of oil/water separator using renewable lignin and pdms. *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, 643, 2022.
 - [11] Baiyi Chen, Rongrong Zhang, Hexuan Fu, Jiadai Xu, Yuan Jing, Guohe Xu, Bin Wang, and Xu Hou. Efficient oil–water separation coating with robust superhydrophobicity and high transparency. *Scientific Reports*, 12:2187, 2022.
 - [12] Charikleia Spathi, Neale Young, Jerry Y.Y Heng, Luc J.M Vandeperre, and Chris R. Cheeseman. A simple method for preparing super-hydrophobic powder from paper sludge ash. *Materials Letters*, 142:80–83, 2015.
 - [13] Ahmer Hussain Shah, Yuqi Zhang, Xiaodong Xu, Abdul Qadeer Dayo, Xiao Li, Shuo Wang, and Wenbin Liu. Reinforcement of stearic acid treated egg shell particles in epoxy thermosets: Structural, thermal, and mechanical characterization. *Materials*, 11(10):1872, 2018.
 - [14] Mingjie Liu, Xingxing Liu, Shaoqin Zheng, Kangle Jia, Longfei Yu, Jinlan Xin, Junhua Ning, Wu Wen, Linjia Huang, and Jinbiao Xie. Environment-friendly superhydrophobic sponge for highly efficient oil/water separation and microplastic removal. *Separation and Purification Technology*, 319, 2023.
 - [15] Junwei Xu, Ya Chen, Lida Shen, Jiantao Zhao, Guibin Lou, Dazhi Huang, and Youwen Yang. Fabrication of superhydrophobic stainless-steel mesh for oil-water separation by jet electrodeposition. *Colloids and Surfaces: Physicochemical and Engineering Aspects*, 649, 2022.
 - [16] Sheng Lei, Zhongqi Shi, Junfei Ou, Fajun Wang, Mingshan Xue, Wen Li, Guanjuan Qiao, Xinhai Guan, and Jie Zhang. Durable superhydrophobic cotton fabric for oil/water separation. *Colloids and Surfaces A*, 533:249–254, 2017.
 - [17] Solomon Tilahun Desisa, Eylen Sema Dalbaşı, and Ahmet Çay. Production of superhydrophobic cotton fabric by layer-by-layer deposition of $\text{SiO}_2/\text{TiO}_2$ -polydimethylsiloxane. *Journal of Natural Fibers*, 2022.
 - [18] Guangming Pan, Xinyan Xiao, Nanlin Yu, and Zhihao Ye. Fabrication of superhydrophobic coatings on cotton fabric using ultrasound-assisted in-situ growth method. *Progress in Organic Coatings*, 125:463–471, 2018.
 - [19] Na Lv, Xiaoli Wang, Shitao Peng, Lei Luo, and Ran Zhou. Superhydrophobic/superoleophilic cotton-oil absorbent: preparation and its application in oil/water separation. *RSC Advances*, 8, 2018.
 - [20] Phuong Nguyen-Tri, Funda Altıparmak, Nam Nguyen, Ludovic Tuduri, Claudiane M. Ouellet-Plamondon, and Robert E. Prud'homme. Robust superhydrophobic cotton fibers prepared by simple dip-coating approach using chemical and plasma-etching pretreatments. *ACS Omega*, 4(4): 7829–7837, 2019.
 - [21] Rafael Garcia Candido. 17 - recycling of textiles and its economic aspects. In Md. Ibrahim H. Mondal, editor, *Fundamentals of Natural Fibres and Textiles*, The Textile Institute Book Series, pages 599–624. Woodhead Publishing, 2021.
 - [22] Alina Peethan, M. Aravind, and Sajan Daniel George. Surface Wettability and Superhydrophobicity. In *Advances in Superhydrophobic Coatings*. Royal Society of Chemistry, 09 2023. ISBN 978-1-83916-786-7.
 - [23] Jeya Jeevahan, M. Chandrasekaran, R.B. Joseph, G. Britto; Durairaj, and G. Mageshwaran. Superhydrophobic surfaces: a review on fundamentals, applications, and challenges. *Journal of Coatings Technology and Research*, 15: 231–250, 2018.
 - [24] O A. Olu-Arotiowa, A.A Odesanmi, Adedotun B.K., Ajibabe O. A., Olasesan I.P., Odofin O.L., and Abass A.O. Review on environmental impact and valorization of waste cooking oil. *LAUTECH Journal of Engineering and Technology*, 16(1):144–163, 2022.
 - [25] M. Mahbulul Bashar, Huie Zhu, Shunsuke Yamamoto, and Masaya Mitsuishi. Superhydrophobic surfaces with fluorinated cellulose nanofiber assemblies for oil–water separation. *RSC Advances*, 7:37168–37174, 2017.
 - [26] Sea temperature. <https://seatemperature.net/>. Accessed: 2023-12-02.
 - [27] Yadong Qin, Yongbin Wu, Pengcheng Liu, Fajun Zhao, and Zhe Yuan. Experimental studies on effects of temperature on oil and water relative permeability in heavy-oil reservoirs. *Sci Rep*, 8, 2018.
 - [28] Yasemin Kismir and Ayse Z. Aroguz. Adsorption characteristics of the hazardous dye brilliant green on saklıkent mud. *Chemical Engineering Journal*, 172(1):199–206, 2011.
 - [29] Tanvi Sheth, Serena Seshadri, Tamas Prileszky, and Matthew E. Helgeson. Multiple nanoemulsions. *Nature Review Materials*, 5:214–228, 2020.
 - [30] Sebastião Mardonio Pereira de Lucena, Daniel Vasconcelos Gonçalves, José Carlos Alexandre de Oliveira, Moises Bastos-Neto, Célio Loureiro Cavalcante, and Diana Cristina Silva de Azevedo. *Activated Carbons for H₂S Capture*, pages 197–215. Springer International Publishing, Cham, 2021.
 - [31] M Cheryan and N Rajagopalan. Membrane processing of oily streams. wastewater treatment and waste reduction. *Journal of Membrane Science*, 151(1):13–28, 1998.
 - [32] S. Abbas Rasouli. Fabrication of superhydrophobic-superoleophilic membranes for oil-water separation applications, 2021.
 - [33] Xiaofeng Huang and Teik-Thye Lim. Performance and mechanism of a hydrophobic–oleophilic kapok filter for oil/water separation. *Desalination*, 190(1-3):295–307, 2006.
 - [34] M.King Hubbert. Darcy's law and the field equations of the flow of underground fluids. *Trans*, 207:222–239, 1956.
 - [35] Despina A. Gkika, Athanasios C. Mitropoulos, and George Z. Kyzas. Why reuse spent adsorbents? the latest challenges and limitations. *Science of The Total Environment*, 822, 2022.
 - [36] Sang Wook Han, Eun Ji Park, Myung-Geun Jeong, Il Hee Kim, Hyun Ook Seo, Ju Hwan Kim, Kwang-Dae Kim, and Young Dok Kim. Fabrication of recyclable superhydrophobic cotton fabrics. *Applied Surface Science*, 400:405–412, 2017.
 - [37] Ali Ashraf Derakhshan, Meghdad Pirsaeheb, and Sirus Zinadini. Synthesis of sustainable poly(s-abietic-co-pinene) through inverse vulcanization of kurdica gum and used to fabricate durable and recyclable super-hydrophobic cotton wool filter: Oil-water separation application. *Progress in Organic Coatings*, 168, 2022.
 - [38] Jintao Wang, Fenlan Han, and Guihong Geng. Hydrothermal fabrication of robustly superhydrophobic cotton fibers for efficient separation of oil/water mixtures and oil-in-water emulsions. *Journal of Industrial and Engineering Chemistry*, 54:174–183, 2017.

Investigating the Electrochemical Behaviour of Graphitic Carbon Cathodes in Aluminium Dual-ion Batteries

James White and Jiahui He

Department of Chemical Engineering, Imperial College London, U.K.

Abstract

Lithium-ion batteries (LIBs) have dominated the majority of battery applications for many years. However, they face a number of growing challenges, including a scarce abundance of chemical components for manufacturing LIBs, safety concerns and the high manufacturing cost. A way to combat and alleviate these challenges is to research and develop alternative battery chemistries such as aluminium dual-ion batteries (ADIBs) with graphitic based cathode materials. In this study, six different graphitic cathode materials were investigated with each material having three different quantities tested. These materials included graphene nano-platelets, pure graphite, pure soft carbon (SC), 75wt%:25wt% of graphite:SC, 25wt%:75wt% of graphite:SC and finally 50wt%:50wt% of graphite:SC. ADIB coin cells for each cathode material were constructed within an inert argon atmosphere using Aluminium as the anode and 1-ethyl-3-methylimidazolium chloride [EMIm]Cl as the electrolyte. The cells were tested which involved 200 galvanostatic cycles of charging/discharging under a constant current density of 500 mA/g. Results revealed that the 75wt%:25wt% of graphite:SC and pure graphite were generally the best performing cathode materials while a greater presence of SC in the cathode was generally inversely proportional to the overall cell performance. Two trends regarding the impact of mass loading were also observed. Graphite revealed a larger initial coulombic efficiency, a better capacity retention with higher amount of active material, while 75wt% graphite and 25wt% soft carbon and pure soft carbon had the opposite trends.

Key words: Aluminium Dual-Ion Batteries (ADIBs), tuneable composite cathode, active material loading, graphitic cathode, soft carbon, galvanostatic cycling.

1. Introduction

Expanding the use of renewable energy resources (RESs) and low-carbon technology, as well as the utilization of microgrids and smart grids, has been considered as the path of decarbonizing the planet and combating climate change ^[1]. The European Union considers the advancement of battery technology as critical in developing the green economy because batteries can be used to maintain a balance between supply and demand within the electricity grid while mitigating the inherent intermittency of RESs thanks to their energy storage capacity and rechargeable characteristics ^[2]. Presently, lithium-ion batteries (LIBs) dominate the battery market because of their ability to have a high voltage and capacity per unit mass and volume. This is down to lithium's small atomic weight and radius ^[3]. From a material perspective, however, manufacturing LIBs that can satisfy energy demands is a challenge. Concerns include the low relative abundance of the chemical elements necessary for electrodes, such as cobalt, and high energy cost associated with battery production, transportation and recycling (400kWh are needed to make 1kWh LIBs, releasing about 75kg CO₂, according to the early integrated LCA estimate) ^[4]. Therefore, potential techniques for developing forefront sustainable storage technologies hinge on the investigation of batteries with alkali-metal-anodes, for example, Na, K, Mg, and Al.

Rechargeable, high-valent aluminium-dual ion batteries (ADIBs) offer possibilities for safe, low cost and high energy density operation. Driven from

its inertness and ease of handling under atmospheric conditions, aluminium has a superior safety characteristic. The high cost-effectiveness of aluminium is attributable to its abundance (it is the most abundant metal in the earth's crust) and the existing established industrial and recycling infrastructure of aluminium. Compared to LIBs, the energy density of AIBs could be further enhanced on a per unit volume basis because the volumetric capacity of aluminium (8.056 Ah/cm³) is four times larger than that of lithium (2.042 Ah/cm³), while its gravimetric capacity (2.981 Ah/g) is also higher than the majority of alkali metals ^[5]. Additionally, Al provides a trivalent ion Al³⁺, which is comparable to three Li⁺ ions in conventional intercalation cathodes, allowing for more electrons and ions to be taken by the cathode with little pulverization ^[5]. For mobile devices, a high energy density is preferable. For instance, an electric vehicle could potentially have two to six times the capacity of LIBs with the same volume AIBs ^[6].

Based on available literature, the AIBs are categorised into both aqueous and non-aqueous operating schemes. It is known that aluminium batteries based on aqueous systems suffer from severe problems such as a passivating oxide layer formation, a hydrogen side reaction and material corrosion. Therefore, a non-aqueous system with a chloroaluminate-based ionic liquid electrolyte is usually implemented for the secondary aluminium battery instead ^[6]. As suggested by Muldoon et al. and Elia et al. in 2016 ^[2], a high-purity aluminium metal is typically used as the anode in novel aluminium dual-ion batteries (ADIBs) with its

ability to exchange three electrons per Al atom. The ionic liquid (IL) electrolyte was made of a mixture of 1-ethyl-3-methylimidazolium chloride ([EMIm]Cl) and aluminium chloride (AlCl₃) in a molar ratio of 1.5:1. Chloroaluminate (AlCl₃) melts dissolved in 1-ethyl-3-methylimidazolium chloride ([EMIm]Cl) ionic liquid as this molar ratio is considered as a promising electrolyte. Despite the fact that the chloroaluminate melts electrolyte has various practical challenges due to its high reactivity, corrosivity, and hygroscopic nature, its benefits become dominant when considering its ability to avoid the surface oxide film formation and its support for reversible Al deposition/stripping at this point.

Finding a high-performance cathode of AIBs has been an unsolved difficulty because the high charge density of Al³⁺ ions cause a high intercalant concentration in the host materials [7]. This presents a challenge for the reversible electrochemical intercalation of Al³⁺ ions, which functioned via the cathode. Drawing from prior research, only a handful of the proposed cathode compounds - which are known from conventional lithium, sodium, or magnesium ion batteries research - appear to reveal an evident reversible intercalation of Al³⁺ ions. Layered TiS₂, various manganese oxides, V₂O₅ and graphite warrant particular focus in recent years as they showed the strongest Al³⁺ intercalation [6]. Graphite materials prevail over alternative cathodes due to its capability to reversibly accommodate Al³⁺ ions between its planar graphene sheets, and its superior stability and low operational potential [23]. The schematic configuration of ADIBs has been shown in Figure 1 below.

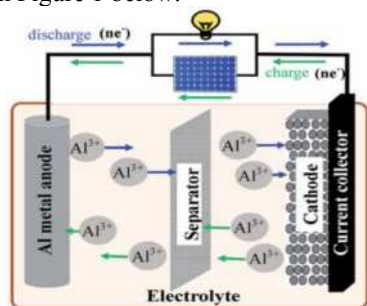
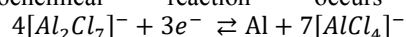


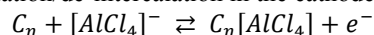
Figure 1: Schematic configuration of aluminium-ion battery. [5]

The electrolyte performs as the transmitter for ions. The redox active chloroaluminate anions AlCl₄⁻ and Al₂Cl₇⁻ can both be provided only when the Lewis acid AlCl₃ is excessive (>50 mol%), forming the acidic electrolyte [5]. These two anions are actively involved in redox reactions at both the anode and the cathode, causing the variation of electrolyte composition upon charging and discharging. Al₂Cl₇⁻ is the main responsible anion to facilitate Al deposition/stripping occurring at the anode and the electrochemical reaction occurs via



As three AlCl₄⁻ anions insert into the graphite, an aluminium atom is deposited simultaneously. Charging halts when only AlCl₄⁻ anions remain in the electrolyte, resulting in a neutral chloroaluminate melt with an AlCl₃ to [EMIm]Cl ratio of 1 [2].

Owing to carbon-based material's loosely bonded layered structure, AlCl₄⁻, the single-charged complex anion, is transmitted through the electrolyte and gets stored among the graphene stacked layers at the cathode during the charging process, which is known as intercalation. While discharging, the AlCl₄⁻ anions de-intercalate from the graphitic cathode and return to the anode. The following half-reaction presents the mechanism of anion intercalation/de-intercalation in the cathode:



Where C is the graphitic carbon, n denotes the molar ratio of carbon atoms to intercalated anions in the graphite.

The structural and electrochemical characteristics of the pure graphene nanoplatelets (GNP) cathode and the pure graphite (G) cathode was investigated and assessed for the first time to study the typical trend and correlations in galvanostatic charge-discharge. Considering the disordered, semi-graphitic nature of mesoporous soft carbon (SC), the project delved into various ratios of G-SC composite cathodes, and the primary objective was to compare the electrochemical performance among these different graphitic carbon cathode types in non-aqueous aluminium dual-ion batteries.

II. Methodology

The three fundamental components for ADIBs (the graphitic carbon cathode, the metallic aluminium anode and the ionic liquid electrolyte) must all be prepared before fabricating the aluminium coin cell.

1. Preparation of Cathode

The cathode slurry is conventionally prepared by mixing the cathode material, a conductive additive such as carbon black, a polymeric binder such as sodium alginate and a liquid solvent such as deionised water. In the first stage, the electrochemically active cathode material was GNP. In the experiment, a 1g solid mixture composed of 75%wt of GNP powder, 15%wt of sodium alginate powder and 10wt% of carbon black powder was mixed with 15ml of deionised water to create a slurry. 15 identical-shaped molybdenum (Mo) disc samples with a thickness of 0.025mm and a diameter of 12mm were weighed, and the well-mixed homogenous slurry was applied to the molybdenum discs with varying quantities to compare the effect of cathode material mass loading on the electrochemical performance of ADIBs. Mo samples 1 to 5 were smeared with 15.8μl of slurry (with an average mass loading of $4.8 \times 10^{-3} \text{ mg/mm}^2$), samples 6 to 10 were smeared with 31.6μl of slurry (with an average mass loading

of $11.6 \times 10^{-3} \text{ mg/mm}^2$) and 11 to 15 were smeared with 47.4 μl of slurry (with an average mass loading of $18.9 \times 10^{-3} \text{ mg/mm}^2$). These GNP cathode samples were dried overnight at 80 °C under vacuum and then re-weighed to calculate the mass loading of each individual sample using the following equation:

$$\text{Mass Loading} = (\text{Mass}_{\text{Dried cathode sample}} - \text{Mass}_{\text{Pure Mo disc}}) \times \text{Composition}_{\text{Cathode material}}$$

All the detailed calculations can be found in the spreadsheet under Supplementary Information.

Cathodes consisting of a blend of graphite and soft carbon (SC) in different compositions were also prepared to compare and assess the electrochemical behaviour of various graphitic carbon cathodes and to investigate which graphite material performed best as a cathode in ADIBs. To prepare graphite cathodes, graphite and mesoporous pitch were weighed in the fume cupboard with specific compositions, followed by grinding with an agate pestle and mortar for 30 minutes. Since the MSC is typically formed by heat treatment of mesoporous pitch, the grounded mixture was therefore heated in the furnace at 1100 °C for 2 hours. The post-heated mixture was then further ground into a homogeneous powder via an agate pestle and mortar. The rest of the steps to make graphite cathodes are the same as the steps to make GNP cathodes with all other quantities the same but substituting the GNP with graphite.

This research investigated 6 various graphite materials for ADIB cathodes: pure GNP, pure graphite, pure soft carbon, a mixture of 75wt% graphite and 25wt% soft carbon, a mixture of 50wt% graphite and 50wt% soft carbon, and a mixture of 25wt% graphite and 75wt% soft carbon. Three different quantities of cathode material were tested for every type of cathode except for the 25wt% graphite and 75wt% soft carbon mixture. This was because of a shortage of Mo discs at the time of cell assembly for this mixture.

2. Preparation of Aluminium Anode

High-grade pure aluminium foil of 99.999% purity and 0.050mm thickness was cut into 16mm diameter discs. These aluminium discs were washed and dried in order to remove any contaminants on the surface. An ultrasonic cleaner was used to clean the surfaces of Al discs under ethanol first, which was followed by immersing the discs in a highly concentrated solution of nitric acid (3M concentration) for exactly five minutes to dissolve any chemical contaminants. To ensure that the leftover nitric acid was completely eliminated, the acid-washed aluminium was rinsed several times with deionised water and acetone before being dried at 80°C using a vacuum oven.

3. Preparation of Ionic Liquid Electrolyte

The preparation of ionic liquid electrolyte was carried out in an inert Argon environment with less than 0.5ppm concentration of oxygen and water at standard room temperature through the use of a glovebox. Before being mixed with anhydrous AlCl_3 inside the glovebox, [EMIm]Cl had been degassed under vacuum conditions and heated at 50°C for 10 hours using a Smart VacPrep system. The AlCl_3 to [EMIm]Cl mixing mole ratio of the commercial ionic liquid electrolyte of AIBs is approximately 1:1.5 with over 98% purity ^[12], resulting a light-yellow transparent liquid.

4. Cells Fabrication

The coin cells were assembled in a glovebox with the concentration of water and oxygen kept lower than 0.5 ppm. The initial stage of assembly involves placing the top cell cap onto the worksurface, then inserting a Molybdenum Spacer of 0.127mm thickness, followed by the Aluminium Anode and then a Whatman Glass Fibre separator into said bottom cap. This separator acts as an insulating layer between the anode and cathode, preventing any potential short circuits. The second stage of assembly involves placing the bottom cell cap onto the workstation, then inserting a spring, followed by a Molybdenum spacer of 0.5mm and the cathode into said top cap. 150 μl electrolyte is pipetted onto the Whatman Glass Fibre Separator ensuring an even distribution and application when discharging the electrolyte from the pipette. Once this has been completed, the Cathode, 0.5mm Molybdenum Spacer, spring and cell top cap are inserted in that order. Finally, the cell was crimped in the crimping device with the top cap of facing upwards.

5. Testing of Cells

The fabricated coin cells were tested in a high-precision battery testing system with LAND battery testing software. The testing protocol entails conducting 200 galvanostatic charging-to-discharging cycles, applying a constant current-constant voltage with current density of 500 mA/g (mass of active material) and voltage ranges from 0.5V to 2.37V. The data processing program of the LAND battery test system displays different plots including voltage vs. specific capacity, voltage vs. cycle number, and voltage vs. efficiency, which are useful in evaluating individual cell performance.

III. Results and Discussion

1. Material Characterization

The morphologies of all six materials were shown in the SEM images (Figure 2). The SEM image of pure GNP (Figure 2a) revealed a fragmented and irregular structure, presented as thin platelets, while the graphite (Figure 2b) is thicker than GNP's and displayed a layered structure with stacked sheets. As the graphite was mixed with SC in a mass ratio of 3:1, the morphology of this mixture (Figure 2c)

became more cohesive with SC coated on the graphite particles. Increasing the proportion of SC in the graphite and soft carbon mixture led to increased particle size and a smoother surface. And for G25-SC75 (Figure 2e), it appears that the SC coated over most of the graphite particles. The SEM images revealed that during the pyrolysis of mesoporous pitch with graphite, the molten mesoporous pitch coated the graphite particles, and leading to a larger particle.

Referring to Xie et al. [29] investigating the hard-soft carbon composite anode in sodium-ion battery, N₂ adsorption, small-angle X-ray scattering (SAXS), as well as BET method were employed to assess the microstructures and porosity of carbon materials. Their research revealed a decreased BET surface area, and a steady decline in a parameter linked to the large pores surface area ratio in the N₂ adsorption isotherm upon the addition of SC to the carbon material. These results indicated that the introduction of SC led to the closure of the open pores and a reduced surface area, and consequently restricted the access to the pore surface [29, 30]. As aforementioned, the SEM image of the G-SC mixture exhibited a more cohesive structure and smoother surface compared to that of pure graphite, visually indicating the pores blocking effect of SC.

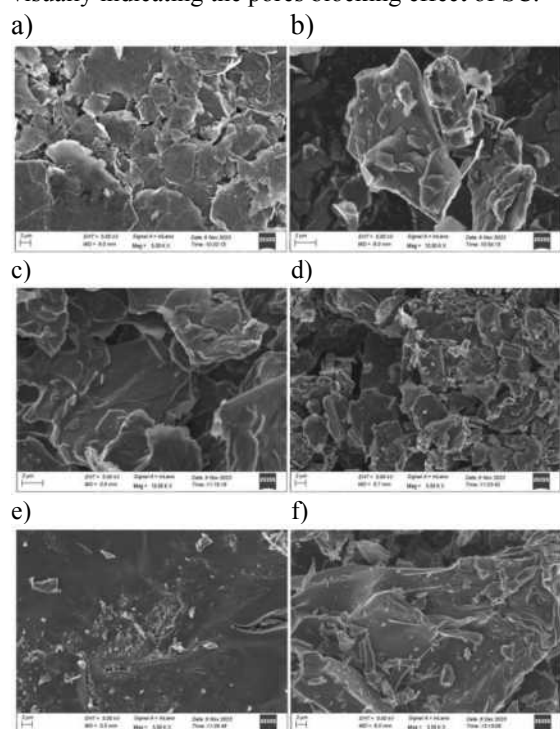


Figure 2: SEM images of pure GNP (a), pure graphite (b), G75-SC25 (c), G50-SC50 (d), G25-SC75 (e) and pure SC (f) active materials used in the research. (Image provided by Anastasia Teck)

SC was prepared through heat treatment of MP. It is noteworthy that varying carbonization temperatures led to different degree of graphitization, and thus, different SC morphologies. Changing the carbonization temperature would significantly affect

capacity of the cathode [30][31][32], and it directly related to the BET surface area, as stated by Li et al. [32]. Magda's research team [30] investigated the optimal carbonization temperature at 1100°C for mesoporous pitch to yield soft carbon as cathode material. Their findings, together with recent related studies, revealed a consistent trend: as the carbonization temperature rises, pore size reduces, which is advantageous in maintaining a high reversible capacity during cycling [31][32]. But excessively high temperatures lead to a huge decline in sloping capacity and the associated extra energy use and cost. Thus, the optimal temperature 1100°C, as proposed by Magda's research group, was obtained from the balance between cost-effectiveness and electrochemical performance.

2. Galvanostatic Performance

The comparison of electrochemical performance among cells with different cathode materials primarily focused on those featuring the average mass loading of $11.6 \times 10^{-3} \text{ mg/mm}^2$ and $18.9 \times 10^{-3} \text{ mg/mm}^2$ of active material on cathodes because of the limited valuable experimental data of the cells featuring an average mass loading of $4.8 \times 10^{-3} \text{ mg/mm}^2$ on cathode.

The galvanostatic charge-discharge (GCD) curves (Figure 3) showed steep slopes with a rapid rise of voltage from 0.6V to 1.8V, followed by a multi-plateau pattern with the working voltage platform ranging from 1.8 V to 2.3 V. These two regions correspond to two different mechanisms where the sloping region refers to fast and reversible faradaic charge-transfer reactions [18][19] occurring on the surface of the cathode material, resulting pseudocapacitance. The plateaus signalling the formation of stage n (n is the stage index and represents the number of graphene layers separating two intercalant layers) ionic graphite intercalation compounds (GICs) [28] and intercalant in staging mechanisms [23], where has been investigated and proved by XRD and Raman spectroscopy analysis in numerous studies.

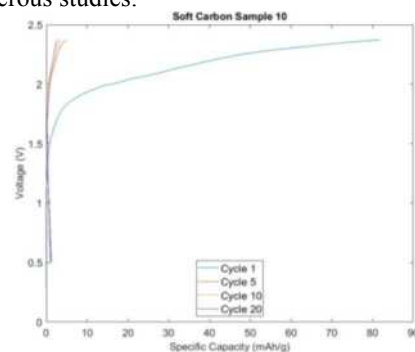


Figure 3: Plot of voltage-specific capacity for cycle 1, 5, 10 and 20 for SC-10.

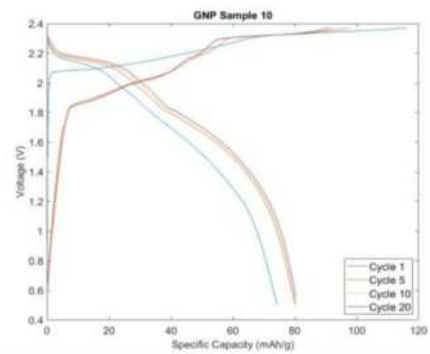
Figure 3 shows the GCD curves SC, while the GCD curve for GNP can be found in Figure 4a, and all GCD curves for cells tested in this research are available in the Supplementary Information section. In general, pure GNP, pure graphite, and the G75-

SC25 composite cathode displayed a similar shape in GCD curve, with a flatter platform ranging from 2V to 2.4V compare with G50-SC50, G25-SC75 and pure SC cathodes. While a steeper sloping region in a broader voltage range from 0.5V to 2.4V was observed for the composite cathode containing a higher ratio of SC (G50-SC50, G25-SC75 and SC), the slope was even steeper for pure SC cathode (Figure 3). This finding indicates that the pseudocapacitance significantly contributes to the capacity in SC, highlighting the dominance of surface reactions in its cycling process, while the capacity of graphite and GNP were mainly driven by ion diffusion.

A notable difference between charge and discharge capacities was observed during the first cycle, leading to a high irreversibility of capacity and thus, low initial coulombic efficiency (ICE) in the cells tested. Figure 4a displays an example of voltage profiles relative to the first, fifth, tenth, and twentieth cycles of the GNP cathode (sample GNP-10). The initial specific charge capacity of GNP was 118mAh/g with a specific discharge capacity of 80mAh/g and it led to a poor initial coulombic efficiency of 69% (can be seen directly from Figure 4b). This observed phenomenon was caused by the formation of the solid-electrolyte interface (SEI) [33] in AIBs, occurring as the ionic species decompose between the electrolyte and the cathode. The irreversible formation of SEI leads to a permanent consumption of a specific amount of electrolyte and anions [14], resulting in a low ICE. In addition to this reason, Elia et al. [2] proposed two additional potential causes: one involving the side interactions between the functional group or defects in the graphite and the intercalated anions, and another linked to the irreversible insertion due to structural changes in the cathodes during cycling.

Over the subsequent few cycles, the gap between charge and discharge capacities decreased, resulting in an improved efficiency. This is largely because, once formed, SEI could resist the interaction among the electrolyte, aluminium anions and cathode [17], thereby stopping further expansion of SEI and preventing further consumption of anions and electrolyte.

a)



b)

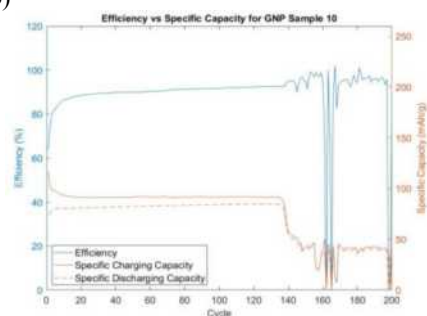


Figure 4: a): Plot of voltage-specific capacity for cycle 1, 5, 10 and 20 for GNP10.

b): Plot of efficiency-specific capacity over 200 cycles for GNP10.

When comparing the initial charge capacity to subsequent cycles, it was noticed that cells always have a higher initial charge capacity and exhibit a slight capacity fading in the following cycles. For instance, in the GNP cathode, the charge capacity declined gradually from 118mAh/g in cycle 1 to about 90mAh/g by cycle 20 ($\approx 98\text{mAh/g}$ at cycle 5 and $\approx 92\text{mAh/g}$ at cycle 10) while still maintaining a 90% CE after cycle 20. There are several possible reasons for the capacity fade, including the volume expansion in both cathode and electrolyte caused by the AlCl_4 anion intercalation, active material dissolution, formation of passivation film and electrolyte decomposition.

To compare the capacity fade across different materials, the capacity retention over 20 cycles was calculated by:

$$\text{Capacity retention} = \frac{\text{Charge Capacity at cycle 50}}{\text{Initial Charge Capacity}}$$

The SC cathode exhibited the lowest capacity retention, falling below 25% after 20 cycles, while the capacity retention data for graphite and G75-SC25 showed inconsistency but could reach over 90% and approximately 85% after 20 cycles, respectively. The data for the G50-SC50 and G25-SC75 composite cathodes was insufficient to state a reliable result, however, the existing data showed that their capacity retention was better than soft carbon but worse than graphite and G75-SC25. This outcome implies that a higher proportion of SC contributes to increased capacity fading. Detailed data can be found in Supplementary Information. Both Wang et al. [27] and Elia et al. [2] highlighted the sensitivity of ADIBs' performance to volumetric expansion, due to the large atomic radius of AlCl_4 anions at approximately 6.09 Å (compared to the 0.76 Å radius of Li ions) [27].

Therefore, the possible reasons for this finding might stem predominantly from the disordered structure of SC and increased number of functional groups in SC, causing more irreversible intercalation and multiple side interactions, subsequently causing volume expansion.

Experimental evidence can be obtained through diffraction techniques to validate this hypothesis. The irreversible insertion results in an increased

interlayer spacing, consequently leading to a decrease in the cathode porosity and the formation of an inactive area within the electrode. This ultimately causes a decrease in the capacity and energy density [13].

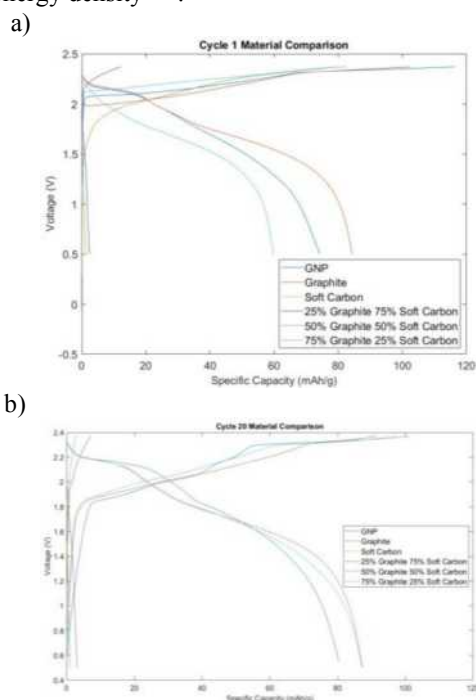


Figure 5: a): Voltage profile for GNP, graphite, MSC, G75-SC25, G50-SC50, G25-SC75 cathodes during the first cycle.

b): Voltage profile for GNP, graphite, MSC, G75-SC26, G50-SC50, G25-SC75 during twentieth cycle.

The outcomes highlight that ADIBs utilising a graphite cathode and G75-SC25 composite cathode displayed the best comprehensive performance among the tested cells when considering the reversible capacities, the ICE, the cycling stability and the cells failure rate across 15 samples. The detailed data comparison across all types of cathode materials can be found in the Supplementary Information. Figure 5a) and 5b) illustrates the voltage plotted against specific capacities of six graphitic materials at cycle 1 and cycle 20, measured at a current density of 500 mA/g ranging from 0.5V to 2.37V. During the first cycle, the GNP cathode revealed a longer platform with high initial capacity (118mAh/g). Despite this, its ICE (65%) was always lower than that of G75-SC25 (72%) and graphite (81%). The initial cycle of the G75-SC25 composite cathode demonstrated a charge capacity of 83mAh/g and a discharge capacity of 60mAh/g, which resulted in a relatively high ICE of 72%. By the twentieth cycle, it achieved a higher CE of 91%, along with a charge capacity of 97mAh/g and a discharge capacity of 88mAh/g. Even after 20 cycles, the charge capacity remained a high capacity of 88mAh/g. These results indicated the exceptional cycling stability of the G75-SC25 composite

cathode, highlighting its stable structure and high electrical conductivity in the GICs.

The relatively high ICE of G75-SC25 indicated less consumption of the electrolyte during cycling. The G75-SC25 characteristics of the blocked open pores and reduced surface area would explain this observation. As the molten mesoporous pitch fills the pore in graphite during pyrolysis, the access of electrolyte to the pore was obstructed, and subsequently preventing excessive electrolyte consumption, enhancing reversible capacity and ICE. In addition to the pore filling mechanism, Winter et al. [36] stated that composite carbon materials, consisting of graphite and disordered carbon (SC), capitalizes on the benefits of both materials to improve the overall performance and cycling stability. The graphite facilitates anion storage, while the disordered carbon (SC) improves long-term stability.

A pattern has emerged where increasing dominance of SC in the cathode active material leads to a worse electrochemical performance. In the case of the pure SC cathode, while it suffered from capacity fading likely due to structural degradation as previously discussed, it also exhibited a much lower capacity and an extremely low ICE (below 10%) throughout the experiments when compared to the other five cathode materials. This observation agrees with what Xie et al. [29] documented: though the addition of a certain amount of SC could prevent further growth of SEI, a higher SC fraction may lead to a reduced reversible capacity due to limited effectiveness in blocking open pores. Instead, the nature of high reactivity and the tendency towards irreversible intercalation and structural degradation during SC charge-discharge cycling would contribute as more prominent factors negatively impacting the cathode's capacity [29]. Additionally, a higher rate of cell failure was noted in cells containing a greater SC content. Specifically, two G50-SC50 samples operated normally in 15 samples and only one G75-SC25 sample underwent successful cycling. The increased failure rate of cathode samples with higher SC proportions suggested a potential increase in unwanted side reactions within the cells. This includes electrode corrosion, while more material dissolution was also likely to have occurred during repeated charge-discharge cycles due to larger particle sizes of SC.

3. Impact of Active Material Mass Loading

Only a limited number of cells completed charging and discharging cycles successfully using the smallest amount of active material mass loading. However, the data from these cells was not included in the comparison of electrochemical performance because insufficient data would not provide reliable conclusions. In addition, batteries with low active material loading might be more susceptible to other factors like human error, electrode detachment, and temperature fluctuations during operations. This can

significantly influence their performance and skew the results. Therefore, the comparison predominantly focused on the cells with 31.6 μ l and 47.4 μ l active material slurries coated on cathodes. Table 1 shows the battery performance variations with distinct mass loadings of graphite, SC, and G75-SC25. However, due to fewer than 5 samples of GNP, G50-SC50 and G25-SC75 being operational, reliable results could not be obtained. Nonetheless, their data are included in the supplementary Excel spreadsheet which illustrates the comparative electrochemical performance of carbon-based cathodes using different qualities of active material slurries.

Sample	Mass Loading	Initial SpeCap (mAh/g)	ICE	SpeCap-20 (mAh/g)	CE-20
G-6	2.55 mg	87	80	68	71
G-8	2.34 mg	400	13	65	75
G-11	3.99 mg	72	86	71	91
G-13	4.31 mg	264	20	18	72
G-15	3.11 mg	23	39	51	82

Table 1: Graphite samples. (Note: SpeCap-20 and CE-20 denotes the specific charge capacity and CE at cycle 20)

Sample	Mass Loading	Initial SpeCap (mAh/g)	ICE	SpeCap-20 (mAh/g)	CE-20
G75-SC25-6	0.90 mg	5672	0.6	163	40
G75-SC25-7	0.85 mg	86	65	73	78
G75-SC25-8	0.71 mg	83	72	97.5	90
G75-SC25-9	1.34 mg	106	26	49.5	53
G75-SC25-11	1.73 mg	389	0.1	48.5	66
G75-SC25-12	2.80 mg	93	40	45.5	73
G75-SC25-13	1.95 mg	93.5	35	50	74
G75-SC25-15	1.08 mg	104	72	122	87

Table 2: G75-SC25 samples.

Sample	Mass Loading	Initial SpeCap (mAh/g)	ICE	SpeCap-20 (mAh/g)	CE-20
SC-6	1.99 mg	26	11.5	6	50
SC-10	1.63 mg	82.5	3.6	3	36.7
SC-13	2.5 mg	336.5	0.9	14	28.6

Table 3: SC samples

The data for graphite cathodes showed that the higher quantity active material loading displayed an increased ICE and better capacity retention but started with lower initial capacities, while lower active material loading began with higher capacities but exhibited noticeable degradation of capacity upon charge-discharge cycling. In the comparison between graphite cathode samples 6 and 11, coated with 31.6 μ l and 47.4 μ l of graphite slurry respectively, sample 6 demonstrated higher initial reversible capacity (87mAh/g) compared to sample 11 (72mAh/g). However, although sample 6 demonstrated a higher initial capacity, it showed a lower ICE at 80% compared to sample 11 at 86%. Sample 6 also exhibited poorer cycling stability and capacity retention, indicated by a decline from 87mAh/g in the first cycle to 68mAh/g after 20 cycles, along with a CE of 71% in the twentieth cycle, while sample 11 maintained a capacity of

71mAh/g with a CE of 91% during the twentieth cycle.

This observation aligns with the findings proposed by Angelopoulou et al. [10], which investigated the impact of electrode loading on lithium-ion battery electrochemical performance. Currently, we are unable to further investigate and prove the cause behind this observed phenomenon in our research. However, based on the previous published literature, the reduced capacity in lower active material mass loading might be connected to the side reaction, especially corrosion, of the current collector. As mentioned previously, Molybdenum was used as current collector that the cathode active material slurry was coated on, and ideally it should maintain good chemical and electrochemical stability while performing strong adhesion to the active material slurry. As documented [16][15][11], the current collector corrosion would lead to reduced contact and increased internal impedance, while the product from corrosion potentially reduces the ion electronic conductivity, and hence decreasing cell capacity. With a thinner layer of active materials slurry deposited on current collector, the passivation film, which acts as the barrier to prevent further reaction and corrosion of the current collector, would be thinner. Consequently, it might be easier for electrolyte ions to penetrate the passivating layer and react with the current collector.

Interestingly, the G75-SC25 cathode demonstrated an opposite trend, where the larger quantity of active material negatively affected the cells' CE and capacity retention. For instance, in the case of G75-SC25 cathode sample 8 (coated with 31.6 μ l slurry), an initial reversible capacity of 83mAh/g was achieved alongside an ICE of 72%. It then displayed a charge capacity of 97.5mAh/g and a CE of 90% by the twentieth cycle. Sample 13 (coated with 47.4 μ l slurry) showcased an initial charge capacity of 93.5mAh/g but presented an ICE of 35% initially. By the twentieth cycle, it exhibited a charge capacity of 50mAh/g and a discharge capacity of 37mAh/g, resulting in a lower CE of 74%. This trend was similarly observed in the SC cathode.

A study conducted operando 3D observations to investigate the evolution of electrochemical reactions during charge and discharge [21]. This research suggests that higher active material mass loading led to slower ion diffusion and more inactive area within aggregated active material areas. While lower mass loading reduces the number of high-resistance interfaces for ions between active materials, and thus decreasing active material aggregation and benefiting ion diffusion [21].

Despite the lack of studies and a relevant literature explaining the opposite trend observed between G75-SC25, SC and the graphite, a possible reason would be related to the morphologies of these materials. Higher active material mass loading and thicker electrode leads to the reduction of ion

diffusion, where this reduction could affect the SC and SC composite cathode more significantly due to their smaller surface areas and less accessibility compared to graphite. Given their limited ion diffusion pathways, any further decrease might apply a greater impact on these materials compared to graphite, which naturally have more diffusion pathways.

4. Discussion about Cell Failure

The data obtained demonstrates the substantial impact of assembly parameters and reveals that active materials containing a higher fraction of SC exhibited a greater occurrence of cell failure when compared to those composed of GNP, graphite and G75-SC25 cathodes. Only one G25-SC75 cell was operational out of 15 prepared samples, while the failure rate of G75-SC25 samples were the lowest, with 8 successful cells out of 15 samples in total. Alongside factors such as the actual quantity of electrolyte and electrode positioning, precise cell stack height emerges as a critical variable in determining the cells' behaviour. The outcomes might be biased if the cells failed to meet high quality and reproducibility criteria.

Some coin cells investigated in this paper failed due to the accidental misaligning of components during cell assembly. It is therefore crucial to ensure that the placement of these components is central to prevent accidental short-circuiting. Additionally, improper crimping of cells causes both component misalignment and electrolyte leakage or exposure which not only compromises the cell's performance, but it also raises safety concerns due to the highly corrosive nature of the electrolyte. Other cells failed due to the characteristics of the active materials at the time of assembly. For example, electrode contamination due to the accidental mishandling of components could significantly impact the overall efficiency and performance of the cells. Additionally, uneven electrolyte distribution on the separator or a slight local variation in electrolyte quantity from the pipette can disrupt the flow of ions which impacts the cell's capacity. Therefore, addressing and minimizing these potential human errors during cell fabrication is crucial to guarantee the reliability and the electrochemical performance of cells across various experiments.

A likely cause of cell failure outside of human error could be due to the presence of side reactions. In a different study [8] which looked at various lithium, sodium, potassium and aluminium ion batteries, it is revealed that side reactions within graphitic electrodes result from surface groups and anion intercalation when at high voltage. These side reactions cause a reduction in the electrochemical performance of cells because of electrolyte decomposition within the carbon and a deterioration of the graphitic structure. Additionally, the side reaction might increase the contact resistance on the current collector, lead to adhesion loss of active

materials, and potentially result in short circuit when the corrosion product penetrates through the separator [16]. This was frequently observed in the cells fabricated during our experiment. A study [8] also mentions the impact of viscosity on the performance of the electrolyte. Ionic liquid electrolytes are known to have a higher viscosity than other electrolytes which results in a lower ionic conductivity and a lower wettability of the separator, anode and cathode materials.

IV. Conclusions and Outlooks

1. Conclusions

This research investigated six variations of graphitic materials employed in cathodes, aiming to compare their electrochemical performances and identify the most effective cathode material for ADIBs. Among the six tested cathode materials, the best-performing cathode material was determined to be the pure graphite and the mixture of 75%wt graphite with 25%wt SC cathode. As the proportion of soft carbon increased to 50%wt, the ADIBs exhibited a reduced charge/discharge capacity and CE. Further decreases in both charge storage capacity and CE were observed as the soft carbon composition reached 75%wt, showing a similar electrochemical behaviour to pure soft carbon. There was also an impact of mass loading on cell performance. For example, with graphite and soft carbon they exhibited a larger initial coulombic efficiency, retained their capacity better and had a lower initial capacity while active material content increased. Conversely, 75wt% graphite with 25wt% soft carbon exhibited a worse initial coulombic capacity, a worse capacity retention and a higher initial capacity.

2. Outlooks

It is worth mentioning that this work only assesses ADIB cells using the capabilities of Land testing system when charging and discharging them. A tried and tested way to further assess ADIB performance would be to utilise BET analysis.[22] This would allow for the generation of quantitative data on the surface area and porosity distribution of cathode materials. BET techniques have been previously utilised in the field of Aluminium Ion batteries such as in Lin et al. [25] and Huang et al. [26]. Such BET data would reveal the differences in surface area for each graphitic material and could help explain their differing levels of performance.

The non-destructive and highly versatile technique of XRD was not used in this paper and has a strong potential for further exploration [35]. It has been used to study chloroaluminate anion-graphite intercalation in aluminium batteries just like the ADIBs studied here. For example, in Pan et al. [17] XRD revealed a surprisingly ordered anion intercalation staging behaviour in graphite despite the large anion size and stable graphite structure

during intercalation and deintercalation. While graphite was analysed in the work by Pan et. al. [17], the other five materials covered in this paper were not and so this opens the window for further analysis.

Raman spectroscopy has been recently employed to assess the performance of ADIBs. Liu et al. [24] applied Raman spectroscopy to characterise graphite flakes (u-GF) on carbon fibre cloth (CFC) under different durations of ultrasonication before being used as a cathode material in ADIBs. This concluded that a graphite intercalation/de-intercalation behaviour of the chloroaluminate ions into the u-GF took place from the analysis of the Al/u-GF at CFC battery's Raman spectra. Such characterisation and analysis using Raman spectroscopy could be applied to the graphitic materials investigated in this paper and potentially lead to further conclusions.

V. References

- [1]: Matuszak, J. (2022). *The Importance of Batteries in Renewable Energy Transition*. [online] KnowHow. Available at: <https://knowhow.distrelec.com/energy-and-power/the-importance-of-batteries-in-renewable-energy-transition/> [Accessed 25 Oct. 2021]
- [2]: Elia, G.A., Kravchyk, K.V., Kovalenko, M.V., Chacón, J., Holland, A. and Wills, R.G.A. (2021). *An overview and prospective on Al and Al-ion battery technologies*. *Journal of Power Sources*, [Online] 481, p.228870. Available at: <https://www.sciencedirect.com/science/article/pii/S0378775320311745?via%3Dihub> [Accessed 11 Dec. 2021]
- [3]: Clean Energy Institute (2020). *Lithium-Ion Battery*. [online] Clean Energy Institute. Available at: <https://www.cei.washington.edu/research/energy-storage/lithium-ion-battery/> [Accessed 11 Dec. 2023].
- [4]: Larcher, D. and Tarascon, J.-M. (2014). *Towards Greener and More Sustainable Batteries for Electrical Energy Storage*. *Nature Chemistry*, [online] 7(1), pp.19–29. Available at: <https://www.nature.com/articles/nchem.2085> [Accessed 11 Dec. 2023].
- [5]: Das, S.K., Mahapatra, S. and Lahan, H. (2017). *Aluminium-ion batteries: Developments and Challenges*. *Journal of Materials Chemistry A*, [online] 5(14), pp.6347–6367. Available at: <https://pubs.rsc.org/en/content/articlelanding/2017/ta/c7ta00228a> [Accessed 11 Dec. 2023].
- [6]: Leisegang, T., Meutzner, F., Zschornak, M., Münchgesang, W., Schmid, R., Nestler, T., Eremin, R.A., Kabanov, A.A., Blatov, V.A. and Meyer, D.C. (2019). *The Aluminium-Ion Battery: a Sustainable and Seminal Concept?* *Frontiers in Chemistry*, [online] 7. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6504778/> [Accessed 13 Dec. 2023].
- [7]: Chen, C.-Y., Tsuda, T., Kuwabata, S. and Hussey, C.L. (2018). *Rechargeable Aluminum Batteries Utilizing a Chloroaluminate Inorganic Ionic Liquid Electrolyte*. *Chemical Communications*, [online] 54(33), pp.4164–4167. Available at: <https://pubs.rsc.org/en/content/articlelanding/2018/cc/c8cc00113h> [Accessed 11 Dec. 2023].
- [8]: Xu, J., Dou, Y., Wei, Z., Ma, J., Deng, Y., Li, Y., Liu, H. and Dou, S. (2017). *Recent Progress in Graphite Intercalation Compounds for Rechargeable Metal (Li, Na, K, Al) Ion Batteries*. *Advanced Science*, [online] 4(10), p.1700146. Available at: <https://onlinelibrary.wiley.com/doi/epdf/10.1002/adv.201700146> [Accessed 11 Dec. 2023].
- [9]: Zhang, E., Cao, W., Wang, B., Yu, X., Wang, L., Xu, Z. and Lu, B. (2018). *A novel aluminium dual-ion battery*. *Energy Storage Materials*, [online] 11, pp.91–99. Available at: https://www.sciencedirect.com/science/article/abs/pii/S2405829717304683?fr=RR-2&ref=pdf_download&rr=8251214db8cad31 [Accessed 25 Oct. 2021].
- [10]: Angelopoulou, P. and Avgouropoulos, G. (2019). *Effect of Electrode Loading on the Electrochemical Performance of LiAlO₂·1Mn₂O₄ Cathode for lithium-ion Batteries*. *Materials Research Bulletin*, [online] 119, p.110562. Available at: <https://www.sciencedirect.com/science/article/pii/S0025540819309626> [Accessed 11 Dec. 2023].
- [11]: Unal, B., Ozlem Sel and Rezan Demir-Çakan (2023). *Current Collectors Corrosion Behaviours and Rechargeability of TiO₂ in Aqueous Electrolyte Aluminium-ion Batteries*. *Research Square*. [online] Available at: <https://doi.org/10.21203/rs.3.rs-3154331/v1> [Accessed 12 Dec. 2023].
- [12]: iolitec.de. (n.d.). *1-Ethyl-3-methylimidazolium Chloride and Aluminum Chloride 1:1.5, >98% | IoLiTec*. [online] Available at: <https://iolitec.de/en/node/276> [Accessed 11 Dec. 2023].
- [13]: Greco, G., Tatchev, D., Hoell, A., Krumrey, M., Raoux, S., Hahn, R. and Elia, G.A. (2018). *Influence of the Electrode nano/microstructure on the Electrochemical Properties of Graphite in Aluminium Batteries*. *Journal of Materials Chemistry A*, [online] 6(45), pp.22673–22680. Available at: <https://pubs.rsc.org/en/content/articlelanding/2018/TA/C8TA08319C> [Accessed 11 Dec. 2023].
- [14]: Momidi, K. (2019). *Understanding Solid Electrolyte Interface (SEI) to Improve Lithium-Ion Battery Performance*. [online] circuitdigest.com. Available at: <https://circuitdigest.com/article/what-is-solid-electrolyte-interface-sei-to-improve-lithium-ion-battery-performance> [Accessed 11 Dec. 2023].
- [15]: Arora, P., White, R. and Doyle, M. (1998). *Capacity Fade Mechanisms and Side Reactions in Lithium-Ion Batteries*. *Journal of the Electrochemical Society*, [online] 145(10), pp.3647–3667. Available at: https://scholarcommons.sc.edu/eche_facpub/379/ [Accessed 12 Dec. 2023].
- [16]: Guo, L., Thornton, D.B., Koronfel, M.A., Stephens, I.E.L. and Ryan, M.P. (2021a). *Degradation in Lithium-Ion Battery Current Collectors*. *Journal of Physics: Energy*, [online] 3(3), p.032015. Available at: <https://iopscience.iop.org/article/10.1088/2515-7655/ac0c04#jpenergvac0c04s1> [Accessed 12 Dec. 2023].
- [17]: Pan, C., Yuan, C., Zhu, G., Zhang, Q., Huang, C., Lin, M., Angell, M., Hwang, B. and Payam Kaghazchi (2018). *An Operando X-ray Diffraction Study of Chloroaluminate anion-graphite Intercalation in Aluminium Batteries*. *Proceedings of the National*

- Academy of Sciences of the United States of America, [online] 115(22), pp.5670–5675. Available at: <https://www.pnas.org/doi/epdf/10.1073/pnas.1803576115> [Accessed 11 Dec. 2023].
- [18]: Zhang, W., Huang, R., Xu, Y., Tian, C., Xiao, Y., Lin, Z., Dai, L., Guo, Z. and Chai, L. (2023). Carbon Electrode Materials for Advanced Potassium-Ion Storage. *Angewandte Chemie International Edition*, [online] 62(43). Available at: <https://onlinelibrary.wiley.com/doi/10.1002/anie.202308891> [Accessed 11 Dec. 2023].
- [19]: Kurzweil, P. (2009). BATTERIES | Nomenclature. [online] ScienceDirect. Available at: <https://www.sciencedirect.com/topics/engineering/pseudocapacitance> [Accessed 11 Dec. 2023].
- [20]: Shen, Y., Zhang, M., Yan, D., Jie Lv, Wu, T., He, B. and Li, W. (2023). Soft Carbon as Cathode with High-Rate Performance for Dual-Ion Batteries via Fast PF6–Intercalation Improved by Surface Effect. *ChemSusChem*, [online] 16(17). Available at: <https://chemistry-europe.onlinelibrary.wiley.com/doi/full/10.1002/cssc.202300493> [Accessed 11 Dec. 2023].
- [21]: Kimura, Y., Mahunnop Fakkao, Nakamura, T., Okumura, T., Ishiguro, N., Oki Sekizawa, Nitta, K., Tomoya Uruga, Tada, M., Yoshiharu Uchimoto and Koji Amezawa (2020). Influence of Active Material Loading on Electrochemical Reactions in Composite Solid-State Battery Electrodes Revealed by Operando 3D CT-XANES Imaging. *ACS Applied Energy Materials*, [online] 3(8), pp.7782–7793. Available at: <https://pubs.acs.org/doi/full/10.1021/acsaem.0c01186> [Accessed 11 Dec. 2023].
- [22]: Measurlabs (n.d.). Brunauer-Emmett-Teller (BET) Analysis | Measurlabs. [online] measurlabs.com. Available at: <https://measurlabs.com/methods/brunauer-emmett-teller-bet-analysis/> [Accessed 11 Dec. 2023].
- [23]: Li, Y., Lu, Y., Adelhelm, P., Titirici, M.-M. and Hu, Y.-S. (2019). Intercalation chemistry of graphite: alkali metal ions and beyond. *Chemical Society Reviews*, [online] 48(17), pp.4655–4687. Available at: <https://pubs.rsc.org/en/content/articlelanding/2019/cs/c9cs00162j#> [Accessed 11 Dec. 2023].
- [24]: Liu, C., Liu, Z., Niu, H., Wang, C., Wang, Z., Gao, B., Liu, J. and Taylor, M. (2019). Preparation and in-situ Raman Characterization of binder-free u-GF@CFC Cathode for Rechargeable aluminium-ion Battery. *MethodsX*, [online] 6, pp.2374–2383. Available at: <https://www.sciencedirect.com/science/article/pii/S2215016119302730> [Accessed 11 Dec. 2023].
- [25]: Lin, Z., Mao, M., Yang, C., Tong, Y., Li, Q., Yue, J., Yang, G., Zhang, Q., Luan, H., Yu, X., Gu, L., Hu, Y., Li, H., Huang, X., Suo, L. and Chen, L. (2021). Amorphous anion-rich Titanium Polysulfides for aluminium-ion Batteries. *Science Advances*, [online] 7(35). Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8386935/> [Accessed 11 Dec. 2023].
- [26]: Huang, Z., Du, X., Ma, M., Wang, S., Xie, Y., Meng, Y., You, W. and Xiong, L. (2023). Organic Cathode Materials for Rechargeable Aluminium-Ion Batteries. *ChemSusChem*. [online] Available at: <https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/cssc.202202358> [Accessed 11 Dec. 2023].
- [27]: Wang, S., Kravchyk, K.V., Krumeich, F. and Kovalenko, M.V. (2017b). Kish Graphite Flakes as a Cathode Material for an Aluminium Chloride–Graphite Battery. *ACS Applied Materials & Interfaces*, [online] 9(34), pp.28478–28485. Available at: <https://pubs.acs.org/doi/10.1021/acsami.7b07499> [Accessed 11 Dec. 2023].
- [28]: Ng, K.L., Dong, T., Anawati, J. and Azimi, G. (2020). High-Performance Aluminium Ion Battery Using Cost-Effective AlCl₃ - Trimethylamine Hydrochloride Ionic Liquid Electrolyte. *Advanced Sustainable Systems*, [online] 4(8), p.2000074. Available at: <https://onlinelibrary.wiley.com/doi/full/10.1002/adss.202000074> [Accessed 11 Dec. 2023].
- [29]: Xie, F., Xu, Z., Jensen, A.C.S., Au, H., Lu, Y., Araullo-Peters, V., Drew, A.J., Hu, Y. and Titirici, M. (2019). Hard–Soft Carbon Composite Anodes with Synergistic Sodium Storage Performance. *Advanced Functional Materials*, [online] 29(24), p.1901072. Available at: <https://onlinelibrary.wiley.com/doi/10.1002/adfm.201901072> [Accessed 11 Dec. 2023].
- [30]: Zhang, S., Teck, A.A., Guo, Z., Xu, Z. and Titirici, M. (2021). Carbon Composite Anodes with Tunable Microstructures for Potassium-Ion Batteries. *Batteries & Supercaps*, [online] 4(4), pp.663–670. Available at: <https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/batt.202000306> [Accessed 11 Dec. 2023].
- [31]: Qi, Y., Lu, Y., Ding, F., Zhang, Q., Li, H., Huang, X., Chen, L. and Hu, Y.-S. (2019). Slope-Dominated Carbon Anode with High Specific Capacity and Superior Rate Capability for High Safety Na-Ion Batteries. *Angewandte Chemie (International Edition)*, [online] 58(13), pp.4361–4365. Available at: <https://onlinelibrary.wiley.com/doi/10.1002/anie.201900005> [Accessed 11 Dec. 2023].
- [32]: Li, Y., Hu, Y.-S., Titirici, M.-M., Chen, L. and Huang, X. (2016). Hard Carbon Microtubes Made from Renewable Cotton as High-Performance Anode Material for Sodium-Ion Batteries. *Advanced Energy Materials*, [online] 6(18), p.1600659. Available at: <https://onlinelibrary.wiley.com/doi/10.1002/aenm.201600659> [Accessed 11 Dec. 2023].
- [33]: Nicolò Canever, Hughson, F.R. and Nann, T. (2020). Solid-Electrolyte Interphases (SEI) in Nonaqueous Aluminium-Ion Batteries. *ACS applied energy materials*, [online] 3(4), pp.3673–3683. Available at: <https://pubs.acs.org/doi/10.1021/acsaem.0c00132> [Accessed 11 Dec. 2023].
- [36]: Winter, M., Besenhard, J.O., Spahr, M.E. and Novák, P. (1998). Insertion Electrode Materials for Rechargeable Lithium Batteries. *Advanced Materials*, [online] 10(10), pp.725–763. Available at: <https://onlinelibrary.wiley.com/doi/epdf/10.1002/%28SICI%291521-4095%28199807%2910%3A10%3C725%3A%3AAID-ADMA725%3E3.0.CO%3B2-Z> [Accessed 11 Dec. 2023].

CFD Modelling of Air Entrainment Mechanisms in a Plunging Jet

Samreen Malik and Yiyao Qiu

Department of Chemical Engineering, Imperial College London, U.K.

Abstract

This study analysed complex dynamics on the process of plunging liquid jet through the computational fluid dynamics (CFD) modelling under the regime of viscous laminar fluid. Different scenario simulations with varying plunging speed and fluid viscosity were created. Numerical modelling was adopted in a hybrid pattern that combines interface-tracking and level-set method in a three-dimensional Cartesian domain. Results were primarily evaluated in aspects of jet evolution, incipient dynamics, and deformation scale, with prediction on further possible plunging behaviour. Jets tend to develop a bulge-like structure with different extent of development and high dependency on falling time. In the scenario of longer falling time, jet breaks up into droplets, indicating an inception mechanism through dripping rather than straight jetting. The initial cavity of the free surface deformation is the proximate cause of the initial bubble formation, whereas higher viscosity would lead to a flatter deformation thus flatter bubble geometry. A growth in deformation size was observed both in terms of width and depth, where the convergency in depth implying a steady plunging state through the impending process. In-depth computational research considering finer computational resolution and capacity, bubble dynamics control, turbulence and factors of disturbance are recommended for future applicable simulations.

Keywords: *Air entrainment; Plunging jet; Air bubble; Free surface deformation; CFD*

1 Introduction

The phenomenon of air entrainment is well-observed in various natural processes, as well as numerous manufacturing and industrial applications. Air entrainment is a mechanism of how insoluble gas is trapped and dispersed when a liquid stream injects to a pool with the same liquid at a certain velocity. This serves as a mean of pressure relief reservoirs in order to prevent ruptures in concrete structures (Panarese and William, 1963). Similar behaviour may be observed in daily and manufacturing operation of water filtrating and filling process, as well as medical blood transfusion process. The air entrainment also takes a crucial role in global climate evolution by facilitating the transportation of oxygen and carbon dioxide through bubbles at the free surface of oceans and rivers. Nevertheless, the understanding of the entrainment mechanism is not yet sufficiently developed for widespread commercial application in industry.

An effective method for simulating the air entrainment mechanism involves utilizing the plunging jet scheme, in which a jet is introduced into the stationary liquid bath at a specific plunging velocity. Extensive research has been conducted on various detailed aspects on plunging jet either experimentally or computationally. Biń 1993, performed the initial comprehensive analysis of the mechanism of gas entrainment and bubble dispersion based on earlier investigations. This work established a robust groundwork for subsequent research. Vast majority of articles focused on studying the mechanisms experimentally, using a similar experimental setup but different jet or fluid parameters. These papers particularly capture the

impinging and bubble dynamics occurring beyond the free surface, and discuss the underlying correlation between different parameters to the entrainment dynamics (Qu et al. 2013; Zhu et al. 2000). Several research categorized the fluid regime of the gas entrainment based on the gas concentration (Biń, 1993; Chirichella, et al., 2002). They also examined the inception conditions under a range of Reynolds and Weber numbers. Various computational fluid dynamics simulations added the credit of aforementioned experimental study by verified results in simulation models. These lead to further development in the quantification on amount and size distribution of the bubbles and the influence of different physical parameters on the aeration behaviour (Salehi et al., 2022; Lopes, 2016; Kendil et al., 2012).

While the present research have been relatively thorough, the majority of studies focused on turbulent fluid regime. There has been limited research conducted on laminar viscous fluid regime, despite the fact that many fluids in industries and nature are significantly more viscous than water. The fact that bubble related analysis is essential is recognized but requires a significant amount of time for simulation running and costly computational resources. Therefore, this project concentrates on the incipient dynamics of the plunging jet and the topology development of deformation with time. Simulations were conducted using different impinging velocities on two distinct viscosities, while keeping the jet configuration uniform.

2 Background

2.1 Entrainment Inception Conditions

Research findings indicate that air entrainment and subsequently bubble formation only occur when the liquid jet impacting the free surface exceeds a critical velocity (El Hammoumi, Achard, Davoust, 2002). The critical entrainment conditions are normally discussed in terms of the dimensionless numbers of Reynolds and Froude number. Zhu et al., 2000's work emphasised the necessity of analysing whether the jet has been built to minimise turbulence, taking into account the consistency of threshold values across different experiments. For the studies that are unable to obtain entrainment results even when critical parameters have been exceeded, it has been recommended by Zhu et al., 2000 to introduce manual disturbances to get further bubble related progress.

2.2 Effect of Jet Configurations

Extensive research has been conducted on the impact the jet configuration has on entrainment behaviour. The studies demonstrate that the increase in jet's diameter leads to an augmentation in the rate of oxygen transfer, the amount of air entrapped, and the depth of penetration (Kumar and Tiwari, 2021). In addition, the likelihood of air entrainment is greater when the jet has a larger diameter, as this increases the extent of contact between the jet and the ambient air (Deswal and Verma, 2007a)

According to Ahmed's study in 1974, enlarging the length of the jet will result in a greater amount of air being entrained. The study also determined that a critical length of 0.66m was identified, beyond which no additional air entrainment occurs for cylindrical jets (Biń, 1993). McKeogh and Ervine 1981 validated this statement and further noted that when the values are higher, the jet undergoes disintegration and ceases to enhance the air entrainment. While the length of jet has less influence relative to its velocity and diameter, it has a greater impact on the mass transfer of smooth jets (Donk, 1981).

The investigation of jet inclination angles was conducted by several researchers, they experimented with various angles while maintaining a constant length for the jet (Biń, 1993). Van de Donk 1981 and Ahmed 1974 discovered that inclination has a minimal impact on the pace at which oxygen is transferred (Biń, 1993). Nevertheless, as stated by Toko et al., 1982, the angle of the jet, ranging from 45 to 90 degrees, exerts a substantial impact on the oxygen transfer rate.

The ratio of the nozzle's length to diameter is another important factor in designing the jet configuration. A low ratio number around 6.23 is preferred according to the study of Ohkawa et al., 1986. Moreover, a circular truncated nozzle is

preferred in optimizing the air entrainment rate since it normally leads to higher bubble penetration depth (Bagatur, 2014)

2.3 Falling jet evolution

The study of Eggers and Villermaux 2008, gave detailed numerical analysis into the liquid jet, which discussed jet evolution into either dripping or jetting as a result of perturbation applied at the nozzle. This kind of perturbation impacts the jet in a directly proportional manner to the jet velocity, making the jet convectively unstable. They commented the dripping scenario as an absolute instability regime, whereas jetting scenario as velocity increases is described as a transit regime. The boundary between these transitable regimes can be characterised by the dimensionless critical Weber number as well as a corresponding criterion equation provided by the research from Clanet and Lasheras 1999.

2.4 Initial Bubble Formation

Investigating the formation of the initial bubble during the plunging process is crucial for comprehending the mechanism of air entrapment. In the early stages of research in this field, Erine et al., 1988 studied the entrained air bubble formation under the assumption of dependence on the enclosure between distorted free surface. Later analysis of Rein's 1998 work revealed that the process of air entrainment is not only determined by the impact of droplets (Han and Ease, 2018). In addition to that a recent study conducted by Wei et al. in 2016 examined the curvature radius of the deformation on the free surface. This study highlighted that the distortion of the free surface is a rapid phenomenon that can be defined by a critical state in shape. As a result, Wei et al. 2019 conducted a further investigation, where a high-speed camera was utilized to capture the quick changes that occurred. The study examined the deformation in the curvature radius with time, using a power-law scaling, and investigated the connection between bubble size and the width of the deformed surface.

2.5 Fluid Regime

The understanding of the air entrainment mechanism typically relied on the categorization of fluid regime. Initial classification, put forward by Laura 1979, suggested that vertical plunging jets can be categorized into two regions based on whether the jet splits into droplets or is being continuous. The boundary between these two regions is determined by jet length and velocity. The fluid regimes were specified as more studies were carried out, leading to an air concentration-based classification. The boundaries between these regimes are dependent on both dimensionless Froude number and the velocity ratio between jet impact and horizontal movement (Chirichella, et al., 2002).

2.6 Computational Fluid Dynamics

Numerous computational fluid dynamics (CFD) simulations were investigated both to verify previous experimental studies and to understand the phenomenon from a computational point of view. Kendil et al., 2012 simulated the jet as a two-phase bubbly flow injecting into the pool, and indicated a limited influence of the lift force on the velocity field but the plume shape would be affected easily by bubble size and void fraction. Different mathematical models were adopted across studies, where the volume of fluid (VOF) method was applied the most due to its reliable accuracy and reasonable computational intensity (Lopes et al., 2016). A hybrid numerical solver was developed and adopted by Xiang et al., 2014, the model takes the advantage of selective interface sharpening to predict the occurrence and spread of bubbles more accurately. A recent study of Salehi et al., 2022 focused on a more applicational problem of how entrainment mechanism is involved in the overflowing problem in hydraulic break columns (HBC). They had the models developed and validated by comparing liquid jet penetration snapshots with experimental figures given in Qu et al., 2013. Results were discussed in a classification of flow patterns in terms of dispersed bubbles, medium air pockets and large air pockets. The bubble case was ideally operational for HBC, where bubbles were able to break into smaller sizes and led to a predictable linear rising pattern. The later two need to be prevented due to the high potential risk of overflow from larger size and amount of air pocket entrained.

2.7 Reviews

In addition to the previously discussed Bin 1993, there are a couple of more noteworthy reviews available from Kiger 2012, and Kumar 2021. Kiger, 2012's review discussed the mechanisms of air entrainment in the category of low and high fluid viscosity, as well as the complicated application of plunging breaking waves. In contrast, Kumar, 2021 relied more on the oxygen mass transfer phenomenon associated with plunging jet.

3 Methods

3.1 Model formulation

The numerical model was built based on the concept of solving Navier-Stokes equations for an incompressible two-phase system in Cartesian coordinates $x = (x, y, z)$. The system can be expressed in a single field formulation with the governing equations of:

$$\nabla \cdot \mathbf{u} = 0 \quad (1)$$

$$\rho \left(\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} \right) = -\nabla P + \rho \mathbf{g} + \nabla \mu (\nabla \mathbf{u} + \nabla \mathbf{u}^T) + \mathbf{F} \quad (2)$$

where \mathbf{u} is the fluid velocity, t the time, P the pressure, \mathbf{g} the gravitational acceleration. The density and viscosity can be further expressed by the formulation of

$$\rho(x, t) = \rho_a + (\rho_l - \rho_a) \mathcal{H}(x, t) \quad (3)$$

$$\mu(x, t) = \mu_a + (\mu_l - \mu_a) \mathcal{H}(x, t) \quad (4)$$

where the subscript of a and l denotes the air and liquid phase respectively. $\mathcal{H}(x, t)$ is an indicator function in the characterization of numerical Heaviside function. It is a binary function showing the value of zero in the air phase and 1 in the liquid phase. The Heaviside function can be solved by the computed vector distance function from the tracked interface (Shin and Juric, 2008).

\mathbf{F} is the interfacial local surface tension force which can be explained and interpreted in a hybrid formulation defined in Shin et al., 2005 as:

$$\mathbf{F} = \sigma \kappa_H \nabla \mathcal{H} \quad (5)$$

where σ is the surface tension coefficient, which is assumed to be constant.

κ_H is twice the mean interface curvature calculated on the Eulerian grid with the expression of:

$$\kappa_H = \frac{\mathbf{F}_L \cdot \mathbf{G}}{\sigma \mathbf{G} \cdot \mathbf{G}} \quad (6)$$

where

$$\mathbf{F}_L = \int_{\Gamma(t)} \sigma \kappa_f \mathbf{n}_f \delta_f(\mathbf{x} - \mathbf{x}_f) ds \quad (7)$$

$$\mathbf{G} = \int_{\Gamma(t)} \mathbf{n}_f \delta_f(\mathbf{x} - \mathbf{x}_f) ds \quad (8)$$

\mathbf{x}_f in equation 5 and 6 is a parameterization of interface, $\Gamma(t)$ and $\delta_f(\mathbf{x} - \mathbf{x}_f)$ is a Dirac distribution that is only accountable when $\mathbf{x} = \mathbf{x}_f$. \mathbf{n}_f represents the unit normal vector to the interface with an interface elemental length ds . κ_f , again, corresponds to twice the mean interface curvature but calculated from the Lagrangian structure. These geometric information were computed and distributed from the Lagrangian interface grid onto a fixed Eulerian grid by applying Peskin and Charles' 1977 immersed boundary method of the surface integral.

The interface is advected in a Lagrangian fashion by integrating

$$\frac{d\mathbf{x}_f}{dt} = \mathbf{V} \quad (9)$$

where V is the interface velocity interpolated at x_f from the Eulerian velocity.

The simulations are dealing with system of high-density ratio approximately up to 1000, the procedures of efficient detailed solution of how simulations handle large density discontinuities can be found in Shin et al. 2017.

3.2 Numerical configuration and physical parameters

The configuration of plunging jet consists of a nozzle and a stationary pool of receiving liquid as shown in Figure 1. The setup is similar to Narendra et al., 2022 with a nozzle diameter of 2.7mm and nozzle thickness of 0.5mm as shown in the magnified details in Figure 1. The receiving pool is in a cubic shape with length of 5cm. The height of nozzle above the liquid surface was set to a lower value of 3cm due to the computational intensity and time limitation under the project framework. The reduction in height enables the liquid jet to reach the pool in a smaller amount of time, resulting in a more efficient collection of data regarding the dynamics beneath the pool induced by the plunging jet. The liquid level in the bath is set to be unchanged and to neglect the influence of overflowing.

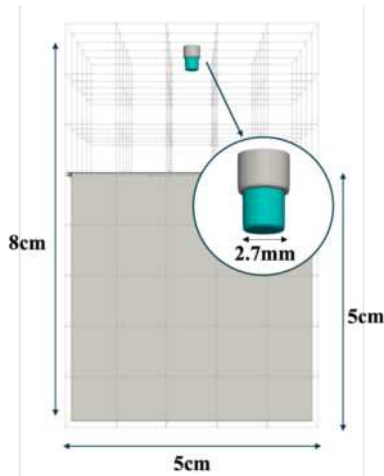


Figure 1. Schematic illustration of computational domain: a cubic tank with a nozzle set at certain height above.

The three-dimensional computational domains are set to be a rectangular box with volume of $5 \times 5 \times 8 \text{ cm}^3$. A sub-domain represents a cubic of $1 \times 1 \times 1 \text{ cm}^3$, with a mesh size of $64 \times 64 \times 64$ for each sub-domain. The time interval between two consecutive snapshots was set to be 0.01 seconds. This number was chosen as a compromise between achieving high resolution and obtaining sufficient amount of data for postprocessing. The postprocessing was conducted by means of code transition and conversion, subsequently followed by integration and analysis using Paraview.

Simulations were run with 6 set of setups with the same jet and bath parameters but different liquid

properties, the parameters of simulations are summarized in Table 1. The viscosity of fluid was chosen to be ten and fifty times the viscosity of water with a constant density of 1000 kgm^{-3} to ensure a low Reynolds number that lies in the laminar flow regime.

	Velocity [m/s] (inlet average)	Liquid Viscosity [Pa·s]	Re (at nozzle)
Lower Viscosity			
PL05	0.50	0.01	135
PL075	0.75		202.5
PL1	1.00		270
Higher Viscosity			
PL05	0.50	0.05	27
PL075	0.75		40.5
PL1	1.00		54

Table 1. Parameters for different cases of simulations

The jet velocity was set in a constant value of U_{av} on the average base of the parabolic fluid profile. The flow was assumed to be fully developed at the outlet of the nozzle. It was also assumed that the liquid jet has the same velocity value when leaving the nozzle and hitting the pool surface.

The nondimensional control parameters of the governing phenomenon applied in the study are:

$$Re = \frac{We}{Ca} = \frac{\rho_l U_{av} D}{\mu_l} \quad (10)$$

$$Ca = \frac{\mu_l U_{av}}{\sigma} \quad (11)$$

$$We = \frac{\rho_l U_{av}^2 D}{\sigma} \quad (12)$$

where μ_l is the viscosity of the liquid, σ is the surface tension.

4 Results and Discussion

4.1 Process overview

The entire plunging process involves a series of complex fluid dynamics in a chronological sequence. These include the evolution of the jet, incipient dynamics and the mechanisms of surface deformation. The simulations showed that specific performance is strongly influenced by the physical characteristics of fluids in terms of velocity and viscosity. It is frequently observed that jets tend to form a balloon-like structure at the leading edge as they fall due to the kinematics. Furthermore, the stationary interface surface of the receiving pool undergoes deformation from the close impact of the approaching jet, resulting in the formation of a conspicuous cavity at its centre, which subsequently evolves into the initial bubble. The initial formation of the bubble's shape differs as a result of differences in viscosity. The bubble would

eventually coalescence during the plunging process, either because to interfaces coinciding or limitations in computing resolution. Quantitative analyses were carried to scale the free surface deformation in terms of width and depth. A prediction that a steady state could be reached in the near further process can be made based on the convergency trend of width within the first 0.20 second of deformation. While the simulations were computed in three dimensional coordinates for the purpose of computational accuracy and higher reliability, schematic figures in 2D were presented in this section for easier geometric illustration and understanding.

4.2 Plunging jet evolution







	0.01 Pas	0.05 Pas
PL 05	 (a)	 (d)
PL 075	 (b)	 (e)
PL 1	 (c)	 (f)

Figure 2. Plunging jet evolution before reaching the receiving surface, time interval between two successive figure of 0.05sec (a) 0 to 0.45sec (b) 0 to 0.35sec (c) 0 to 0.25sec (d) 0 to 0.43sec (e) 0 to 0.35sec (f) 0 to 0.25sec

This section focuses on the geometrical development of liquid jets before to impact with the receiving pool. Figure 2. presents a summary table for the jet under all running conditions in every 0.05s. Simulations with higher plunging velocity reach the pool faster with less time taken. It took 0.45sec for plunging jet with the velocity of 0.5m/s, 0.35sec for 0.75m/s jet and 0.25sec for 1m/s jet to travel the same distance. The jets show a general trend of the formation of a balloon-like structure at the jet leading edge, which can also be referred as bulge (Zhu et al. 2000). By comparing Figures 2 (b), (e) and Figures 2 (c), (f), it can be noticed that changing fluid viscosity would not have significant variation on jet evolution. It is expected that jet with higher viscosity could have slower formation of the bulge due to higher fluid surface tension. This could be more apparent at higher nozzle and thus longer evolution time. Similar bulge phenomenon is not observed for turbulent jet according to the literatures.

It is worth to notice that simulations conducted with a velocity of 0.5m/s and viscosity of 0.01Pas (Figure 2. (a)) demonstrates the phenomenon of the liquid jet breaking up into droplets and impacting the surface through dripping rather than jetting. This phenomenon could be explained by the Rayleigh instability, which states that a liquid jet tends to breakup into smaller droplets due to the surface tension acting upon it. The system is in excess of surface energy when there is a larger surface area than the minimum necessary to hold a specific volume. Consequently, the system reorganizes itself to prioritize the state with the minimum energy, resulting in the fragmentation of the continuing stream into smaller droplets. As the bulge has been fully developed through the falling process, the breakup occurs at the time of 0.38sec with length of 18mm. A droplet with a radius of 4.7mm was formed simultaneously, which proceeds to fall into the receiving pool. In the case of the situation where the velocity is the same, but higher viscosity of 0.05Pas (Figure 2. (b)), no breakage is observed. The observed distinction is inherent, as fluids with higher viscosity exhibit more internal frictional forces, resulting in a reduced rate of deformation.

A vortex behavior can be observed when taking the closer examination of the velocity field in the Z-direction. Figure 3 (below) illustrates a schematic representation of jetting snapshot, with velocity of 0.75m/s and viscosity of 0.01Pas at the time of 0.35sec as an example. The streamlines within the diameter of descended neck flow continuously downwards, while the peripheral streamlines were redirected into two distinct swirling area with relatively lower velocity.

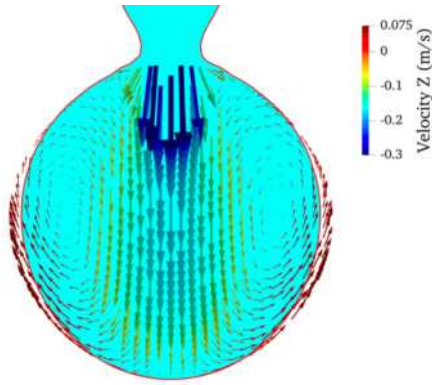


Figure 3. Z-direction velocity vector field for p1075, 0.01Pas at time of 0.35sec

4.3 Inception dynamics

This section primarily discusses the incipient dynamics on the free surface caused by the plunging activity. Specific analyses were carried to understand the mechanism of early bubble formation and curvature of deformation. Inception occurred via either dripping or jetting, resulting in an initial convexity which subsequently evolved into the bubble. Upon contact with the plunging liquid, the surface underwent a deformation characterized by a radial short crest and an air void cavity. The detail parameters were found to depend on the physical properties of the fluid, primarily the velocity and viscosity.

4.3.1 Incept via dripping

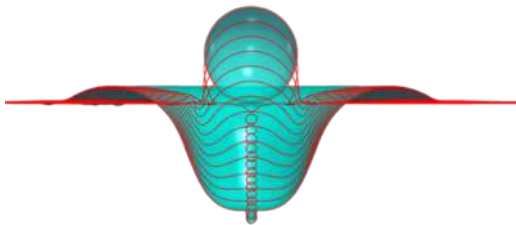


Figure 4. The inception process from 0.48sec to 0.62sec, for 0.5m/s velocity and 0.01Pas viscosity

As mentioned in the previous section of the jet evolution, the only scenario in which the jet developed and broke up into droplet was during the simulation with a velocity of 0.5m/s and a viscosity of 0.01Pas. The droplet initially made contact with the free surface at 0.48 seconds, as shown in Figure 4, resulting in the formation of a little convex protrusion in the centre. The bubble diameter was estimated to be about 0.6mm due to the constraints of simulation resolution. The convex shape of the droplet is a result of the rapid flow of liquid surrounding it, which is discussed in detail in section 4.3.2. This observation is similar as Hendrix. et al., 2016 reported in their study. Air was dragged and entrapped at the convexity point from the radial

closure of contact between the droplet and receiving liquid.

Figure 5. shows the time dependent development in depth between the top of the bubble and the surface deformational interface. The initial decrease in depth of the surface represents the general rise in liquid level due to the addition of the droplet. The depth of surface deformation increases in a faster rate than the travel depth of bubble.

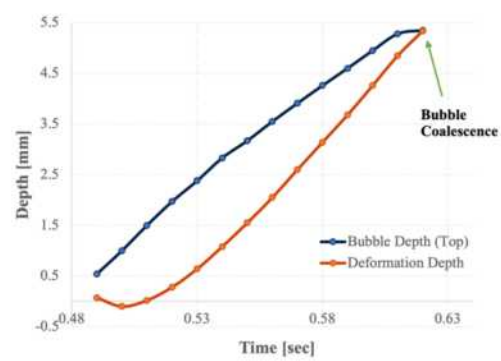


Figure 5. Depth development comparison between the top of the bubble and the surface deformation

Ultimately as time reaching the 0.62 second, the deformational interface coincided with the top of the bubble at a depth about 5.35mm. These overlapping interfaces lead to bubble coalescence, as shown as the modest increase in depth from the last two data point.

The further evolution of the low velocity (0.5m/s) scenario will not be discussed since the primary focus of this study is the mechanism associated with plunging jet. Moreover, the jetting processes are more preferred in efficiency and more widely applied in industries for convenience.

4.3.2 Incept via jetting

At higher velocities, the liquid surface interacted in the form of liquid jet rather than droplets. Figure 6. (a) depicts the schematic progression during the first 0.10sec of the inception. It is easy to observe the formation of a similar little convexity right below the centre of the jet. This is due to the compact high air velocity as the bulge is approaching the surface. This is indicated by the red arrow pointing downwards from the interface in Figure 6. (b), representing an impact velocity magnitude of 1.5m/s and a corresponding Reynolds number about 700. The convexity was considered as the proximate reason for air entrainment, subsequently leading to the formation of the small bubble underneath fluid surface. Because there is a constant contact region between the liquid jet and bath, the bubble will not collapse as a result of the coincidence of interfaces. The deformation of the surface is comparable as the dripping case but owns a shallower and more trapezoidal shape.

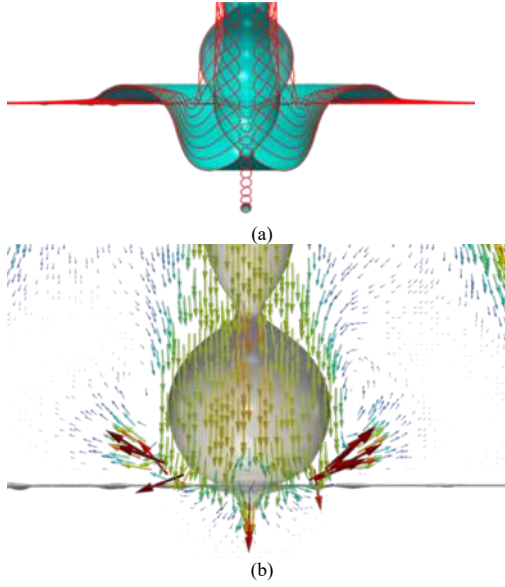


Figure 6. (a) The inception process from 0.38sec to 0.49sec, for 0.75m/s velocity and 0.01Pas viscosity (b) Velocity vector field at time 0.38sec

Although a convex deformation was also observed for higher viscosity scenario, the convexity is noticeably flatter. The flatter morphology is due to a more even velocity magnitude distribution as shown by the mint-blue colored arrow crossing the interface in Figure 7. (b) with the impacting velocity magnitude approximately 0.9m/s and a corresponding Reynolds number of about 87.

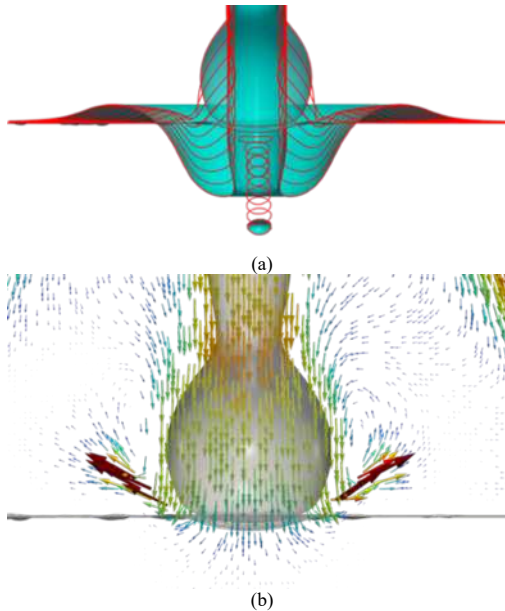


Figure 7. (a) The inception process from 0.39sec to 0.50sec, for 0.75m/s velocity and 0.05Pas viscosity (b) Velocity vector field at time 0.39sec

This kind of air layer instead of air void cavity results in the formation of a flat bubble, which is gradually developed into a more ellipsoid shape

through the continuous plunging progress. The skewness of the surface deformation is less steep compared to the lower viscosity, as higher viscosity means a higher coherence within molecules.

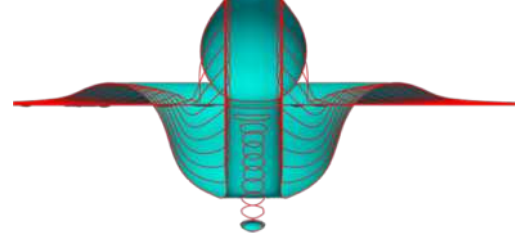


Figure 8. The inception process from 0.26sec to 0.37sec for 1m/s velocity and 0.05Pas viscosity

The primary distinction of a higher velocity acting on the jetting inception is that the liquid is plunged in a straighter pattern into the receiving liquid, as seen in Figure 8 above. This happens because when the falling distance is limited, rapid jetting experiences less time for deformation. Therefore, the shape of the jet remains mostly unaffected from its original configuration near the nozzle.

4.3.3 Curvature Analysis

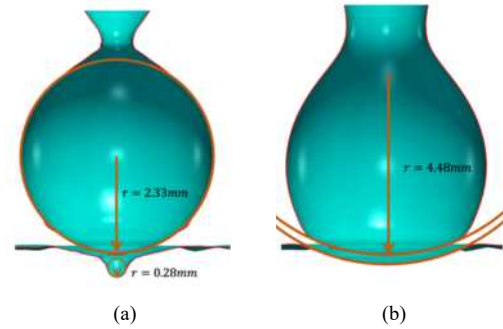


Figure 9. Zoomed detailed Comparison of curvature for impacted surface and approaching jet in the plunging speed of 0.75m/s (a) viscosity of 0.01Pas at time 0.38sec (b) viscosity of 0.05Pas at time 0.39sec

A further analysis of the curvature was conducted in this section for a closer look into the cavity formed when the jet approaches the receiving pool. The void is characterised by the curvature of an equivalent sphere with a specific radius. Figure 9 demonstrates the comparison of the impacted curvature for both jet and pool surface for the same plunging velocity at 0.75m/s at the viscosity of 0.01Pas (Figure 9. (a)) and 0.05Pas (Figure 9. (b)) respectively. The reduced viscosity results in a significant disparity in curvature up to 10 times difference, with a curvature similar to that of a jet with radius 2.33mm and an interface with radius 0.28mm. When the viscosity is increased, both the jet and interface show the same curvatures, which can be represented by an equivalent sphere with a radius of 4.48mm. This

phenomenon is intuitive since viscosity is a measure of fluid to resist deformation. The higher viscosity corresponds to higher inter-molecular frictional force and resistive to external force.

The curvature deformation of the free surface in Figure 9. (a) with the viscosity of 0.01Pas shows a geometric similarity to Gaussian function. By plotting out the curvature in terms of width and depth, a comparison between simulation data and Gaussian distribution can be shown as in Figure 10. The high extent of overlap between the two curves verified the idea that the deformation of the free surface can be approximated as a Gaussian-type curve with a correlation coefficient R^2 value up to 0.953. This finding is consistent with the discussion made in Wei et al., 2019, which stated the assumption of describing the deformation shape as a Gaussian-type curve as long as the correlation coefficient $R^2 > 0.95$ for incipient period.

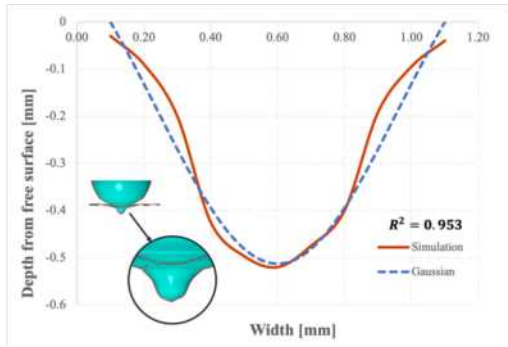


Figure 10. Comparison of surface deformation simulation data (time at 0.38sec) with the Gaussian-type curve

Nevertheless, the simulation result in this study could not make further clarification on the precise stage within the deformation process. The study conducted by Wei et al. in 2019 utilized a high-speed camera to capture images at a rate of one millisecond, resulting in a tenfold increase in precision for process analysis. Having the same resolution would result in a data size that is 1000 times bigger, which is impractical given the limited time and computer resources available for this project.

4.4 Deformation Scale

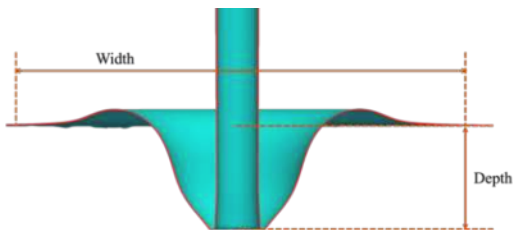


Figure 11. Illustration of measurement in terms of width and depth for deformation. Example snapshot adopted from velocity 1m/s with viscosity 0.05Pas at 0.40sec.

The deformation in the pool's free surface gives direct topological results when the liquid jet is striking the free surface. This section quantifies the deformation by the measurements of width and depth, as illustrated in Figure 11. The trendline correlation figures were plotted for three example simulation results, with one fluid parameter same to the other. Due to the project's time constraints and rounds of simulations completed, the trendlines were only plotted and examined inside the initial 0.20 second.

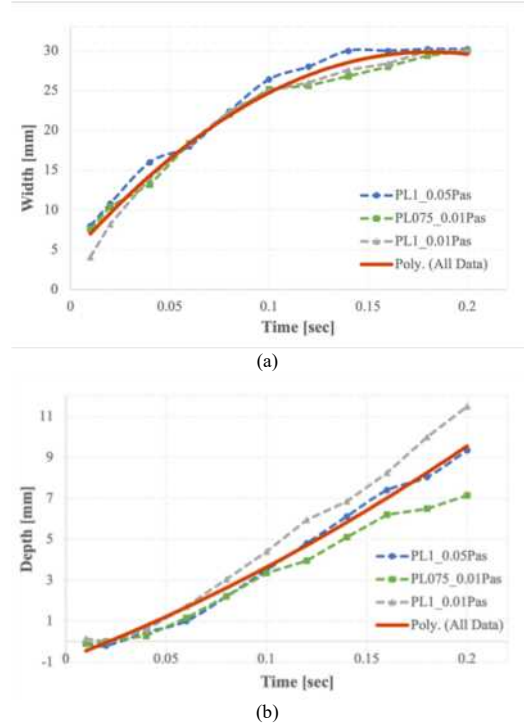


Figure 12. Trendlines for the first 0.20sec of free surface deformation in terms of width and depth along time

Both the width and depth exhibit a constant increasing trend over time, owing to the continuous transformation of the kinetic energy from the jetting liquid into the potential energy of the receiving pool, especially the pool's free surface. The difference in physical properties from fluid made no apparent difference in the rate of increasement in width, however, the cases with higher plunging velocities show a dominated effect with larger gradient for the depth. Figure 12 (a) demonstrates a convergence relationship in the width of deformation over time. The change in width steadily decreases within each cumulative time snapshot and stabilises at a consistent level of around 30mm across all scenarios. A dashed polynomial best-fitted line reinforces the idea of convergency. It can thus be concluded that the free surface deformation would reach a steady state in the width within the first 0.20 seconds of deformation, but more rounds of simulations with more results would be required to reach a steady plunging state with constant deformation topology.

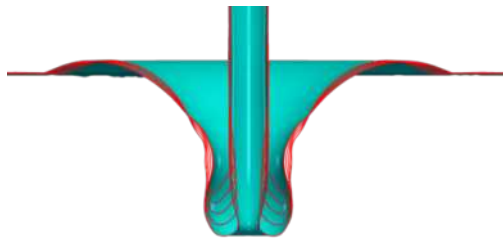


Figure 13. The process from 0.58sec to 0.62sec for 0.75m/s velocity and 0.01Pas viscosity

Figure 13 displays the farthest simulation results that this study was able to achieve. The cylindrical cavity gradually shrinks in the middle, forming a narrowing geometry. This narrowing pattern is described as neck according to Qu et al., 2013 and Salehi 2022. The further process of similar neck formation is discussed in their study, where the cylindrical cavity collapses at the neck rim because of the pressure equilibrium on the cavity wall. The lower part would undergo a transformation into toroidal bubbles, which then proceed to collapse even further into an air plume. It is expected that the simulation of this study would continue and give similar kind of process as discussed in Qu et al., 2013 and Salehi 2022. It is worth to notice that this type of process is a rapid dynamic with enormous amount of interface singularity, which would require larger amount of time and computational resources.

5 Conclusion

The project investigated the incipient air entrainment mechanisms of plunging jet, by applying CFD modelling under the scenarios of same jet configuration but varying viscosities and velocities under laminar flow regime. Due to the properties of laminar viscous flow, the jet evolved into a geometry of a bulge at the leading falling edge. The study found that the detailed development of the bulge formation highly depends on the plunging velocity and thus travelling time. The simulations with low velocity and viscosity tend to breakup from jetting into dripping due the Rayleigh instability of fluid. The closer examination of the bulge structure shows a swirling behaviour of the velocity streamlines in the Z-direction. As generally observed through all simulations, a convexity is formed on the pool centre, which is the proximate factor of the initial bubble formation due to air entrainment. Detailed analysis into the convexity illustrated a relationship between cavity geometry and fluid viscosity. The higher viscosity leads to a formation of air layer instead of dragged air void. The air layer thus evolves into the flatter bubble, which gradually develops into a more ellipsoid shape as travelling deeper in the pool. A high correlation was found between the small convexity

and Gaussian type curve for lower viscosity case. By characterising the deformation into parameters of width and depth development with time, a general increase trend was found for both with no significant difference from velocity and viscosity. The simulations are expected to demonstrate a similar progress of the jet to those seen in existing literature with extra computational time provided.

The results of the analysis can be used as a reliable preliminary step for further research, especially in the area related to small scale with precise operation. However, further entrainment process from incipient period, for instance the bubble penetration depth, bubble size and amount distribution in the steady plunging state have not been determined due to time and computational resources constraints. Since plunging jet is recognized as an efficient role for mixing temperature-sensitive fluids, further research in quantifying and controlling the air entrainment process in terms of bubble dynamics are recommended. Industrial operation with air entrainment phenomenon normally have high plunging velocities, therefore simulations running under turbulent regime are suggested for more extensive research. Disturbance's factor such as oscillating jet, jetting with angles and jetting in fluctuating velocities are also advised for future research for industrial flow assurance. Moreover, other jets related to fluid dynamics are recommended under the topics of bouncing jet, Kaye effect and coiling jet.

Acknowledgement

The authors would like to express their gratitude to Dr. Lyes Kahouadji for his valuable technical assistance and insightful discussions. They would also like to express appreciation to Shin 2017 for providing the numerical simulations code BLUE and to the Research Computing Service of Imperial College London for the High-Performance Computing resources. Furthermore, give thanks for the analytical visualisation created by Paraview.

References

- A. Ohkawa, D. Kusabiraki, Y. Kawai, N. Sakai, K. Endoh, 1986. Some flow characteristics of a vertical liquid jet system having downcomers. *Chemical Engineering Science*, 41(9), pp. 2347-2361.
- Bagatur, T., 2014. Experimental Analysis of Flow Characteristics from Different Circular Nozzles at Plunging Water Jets. *Arabian Journal for Science and Engineering*, Volume 39, p. 2707-2719.
- Biñ, A. K., 1993. Gas entrainment by plunging liquid jets. *Chemical Engineering Science*, 48(21), pp. 3585-3630.
- Chirichella, D., Ledesma, R. G., Kiger, K. T. & Duncan, J. H., 2002. Incipient air entrainment in a

- translating axisymmetric plunging laminar jet. *Physics of Fluids*, Volume 14, p. 781–790.
- Christophe Clanet, Juan C. Lasheras, 1999. Transition from dripping to jetting. *Journal of fluid mechanics*, Volume 283, pp. 307-326.
- D. Ervine, M. Ice, M. Iwes, 1988. Aeration in jets and high velocity flows. In: s.l.:s.n.
- EJ McKeogh, DA Ervine, 1981. Air entrainment rate and diffusion pattern of plunging liquid jets. *Chemical engineering science*, Volume 36, pp. 1161-1172.
- Élise Lorenceau, David Quéré, Jens Eggers, 2004. Air entrainment by a viscous jet plunging into a bath. *Physical review letters*, 93(25).
- Faiza Zidouni Kendil, Dana V. Danciu, Martin Schmidtke, Anis Bousbia Salah, Dirk Lucas, Eckhard Krepper, Amina Mataoui, 2012. Flow field assessment under a plunging liquid jet. *Progress in Nuclear Energy*, Volume 56, pp. 100-110.
- Fatemeh Salehi, Esmaeil Ajdehak, Yannis Hardalupas, 2022. Computational fluid dynamics modelling of air entrainment for a plunging jet. *Chemical Engineering Research and design*, Volume 179, pp. 319-330.
- Jens Eggers, Emmanuel Villermaux, 2008. Physics of liquid jets. *Reports on progress in physics*, 71(3).
- K. Tojo, N. Naruko, K. Miyanami, 1982. Oxygen transfer and liquid mixing characteristics of plunging jet reactors. *The chemical engineering journal*, 25(1), pp. 107-109.
- Laura, P., 1979. Onset of air entrainment for a water jet impinging vertically on a water surface. *Chemical Engineering Science*, 34(9), pp. 1164-1165.
- M. El Hammoumi, J. L. Achard, L. Davoust, 2002. Measurements of air entrainment by vertical plunging liquid jets. *Experiments in Fluids*, Volume 32, pp. 624-638.
- M. Xiang, S.C.P. Cheung, J.Y. Tu, W.H. Zhang, 2014. A multi-fluid modelling approach for the air entrainment and internal bubbly flow region in hydraulic jumps. *Ocean Engineering*, Volume 91, pp. 51-63.
- Maurice H. W. Hendrix, Wilco Bouwhuis, Devaraj van der Meer, Detlef Lohse, Jacco H Snoeijer, 2016. Universal mechanism for air entrainment during liquid impact. *Journal of Fluid Mechanics*, Volume 789, pp. 708-725.
- Munish Kumar, N. K. Tiwari, 2021. Oxygenation by Plunging Jet Aerators: A Review. *Iranian Journal of Science and Technology, Transactions of Civil Engineering*, Volume 45, pp. 1329-1348.
- Narendra Dev, J John Soundar Jerome, Hélène Scolan, Jean-Philippe Matas, 2022. *Air entrainment by plunging jets*. [Online] Available at: <https://gfm.aps.org/meetings/dfd-2022/6323426c199e4c2c0873f98d> [Accessed 20 October 2023].
- Panarese, William C., 1963. *how to use air-entrained concrete... and why you should use it*. [Online] Available at: https://www.concreteconstruction.net/how-to/materials/how-to-use-air-entrained-concrete-and-why-you-should-use-it_o [Accessed 15 October 2023].
- Pedro Lopes, Gavin Tabor, Rita F. Carvalho, Jorge Leandro, 2016. Explicit calculation of natural aeration using a Volume-of-Fluid model. *Applied Mathematical Modelling*, 40(17-18), pp. 7504-7515.
- Peskin, Charles D, 1977. Numerical analysis of blood flow in the heart. *Journal of Computational Physics*, 25(3), pp. 220-252.
- Seungwon Shin, Damir Juric, 2008. A hybrid interface method for three-dimensional multiphase flows based on front tracking and level set techniques. *International Journal for Numerical Methods in Fluids*, 60(7), pp. 753-778.
- Seungwon Shin, Jalel Chergui, Damir Juric, 2017. A solver for massively parallel direct numerical simulation of three-dimensional multiphase flows. *Journal of Mechanical Science and Technology*, 31(4), pp. 1739-1751.
- Seungwon Shin, S.I. Abdel-Khalik, Virginie Daru, Damir Juric, 2005. Accurate representation of surface tension using the level contour reconstruction method. *Journal of computational physics*, 203(2), pp. 493-516.
- Surinder Deswal, V. S. Verma, 2007a. Air-water oxygen transfer with multiple. *Journal of Indian association for environmental management*, Volume 31, pp. 141-146.
- Wangru Wei, Weilin Xu, Jun Deng, Zhilong Tian, Faxing Zhang, 2016. Free-surface air entrainment in open-channel flows. *Science China Technological Sciences*, 60(6), pp. 893-901.
- Wangry Wei, Weilin Xu, Jund Deng, Zhing Tian, Faxing Zhang, 2019. Bubble formation and scale dependence in free-surface air entrainment. *Scientific Reports*, 9(1), p. 110008.
- Xiaoliang Qu, Afshin Goharzadeh, Lyes Khezzar, Arman Molki, 2013. Experimental characterization of air-entrainment in a plunging jet. *Experimental Thermal and Fluid Science*, Volume 44, pp. 51-56.
- Yan-Cheng Han, Said M. Easa, 2018. Exact Solution of Optimum Hydraulic Power-Law Section with General Exponent Parameter. *Journal of Irrigation and Drainage Engineering*, 144(12).
- Yonggang Zhu, Hassan Ogoz and Andrea Prosperetti, 2000. On the mechanism of air entrainment by liquid. *Journal of Fluid Mechanics*, Volume 404, pp. 151 -177.

Filling in the Cracks: An Investigation into Surface Patterning via Wrinkling

Begoña Parías Moreno de los Ríos and Helen Barr

Department of Chemical Engineering, Imperial College London, London, SW7 2AZ, U.K.

13th December 2023

Abstract

Surface patterning through wrinkling has attracted a lot of attention with its applicability in many areas, including structural colour, drag reduction, mitigating biofouling, and controlled wetting and spreading. A facile way to achieve surface wrinkling is through the plasma oxidation of an elastomer substrate, such as polydimethylsiloxane (PDMS), to alter the surface chemistry of the substrate, creating a thin stiff film as a result. This bilayer system, thin stiff film on soft substrate, when exposed to external stress, gives rise to surface wrinkles due to the difference in their mechanical properties. However, one of the challenges associated with surface wrinkling is the formation of cracks. This report investigates the effect of four variables on crack formation, with the key variables found to be the thermal insulation of the material upon which the sample is placed, and the pressure at which the plasma oxidation chamber is operated. It was concluded that, in order to achieve crack-free wrinkling, materials with *U-values*, which is a measure of the thermal transmittance of a material, above $950\text{ W}/(\text{m}^2\text{K})$ should be used as the sample support during treatment in the oxidation chamber for elastomer thicknesses greater than $50\text{ }\mu\text{m}$. In addition, the plasma chamber should be run above 0.05 mbar to avoid cracking.

Key words: plasma oxidation, PDMS, wrinkling, cracking, AFM, SALS.

1 Introduction

The wrinkling of thin stiff films on soft substrates has a myriad of applications. These include drag reduction [1], antibacterial coatings [2], controlled wetting and spreading [3], and biofouling mitigation [4]. In addition, the scale-up potential of micro- and nano- scale wrinkling is huge. For example, the inside surface of pipes could be covered with nanoscale wrinkles, reducing drag and thus increasing flow rate without increasing pump duty. However, the use of surface patterning through wrinkling is limited due to the emergence of cracks on the surface, which is very often neglected. It is important to understand and mitigate the formation of cracks as they can render the wrinkled surface useless. For example, although wrinkles inhibit the growth of microbes in applications where wrinkles are

used as antimicrobial coatings, these microbes proliferate on the cracks, thus making the wrinkled surface ineffective [5]. This report focuses on the effect of four different variables on crack formation, resulting from the use of plasma oxidation to alter the topography of a soft elastomer substrate.

Surface wrinkles are driven by the mechanical instability of the elastomer substrate brought about by an applied stress. In this report, the temperature of the plasma chamber caused the samples to heat up and expand. After the treatment, the samples were removed from the chamber and cooled to room temperature, contracting in the process. This expansion and contraction of the samples acted as the applied stress needed to drive the mechanical instability of the polymer network, bringing about surface wrin-

kles.

Polydimethylsiloxane (PDMS) is an optically clear and inert elastomer extensively studied in surface patterning and microfluidics. In addition, its incompressibility, indicated by a Poisson's ratio, ν_{PDMS} , of 0.5 [6], makes it the ideal soft substrate upon which to perform the experiments outlined in this report. Moreover, PDMS readily undergoes plasma oxidation, forming a thin, glassy film in the order of 10 nm as a result. The mechanical instability of the PDMS gives rise to buckling and subsequent wrinkling of this thin film, with the wavelength, λ , and amplitude, A , of the wrinkles described by Equations 1 and 2 below:

$$\lambda = 2\pi h_f \left(\frac{\bar{E}_f}{3\bar{E}_s} \right)^{1/3} \quad (1)$$

$$A = h_f \left(\frac{\varepsilon}{\varepsilon_c} - 1 \right)^{1/2} \quad (2)$$

where h_f is the thickness of the glassy film, \bar{E}_f and \bar{E}_s are the elastic moduli of the thin film and thick substrate respectively, and ε_c is the critical strain that must be exceeded to achieve buckling of the thin film. The elastic modulus of the thin film, \bar{E}_f , was considered to be 30 GPa [6]. Given that a typical value for the elastic modulus of PDMS, E_{PDMS} , is 1.6 MPa [6], the elastic modulus of the thick substrate can be calculated using Equation 3, resulting in a thick substrate elastic modulus of $\bar{E}_s \approx 2.1$ MPa.

$$\bar{E}_s = \frac{E_{PDMS}}{(1 - \nu_{PDMS}^2)} \quad (3)$$

As aforementioned, since the temperature difference inside and outside the plasma chamber is being used as the mechanism for applying strain to the sample, it is assumed that each sample undergoes low deformation (strain $\varepsilon \leq 10\%$). This means that amplitude alone depends on strain, allowing λ and A to be decoupled [7].

2 Methods

2.1 Materials & Equipment

2.1.1 Plasma Oxidation Chamber

Samples underwent plasma oxidation in a Diener Femto vacuum chamber to create a thin glassy film on top of a soft elastomer substrate. Plasma is ionised gas with equal amounts of positive and negative ions [8] that alter the chemistry of the elastomer surface. This is created by introducing a gas into the chamber which is then ionised. The excited gas molecules emit UV light, causing the plasma to glow.

2.1.2 Atomic Force Microscopy (AFM)

Bruker (Veeco) AFM in tapping mode was used to analyse the topography of the samples discussed in this report. This technology works by touching a nanoscale tip repeatedly off the samples surface to produce a 3D image of the surface topography. The images produced were then analysed to determine the wavelength and amplitude of the wrinkles formed. Knowing λ , Equation 1 was used to determine the thickness of the glassy film, h_f . These three parameters, λ , A , and h_f , will define the wrinkling characteristics of each sample.

2.1.3 Small-Angle Light Scattering (SALS)

The SALS set-up consisted of a green laser beam (wavelength of 532 nm) being shone on an X-Y scanning mirror. This reflected the beam at 90° which then passed perpendicularly through the sample. The resulting diffraction pattern was observed on a screen and recorded using a Hamamatsu Orca camera mounted vertically above and was controlled using Wasabi, an in-built software.

2.2 Sample Preparation

Dow SYLGARD 184 silicone elastomer base was mixed with 184 silicone elastomer curing agent in a 10:1 ratio. This was stirred vigorously with a spatula and placed in

a vacuum to remove air bubbles. The resulting mixture was polydimethylsiloxane (PDMS). Silicon wafers were cut into 1cm x 1cm squares using a diamond tip pen. Using tweezers and without touching the surface, a square wafer was placed on the spin coater and secured onto the chuck through vacuum. The wafer was cleaned first with a compressed air gun, sprayed with isopropanol, and then dried with the compressed air gun. Using a pipette, 0.2 mL of PDMS was deposited onto the surface of the cleaned wafer. This was then spin coated at a certain speed, depending on the experiment being carried out, for 1.5 minutes. Once coated with PDMS, the samples were cured in a 75 °C oven for 1 hour. The thickness of the substrate on the wafer was controlled through the spin coating speed. A summary of the spin coating speed and resulting PDMS substrate thickness is presented in Table 1. The substrate thicknesses were determined using a calliper.

Table 1: Spin coating speed and resulting PDMS substrate thickness.

Spin coating speed [rpm]	PDMS substrate thickness [μm]
200	330 \pm 0.01
2000	50 \pm 0.01
6000	10 \pm 0.00

2.3 Plasma Oxidation

The samples coated with PDMS were placed on the centre of a plate and placed in the plasma oxidation chamber. Air was pumped out of the chamber until a pressure of 0.07 mbar was achieved. Gas was then pumped into the chamber to reach a certain pressure. Depending on the experiment being conducted, the gas type being pumped into the chamber and the final operating pressure were varied. Ensuring the power of the chamber was set to 99 W, the generator was started, initiating the oxidation process. This was allowed run for a certain amount of time before the chamber was ventilated and returned to ambient con-

ditions. This process varied slightly depending on the experiment being conducted but will be clearly stated.

2.4 Experimental Studies

2.4.1 Exposure Time

Following the findings in relation to the effect of plasma dose and pressure on surface wrinkling and cracking [9], it was decided to investigate the impact of plasma dose, via exposure time, on the resulting surface topography. Three samples were prepared according to Section 2.2, using a spin coating speed of 6000 rpm. One sample was placed on the middle of a metal plate and put in the plasma chamber. The plasma oxidation treatment was carried out as outlined in Section 2.3, using oxygen to fill the chamber to a pressure of 0.2 mbar. The sample was then treated for 1 minute. This was repeated twice more for plasma exposure times of 5 and 10 minutes. For a constant plasma power of 99 W, the plasma exposure time and corresponding plasma dose are summarised in Table 2. After each plasma exposure, the treated sample was transferred to a glass Petri dish and analysed using AFM. Results are outlined in Section 3.1.

Table 2: Calculated plasma dose for different plasma exposure time.

Plasma exposure time [mins]	Plasma dose [kJ]
1	5.94
5	29.70
10	59.40

2.4.2 Gas Type & Pressure

Twelve samples were prepared according to Section 2.2 and spin coated using a speed of 6000 rpm. Six samples were used to investigate the use of oxygen to fill the oxidation chamber, and six were used to investigate the use of air. In every case, the plasma exposure time and power were 1 minute and 99 W respectively, resulting in a plasma dose of 5.94 kJ. After removal from the oxidation chamber,

samples were left to cool at the ambient rate.

One sample was placed on the middle of the metal plate and put into the oxidation chamber. The method outlined in Section 2.3 was followed, using oxygen as the gas to fill the chamber to reach the desired operating pressure. The operating pressure was varied for the six different samples using the variable area flow meter on the plasma chamber. The pressures investigated were the pressures corresponding to a reading on the variable area flow meter of 0, 0.5, 1, 5, 10, and 15. The corresponding operating pressures of the chamber are summarised in Table 3.

This was repeated using the remaining six samples, but filling the chamber with air instead of oxygen. Again, the operating pressures of the chamber and the corresponding readings on the variable area flow meter are summarised in Table 3.

Table 3: Gas studies operating pressures and corresponding variable area flow meter settings.

Sample number	Variable area flow meter setting	Pressure Study 1 Oxygen [mbar]	Pressure Study 2 Air [mbar]
1	0	0.017	0.017
2	0.5	0.053	0.037
3	1	0.093	0.050
4	5	0.200	0.120
5	10	0.350	0.240
6	15	0.540	0.310

2.4.3 Cooling Rate

Initially it was thought that the cooling rate of the samples after plasma exposure could cause the samples to crack. The effect of three cooling rates on sample cracking were investigated. These cooling rates were 0.1 °C/min, ambient, and 50 °C/min cooling. In each experiment, the plasma power was maintained at 99 W with an exposure time of 5 minutes, resulting in a plasma dose of 29.7 kJ.

Three samples were prepared as outlined in Section 2.2 and were spin coated with PDMS at 6000 rpm for 1

minute. For each experiment, one sample was placed on a metal plate and was pre-heated, using a hot plate, to 75 °C. This was then placed in the plasma oxidation chamber and the method outlined in Section 2.3 was followed, filling the chamber with oxygen to a pressure of 0.2 mbar.

Once the plasma treatment was completed, the sample was quickly transferred to a Linkam temperature controller where it was maintained at 75 °C. This was placed under the SALS set-up, where a time-lapse of the sample was recorded. On the Linkam controller, the cooling rate was set to 0.1 °C/min. The sample was cooled at this rate until it reached 25 °C at which it was held for 2 hours. This process was repeated for the 50 °C/min cooling rate. In the case of the ambient cooling rate, the Linkam temperature controller was not switched on, allowing the sample to cool according to Newton's law of cooling. All samples were analysed using AFM and results are presented in Section 3.3.

2.4.4 Thermal Insulation

The effect of placing materials with different thermal insulation properties between the sample and its supporting plate in the oxidation chamber was investigated. It was thought that the amount of heat dissipated from the sample during the oxidation treatment would vary depending on the conductivity of the material it was in contact with. This in turn would have an impact on the amount of strain, and hence potential cracking, of the sample. Since the effect of substrate thickness on sample cracking was unknown, three samples, one of each 10, 50, and 330 µm, were prepared according to Section 2.2 for each study. The study numbers and corresponding contacting materials are summarised in Table 4, and schematics of each study are shown in Figure 1.

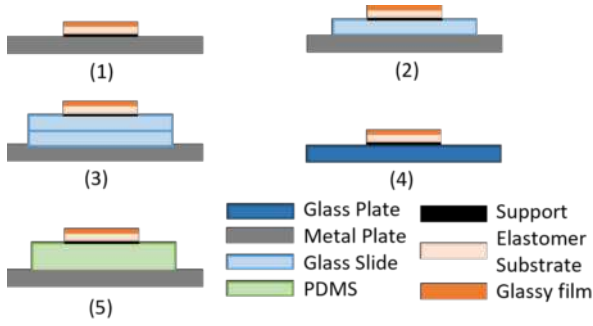


Figure 1: Thermal insulation studies set-up. Set-ups (1) to (5) are in order of increasing thermal insulation. In (1), the sample is placed directly on a metal plate. (2) and (3) involve placing 1 and 2 glass slides respectively between the sample and metal plate. In (4), the sample is placed directly on a glass plate. In (5), 2 mm of PDMS is placed between the sample and metal plate.

The glass slides used were soda-lime glass, the glass plate was borosilicate glass, and the PDMS was made of Dow SYLGARD 184 silicone elastomer base and 184 silicone elastomer curing agent in a 10:1 ratio.

In order to quantify the thermal insulation performance, the U -values of each set-up in Table 4 were calculated using Equation 4 [10]. The U -value is defined as the reciprocal of the total thermal resistance, R_T , which is the sum of the ratio of the material thickness, d_i , to its thermal conductivity, λ_i , for every layer i [11], [12], [13], [14]. This U -value is also known as the thermal transmittance which quantifies the amount of heat that passes through a material. Therefore, a lower U -value indicates a better insulator, as the material resists the transfer of heat more effectively.

$$U\text{-value} = \frac{1}{R_T} = \frac{1}{\sum_{i=1}^n \frac{d_i}{\lambda_i}} \quad (4)$$

Table 4: Thermal insulation study number, corresponding contacting material and U -value.

Study number	Contacting material	U -value [W/(m ² K)]
1	aluminium plate	79,000
2	aluminium plate + 1 glass slide	952.024
3	aluminium plate + 2 glass slides	478.897
4	glass plate	240.000
5	aluminium plate + 2 mm PDMS	79.919

The U -values corresponding to each thermal insulation study set-up are summarised in Table 4, and are shown visually in Figure 2.

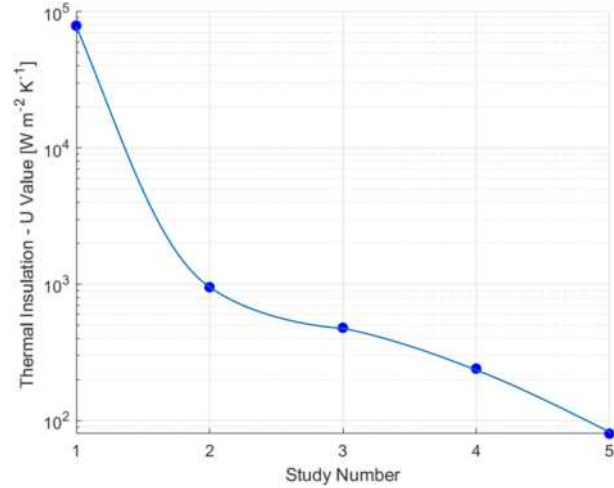


Figure 2: Thermal insulation performance (U -values) of each thermal insulation study set-up.

For each study, the samples were placed on the middle of their contacting materials and inserted into the plasma chamber. The plasma chamber was run according to Section 2.1.1, filling the plasma chamber with oxygen to 0.2 mbar and treating the samples for 5 minutes with plasma power 99 W. This resulted in a plasma dose of 29.7 kJ. Using tweezers, the treated samples were removed from the plasma chamber and transferred to a labelled glass Petri dish. These samples were then analysed using the SALS set-up, AFM, and optical microscope, and the results are summarised in Section 3.4.

3 Results & Discussion

3.1 Exposure Time

Figure 3 shows the AFM images that were analysed to determine the wrinkling characteristics for varying exposure time, shown in Figure 4.

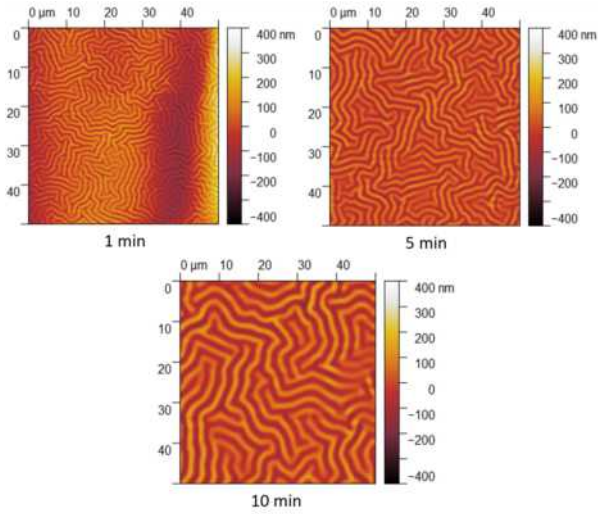


Figure 3: AFM images of samples realised when exposed to plasma for 1, 5, and 10 minutes.

It can be seen that an increase in exposure time results in an increase in both h_f and λ . This is unsurprising since the longer the sample spends in the oxidation chamber, the further into the substrate the plasma can penetrate, resulting in the formation of a thicker glassy film. Since λ is proportional to h_f , as shown by Equation 1, it follows that the λ also increases with exposure time.

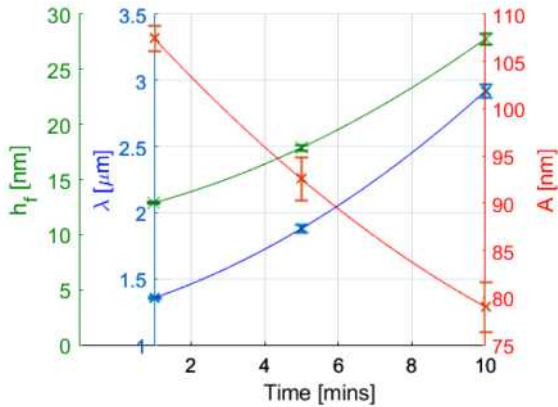


Figure 4: Wrinkling characteristics for varying exposure time.

The amplitude of the wrinkles decreases with increasing exposure time. This can be explained by Equation 2, showing that A increases as a function of the square root

of the applied strain, along with the fact that longer exposure times result in larger h_f . For thinner films, observed at shorter exposure times, the elastic modulus of the soft substrate dominates, resulting in high strain and therefore high amplitude. For thicker glassy films, there is less strain on the thin film as the film and substrate contribute similar amounts to the overall strain of the sample. No cracking was observed in any samples.

To determine the wrinkling characteristics of the samples, three measurements were made on the AFM image in different locations for both λ and A and the average taken. From this, the standard deviation, and subsequent standard error, of the data was calculated, with the standard error being used to form the error bars shown in Figure 4. The standard error of λ , σ_λ , was propagated according to Equation 5 [15], where k is a constant multiplier shown in Equation 1, to calculate the standard error, and hence error bars, of h_f . This method was used throughout this report to determine the error in each experiment.

$$\sigma_{h_f} = k \sigma_\lambda \quad (5)$$

3.2 Gas Type & Pressure

Figure 5 shows the AFM images of the wrinkles formed when operating the plasma chamber at 0.1 mbar, in the presence of oxygen or air.

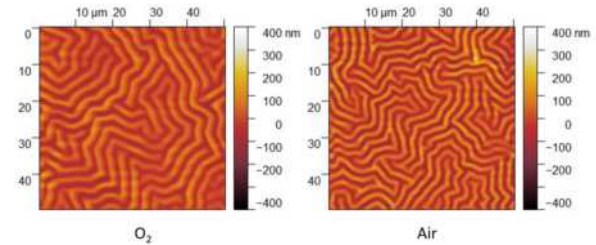


Figure 5: AFM images for 0.1 mbar chamber pressure, using oxygen and air respectively.

From AFM analysis of all samples, it was possible to determine the wrinkling characteristics with varying pressure

for both oxygen and air. These are depicted in Figures 6(a) and 6(b) respectively.

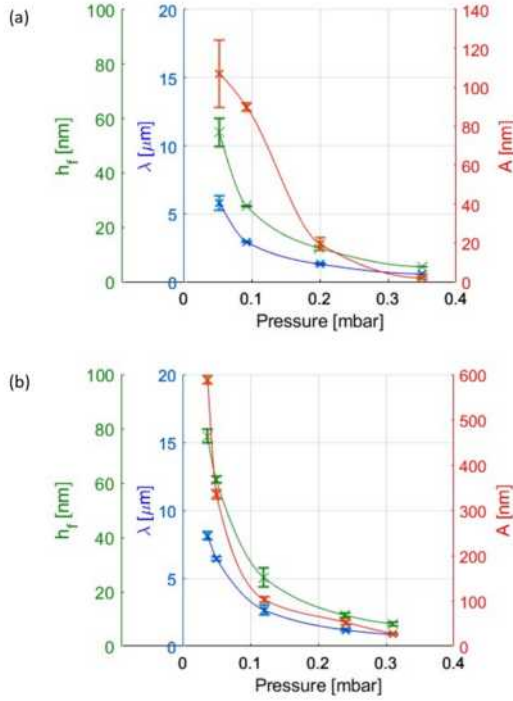


Figure 6: Wrinkling characteristics for (a) oxygen and (b) air with varying pressure.

It can be seen that, for both gas types, λ , A , and h_f decrease with increasing chamber pressure above 0.05 mbar. For both gas types, samples made using a chamber pressure of less than 0.05 mbar during treatment cracked, and samples made using pressures below 0.02 mbar did not undergo any treatment as there was not enough plasma present. As a result, it was possible to estimate a pressure range in which samples would be expected to crack. This is between 0.02 and 0.05 mbar.

The crack densities of the three cracked samples were investigated using the same method outlined in Section 3.4. Similar crack densities were observed across the three samples, meaning if the plasma chamber is operated within the cracking pressure range, the crack density can be accurately predicted to be about 1% as shown in Table 5.

This confirms that the results are robust and non-specific to the plasma chamber.

Table 5: Crack densities for different conditions.

Gas type	Pressure [mbar]	Crack density [%]
Oxygen	0.053	0.92
Air	0.050	1.13
Air	0.037	1.37

3.3 Cooling Rate

The wrinkling characteristics and corresponding cooling rates are plotted in Figure 7. No cracking was observed in any of the samples. The ambient and 0.1 °C/min cooling rates were very similar in terms of λ and A .

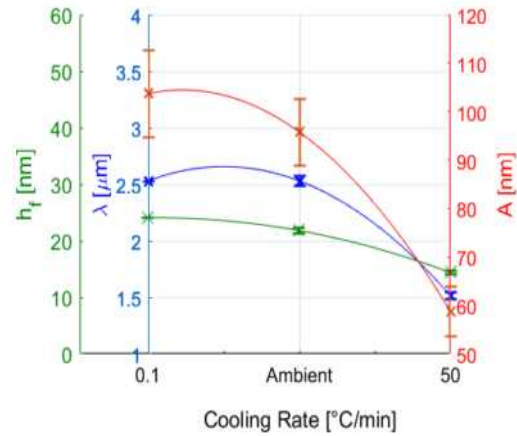


Figure 7: Wrinkling characteristics for various cooling rates.

Figure 8 shows the change in sample temperature, T , and formation of the wrinkles, $I(q^*)$, with time for the different cooling rates. It can be seen that, for 0.1 °C/min and ambient cooling rates, the wrinkles form once the sample temperature has dropped to about 30 °C. For a cooling rate of 50 °C/min, since the temperature is reduced rapidly, the wrinkle formation is instant. As 0.1 °C/min and ambient cooling rates result in similar wrinkle intensities, both of which are higher than that produced from 50 °C/min cooling, it was confirmed that ambient cooling could be used to avoid the need for temperature control.

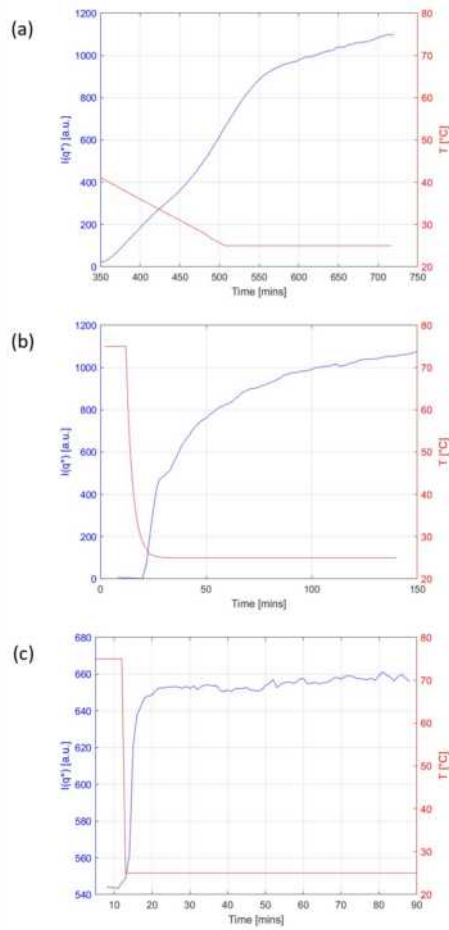


Figure 8: Temperature and wrinkling intensity profiles for (a) 0.1°C/min, (b) ambient and (c) 50°C/min cooling rates.

3.4 Thermal Insulation

Figure 9 shows the SALS images for 50 μm substrate thickness samples of thermal insulation studies 2 and 4 which formed wrinkles and cracks respectively. A summary of set-ups 2 and 4 can be found in Figure 1 and Table 4.

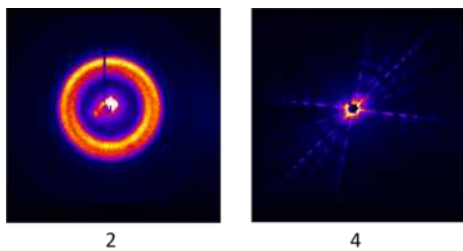


Figure 9: SALS images of thermal insulation studies 2 and 4 for 50 μm PDMS substrate thickness (2000 rpm spin coating speed).

The analysis of these images, along with the use of the optical microscope and AFM, led to the development of the morphology map shown in Figure 10. This depicts the regions of wrinkles, cracks, and wrinkles and cracks that were observed for various substrate thicknesses and U -values of the contacting materials on which the samples were placed.

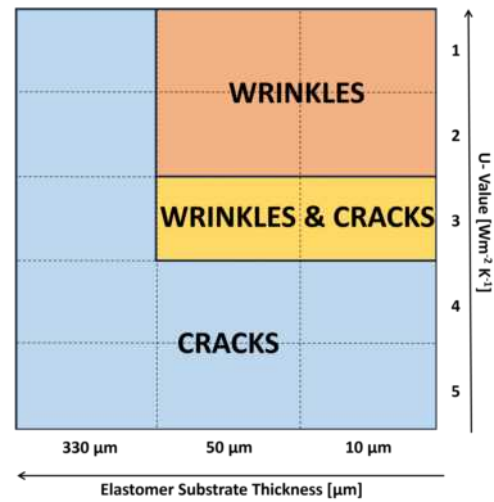


Figure 10: Morphology map depicting the zones of wrinkling, wrinkling and cracking, and cracking for different contacting material U -values and substrate thicknesses.

Figure 2 shows the U -values for the various thermal insulation studies that were carried out. As mentioned in Section 2.4.4, schematics of each thermal insulation study are shown in Figure 1. It was concluded that for contacting materials with U -values greater than $950 \text{ W}/(\text{m}^2\text{K})$, the sample would be expected to form wrinkles. This will be defined as the wrinkling zone. For contacting materials with U -values between 300 and $950 \text{ W}/(\text{m}^2\text{K})$, the sample would be expected to form wrinkles and cracks. This will be defined as the wrinkling and cracking zone. Finally, for contacting materials with U -values less than $300 \text{ W}/(\text{m}^2\text{K})$, the sample would be expected to crack.

Figure 11 shows the wrinkling characteristics in the wrinkling, and wrinkling and cracking zones for two sub-

strate thicknesses. It can be seen that, for a given substrate thickness, h_f is larger in the wrinkling and cracking zone than in the wrinkling zone. It follows that the wrinkle wavelength, λ , is also larger in the wrinkling and cracking zone than in the wrinkling zone. The opposite is observed when looking at A , which is smaller in the wrinkling and cracking zone than in the wrinkling zone. It can also be deduced that thinner substrates result in higher h_f , λ , and A than thicker substrates for a given morphology zone. Samples made using a spin coating speed of 200 rpm corresponding to a substrate thickness of 330 μm are not shown on Figure 11 as they all cracked without forming wrinkles.

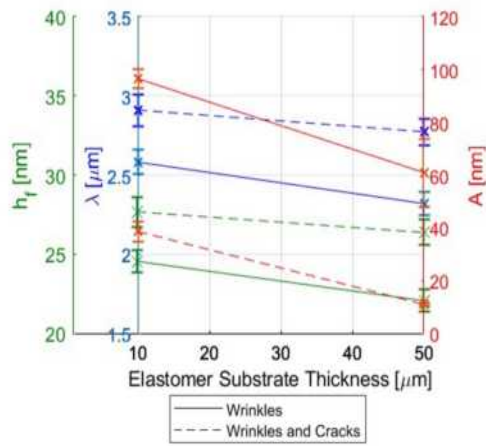


Figure 11: Wrinkling characteristics vs substrate thickness for materials with U -values in the wrinkling and wrinkling and cracking zones.

This understanding, coupled with the crack density results shown in Figure 14, resulted in the development of Figure 12. Figure 12(a) depicts why, for contacting materials with high U -values (low thermal insulation), samples do not crack since the heat, Q , dissipates from the sample more easily. On the other hand, Figure 12(b) shows that for contacting materials with low U -values (high thermal insulation), samples will crack since the heat, Q , cannot easily dissipate from the sample. Therefore, the build up of energy supplied by the plasma to the sample is released via cracking.

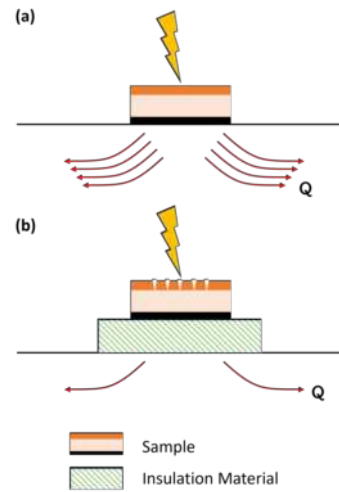


Figure 12: Crack formation resulting from thermal insulation. The yellow lightning bolt represents the energy supplied to the sample from the plasma. In set-up (a), the sample is placed directly on a supporting plate in the plasma oxidation chamber. In set-up (b), insulating material is placed between the sample and supporting plate. The presence of the insulating material means that the heat energy, Q , supplied by the plasma cannot easily dissipate from the sample, thus causing the sample to crack.

The crack density of each sample was calculated from images taken using the optical microscope. These images were processed using MATLAB, making the cracks appear black on a white background. This made it possible to calculate the crack density by calculating the area fraction of the image that was black compared to white. Figure 13 shows an image taken from the optical microscope before and after it was processed using MATLAB to determine the crack density.

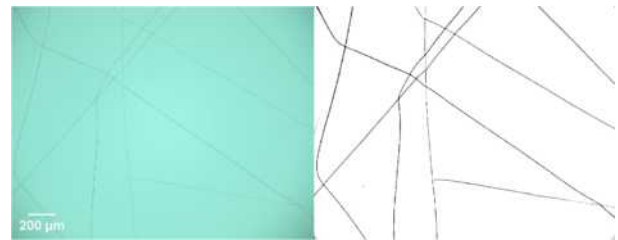


Figure 13: Crack density analysis using MATLAB for 10 μm substrate thickness, thermal insulation study 5.

The results are depicted in Figure 14, showing the crack density vs substrate thickness for the five thermal insulation studies. It shows that, for a given thermal insula-

tion, the crack density is lower for thicker substrates. This can be explained by the fact that there is less substrate present in thinner samples, so more energy is released in the form of cracks rather than dissipating through the sample. It also shows that, for a given substrate thickness, the crack density is higher for contacting materials with lower U -values (better insulators).

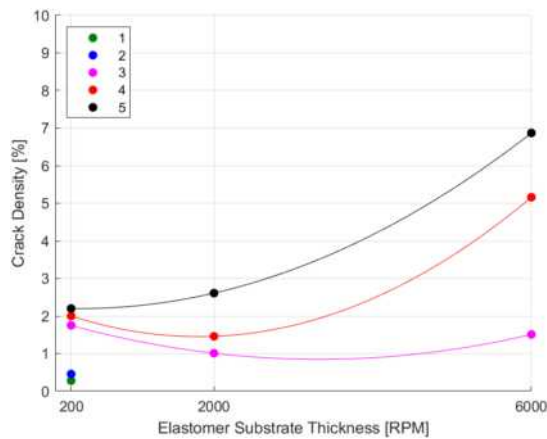


Figure 14: Crack density vs substrate thickness for the different thermal insulation studies.

4 Conclusion & Outlook

The results presented in this report are an important step to understanding the factors affecting crack formation in thin stiff films on soft substrates. Two key variables were identified in relation to unwanted crack formation during wrinkling of a thin elastomeric substrate through plasma oxidation. These were the thermal conductivity of the material on which the sample is placed in the oxidation chamber, and the gas pressure at which the plasma oxidation chamber was operated.

It was concluded that, in order to achieve crack-free wrinkling, contacting materials with U -values above $950 \text{ W}/(\text{m}^2 \text{K})$ should be used as the support during treatment in the oxidation chamber for samples with elastomer thickness greater than $50 \mu\text{m}$. In addition, the plasma chamber should be run above 0.05 mbar to avoid cracking. Adhering to these conditions facilitated the control and

achievement of crack-free wrinkling.

Further testing should be done to refine the range of U -values of the contacting materials and substrate thicknesses that define the wrinkling, wrinkling and cracking, and cracking zones. Moreover, the translation of these parameters into the use of atmospheric plasma for scale-up should be investigated.

5 Acknowledgements

We would like to express our sincere thanks to Zain Ahmed for his invaluable support and guidance over the duration of this study. Additionally, we extend our gratitude to the Polymers and Microfluidics Research Group.

References

- [1] Shabnam Raayai-Ardakani and Gareth H. McKinley. Drag reduction using wrinkled surfaces in high reynolds number laminar boundary layer flows, Sep 2017.
- [2] Zou F; Zhou H; Jeong DY; Kwon J; Eom SU; Park TJ; Hong SW; Lee J.: Wrinkled surface-mediated antibacterial activity of graphene oxide nanosheets.
- [3] Henniker Plasma. Plasma treatment of pdms for microfluidics.
- [4] Mary Graham and Nathaniel Cady. Nano and microscale topographies for the prevention of bacterial surface fouling. *Coatings*, 4(1), 2014. doi: 10.3390/coatings4010037.
- [5] Desmond van den Berg, Dalal Asker, Tarek S. Awad, Nicolas Lavielle, and Benjamin D. Hatton. Mechanical deformation of elastomer medical devices can enable microbial surface colonization. *Scientific Reports*, 13(1), 2023. doi: 10.1038/s41598-023-34217-5.
- [6] Manuela Nania, Fabrizia Foglia, Omar K. Matar, and João T. Cabral. Sub-100 nm wrinkling of polydimethylsiloxane by double frontal oxidation, Jan 2017.
- [7] Annabelle Tan, Luca Pellegrino, and João T. Cabral. Tunable phase gratings by wrinkling of plasma-oxidized pdms: gradient skins and multiaxial patterns, 2021.
- [8] Plasma etch, inc, plasma treatment, 2022. Accessed in Dec 2023.
- [9] Patrick Gamp;ouml;rnn and Sigurd Wagner. Topographies of plasma-hardened surfaces of poly(dimethylsiloxane), Nov 2010.
- [10] Tawfeeq Mohammed Salih. *Insulation materials: Fundamentals and Applications*. 04 2021.
- [11] Shadi Karazi, Inam Ahad, and K. Benyounis. *Laser Micromachining for Transparent Materials*. 12 2017. ISBN 9780128035818. doi: 10.1016/B978-0-12-803581-8.04149-7.
- [12] C, X. Yang, J. Gong, and et al. Enhanced thermal conductivity of polydimethylsiloxane composites with carbon fiber, Dec 2019.
- [13] Ailing Zhang and Yanxiang Li. Thermal conductivity of aluminum alloys-a review, Apr 2023.
- [14] Scientific Glass Laboratories Limited. Physical properties of borosilicate glass.
- [15] Georg Fantner. A brief introduction to error analysis and propagation - epfl.

Exploring the Influences of Impurities and Silica Nano-templates on Diglycine Crystallisation: A Comprehensive Study for Innovative Crystal Engineering

Samson Lai and Zen Wong

Department of Chemical Engineering, Imperial College London, U.K.

Abstract

Crystalline pharmaceutical peptides offer a plethora of advantages such as producing higher stability and purity products while allowing controlled drug release. However, the presence of impurities can hinder crystallisation processes, thus this study aimed to investigate the effect of glycine impurities on diglycine crystallisation. Although diglycine does not have specific therapeutic functions, it can be used as an intermediate reactant to generate complex peptides. In the first phase of this study, homogenous crystallisation experiments with various glycine impurity levels (0 to 2.5% w/w) were conducted at the constant initial supersaturation ratio of 1.35. Findings revealed that increasing glycine impurity level increased induction time and reduced crystal growth rate. In the second stage of the study, 1% w/w silica nanoparticles with 6nm pore size were employed to induce heterogenous nucleation. This led to a reduction in induction time for all glycine impurity levels, and by over 4-fold for the experimental run with 2.5% w/w glycine impurity. In addition, solid characterisation methods such as PXRD and microscopy were utilised to analyse the resulting solid products. It was found that homogenous crystallisation yielded α -diglycine, while the heterogenous case produced diglycine sesquihydrate. The solubility of the hydrate at 5°C was found to be lower at 87 mg/mL, thus generating approximately two-fold higher crystal yield at 7.5g in comparison to 4.0g for homogenous crystallisation. In conclusion, it was found that certain impurities greatly hinder crystallisation processes, but the presence of heterosurfaces can mitigate the increase in induction times and should be employed in industrial processes for improved cost-competitiveness.

Keywords: crystallisation; homogenous nucleation; heterogenous nucleation; impurities; nano-templates; peptides

1. Introduction

Therapeutic drugs often comprise peptides with molecular weights between 500-5000 Da. Since the first therapeutic peptide, insulin was synthesised, there have been over 80 peptide drugs approved worldwide. Demand for therapeutic peptides is expected to increase at a CAGR of 10.8% from 2023 to 2033, potentially reaching US\$ 106 billion in market size by 2033 [1]. The main techniques employed for peptide purification are chromatography methods such as RP-HPLC and ion-exchange chromatography, but these methods often require large volumes of solvent which generate significant liquid waste [2]. Alternative purification techniques include ultrafiltration, fractional precipitation, centrifugation and crystallisation, where crystallisation is mainly utilised as the final purification step to achieve >99% purity [3]. Crystalline pharmaceutical peptides offer numerous advantages such as higher stability for longer shelf life, possibility for high dosage due to higher purity levels, and allows controlled drug release [4]. However, crystallisation processes occur very slowly and require precise optimisation of temperature, pH, impurity levels, etc. Thus, it is imperative to investigate factors that might hinder crystallisation and ensure crystallisation processes occur as quickly as possible while yielding high quality crystal products.

Upstream peptide products often contain impurities such as solvents and unreacted reactants, thus it is crucial to understand the impact of impurities on key performance indicators of crystallisation such as crystallisation rate and crystal product quality. In order to reflect industrial conditions, this study investigates

the crystallisation of diglycine with the presence of glycine impurities up to 2.5% w/w as diglycine is produced via the reflux reaction between glycine and glycerol. Although diglycine does not have specific therapeutic functions on its own, it can be used as a short-chain intermediate reactant to produce complex therapeutic peptides or proteins [5]. Additionally, heterogenous crystallisation of diglycine was also investigated by employing 1% w/w silica nanoparticles with 6 nm pore diameter as hard templates to identify any potential improvements to the crystallisation process.

2. Background

2.1. Crystallisation

Crystals are solid materials comprising constituent atoms, ions or molecules that are orderly arranged in three dimensional arrays, exhibiting a repeating and symmetric pattern known as a crystal lattice. Angles between crystal faces of the same compound are identical and characteristic for that material. The driving force of crystallisation is supersaturation, which occurs when a solution contains a solute concentration that exceeds its equilibrium solubility for a specified temperature and pressure, thus surpassing its thermodynamically stability. A saturated solution can become supersaturated through two routes: the first is to increase the amount of solute dissolved in solution, and the second method is to induce a temperature change in solution through means of evaporation or cooling.

According to Classical Nucleation Theory (CNT), crystallisation occurs in two stages. Nucleation is the

initial step in crystallisation process where genesis of crystal nuclei occurs and is the rate limiting step. It occurs a period of time after the solution is left in supersaturated state, otherwise known as induction time, t_{ind} . Nucleation rate J is the inverse of induction time ($J = 1/t_{ind}$), and can be determined as follows:

$$J = A \cdot S \cdot \exp\left(-\frac{B}{\ln^2 S}\right) \quad (1)$$

where A is the pre-exponential constant dependant on nucleation kinetics, S is supersaturation ratio and B is a thermodynamic parameter relating to the interfacial energy between solid crystal and the solution, quantifying the energy barrier to nucleation. [6] As proven in Equation 1, the driving force for crystallisation is indeed the supersaturation ratio.

Crystal growth is the second step and occurs when solute particles adsorb onto kink sites of step lines on the crystal surface, the step lines then move across the crystal surface via step displacement. The types of growth regimes include continuous growth, spiral growth and surface nucleation. [7]

There are two types of nucleation mechanisms: primary and secondary. Primary nucleation occurs when crystal formation is driven by solution properties in the absence of crystals of the same material and can be further classified into homogeneous and heterogeneous routes. Homogeneous nucleation is when crystal formation is driven solely by supersaturation and contains no foreign particles or crystals of its own type in solution. Heterogeneous nucleation results from the presence of foreign particles or growth of a pre-existing surface present in the system, inclusive of mixing equipment, vessel surfaces, gas-liquid interfaces or heterogeneous templates as seen later in this report. Secondary nucleation occurs if crystals of the same material are present in solution and involves attrition where the existing crystals are broken up into smaller crystal sizes through contact or shear forces, increasing the number of crystal fragments available to act as nuclei and thereby increasing crystal formation rate.

Control of nucleation is critical in determining physical characteristics of solid products crystallised from solution, inclusive of crystal habit, morphology and particle size distribution. It is therefore paramount to understand existing nucleation mechanisms which are useful in predicting crystal properties: CNT and the recently adapted Two-Step Nucleation Model (2SN). The mechanism described by CNT attributes the genesis of nuclei to solute particles aggregating together, progressively forming clusters until a critical radius is reached, such that both free energy contributions for phase transformation and interfacial formation are in equilibrium. Upon reaching this equilibrium, the cluster is thermodynamically stable to act as a nucleus and enable subsequent crystal growth. CNT assumes spherical nuclei, and isotropic interfacial tension. In contrast, 2SN introduces an additional step in nucleation, described as follows: the first step is the formation of pre-nucleation nanodroplet clusters in the dense liquid phase which is rate-determining, and the second step involves the faster formation of crystal

nuclei within the nanodroplet clusters. In the context of 2SN, the mesoscopic liquid structures can also act as heterosurfaces, introducing template effects similar to heterogeneous nucleation. 2SN has provided breakthroughs in crystallisation processes by resolving underlying complexities within nucleation mechanisms, most notably clarifying why experimental nucleation rates were several orders lower than the rates predicted with CNT. By considering the second step of 2SN on its own and omitting the rate-determining step, there is a close match between predicted and experimentally measured nucleation rates. This recent advancement accentuates the need for a more thorough comprehension of the mechanisms that strongly influence nucleation, namely the presence of impurities and surface chemistry of solute particles. [8]

2.2. Effect of Impurities on Crystallisation

The presence of impurities in solution can influence crystallisation rate either positively or negatively. Impurities that enhance crystallisation processes are also known as templates; where hard templates are rigid and insoluble while soft templates are soluble. Link and Heng have investigated employing various amino acids as soft templates to enhance insulin crystallisation, where L-arginine and L-leucine were found to improve nucleation rates due to favourable intermolecular interactions between insulin and the amino acids [9]. However, Keshavarz, L. *et al.* has found that certain impurities such as 4-nitrophenol and 4'-chloroacetanilide acted as nucleation inhibitors, effectively slowing nucleation rate J [10]. Therefore, this study aims to investigate the effect of glycine, the main impurity in diglycine production on its crystallisation process. Generally, the presence of hard templates will enhance nucleation via favourable intermolecular interactions between solutes and heterosurfaces as seen in the study by Verma, V. *et al.* [11], leading to the second phase of this study which investigates the potential of silica nanoparticles on improving induction times of diglycine crystallisation.

2.3. Surface Chemistry

Verma, V. *et al.* has investigated the crystallisation of glycine and diglycine in the presence of glass beads as heterosurfaces and found that the pre-exponential constant of nucleation has increased significantly due to complimentary hydrogen bonds between solutes and the heterosurfaces, contrary to traditional CNT that heterosurfaces reduce the interfacial energy between solute molecules and the solution. Each diglycine molecule contains 2 hydrogen bond donor (HBD) and 3 hydrogen bond acceptor (HBA) sites; where the lifetime of hydrogen bonds between solutes and heterosurfaces is around 30 ns, astronomically greater than the time required to attach a diglycine molecule to a growing crystal (2.01 ps). This allows the solute to adsorb to the heterosurface for sufficient time to allow subsequent solute molecules to attach and form stable nuclei. [12]

3. Materials & Methodology

3.1. Materials

Diglycine (Digly, $\geq 99\%$ by titration) and glycine ($\geq 99\%$ by HPLC) were both procured from Sigma-Aldrich Co. Ltd as the main peptides used in this investigation. The salts KH_2PO_4 (purity $\geq 99\%$, white crystals, anhydrous form) and $\text{C}_6\text{H}_{13}\text{O}_3\text{SNa}$ (purity $\geq 98\%$, white crystals, anhydrous form) were also supplied by Sigma-Aldrich Co. Ltd. Porous silica-OH particles (particle diameter $50\mu\text{m}$, pore size 6nm) purchased from SiliCycle were used in the heterogeneous experiments. As for solvents, deionised water was attained from the analytical laboratory.

3.2. Diglycine Solubility Test

100 mg/mL diglycine solution was first prepared by weighing the required masses and dissolving in deionised water using a magnetic stirrer and water bath. 0.5 mL samples of diglycine (10 mg/mL, 20 mg/mL, 30 mg/mL, 40 mg/mL, 50 mg/mL) were then prepared via dilution. The samples were utilised to generate a calibration curve, depicted in Fig.1 below using the Shimadzu LC-2030C High-Performance Liquid Chromatography (HPLC) system, which uses UV-vis spectra for analysis. It made use of hydrophobicity differences to separate and identify various components in samples. 15mM $\text{C}_6\text{H}_{13}\text{O}_3\text{SNa}$ solution was used as the mobile phase with hydrophobic C18 silica packing as the stationary phase. 50mM KH_2PO_4 was used as the buffer solution to maintain acidic condition at pH of 2 and ensure positive charge of amino groups on analytes. $\text{C}_6\text{H}_{13}\text{O}_3\text{SNa}$ forms ion-pairs with the analytes through interactions between negatively charged sulfonate groups and positively charged amino groups [13], leading to the formation of hydrophobic complexes which can interact with the C18 stationary phase. The additional amino acid in diglycine increases its hydrophobicity, thereby allowing diglycine to exhibit a longer retention time, separating glycine and diglycine into distinct peaks to be analysed.

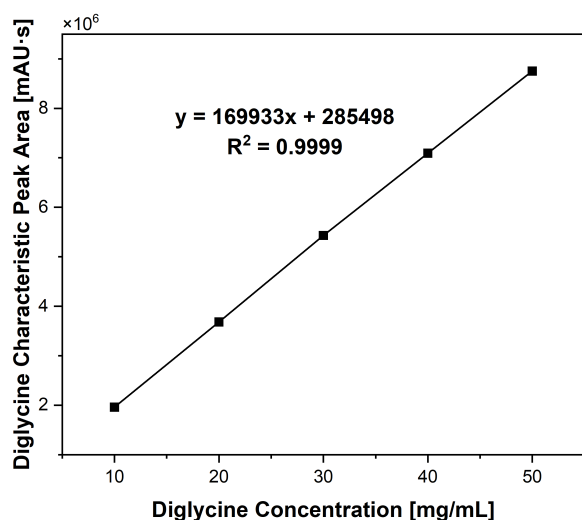


Figure 1. HPLC Calibration Curve to Determine Diglycine Concentrations from Characteristic Peak Areas

For the solubility test, excess diglycine solids were dissolved in 1 mL of deionised water containing glycine impurities of 0 mg/mL, 1 mg/mL and 5 mg/mL. The samples were then placed in a shaker with temperatures equilibrated at 5°C , 15°C and 25°C for 48 hours and allowed to settle for 2 hours. The supernatant from the samples were pipetted into vials and diluted by the factor of 1/5 to be analysed by the HPLC system as the diglycine characteristic peak area curve plateaus at high concentrations. The generated calibration curve shown in Fig.1 was used to determine diglycine concentrations in the vials, which were multiplied by 5 to reflect true diglycine concentrations in the samples. Diglycine solubility curves between 5°C and 25°C with various glycine impurity levels could then be generated as shown in Fig.2 below, showing a linear relationship between solubility and temperature for all impurity levels. Note that repeats of diglycine solubility test at 15°C in the absence of glycine did not show consistent concentrations, thus it was elected to use the literature concentration of 173 mg/mL from the study by Guo, M. *et al.* instead [14]. As shown in Fig.2, diglycine solubilities between 5°C and 25°C are relatively constant with various glycine impurity levels, and the constant diglycine solubility at 5°C (148 mg/mL) is the theoretical end-point concentration of experimental runs, allowing the determination of theoretical diglycine crystal yields.

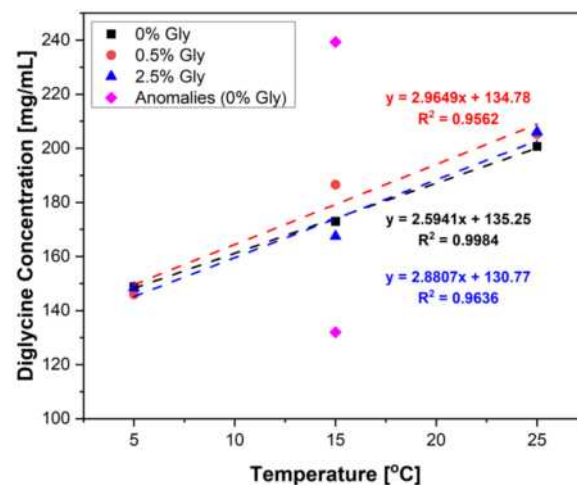


Figure 2. Diglycine Solubility Curves with Various Glycine Impurity Levels

3.3. Set-up for Homogeneous Nucleation Experiments

The experimental setup consisted of Mettler-Toledo EasyMax 102, which were able to fit two glass reactors of 50mL and 100mL. The EasyMax allowed for precise control over experimental variables inclusive of reactor temperature, stirring speed and heating/cooling rate. A HC-22 4-blade pitch-down impeller was fitted to an overhead stirrer and inserted into the vessel head, with the impeller immersed in solution. A temperature probe, as well as a Mettler-Toledo ReactIR15 (FTIR) probe were also inserted into the vessel head, taking care to place both probes well above the impeller to avoid any collision. Both probes allowed for automated in situ measurements of temperature and concentration; thus,

eliminating the need for extracting in-process samples for measurements.

Prior to every experimental run, the IR probe was calibrated to obtain a spectrum for the background atmosphere; doing so ensured the probe head was clean, measuring only the sample and not any contaminants. Furthermore, the FTIR required liquid nitrogen refilling every 24 hours to cool the internal optical fibres to reduce thermal noise and improve apparatus sensitivity, while preventing potential malfunctions from overheating.

For every experimental run, a saturated 80mL diglycine solution was prepared at 25°C in a 100mL reagent bottle, with the concentration of 200 mg/mL as per previously determined solubility data. To negate solution volume increase upon solute dissolution, the solution was first prepared to 60mL with appropriate diglycine mass added, then transferred to a water bath at 35°C and mixed with a magnetic stirrer bar. Once most of the diglycine solids had dissolved, the solution was then topped up to 80mL with deionised water and further mixed until complete dissolution. Subsequently, the solution was transferred to a cleaned 100mL reactor vessel with a blunt needle syringe fitted with a 200nm Nylon membrane syringe to filter out particulate matters and contaminants as a preventive measure against undesired secondary nucleation. The 100mL reactor was then fitted with its vessel head, the overhead stirrer as well as both temperature and IR probes, clamped and placed into the EasyMax.

In each run with glycine impurity presence, glycine mass was measured according to the concentration of interest and added to the saturated diglycine solution prior to placement into the EasyMax.

3.4. Script for Crystallisation Control

The use of iControl software allows for controlled crystallisation process, with a script outlining a sequence of steps to vary operating conditions of the EasyMax apparatus over time. Initially, the 80mL solution was stirred up to 300rpm over 10 seconds and heated up 40°C over 10 minutes. The solution was maintained at 40°C for 60 minutes to allow the system to equilibrate and ensure complete solute dissolution. The solution was then cooled to 5°C over 5 minutes and maintained until experiment completion. With the solution supersaturated at 5°C, the induction time marking the onset of nucleation was obtained based on diglycine concentration readings obtained from the FTIR. The process would continue to run allowing for diglycine crystal growth until no further changes in diglycine concentration in solution were observed, upon which the solution would be reheated to 40°C to ensure complete dissolution of solutions in preparation for repeats if necessitated. Once experiments were completed, the crystals formed would be extracted and filtered for analysis, and the solution sampled to validate theoretical endpoint concentration with HPLC analysis.

3.5. Interpretation of the Infrared Spectrum for Transmittance

The FTIR was calibrated using 5 mL of 200 mg/mL diglycine solution diluted successively to 50 mg/mL. The IR spectra intensity from FTIR was recorded for each concentration, and the assumption of linear relationship between IR spectra intensity and concentration was subsequently validated. As a result, linear interpolation could be used using initial and final diglycine concentrations obtained from solubility data in Section 3.2 to translate IR spectra intensity curves into diglycine concentration curves.

3.6. Induction Time Determination for Homogeneous Nucleation Experiments

Fig.3 below shows the determination of induction time t_{ind} for crystal nuclei formation from the concentration curve of diglycine with 0.5% w/w glycine impurity. The induction time is defined as the time period between the initiation of crystallisation and the formation of nuclei; thus the tangent method could be used to determine the induction time as it accurately determines the point where solution concentration falls. The gradient of the curve was first evaluated at 30 minutes intervals to find the steepest point, a tangent line was then plotted at the steepest point. The induction time was determined at the interception point between initial concentration and tangent line.

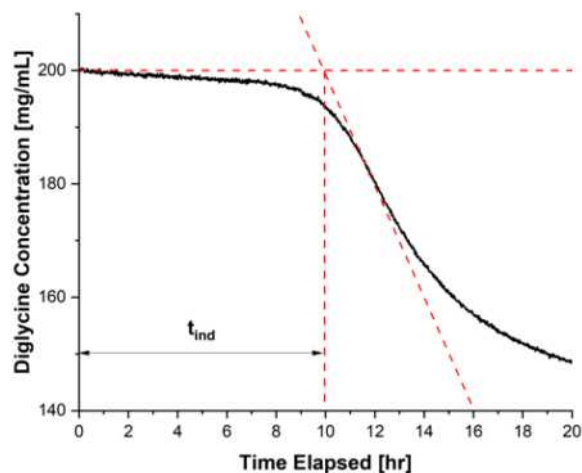


Figure 3. Induction Time Determination Using the Tangent Method (Example Shown for 0.5% Glycine Impurity, $t_{ind} = 10$ hours)

3.7. Set-up for Heterogeneous Nucleation Experiments

80mL solutions were prepared with the same diglycine and glycine concentrations investigated in the homogeneous nucleation experiments (Section 3.3), and 1% w/w silica nanoparticles with 6nm pore size were added to the solutions. Using the maximum mass of diglycine crystals obtained from solution (m_{crys}) determined from homogeneous nucleation experiments, initial diglycine concentration (c_0) of 200 mg/mL and final concentration at 5°C (c^*), as well as solution volume (V), the silica mass loading m_L was determined through the following equations:

$$m_{crys} = (c_0 - c^*) \cdot V \quad (2)$$

$$m_{crys} = (200.00 - 148.00) \cdot 80 = 4160mg \quad (3)$$

$$m_L = 0.01 \times m_{crys} = 41.6mg \quad (4)$$

Heterogenous nucleation experiments were initially planned to follow the same setup as the one used for homogenous nucleation; it was however modified due to the FTIR malfunctioning during this study. As such, a camera set-up was used to visually observe changes in solution appearance occurring in the reactor as shown in Fig. 4, capturing images at 5-minute intervals. Induction time in this case was evaluated by analysing reactor temperature profiles to obtain the relative start time of the experiment, and evaluating the time taken for the solution to go from clear to cloudy once the nucleation of diglycine solid composites had occurred.



Figure 4. Camera Imaging of Vessel, with Solution Before and After the Occurrence of Nucleation

3.8. Powder X-Ray Diffraction (PXRD)

Powder X-ray diffractograms were recorded for diglycine composites formed in the absence and presence of silica, which were vacuum filtered and dried after each crystallisation experiment. A PANalytical Empyrean diffractometer with a copper radiation source ($\lambda = 1.541 \text{ nm}$) at 40 mA and 30 kV was used in conjunction with XPRT-PRO diffractometer system. Scans were performed at a rate of $0.107815^\circ 2\theta \text{ min}^{-1}$ in the range from 9° up to 40° , generating patterns of intensity (a.u.) against 2θ ($^\circ$) to identify the diglycine polymorphs formed by comparing them with existing polymorphic patterns sourced from literature.

3.9. Microscopic Imaging

Diglycine crystals obtained from crystallisation experiments were visually inspected with Olympus CX-41 microscope under the magnifications of 5x and 10x. Images from the microscope were captured using GT Vision GXCAM HiChrome MET for crystal habit, size and morphology analysis, and compared against PXRD results to verify solid characteristics.

4. Results and Discussion

4.1. Diglycine De-supersaturation (S-Shaped) Curves

Crystallisation experiments commenced with the identical initial diglycine concentration c_0 of 200 mg/mL. Supersaturation ratio S could then be determined with Equation 5 using c_0 and saturated diglycine concentration at 5°C , c^* obtained from the solubility curves Fig.2 in Section 3.2 above.

$$S = \frac{c_0}{c^*} \quad (5)$$

As diglycine solubility curves are identical with various glycine impurities, with the diglycine concentration of 148 mg/mL at 5°C , all experimental runs commenced with the same initial supersaturation ratio S of 1.35. In order to determine crystallisation rates, solution concentration c was measured continuously via the FTIR probe and used to calculate % de-supersaturation which is essentially the percentage of crystallisation completion.

$$\text{Desupersaturation \%} = (c_0 - c)/(c_0 - c^*) \quad (6)$$

Fig.5 shows the de-supersaturation S-shaped curves of various glycine impurities starting from 0% where the solution is supersaturated at 5°C and ending at 100% where the solution reaches saturation. As shown in Fig.5, crystallisation experiments take longer to complete with increasing glycine impurity levels. This is attributed to both the increase in induction times and the reduction in crystal growth rates, further comparisons and explanations can be found in Sections 4.2 and 4.3.1 below. From these curves, it can be deduced that the presence of impurities such as glycine can hinder the crystallisation process significantly.

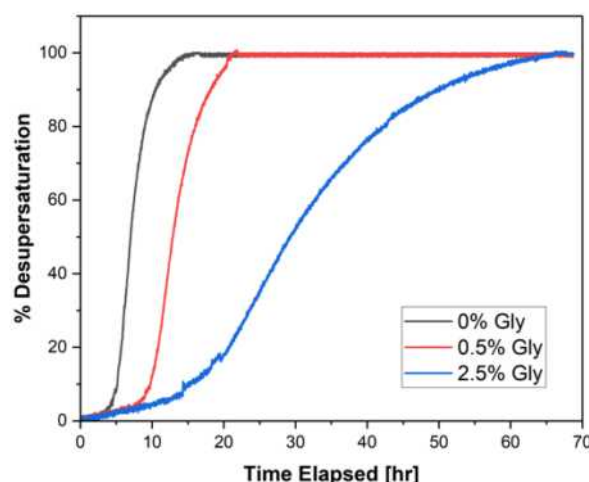


Figure 5. Diglycine De-supersaturation (S-shaped) Curves with Various Glycine Impurities from 0% to 2.5% w/w

4.2. Effect of Glycine Impurities on Crystal Growth Rate

Fig.6 demonstrates the effect of glycine impurities on the crystal growth step of homogenous nucleation crystallisation. Growth rates were determined by normalising the linear regions' gradients of the S-shaped curves with the gradient with 0% impurities (G/G_0).

The increase in glycine impurity level from 0% to 2.5% w/w appears to cause over 80% reduction in growth rate, this can also be observed in Fig.5 where the gradients of the linear regions in S-shaped curves decrease with increasing glycine impurity levels. A possible explanation for the reduction in crystal growth rate is deduced in the study by Cabrera and Vermilyea, which proposed that impurity species present in the

solution would adsorb onto kink sites of step lines, pinning step displacement and forcing it to be curved, thereby reducing step advancement velocity and subsequently crystal growth rate [15].

However, the evidence pointing towards the reduction in growth rate is not concrete; an alternative explanation for the reduction in gradients of S-shaped curves is that slower nuclei formation inadvertently led to less solutes exiting the liquid phase and contributing to crystal growth. Thus, the gradients of S-shaped curves may appear to decrease but the growth rate of each specific crystal might not be significantly different.

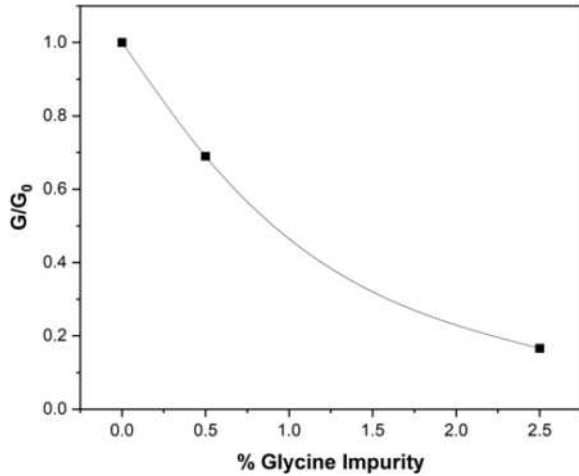


Figure 6. Normalised Growth Rates with Varying Glycine Impurity Levels from 0% to 2.5% w/w

4.3. Effect of Glycine Impurities on Induction Time

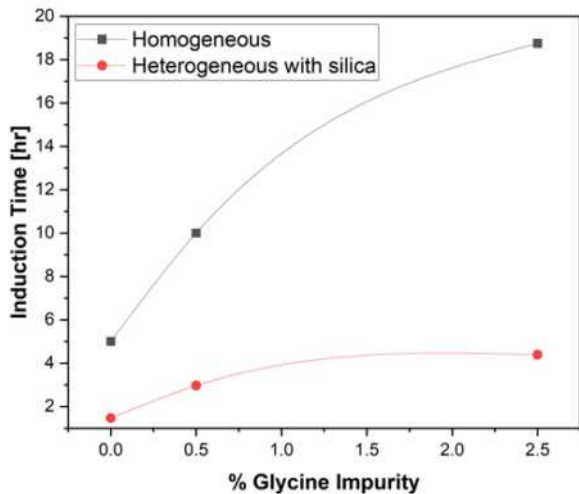


Figure 7. Diglycine Crystallisation Induction Time with Varying Glycine Impurity from 0% to 2.5% w/w

4.3.1. Homogeneous Case

From Fig.7, it can be observed that pure diglycine with the initial concentration of 200 mg/mL had an induction time of 5 hours. By introducing 0.5% w/w glycine impurity, the induction time doubles to 10 hours and increasing to 2.5% w/w glycine impurity causes the induction time to again increase by around two-fold to 19 hours. From this observation, it can be deduced that a miniscule amount of glycine impurity induces a

significant impact on induction time of diglycine crystallisation, thus the presence of impurities can be incredibly detrimental to crystallisation processes. The effect of impurities on nucleation kinetics is very hard to be explained due to the lack of understanding on crystal nucleation. Keshavarz, L. *et al.* investigated the effect of impurities on paracetamol crystallisation and found that the presence of impurities exhibited negligible effect on the interfacial energy of the crystals, but instead significantly reduced the pre-exponential constant of nucleation rate J in Equation 1. One possibility is that impurities might form interactions with nucleation interfaces, reducing the amount of nucleation sites for key solutes to bind to. Another possibility is the increase in activation energy barrier due to additional energy required to eliminate impurities from nuclei. [16]

4.3.2. Heterogeneous Case

6nm pore size silica nanoparticles at 1% w/w were introduced to identical solutions as the homogenous case in order to investigate the potential of hard templates in mitigating the increase in induction times caused by the presence of glycine impurities. Verma, V. *et al.* has investigated the effect of porous silica with various pore sizes on induction times of diglycine crystallisation and determined that 6 nm is the ideal pore size as it closely matches with diglycine cluster size [17]. Thus, it was elected to utilise 6 nm silica nanoparticles to ensure that a significant effect on induction times could be observed. Fig.7 shows the induction times of diglycine crystallisation with the presence of glycine impurities and silica nanoparticles. The general trend still remains consistent with the homogenous case, where the absence of glycine impurity saw the induction time of 1.5 hours and increasing glycine impurity to 2.5% w/w increased the induction time to 4.5 hours. However, the timescale was much smaller with the presence of silica nanoparticles; for the case of pure diglycine, there was a reduction of 70% in induction time and for the case with the presence of 2.5% w/w glycine impurity, there was a reduction of 76% in induction time. This result can be quantified using the improvement factor (i), where:

$$i = \frac{\text{Homogenous } t_{ind}}{\text{Heterogenous } t_{ind}} \quad (7)$$

Table 1. Improvement Factors of Various Glycine Impurity Levels with Heterogenous Crystallisation

Glycine Impurity Level (w/w)	Homo t_{ind} [hr]	Hetero t_{ind} [hr]	Improvement factor (i)
0%	5.0	1.5	3.33
0.5%	10.0	3.0	3.33
2.5%	19.0	4.5	4.22

The improvement factors show that the effect of silica nanoparticles in reducing induction time is slightly more profound with higher glycine impurity concentration. It was also noted that in the case of heterogenous nucleation, the induction time curve plateaus earlier, suggesting that further increase in

impurity concentration will no longer cause an increase in induction time. Contrary to the CNT, Verma, V. *et al.* suggested that the presence of glass beads has insignificant influence on the interfacial energy of diglycine crystals, but instead increased the pre-exponential constant of diglycine nucleation kinetics equation by two-fold, leading to increased nucleation rate. This is due to hydrogen bond donor (HBD) and acceptor (HBA) sites on diglycine molecules forming hydrogen bonds with heterosurfaces which possess astronomically longer lifetime than the time required to attach a single solute to a growing crystal, allowing the heterosurfaces to act as stable nucleating surfaces and induce a reduction in induction times. [18]

4.4. Solid-State Characterisation

Fig.8 presents a solid-state analysis pertaining to

diglycine solid products formed from the crystallisation experiments performed in water with glycine impurities of 0% w/w and 2.5% w/w for both homogeneous and heterogeneous crystallisation. Upon inspection of both PXRDs for the homogeneous experiments in Fig. 8(A), most peaks are present at similar 2θ values and correspond to α -diglycine, confirmed after comparisons against a PXRD from a publication by Verma, V. *et al.* [17]. It can therefore be inferred that morphology of the diglycine crystals formed from homogeneous nucleation remains unaffected even in the presence of impurities such as glycine. Variance in relative peak intensities were also observed, attributed to the non-uniformity in the powder sample and could be mitigated in future work by further grinding the powder into fines. Nevertheless, small discrepancies in peak intensities do not have any effects on diglycine morphology.

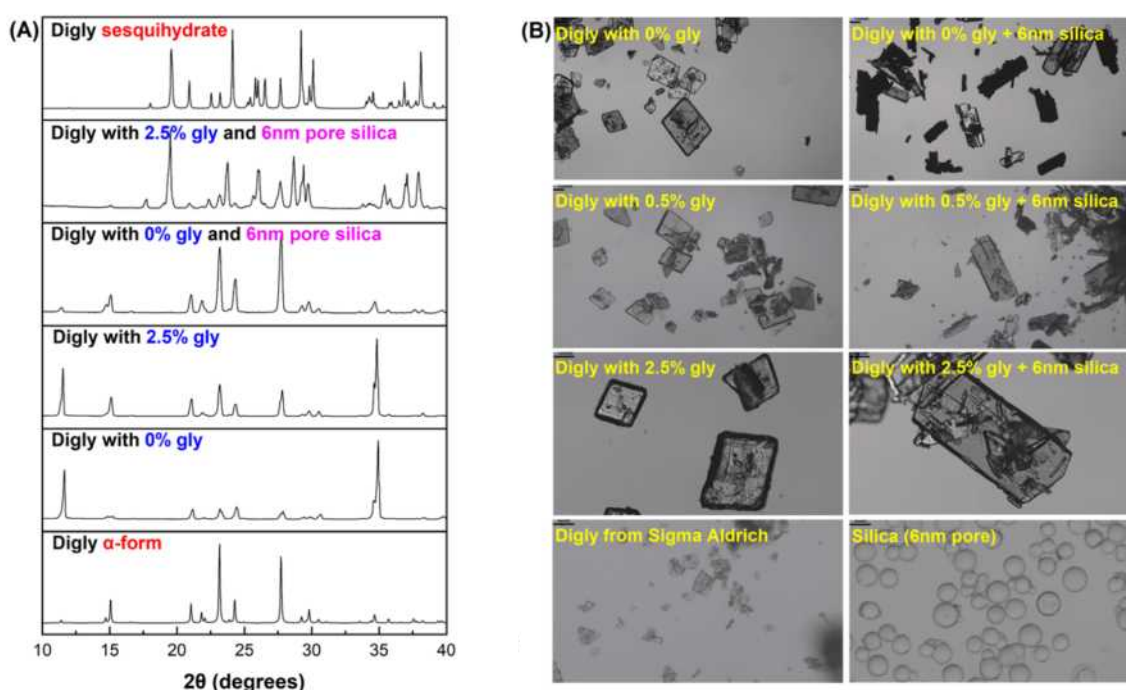


Figure 8. (A) Powder X-ray Diffraction Spectra of the Diglycine Solid Products Isolated from Experiments as Labelled (in blue and magenta), along with Patterns for α -diglycine and Diglycine Sesquihydrate from Literature (in red); (B) Microscopy images of Diglycine Procured from Sigma Aldrich, Porous Silica and Isolated Solids from Crystallisation Experiments.

However, it was found that the diglycine solid formed in the presence of silica and 2.5% glycine impurity had significant peaks at 18° and 19.5° on its PXRD, which were not found on α -diglycine. Instead, the solids were identified to be diglycine sesquihydrate (1.5 H_2O molecules per diglycine molecule) after comparisons with existing PXRDs obtained from a study by Drebuschak, T.N. *et al.*, which investigated the effects of cooling on intermolecular hydrogen bonds and molecular confirmations within anhydrous α -diglycine crystals as well as the hydrated form [19].

While diglycine crystals formed in the presence of silica and 0% w/w glycine impurity did not match the PXRD patterns of the sesquihydrate and resembled more of the anhydrous α -form, this was eventually confirmed to have originally been a sesquihydrate as well using the

micrographs of isolated diglycine solids depicted in Fig.8 (B). The diglycine crystals from all homogeneous nucleation experiments appear to be more rhombical in shape, with sharp and well-defined edges and corners; whereas for diglycine crystallised through heterogeneous nucleation, the solids are more elongated and irregular in shape. As the sample crystallised in 2.5% w/w glycine impurity is likely confirmed to be diglycine sesquihydrate through PXRD, it is thus likely that the remaining samples crystallised in the presence of silica are diglycine sesquihydrates as well based on the micrographs. Furthermore, for heterogeneous experiments, the solid samples crystallised in the presence of 0.5% w/w and 2.5% w/w glycine impurity levels are semi-transparent; this is however not the case for the crystals with 0% w/w glycine impurity as a significant number of crystals are opaque. A possible

cause for the appearance of the latter solid could be its inadequate storage (kept in a capped glass vessel and left in a fume hood), where water may have evaporated out of the solid over time due to the cap being loose, thus causing the crystals to lose their semi-transparent nature. This therefore reflects the results depicted in Fig.8 as the PXRD and microscopic imaging were only performed after all the crystallisation experiments were completed.

Further evidence pointing towards diglycine sesquihydrate as the crystal formed during heterogenous nucleation is the spike in solution temperature around the time when nucleation occurs. It is postulated that the formation of hydrates in porous media is more exothermic than anhydrous crystals due to stronger bonds formed between solutes and water molecules releasing more energy, thus inducing a temperature increase in the solution. This phenomenon can be seen in Fig.9 where the temperature reading of 0.5% w/w glycine impurity experimental run experienced a spike at t_{ind} of 3 hours above 5°C in the presence of silica nanoparticles, but in the case of homogenous nucleation no temperature increase could be observed. This observation can also be seen in the study by Zhang, L. *et al.* where temperature increase could be observed during the formation of THF crystal hydrates, further highlighting the exothermic nature of hydrate formation. [20]

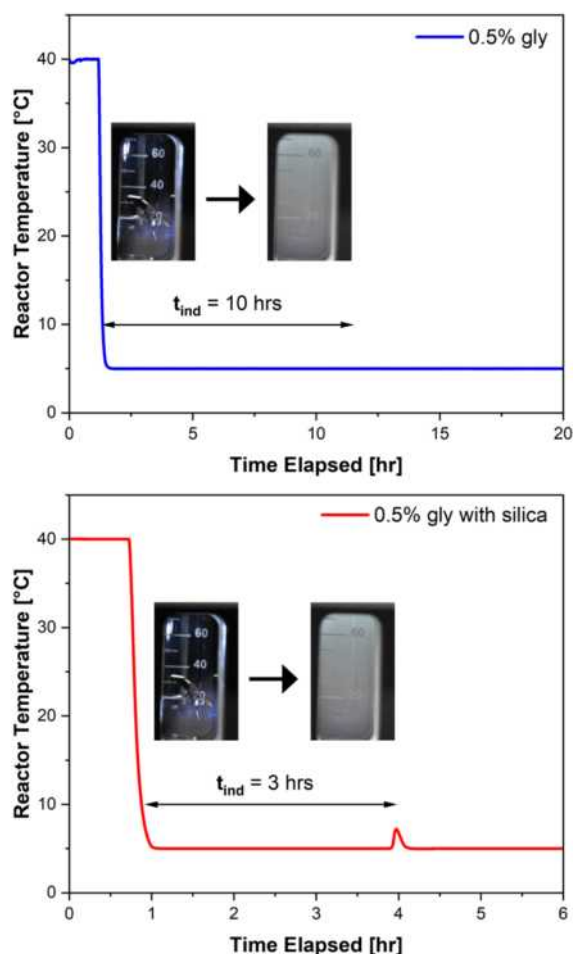


Figure 9. Diglycine with 0.5% w/w Glycine Impurity Solution Temperature Profiles, with and without the Presence of Silica Nanoparticles

4.5. Crystal Yield

The theoretical crystal yields y of α -diglycine were evaluated using the total solution volume V , initial diglycine concentration c_0 of 200 mg/mL and final diglycine concentration c^* corresponding to diglycine solubility at 5°C using Equation 8 below.

$$y = V(c_0 - c^*) \quad (8)$$

As shown in Fig.2 in Section 3.2, final diglycine concentrations of homogenous nucleation at 5°C are identical at 148 mg/mL with various glycine impurity concentrations. The theoretical crystal yield of α -diglycine is therefore constant at approximately 4.0g.

For experiments with the addition of silica nanoparticles, the supernatant was extracted at the end of each run and analysed with HPLC, yielding an identical concentration of around 87 mg/mL across all glycine impurity concentrations. As postulated in Section 4.4 above, heterogenous nucleation in the presence of silica nanoparticles produced diglycine sesquihydrate, thus the measured final concentration should correspond to the solubility of the hydrate at 5°C as the experiments were left to run for sufficient time to achieve steady state final concentrations. The yields of diglycine sesquihydrate crystals were calculated using Equation 8 and they were found to be constant at approximately 9.0g, deviating by more than two-fold from the yield of α -diglycine. Upon discounting the water mass within the hydrate using Equation 9 below, diglycine crystal yield is determined to be 7.5g, which is still under two-fold higher than homogenous crystallisation forming α -diglycine.

$$\begin{aligned} & \text{Diglycine mass without } H_2O \\ &= \left(\frac{MW_{digly}}{MW_{digly} + 1.5MW_{H_2O}} \right) \times y_{digly\ hyd} \end{aligned} \quad (9)$$

Where MW_i corresponds to the molecular mass of species i .

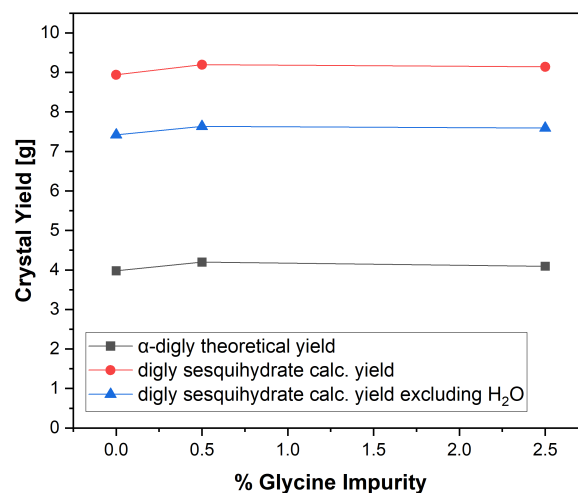


Figure 10. Crystal Yields for Homogeneous and Heterogeneous Nucleation Experiments with Varying Glycine Impurity Levels from 0 to 2.5% w/w

4.6. Diglycine Sesquihydrate Solubility

In order to validate the assumption that the measured final diglycine sesquihydrate concentration of 87 mg/mL corresponds to its solubility at 5°C, a solubility test was performed at 25°C, 15°C and 5°C with 0%, 0.5% and 2.5% w/w glycine impurity levels using the identical method as Section 3.2 above for α -diglycine. As shown in Fig.11 and Table.2 below, the solubilities of diglycine sesquihydrate are linear with temperature and possess negligible differences with various glycine impurity levels. Therefore, it can be concluded that glycine impurity level does not have an impact on the final diglycine sesquihydrate crystal yield obtained from heterogenous crystallisation. It can also be inferred that the solubilities of diglycine sesquihydrate are consistently lower than α -diglycine at all glycine impurity levels and temperatures between 5°C and 25°C. This lower solubility is due to higher thermodynamic stability of hydrates in comparison to their anhydrous forms [21].

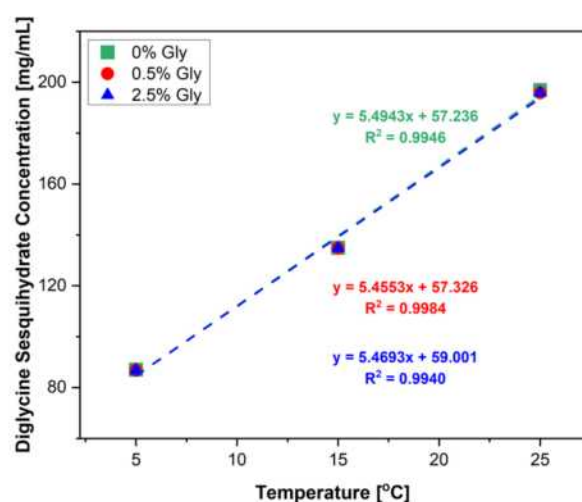


Figure 11. Diglycine Sesquihydrate Solubility Curves with Various Glycine Impurity Levels

Table 2. Measured Diglycine Sesquihydrate Solubilities with Varying Temperatures and Glycine Impurity Levels

Temp. [°C]	Glycine Impurity Level (w/w)		
	0%	0.5%	2.5%
5	87.05	86.78	88.81
15	134.97	134.81	136.13
25	196.94	195.88	198.19

5. Conclusion

The main parameter explored in this study is the effect of various glycine impurity levels (0 to 2.5% w/w) on key performance indicators of diglycine crystallisation such as crystallisation rate and crystal product yield. It was found that increasing glycine impurity level induces significant increase in induction time and reduction in growth rate, thereby slowing the entire crystallisation process. Therefore, it is confirmed that the presence of certain impurities can greatly hinder crystallisation, and it is imperative to eliminate impurities as much as possible prior to the crystallisation step using methods such as distillation and chromatography. It was also noted that crystal yield remains consistent at around 4.0g

with various impurity levels, owing to the constant diglycine solubility at 5°C. Potential benefits of employing 6nm pore diameter silica nanoparticles at 1% w/w loading as hard templates for diglycine crystallisation were also investigated; it was found that upon addition of silica nanoparticles, induction times of crystallisation experiments reduced significantly. This reduction was observed for all glycine impurity levels and by over four-fold in the case with 2.5% w/w glycine. However, the general trend of increasing induction time with the increase in impurity level still holds true.

Additionally, solid crystal products from the study were characterised using powder X-ray diffraction (PXRD) and light microscopy. Interestingly, it was observed that crystallisation experiments with silica nanoparticles generated solid products corresponding to diglycine sesquihydrate for all glycine impurities, where the formation of hydrates led to under two-fold higher crystal yield of 7.5g due to the lower solubility of diglycine sesquihydrate (87 mg/mL). Thus, it can be concluded that the addition of silica nanoparticles as hard templates for diglycine crystallisation is incredibly beneficial, improving both crystallisation rate and final crystal product yield.

6. Outlook

Further work can be conducted to gain a more comprehensive understanding on peptide crystallisation processes in the presence of impurities. Firstly, the experiments with heterogeneous silica-OH nanoparticles should be reconducted with solution concentrations continuously measured using FTIR, which malfunctioned in the midst of the study preventing S-shaped curves to be generated for heterogenous crystallisation runs. This led to the use of camera imaging which caused possible discrepancies between the different methods of deducing induction times. Reconducting the experiments using FTIR will generate greater accuracies for induction time comparisons along with allowing growth rates to be determined for the heterogenous crystallisation case, subsequently allowing comparisons of growth rates to be made between the two nucleation methods. Furthermore, it was postulated in Section 4.3.2 that the induction time curve in the presence of silica nanoparticles plateaus above 2.5% w/w glycine impurity level, thus further work should be conducted with higher glycine impurity levels in order to verify this claim.

The influence of impurities on other peptides can also be investigated; for instance, the impact of glycine impurities on engineering homopeptide crystals comprising longer glycine chains (e.g. triglycine, tetraglycine). Silica nanomaterials at different weight loadings or different functional groups can also be explored to determine their efficiencies in negating unfavourable effects from impurities in peptide crystallisation processes.

A significant proportion of the crystallisation experimental runs were performed without repeats due to time constraints, consequently reducing their reliability. It would be ideal to conduct repeats for all

experimental runs in this study and perform appropriate statistical analysis to obtain statistical metrics such as standard deviation for verifying the accuracy of reported findings.

Currently, there are little to no reported studies that highlight key factors leading to the formation of diglycine sesquihydrate through cooling crystallisation. Studies can be conducted to investigate possible conditions that lead to higher selectivity towards hydrated peptide formation; for instance, varying silica nanoparticle properties as aforementioned, or exploring other types of hard templates. Examining the recovery and recyclability of templates is also crucial in evaluating their applicability in scaled-up peptide crystallisation processes to maximise economic potential and minimise environmental implications.

Acknowledgements

The authors of this report would like to extend their sincerest gratitude to the members of the Heng Research Group for providing continuous support throughout this study. In particular, they are especially grateful to have been under the tutelage of Enshu Liang and Vivek Verma, and truly appreciate the opportunity to contribute to their research on investigating peptide behaviour and pioneering innovative methods towards improving peptide crystallisation processes.

References

- [1] Wang, L. *et al.* (2022) *Therapeutic Peptides, current applications and future directions*. Sig Transduct Ther 7, 48. [\[CrossRef\]](#)
- [2] Insuasty Cepeda, D.S. *et al.* (2019) *Synthetic Peptide Purification via Solid-Phase Extraction and Gradient Elution: A Simple, Economical, Fast, and Efficient Methodology*. Molecules, 24, 1215 [\[CrossRef\]](#)
- [3] Mehta, A. (2019) *Downstream Processing for Biopharmaceuticals Recovery*. In: Arora, D. *et al.* Pharmaceuticals from Microbes. Environmental Chemistry for a Sustainable World, vol 26. Springer, Cham. [\[CrossRef\]](#)
- [4] Basu, S. *et al.* (2005) *Protein crystals for the delivery of biopharmaceuticals*. Expert Opinion on Biological Therapy, 4:3, 301-317 [\[CrossRef\]](#)
- [5] ChEBI. (2015) *CHEBI:17201 – glycylglycine* [\[CrossRef\]](#)
- [6] Keshavarz, L. *et al.* (2019) *Influence of Impurities on the Solubility, Nucleation, Crystallization, and Compressibility of Paracetamol*. Cryst. Growth Des. 19, 4193-4201 [\[CrossRef\]](#)
- [7] Carbera, N. Vermilyea, D.A. (1958) *Growth and Perfection of Crystals*, Eds. R. H. Doremus. *et al.* Wiley, New York, p. 393
- [8] Erdemir, D. *et al.* (2009) *Nucleation of Crystals from Solution: Classical and Two-Step Models*. Acc. Chem. Res. 42, 5, 621-629 [\[CrossRef\]](#)
- [9] Link, F. Heng, J. (2021) *Enhancing the crystallisation of insulin using amino acids as soft-templates to control nucleation*. CrystEngComm, 2021, 23, 3951 [\[CrossRef\]](#)
- [10] Keshavarz, L. *et al.* (2019) *Influence of Impurities on the Solubility, Nucleation, Crystallization, and Compressibility of Paracetamol*. Cryst. Growth Des. 19, 4193-4201 [\[CrossRef\]](#)
- [11] Verma, V. *et al.* (2021) *Studying the impact of the pre-exponential factor on templated nucleation*. Faraday Discuss., 2022, 235, 199 [\[CrossRef\]](#)
- [12] Verma, V. *et al.* (2021) *Studying the impact of the pre-exponential factor on templated nucleation*. Faraday Discuss., 2022, 235, 199 [\[CrossRef\]](#)
- [13] TCI Chemicals. (no date) *Ion-Pair Reagents for HPLC*. [\[CrossRef\]](#)
- [14] Guo, M. *et al.* (2022) *The effect of chain length and side chains on the solubility of peptides in water from 278.15K to 313.15K: A case study in glycine homopeptides and dipeptides*. Elsevier B.V [\[CrossRef\]](#)
- [15] Carbera, N. Vermilyea, D.A. (1958) *Growth and Perfection of Crystals*, Eds. R. H. Doremus. *et al.* Wiley, New York, p. 393
- [16] Keshavarz, L. *et al.* (2019) *Influence of Impurities on the Solubility, Nucleation, Crystallization, and Compressibility of Paracetamol*. Cryst. Growth Des. 19, 4193-4201 [\[CrossRef\]](#)
- [17] Verma, V. *et al.* (2023) *Experimental Elucidation of Templated Crystallization and Secondary Processing of Peptides*. Pharmaceutics 2023, 15(4), 1288 [\[CrossRef\]](#)
- [18] Verma, V. *et al.* (2021) *Studying the impact of the pre-exponential factor on templated nucleation*. Faraday Discuss., 2022, 235, 199 [\[CrossRef\]](#)
- [19] Drebuschak, T.N. *et al.* (2006) *Variable temperature (100-295K) single-crystal X-ray diffraction study of the α -polymorph of glycylglycine and a glycylglycine hydrate*. Zeitschrift für Kristallographie – Crystalline Materials, vol. 221, no. 2, 2006, pp. 128-138 [\[CrossRef\]](#)
- [20] Zhang, L. *et al.* (2022) *An In-Situ MRI Method for Quantifying Temperature Changes during Crystal Hydrate Growths in Porous Medium*. Journal of Thermal Science. 31, 1542-1550 [\[CrossRef\]](#)
- [21] Braun, D. *et al.* (2015) *Navigating the Waters of Unconventional Crystalline Hydrates*. Mol. Pharmaceutics. 2015, 12, 8, 3069-3088 [\[CrossRef\]](#)

Catalytic Performances of UiO-66(Hf) under Various Synthesis Conditions for the Methylation of DHA

Luen Yu and Shuxuan Liang

Department of Chemical Engineering, Imperial College London, U.K.

Abstract UiO-66 is a metal-organic framework (MOF) with high stability and surface area that has been the focus of extensive research. This study focused on synthesising UiO-66(Hf) under various conditions—changing the ratios of metal ion, solvent and modulators (acetic acid, formic acid, and oxalic acid) added during synthesis, and changing the post-crystallisation treatment—to evaluate their effect on the structure of synthesised materials and thus their activity for catalysis on the methylation reaction of DHA to produce methyl lactate. Three characterisation techniques—TGA, XRD, and DRIFTS—were used to identify the structural features of synthesised UiO-66(Hf). The results demonstrated that when oxalic acid was added during synthesis, UiO-66(Hf) modified its bonding structure and had more defects. The catalytic performance of UiO-66(Hf) increased with the amount of modulator added during synthesis, especially oxalic acid, resulting in a dramatic boost in activity. Other synthesis conditions investigated had negligible effects on UiO-66 (Hf) catalytic activity. The best proposed catalyst structure was synthesised with a molar ratio of 1:2:370:56:1.03 (Hf ion: linker: DMF: acetic acid: oxalic acid) and with ethanol wash treatment.

Keywords: Metal-organic framework, UiO-66, Synthesis, Catalysis

1. Introduction

1.1 MOF

Metal-organic frameworks (MOFs) are a growing topic of interest in both academic and industrial settings, and they have been the subject of extensive research in recent years.¹ MOFs are porous coordination polymers (PCPs) synthesised from metal cation salts and organic ligands linked together by coordination-type bonds, resulting in two- or three-dimensional porous crystalline solids with infinite lattices.² They have an extremely large surface area and a highly tuneable structure - post-synthesis modification (PSM) can be done to tune the structure by changing the nature of the metal cations and ligands. Because the atomic-level control of their pore structures, MOF structures are highly flexible and have been used in many different applications.³ For example, a highly porous structure has a high surface adsorption potential, which can be used for hydrogen storage, CO₂ capture, and the removal of hazardous chemicals from the environment.⁴ More recently, MOF has been put on a larger scale of production and used industrially, with rising importance in the transportation, textile, and food packaging industries.⁴ MOF has also been shown in other studies to have functionality in drug transport and biomedicine,⁵ as well as the ability for heterogeneous catalysis.⁶

More than 20,000 MOF structures have been stated

in the literature to date.² According to the Web of Science, the structures that are undergoing the most investigation are ZIF-8, MIL-101, UiO-66, MOF-5, and so forth. This report will focus on the UiO-66 structure. UiO-66 was first synthesised in 2008 at the University of Oslo, which also gave it its name.⁷ The classic structure of UiO-66 is a crystal with zirconium oxide as the metal node and terephthalic acid as ligands, with the addition of solvents and modulators during the synthesis. The metal node (zirconium oxide) is referred to as the secondary building unit (SBU), and by developing novel SBUs, MOFs can attain greater stability.⁸ Apart from Zr, other metals like Hf are also possible to use as metal nodes. The UiO-66 SBU features 12 points of extension, allowing linkers to be bonded. Terephthalic acid, also known as 1,4-benzenedicarboxylic acid (BDC), serves as the linker for UiO-66. The ideal composition of UiO-66 is Zr₆O₆(BDC)₆, with each Zr₆O₆ cluster bonding to six BDC linkers.⁹ As BDC is a bidentate ligand, there will be 12 coordination bonds for the metal node. (Structure may be seen in Figure 1 middle part)

UiO-66 possesses various remarkable properties, including great mechanical stability, excellent thermal stability due to the strong metal-oxygen bond, and excellent acidic, aqueous, and water vapour stability.¹⁰ Furthermore, UiO-66 can be synthesised on a lab scale, has high catalytic properties, and has reproducible adsorption qualities.³

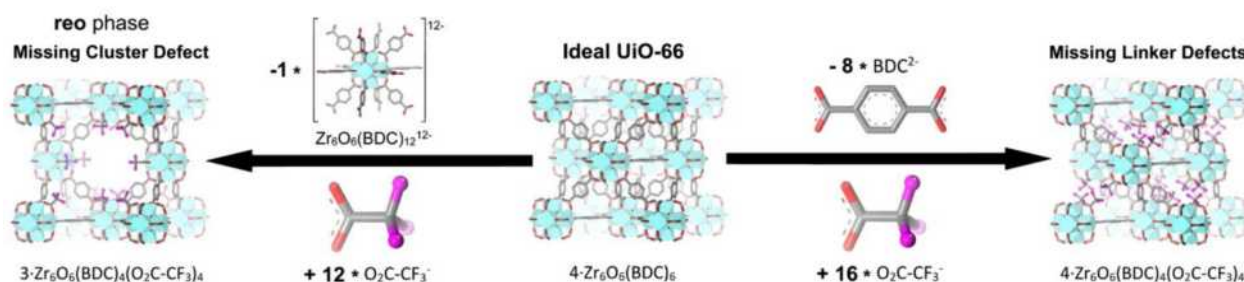


Figure 1: The structure of UiO-66 with defect engineering⁹

However, because of the high coordination number that leads to the high stability of UiO-66, it also has the disadvantage of inertness, which limits its functionality.³ Consequently, it is critical to present strategies for enhancing the functionality of UiO-66. One effective method is to perform defect engineering on the structure. The introduction of defects results in larger pores that modify mass-transport pathways, hence improving adsorption and catalytic properties.¹¹ Nevertheless, increasing the number of defects will lead to lowering stability and crystallinity and increasing the heat of adsorption.¹² As a result, defect control must be done carefully. There are two types of defects for UiO-66: missing linkers and missing clusters (Figure 1). Missing linkers require the removal of multiple ligands while maintaining structural integrity,¹³ whereas missing clusters require the removal of an entire cluster.¹⁴

Adjusting the synthesis conditions¹⁵ can be carried out as a means to control defects. Additionally, different synthesis conditions will result in diverse attributes in the crystallised product,¹⁶ such as different thermal stability, different weight loss during thermal decomposition, different catalytic performance, etc. Possible changes in the synthesis parameters include changing the component ratio. Modulators, which are among the parameters modified during synthesis, have a significant impact on defect control. Modulators are usually organic carboxylic acids that attach to the metal node and then influence crystal growth.¹⁷ There is competitive coordination between the ligand and the modulator, and adding more modulators with low pK_a values introduces more defects.¹² It is worthwhile to investigate the relationship between synthesis conditions and product structure, subsequently affecting its performance and properties.

1.2 Catalysis

One major research field for MOF is catalysis, as MOF offers excellent controllability, tunability, and catalytic activity as a catalyst. In general, approximately 85–90% of the products produced in chemical industries are produced through catalytic processes, making catalysis an important topic. Catalysis is used extensively in the synthesis of bulk and fine chemicals, as well as in the prevention and abatement of pollution.¹⁸ The material is particularly appealing for a variety of catalytic applications because to its high specific surface areas.¹⁹ When MOF is compared to another well-known porous catalyst, zeolites, it has a higher surface area and porosity, a more predictable design, and a high metal site density. These advantages give MOFs considerable potential for further study,²⁰ even though it has lower chemical stability and vapour phase catalytic activity than zeolite. Possible MOF-catalysed reactions include redox reactions,²¹ methylation reactions,²² polymerisation reactions,²³ etc.

Previous research demonstrated that hafnium (IV) analogues of zirconium (IV)-containing MOFs are excellent heterogeneous catalysis catalysts due to their robust and highly tuneable character.²⁴ Additionally,

hafnium (Hf)-based MOFs have been shown to be highly promising for practical use due to their exceptional chemical, thermal, and mechanical stability, as well as their acidic nature.²⁵ The catalytic characteristics of Hf-based MOFs can be examined by synthesising UiO-66 with Hf instead of Zr (UiO-66(Hf)) and testing the catalyst with reactions.

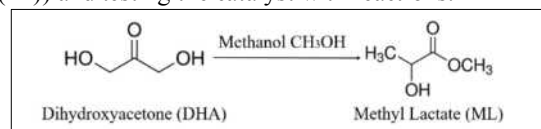


Figure 2: The reaction for the methylation of DHA

Figure 2 shows one reaction that UiO-66(Hf) can catalyse. The methylation reaction with dihydroxyacetone (DHA) to form methyl lactate (ML). ML is a significant green solvent for dealing with chemical compounds that are harmful to humans and the environment, and hence for reducing pollution at its source whenever possible.²⁶ Therefore, it is highly motivated to develop a better catalyst structure that leads to a higher production of methyl lactate.

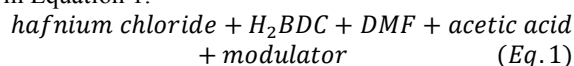
1.3 The aim of this research

With all of the interesting facts about MOFs and the motives for producing methyl lactate, the goal of this study is to investigate the catalytic activity of various UiO-66(Hf) on the DHA methylation reaction. The variation in UiO-66(Hf) performance is raised by different ratios of Hf metal ion/BDC linker/DMF solvent/modulator during the synthesis process. Aside from adjusting the ratios, the effects of different post-crystallisation treatments and several modulators—acetic acid, formic acid, and oxalic acid—were also investigated. Characterisation techniques such as Thermogravimetric Analysis (TGA), X-ray Diffraction (XRD), and Diffuse Reflectance Infrared Fourier Transform Spectroscopy (DRIFTS) were used to gain information on crystal structure and to quantify the number of defects in the synthesised products. The catalytic performances of the different materials prepared in this study, were investigated by doing reactions at different lengths of time using synthesised UiO-66(Hf) catalysts, the ML yield time-on-line profile was obtained for each catalyst. The ultimate objective was to identify a synthesis condition that would provide a catalyst with the optimum catalytic performance during the reaction.

2. Methodology

2.1 Synthesis of UiO-66(Hf)

The components for UiO-66(Hf) synthesis are shown in Equation 1.



In this project, 14 different UiO-66(Hf) samples were synthesised, each with different compositions of the synthesis mixtures or variations in post-crystallisation treatments. The compositions and post-crystallisation treatments of each generated UiO-66(Hf) are summarised in Table 1.

Table 1: Summary of the synthesis results of synthesised UiO-66(Hf)

Sample Code	mMoles of HfCl ₄	Relative molar ratios respect to moles of HfCl ₄						EtOH Wash	Yield / mg	Yield / g/mol Hf
		Linker/HfCl ₄	DMF/HfCl ₄	Acetic Acid/HfCl ₄	Oxalic Acid/HfCl ₄	Formic Acid/HfCl ₄				
A	2.172	1.55	523	79	/	/	Yes	/	/	/
B	2.172	1.55	523	79	/	/	No	/	/	/
C	1.678	2.00	677	102	/	/	Yes	281.6	366	
D	1.678	2.00	677	102	/	/	No	311.0	371	
E	0.838	2.00	677	56	/	/	Yes	/	/	/
F	1.677	2.00	370	102	/	/	Yes	532.0	317	
G	1.677	2.00	370	102	/	/	Yes	547.0	326	
H	1.679	2.00	370	56	/	/	Yes	630.5	375	
I	1.681	2.00	677	102	/	102	Yes	536.2	319	
J	1.682	2.00	677	102	/	51	Yes	480.8	286	
K	1.680	2.00	370	56	1.03	/	Yes	422.4	251	
L	1.681	2.00	370	56	0.52	/	Yes	544.8	324	
M	1.681	2.00	370	56	0.26	/	Yes	603.2	359	
N	1.680	2.00	677	102	0.103	/	Yes	547.0	326	

In a typical synthesis procedure, the synthesis mixtures were prepared in a Teflon liner followed the equivalent molar ratios with respect to moles of hafnium (shown in Table 1). Two different sizes of liner were used to account for different DMF ratios. The mixtures were stirred at 300 rpm for 30 minutes until the solid completely dissolved in DMF solvent. Once the solution was thoroughly mixed, placed the liner in a suitable-size stainless-steel autoclave (Figure 3a) and transferred it into an oven which preheated to 120°C for 48 hours.

Two types of post-crystallisation treatment were introduced in this project: ethanol washing or only centrifugation. For both treatments, the resulting microcrystalline powders were centrifuged for 15 minutes at 4000 rpm as the first step to separate from the solvent. For the materials which underwent ethanol washing, samples were transferred into a round bottle flask and stirred with appropriate amount of ethanol (90 ml for the large liner and 45 ml for the small liner) at 60°C, 300 rpm for 16 hours. Subsequent to the ethanol wash, the samples were separated from the solvent through centrifugation with the same condition, and then left to dry in an oven at 90°C overnight. For materials without ethanol washing, the samples underwent an additional three rounds of centrifugation under identical conditions, with ethanol being added at the start of each cycle and followed by a consistent drying method as part of the treatment.



Figure 3: Equipment used for experiments. a) Two sizes of autoclaves and linear for synthesis. b) Microwave reactor

The dried solids were transferred into a ceramic mortar and gently ground into fine powders. Powders were moved into a sealed vial for subsequent utilisation.

2.2 Characterisation for synthesised materials

2.2.1 Thermogravimetric Analysis (TGA)

Thermogravimetric Analysis is used to convey information about the thermal and oxidative stability of materials, composition of multiple components, product longevity, kinetics of decomposition, as well as the content of moisture and volatiles.²⁷ In this project, TGA was employed to analyse missing linker defects in the synthesised materials, considering the thermal instability of H₂BDC at elevated temperatures. A critical assumption underpinning the use of TGA for quantitative analysis in this study was that the residue after each TGA experiment consists pure HfO₂. Under standard TGA conditions, samples were initially loaded at 30°C and maintained at this temperature for 10 minutes. Subsequently, they were heated to 800°C under an air flow of 40 ml/min with a gradual temperature increase of 10°C/min. The sample was then held at the peak temperature for an additional 10 minutes to ensure complete combustion of all components other than HfO₂. Thermograms were obtained which graphically illustrating the variation in mass of each material in response to changes in temperature.

2.2.2 X-ray Diffraction (XRD)

X-ray Diffraction (XRD) is widely used for both qualitative and quantitative analysis of solid samples, providing insights into their crystal structure, degree of crystallinity, crystallite size, atomic spacing, and other significant characteristics. Each crystalline substance possesses a distinct diffraction pattern, which functions as a fingerprint, uniquely identifying each material.²⁸ In this project, XRD analysis was used to evaluate the degree of crystallinity in each of the UiO-66(Hf)

samples, which reflects the missing cluster defect in the material structure.²⁹ During an XRD analysis operation, $K\alpha$ radiation was utilised, and the scanning range was set to cover 2θ angles from 5° to 60° . The scan step size was incrementally increased by 0.033° every 40 seconds. The resulting spectra was plotted by intensity versus 2θ angles.

2.2.3 Diffuse Reflectance Infrared Fourier Transform Spectroscopy (DRIFTS)

DRIFTS is an infrared spectroscopy technique for sampling powders, employed to identify various characteristics of the sample, including the particle shape, density, refractive index, reflectivity, and absorption properties of the sample.³⁰ This technique is widely used to analyse the properties of solid powders due to the advantage that no prior preparation of samples is required. Nevertheless, the results obtained from DRIFTS cannot be directly applied for quantitative analysis of sample properties, as some standards are necessary to interpret the data accurately. In this project, DRIFTS analysis was utilised to comprehend the differences in MOF structures from the molecular bonding point of view for various samples, achieved through comparing the spectra that were plotted by absorbance versus wavenumber.

2.3 Kinetic investigation of UiO-66(Hf)

UiO-66(Hf) can be used as a catalyst for the methylation reaction of DHA to form ML. In a standard reaction procedure, 4g of 1% weight DHA in methanol were added into a reaction vial as the reactant, accompanied by 10 mg of UiO-66(Hf) serving as the catalyst. Placed reaction vials in the microwave reactor (Figure 3.b). The reaction temperature was set at 160°C so that the reaction could be carried out in a reasonable amount of time while maintaining stable reaction performance. The reaction was conducted for intervals of 5, 15, and 30 minutes respectively, after a two-minute ramping time to allow the reactor to achieve the setting temperature. Examine the yield of ML through the utilisation of Gas Chromatography (GC).

GC is a technique to analyse mixtures by using a mobile gas phase and a stationary liquid phase to separate different components within the mixture.³¹ A GC calibration was established to ascertain the retention time for each target component. A calibration curve was constructed to assist in the quantification of the yield of methyl lactate produced from the reaction. The standard procedure for GC analysis involved transferring 1 mL of the reaction mixture into a centrifuge tube, followed by centrifugation for 1 minute to separate the suspended catalyst. Transferred 0.5 mL of the centrifuged solution and 0.5 mL of a 0.1wt% biphenyl solution in methanol into a GC vial. Inserted the prepared vial into the GC and initiated the analysis. Using the GC calibration curve, the yield of methyl lactate was calculated by evaluating the peak areas for methyl lactate and biphenyl from the resulting chromatogram based on their retention times.

3. Results and Discussions

3.1 Characterisation

In this study, the analysis of characterisation outcomes enhanced comprehension of the structural features of the samples. As a result, it is feasible to identify the specific structural characteristics that potentially enhance the UiO-66(Hf) catalytic performance.

3.1.1 TGA Result

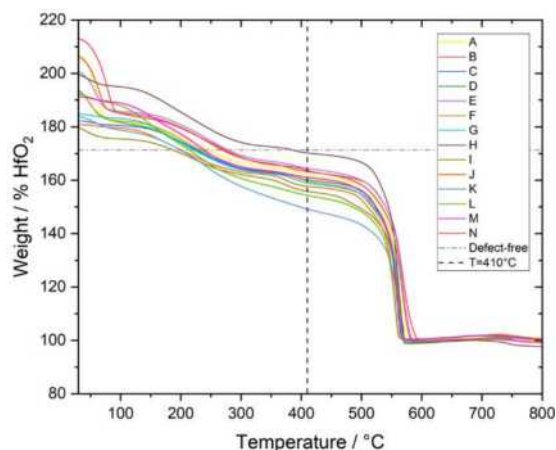
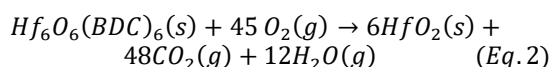


Figure 4: TGA graphical results for all 14 synthesised materials

TGA is a combustion process for the samples. The results of the TGA were represented graphically, depicting weight loss (normalised by mass at the end of TGA which was assumed as the weight of HfO_2) in relation to the variation in temperature, shown as Figure 4. Three distinct weight losses were noticed in this TGA plot. Based on prior research, these losses correspond to specific combustion processes: 1) The volatilisation of absorbates, in this study is ethanol, occurring from 30°C to 100°C . 2) The elimination of excess monocarboxylate ligands and the dihydroxylation of Hf_6 cornerstones, taking place from 200°C to 390°C . 3) The decomposition of the framework, involving the completely combustion of the BDC linkers, from 410°C to 570°C .⁹ This study used TGA to examine the extent of missing linker defects across various materials, therefore leading to a concentrated analysis on the third weight loss step.

The theoretical composition of a defect-free UiO-66(Hf) is represented as $\text{Hf}_6\text{O}_6(\text{BDC})_6$. However, achieving a defect-free MOF structure is unattainable, and varying degrees of linker loss are observed in MOF structures, influenced by diverse synthetic conditions. In order to quantify the precise coordinating linker number in each synthesised material, the following calculation method was used to interpret the TGA plot. The combustion process for defect-free UiO-66(Hf) is shown as Equation 2.



It was crucially assumed that HfO_2 is the only remaining component at the end of the TGA. In the calculation, normalised percentage weight loss per BDC linker ($\text{Wt.PL}_{\text{Theor}}$) was derived from the weight

percentage of a defect-free UiO-66(Hf) structure, shown in Equation 3. Then, the difference between the normalised weight percentage of the sample at the start of the third-stage of weight loss (at 410°C, $W_{exp,plat}$) and its weight percentage at the end of the process (W_{End}) was calculated and divided by $Wt.PL_{Theo}$ to get the coordination number (Equation 4). The coordinating linker number for all synthesised materials was quantified and summarised in Table 2.

$$Wt.PL_{Theo} = \frac{W_{ideal,plat} - W_{End}}{NL_{ideal}} \quad (Eq. 3)$$

$$NL_{Exp} = \frac{W_{Exp,plat} - W_{End}}{Wt.PL_{Theo}} \quad (Eq. 4)$$

Table 2: Summary of coordinating linker number for all 14 synthesised materials, and the molar ratio of additional modulators

Sample Code	Coordination Number	Sample Code	Coordination Number	Additional Modulator
A	5.205	H	5.905	-
B	5.359	I	4.699	FA-102
C	5.057	J	4.859	FA-51
D	5.061	K	4.139	OA-1.03
E	5.421	L	4.571	OA-0.52
F	5.022	M	5.137	OA-0.26
G	4.962	N	5.318	OA-0.103

The analysis of the data revealed that sample H (prepared with only acetic acid as modulator in a ratio of 56) exhibited the most similar structure to the defect-free UiO-66(Hf), which is characterised by a coordination number of six. In contrast, Sample K (prepared with an additional modulator - oxalic acid in a ratio of 1.03) displayed the greatest extent of missing linker defects. This observation could be attributed to the synthetic conditions, particularly the use of different types of modulators. Materials synthesised with oxalic acid were more likely to process remarkable linker deficiencies, which could be linked to its lower pK_a value ($pK_{a1}=1.25$ and $pK_{a2}=3.81$) compared to other modulators (pK_a for acetic acid =4.76 and pK_a for formic acid =3.77). The higher acidity of the modulator used in the synthesis process tended to result in more significant linker defects, which led to a reduced number of linkers attached to the Hf metal.⁹ In addition, a noticeable trend was observed in materials synthesised using the same modulators: an increased equivalent molar ratio of the modulator in the synthesis process generally encouraged the formation of missing linker defects. However, this trend did not apply to acetic acid.

It is noteworthy to observe a general relationship between the extent of linker defects and the mass of dried samples obtained from one mole of Hf used in the synthesis (normalised yield). As indicated in Figure 5, materials such as K and L with significant linker loss in their structure tended to have a lower normalised yield. This phenomenon could be explained by the reduced number of linkers attaching to the Hf centre, which resulted in a smaller unit mass for each single Hf-centred framework.

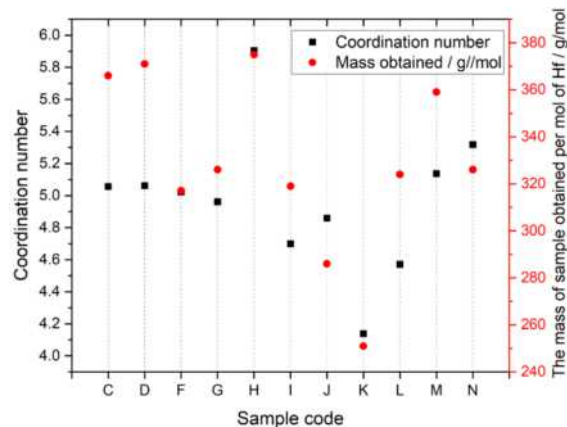


Figure 5: Relationship between coordination number and the obtained mass of UiO-66(Hf) per mole of Hf

3.1.2 XRD Result

Based on the results obtained from TGA, it concluded that the use of oxalic acid as modulator resulted in the most significant missing linker defect among the synthesised materials. Therefore, further investigations were carried out for these materials utilising XRD to determine the presence of missing cluster defects, another type of potential structural defect in MOFs. XRD analysis was applied to two materials that possessed the highest degree of linker loss in their structure (samples K and L), along with sample H, which demonstrated the closest resemblance to a defect-free UiO-66(Hf) structure according to the TGA results. XRD results were plotted by intensity against 2θ , shown as Figure 6a. The areas of an intensive peak (between 25° and 28°) were generated to analyse and compare the crystallinity of different samples (Table 3).

Table 3: Generated peak area for an intensive peak from XRD result

Sample code	Peak Area	Relative Area
H	7619.23	100%
K	3905.73	51.3%
L	4721.82	62.0%

A significant difference in the peak areas was observed. It indicates changes in the long-range order within the structures of the analysed samples. The observed decrease in peak area showed that the crystal structures were more disordered, indicating the presence of missing cluster defects within these structures.¹¹ By comparing the peak areas, it clearly showed that sample H possesses a well-ordered crystal structure, in contrast to samples K and L, which exhibit poor crystallinity. Therefore, it can be stated that the use of oxalic acid as a modulator in the synthesis process promotes the occurrence of missing cluster defects.

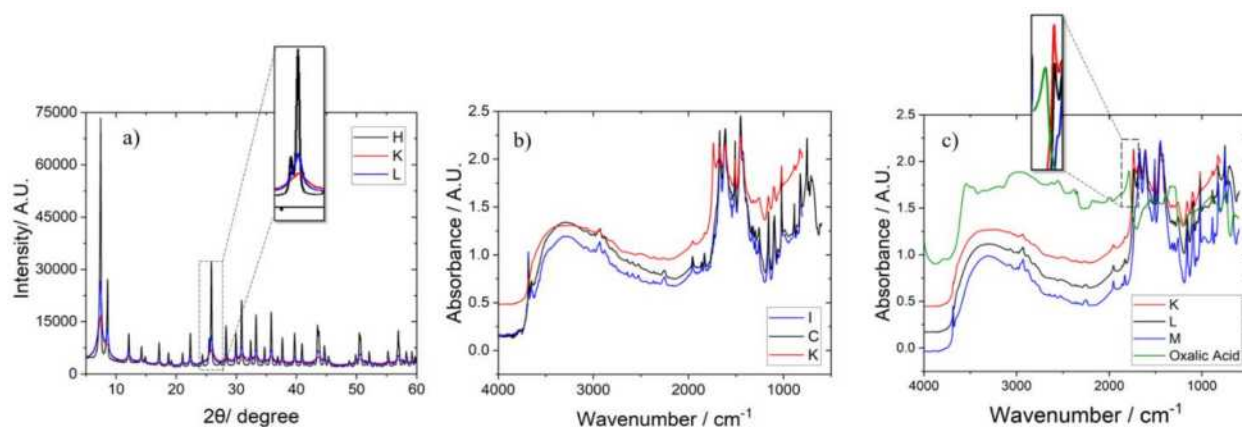


Figure 6: a) XRD results for sample H, K b) DRIFTS for sample C, I and K which respect to use acetic acid, formic acid and oxalic acid as modulator c) DRIFTS for sample K, L, M and pure oxalic acid

3.1.3 DRIFTS Results

The previous two techniques analysed the macrostructure of the synthesised UiO-66(Hf), and hence DRIFTS was applied in this study to get a better understanding of the near-metal structure.

DRIFTS was carried out for all the synthesised materials, and the results were plotted as absorbance versus wavenumber. Interestingly, most of the parameters varied during the synthesis process, such as the equivalent molar ratio of the solvent and post-crystallisation treatments, generated similar patterns in the plot. This could be due to a couple of reasons: either these varying parameters did not significantly impact the microstructure of UiO-66(Hf), or it may be attributed to the insensitivity of DRIFTS to detect changes in the bonding structure among UiO-66(Hf) samples.

However, the DRIFTS curves for samples synthesised using oxalic acid showed a notable difference compared to other materials. As Figure 6b shows, an additional peak (at wavenumber=1736.82 cm⁻¹) was distinctly observed in Sample K. To ensure this peak was not due to residual oxalic acid in the synthesised materials, DRIFTS was also applied to pure oxalic acid for comparison. The pattern for pure oxalic acid, shown in Figure 6c, gave a very different shape, with the closest peak to UiO-66(Hf) being at wavenumber=1780.97cm⁻¹. This confirmed that no excess oxalic acid presented in the synthesised materials. Additionally, all three samples synthesised using oxalic acid as modulator possessed this additional peak at wavenumber=1736.82cm⁻¹. Therefore, it could be concluded that the utilisation of oxalic acid in the synthesis altered the bonding structure between the Hf metal and the linkers.

3.2 Kinetics

3.2.1 Inaccuracies in kinetics

Inaccuracies may arise in the kinetic results. Firstly, inaccuracy may be caused by reactions. To quantify this, reactions were repeated multiple times, the

standard deviation was computed based on yield differences, and the errors were displayed as error bars on time-on-line plots. Occasionally, the error bars were narrow due to low standard deviation values, so the error bars may be too small to be seen.

Secondly, the extent of reproducibility of the synthesised material had an influence on the accuracy of the result. The variation between the reproduced material may arise from inaccuracies in measurement, loss of material during transfer, etc. Sample G was synthesised under the identical conditions as Sample F, with the objective of testing the reproducibility of the materials. The variance between these two samples was evaluated by comparing their TGA results and their kinetic performance. TGA results showed the coordination number for the repeated sample was similar (Table 2). When comparing the kinetic performance between Sample G and F (Figure 7), a deviation of 13% was observed in their k values. Consequently, in this study, a variation in k

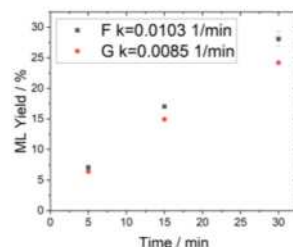


Figure 7: Variance between the reaction rates among the reproduced materials

values up to 13% can be attributed to synthesis errors.

The reaction rate constant – k value (unit: min⁻¹) was determined for each sample and included in the plot legend by assuming the reaction is first-order irreversible. By plotting Ln (1-ML yield) against the reaction time, the k value was obtained from the gradient for this plot. This plot should have a linear shape. By conducting a linear fit for the experimental data, relatively high R² values (>0.99 in most cases) were achieved, indicating that this was a reliable approach for determining the k values. However, at longer reaction times for the fast-reacted reaction conditions, the observed trend was non-linear, which invalidated the first-order reaction assumption due to the fact that the concentration of reactant decreased throughout the reaction.

3.2.2 Effect of modulator ratio – oxalic acid (OA)

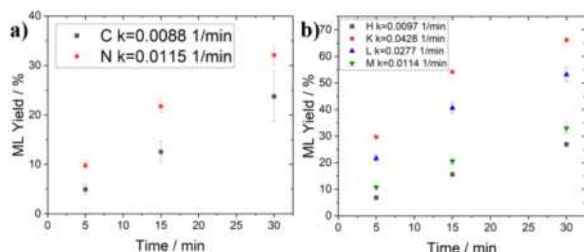


Figure 8: Effect of the OA ratio on reaction rates

Figure 8 illustrates the influence of OA on catalytic performance. Figure 8a and Figure 8b had varying amounts of solvent and acetic acid added during synthesis, however when focused on one time-on-line plot, the only variation during synthesis was the amount of OA added. Figure 8 shows that using more OA during synthesis resulted in a higher ML yield at each time step. The addition of modulators during synthesis had an effect on defects and thus changed the structure of MOF, giving it the opportunity to enhance its catalytic performance.

Although DRIFTS results showed there was no excess OA in the structure, it is worthwhile to further confirm that the better performance is due to the better synthesised structure of MOF. So, it is crucial to investigate whether OA on its own has the ability to catalyse the methylation reaction that produces ML. A reaction was done using only OA as a catalyst; the GC chromatogram showed no ML was produced from the reaction, but other by-products were formed. The by-products seen on the chromatogram were different from the by-products formed during the MOF-catalysed methylation reaction. By-products in an OA-catalysed reaction were possible to be OA degradation products at high temperatures, or OA could act as a Bronsted acid that catalysed another reaction pathway. Due to the time limitation, it is not possible to further analyse the by-products, but a conclusion could be drawn that OA does not have the ability to catalyse the production of ML.

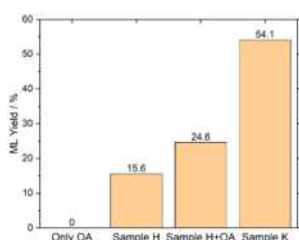


Figure 9: Comparison of the effects of different forms of OA on the ML yields

Another way to study the effect of OA is by doing the reaction with sample H and an additional amount of OA, and then comparing it to sample K. Sample H had the same synthesis conditions as sample K, with the exception that sample K contained OA during synthesis. In sample K, assuming that all of the OA added during the synthesis did not link into the structure, there should be 3.7mg of OA in the catalyst for each reaction. For assurance, an excess of OA (4.4 mg) was added to sample H as a co-catalyst to perform a reaction for 15 minutes. A comparison of ML yields is shown in Figure 9.

With sample H and the co-catalyst OA, the yield of ML increased compared to using only sample H as a catalyst. The reason for this could be explained by the

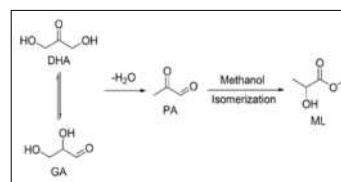


Figure 10: The reaction scheme for the methylation of DHA³²

reaction scheme. The conversion from DHA to ML involved two steps: (1) dehydration of DHA to

pyruvaldehyde (PA), (2) alcohol

addition of PA and isomerisation to ML (Figure 10). With OA providing more Lewis acid sites for catalysis, these two reaction steps may increase their rates.³² Unlike utilising solely OA as a catalyst which may have changed the entire reaction mechanism, using MOF with OA followed the same reaction pathway as MOF-catalysed reactions that produced ML.

The increase in yield by using OA as a co-catalyst was much lower compared to sample K (Figure 9). This proved that the high ML yield obtained with sample K was due to a new structure of MOF formed from synthesis with OA bonded to the structure, not because of free OA presented in the structure which promoted the reaction rate as a co-catalyst.

3.2.3 Effect of modulator ratio – formic acid (FA)

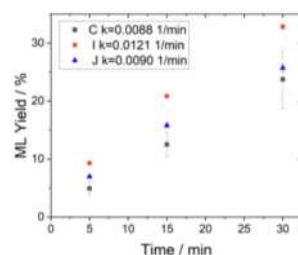


Figure 11: Effect of the FA ratio on reaction rates

Figure 11 shows the yields of ML obtained using catalysts with different amounts of FA during synthesis. There was no FA during synthesis in sample C and a FA molar ratio of 51 and 102 in samples J and I, respectively.

More FA led to a higher reaction rate, but the extent of the increase was much smaller compared to OA. An increase in FA amount from ratio 51 to 102 had a larger promotion for the reaction rate than an increase in FA from none to ratio 51, indicating that with a higher amount of FA added, the effect started to become more obvious.

3.2.4 Effect of modulator ratio – acetic acid (AA)

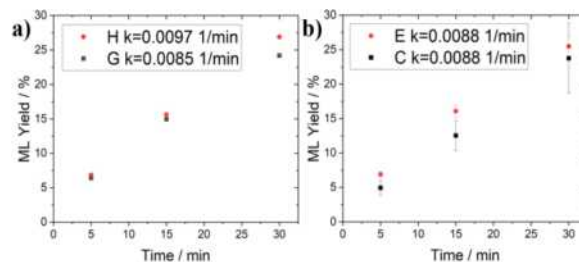


Figure 12: Effect of the AA ratio on reaction rates

Figure 12 shows the effect of changing the AA ratio on reaction kinetics. Red dots in the plots represent the 56 AA ratio during synthesis, while black squares represent the 102 AA ratio. With doubling the amount of AA, there was no significant change in the reaction rate. Past research showed that the structure of the synthesised MOF (linker deficiencies) was constant unless a very large amount of AA (ratio much greater than 100) was added,⁹ which confirmed that the results

obtained were reasonable. Overall, AA had the lowest acidity (highest pK_a value) of the three distinct modulators and was shown to be the least effective.

3.2.5 Effect of solvent ratio – DMF

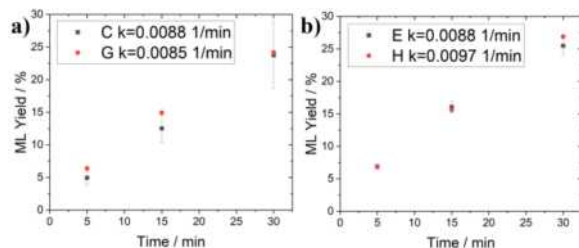


Figure 13: Effect of the DMF ratio on reaction rates

Figure 13 shows the effect of changing the solvent ratio on reaction kinetics. The solvent used during synthesis is N,N-dimethylformamide (DMF). The red dots in the plots represent a lower DMF ratio during synthesis. No difference in catalytic performance with a changing DMF ratio could be concluded from Figure 13. Although the points in the time-on-line plots were not perfectly aligned within the error bars, the deviation was still within the error of synthesis. To account for the role of solvent during crystallisation, increasing the amount of solvent decreased the concentration of solute, thus affecting the nucleation rate. However, the amount of solvent did not have an impact on the structure of the MOF formed after crystallisation, so the catalytic performance was not changed. Furthermore, changing the types of solvent will have an effect on the MOF structure,³³ but in this project, the type of solvent was kept consistent.

3.2.6 Effect of the amount of Hf

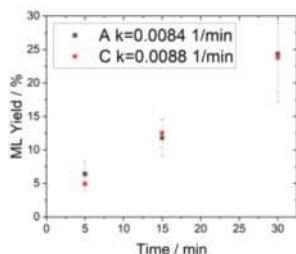


Figure 14: Effect of the amount of Hf on reaction rates

were different because the ratios were normalised to the Hf amount, which was different in this case. The Hf ion was one of the solutes for crystallisation, so changing the amount of Hf changed the solute concentration and hence the nucleation rate for crystallisation. Similar to changing the ratio for DMF, changing the Hf amount only influenced the crystallisation process but did not affect the synthesised MOF structure, so there was no impact on the catalytic performance.

3.2.7 Effect of post-crystallisation treatment

Figure 15 shows the effect of using different post-crystallisation treatments on reaction kinetics. The red dots in the plots represent using the ethanol (EtOH) wash treatment, and the black squares represent using only centrifuges. The purpose of the post-crystallisation treatment was to eliminate the hazardous

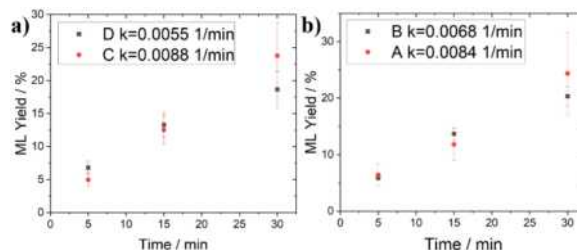


Figure 15: Effect of the post-crystallisation treatment on reaction rates

and redundant chemicals from the crystallised MOF and produce a more purified MOF. DMF is the main chemical that the treatment was designed to remove. The EtOH wash treatment resulted in higher k values, indicating a faster reaction rate. One probable conclusion was that using EtOH wash gave rise to improved DMF removal and a purer catalyst, hence boosting catalytic activity. However, due to synthesis and reaction errors, this conclusion was not very reliable.

3.2.8 Relationship between defects and catalytic performances

To summarise the impact of synthesis changes on kinetics, a significant increase in catalytic performance was seen when the synthesised material had structural changes. The presence of defects in the structure was the major structural change. A scatter plot could be constructed using the TGA and kinetics results to demonstrate the relationship between the missing linker defect (linker loss) and catalytic performance (k value, unit: min^{-1}) (Figure 16).

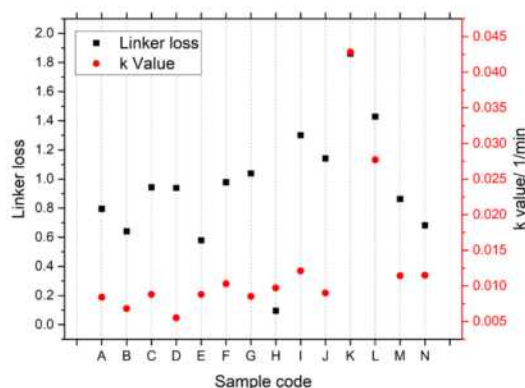


Figure 16: Relationship between linker loss and the k value

A clear trend could be seen from sample K to sample M (samples had decreasing amounts of OA in synthesis). With the most OA added (sample K), the greatest extent of defects was created, and the highest reaction rate was observed. More defects increased the porosity of MOF, resulting in a larger surface area and more adsorption sites for catalysis. The extent of defects and reaction rates both dropped as the amount of OA decreased from sample K to sample M. A similar trend was found in the decrease of FA amounts from samples I to J.

Not all of the samples followed the trend of faster reactions as the extent of defects increased. Apart from the missing linker defect, there might be other reasons

that caused the differences in catalytic performance. For example, missing cluster defects, which had not been quantified but have great potential to impact sample catalytic activity, were likely to be present in the MOF structures.

4. Conclusion

In this work, 14 materials were investigated for their catalytic performances under various synthesis conditions. Three characterisation techniques were used to identify the structural features of the samples. TGA revealed there were more missing linker defects as the amount of OA and FA increased during synthesis. Sample K, which proved to be the most defective structure, had the lowest coordination number of 4.139. Other modifications to synthesis conditions did not result in a comparable number of missing linker defects as OA. XRD indicated that when the amount of OA added during synthesis increased, the structure became more disordered, implying more missing cluster defects. DRIFTS showed that the bonding structure of materials synthesised without OA followed a similar pattern, whereas with OA, an extra peak at wavenumber 1736.82 cm^{-1} was seen in the spectrum. This new peak has been demonstrated to be evidence of modifying the near-metal bonding structure due to OA introduced during synthesis, rather than excess OA in the structure. In summary, the largest structural change occurred during synthesis when an OA modulator was added, which affected both the macrostructure and microstructure of the synthesised material.

From a kinetics perspective, alterations to the ratio of AA, DMF, and Hf and the post-crystallisation treatment had minor impacts on the catalytic activity of synthesised UiO-66(Hf). Because these changes in the synthesis conditions did not modify the structure of materials as shown by characterisation techniques. Changing the FA ratio during synthesis has only a small effect on catalytic activity, due to a slight increase in missing linker defects. The determining factor in catalytic activity was OA, which raised the reaction rate by the most as its ratio increased during synthesis. The best proposed catalyst was sample K, which was synthesised with a molar ratio of 1:2:370:56:1.03 (Hf ion: linker: DMF: acetic acid: oxalic acid) and with ethanol wash treatment. Its k value was 0.428 min^{-1} , which was 75.1% greater than the average k value of the other catalysts.

Overall, adding OA during synthesis introduced defects into the structure, which promoted the catalytic performance of UiO-66(Hf). All other changes during synthesis have minor effects on catalytic performances when compared with OA.

5. Outlook

Although the good progress made in this project, several questions still remain, and future works should be performed. To achieve enhanced catalytic performance, it requires to gain deeper insights into the structural attributes of the synthesised materials and

comprehend the reaction mechanisms involved in the DHA methylation process under the reaction conditions.

In order to get a more comprehensive understanding of the macrostructure of the materials, XRD analysis can be conducted on all synthesised materials. This will help determine if the other varied parameters in the synthesis process also led to the occurrence of missing cluster defects and the extent of these defects. Additionally, techniques like Extended X-ray Absorption Fine Structure (EXAFS) can be utilised to gain a better understanding of the local environment around hafnium oxide in the synthesised samples.³⁴

On the other hand, from kinetic aspect, the analysis of all the obtained GC results revealed several peaks with noticeable areas (not ML), indicating the formation of potential by-products. To enhance the yield of the desired product, ML, it is vital to analyse the possible factors leading to the formation of these by-products, and also to identify them. Investigations can be conducted from various aspects, such as considering side reactions, especially since they might arise from the presence of excess acetic acid in the synthesised sample. The surplus acetic acid could potentially act as a co-catalyst, leading to the activation of alternative reaction pathways. Therefore, it is important to examine whether acetic acid behaves as a co-catalyst during the reactions that result in the production of by-products. In addition, to get a better understanding of the existing by-products, evaluate the degradation performance of the feedstock under the specified reaction conditions and employ GC to ascertain if intermediates of the DHA methylation process are contributing to the formation of these by-products.

To further improve the catalytic performance, several approaches can be implemented to try to optimise the synthesis process of UiO-66 (Hf). Building on insights from previous studies, it is beneficial to focus further analysis on OA, given it demonstrated the best catalytic performance. By adjusting the equivalent molar ratio of OA used in the synthesis process, it is possible to achieve an optimal synthesis composition for UiO-66(Hf) that results in optimised catalytic performance. Furthermore, since modulators typically have a significant impact on catalytic performance, it would be beneficial to explore different types of modulators, such as difluoroacetic acid, trifluoroacetic acid etc.⁹ Testing the performance of materials synthesised with these modulators could provide valuable insights into optimising the catalytic performance.

Acknowledgements

We wish to extend our profound gratitude and appreciation to Hammond Lab, for the invaluable guidance and assistance throughout the duration of this project. We are particularly indebted to Dr. Giulia Tarantino for her generous dedication of time in training, guiding, and assisting us in this project.

References

- [1] Gaab, M., Trukhan, N., Maurer, S., Gummaraju, R. and Müller, U. (2012). The progression of Al-based metal-organic frameworks – From academic research to industrial production and applications. *Microporous and Mesoporous Materials*, 157, pp.131–136. doi:https://doi.org/10.1016/j.micromeso.2011.08.016.
- [2] Wang, Q. and Astruc, D. (2019). State of the Art and Prospects in Metal–Organic Framework (MOF)-Based and MOF-Derived Nanocatalysis. *Chemical Reviews*, 120(2), pp.1438–1511. doi:https://doi.org/10.1021/acs.chemrev.9b00223.
- [3] Winarta, J., Shan, B., McIntyre, S.M., Ye, L., Wang, C., Liu, J. and Mu, B. (2019). A Decade of UiO-66 Research: A Historic Review of Dynamic Structure, Synthesis Mechanisms, and Characterization Techniques of an Archetypal Metal–Organic Framework. *Crystal Growth & Design*, 20(2), pp.1347–1362. doi:https://doi.org/10.1021/acs.cgd.9b00955.
- [4] Silva, P., Vilela, S.M.F., Tomé, J.P.C. and Almeida Paz, F.A. (2015). Multifunctional metal–organic frameworks: from academia to industrial applications. *Chemical Society Reviews*, [online] 44(19), pp.6774–6803. doi:https://doi.org/10.1039/c5cs00307e.
- [5] Abánades Lázaro, I. and Forgan, R.S. (2019). Application of zirconium MOFs in drug delivery and biomedicine. *Coordination Chemistry Reviews*, 380, pp.230–259. doi:https://doi.org/10.1016/j.ccr.2018.09.009.
- [6] Dhakshinamoorthy, A., Santiago-Portillo, A., Asiri, A.M. and Garcia, H. (2019). Engineering UiO-66 Metal Organic Framework for Heterogeneous Catalysis. *ChemCatChem*, 11(3), pp.899–923. doi:https://doi.org/10.1002/cctc.201801452.
- [7] Cavka, J.H., Jakobsen, S., Olsbye, U., Guillou, N., Lamberti, C., Bordiga, S. and Lillerud, K.P. (2008). A New Zirconium Inorganic Building Brick Forming Metal Organic Frameworks with Exceptional Stability. *Journal of the American Chemical Society*, 130(42), pp.13850–13851. doi:https://doi.org/10.1021/ja8057953.
- [8] Ha, J., Lee, J.H. and Moon, H.R. (2020). Alterations to secondary building units of metal–organic frameworks for the development of new functions. *Inorganic Chemistry Frontiers*, 7(1), pp.12–27. doi:https://doi.org/10.1039/c9q01119f.
- [9] Shearer, G.C., Chavan, S., Bordiga, S., Svelle, S., Olsbye, U. and Lillerud, K.P. (2016). Defect Engineering: Tuning the Porosity and Composition of the Metal–Organic Framework UiO-66 via Modulated Synthesis. *Chemistry of Materials*, 28(11), pp.3749–3761. doi:https://doi.org/10.1021/acs.chemmater.6b00602.
- [10] Wu, H., Yildirim, T. and Zhou, W. (2013). Exceptional Mechanical Stability of Highly Porous Zirconium Metal–Organic Framework UiO-66 and Its Important Implications. *The Journal of Physical Chemistry Letters*, 4(6), pp.925–930. doi:https://doi.org/10.1021/jz4002345.
- [11] Feng, X., Himanshu Sekhar Jena, Chidharth Krishnaraj, Leus, K., Wang, G., Chen, H.S., Jia, C. and Der, V. (2021). Generating Catalytic Sites in UiO-66 through Defect Engineering. *ACS Applied Materials & Interfaces*, 13(51), pp.60715–60735. doi:https://doi.org/10.1021/acsami.1c13525.
- [12] Shan, B., McIntyre, S.M., Armstrong, M.R., Shen, Y. and Mu, B. (2018). Investigation of Missing-Cluster Defects in UiO-66 and Ferrocene Deposition into Defect-Induced Cavities. *Industrial & Engineering Chemistry Research*, 57(42), pp.14233–14241. doi:https://doi.org/10.1021/acs.iecr.8b03516.
- [13] Vandichel, M., Hajek, J., Vermoortele, F., Waroquier, M., De Vos, D.E. and Van Speybroeck, V. (2015). Active site engineering in UiO-66 type metal–organic frameworks by intentional creation of defects: a theoretical rationalization. *CrystEngComm*, 17(2), pp.395–406. doi:https://doi.org/10.1039/c4ce01672f.
- [14] Cliffe, M.J., Wan, W., Zou, X., Chater, P.A., Kleppe, A.K., Tucker, M.G., Wilhelm, H., Funnell, N.P., Coudert, F.-X. and Goodwin, A.L. (2014). Correlated defect nanoregions in a metal–organic framework. *Nature Communications*, 5(1). doi:https://doi.org/10.1038/ncomms5176.
- [15] Feng, Y., Chen, Q., Jiang, M. and Yao, J. (2019). Tailoring the Properties of UiO-66 through Defect Engineering: A Review. *Industrial & Engineering Chemistry Research*, 58(38), pp.17646–17659. doi:https://doi.org/10.1021/acs.iecr.9b03188.
- [16] Shearer, G.C., Chavan, S., Ethiraj, J., Vitillo, J.G., Svelle, S., Olsbye, U., Lamberti, C., Bordiga, S. and Lillerud, K.P. (2014). Tuned to Perfection: Ironing Out the Defects in Metal–Organic Framework UiO-66. *Chemistry of Materials*, 26(14), pp.4068–4071. doi:https://doi.org/10.1021/cm501859p.
- [17] Vermoortele, F., Bueken, B., Le Bars, G., Van de Voorde, B., Vandichel, M., Houthoofd, K., Vimont, A., Daturi, M., Waroquier, M., Van Speybroeck, V., Kirschhock, C. and De Vos, D.E. (2013). Synthesis Modulation as a Tool To Increase the Catalytic Activity of Metal–Organic Frameworks: The Unique Case of UiO-66(Zr). *Journal of the American Chemical Society*, 135(31), pp.11465–11468. doi:https://doi.org/10.1021/ja405078u.
- [18] I Chorkendorff and Niemantsverdriet, J.W. (2007). *Concepts of modern catalysis and kinetics*. Weinheim: Wiley-Vch.
- [19] Konnerth, H., Matsagar, B.M., Chen, S.S., Precht, M.H.G., Shieh, F.-K. and Wu, K.C.-W. (2020). Metal-organic framework (MOF)-derived catalysts for fine chemical production. *Coordination Chemistry Reviews*, 416, p.213319. doi:https://doi.org/10.1016/j.ccr.2020.213319.
- [20] Herbst, A. and Janiak, C. (2017). MOF catalysts in biomass upgrading towards value-added fine chemicals. *CrystEngComm*, 19(29), pp.4092–4117. doi:https://doi.org/10.1039/c6ce01782g.
- [21] Dhakshinamoorthy, A., Asiri, A.M. and Garcia, H. (2016). Metal-Organic Framework (MOF) Compounds: Photocatalysts for Redox Reactions and Solar Fuel Production. *Angewandte Chemie International Edition*, 55(18), pp.5414–5445. doi:https://doi.org/10.1002/anie.201505581.
- [22] Mu, J., Liu, J., Ran Zhenzhen, Arif, M., Gao, M., Wang, C. and Ji, S. (2020). Critical Role of CUS in the Au/MOF-808(Zr) Catalyst for Reaction of CO₂ with Amine/H₂ via N-Methylation and N-Formylation. *Industrial & Engineering Chemistry Research*, 59(14), pp.6543–6555. doi:https://doi.org/10.1021/acs.iecr.0c00242.
- [23] Goetjen, T.A., Liu, J., Wu, Y., Sui, J., Zhang, X., Hupp, J.T. and Farha, O.K. (2020). Metal–organic framework (MOF) materials as polymerization catalysts: a review and recent advances. *Chemical Communications*, 56(72), pp.10409–10418. doi:https://doi.org/10.1039/d0cc03790g.
- [24] Rimoldi, M., Howarth, A.J., DeStefano, M.R., Lin, L., Goswami, S., Li, P., Hupp, J.T. and Farha, O.K. (2016). Catalytic Zirconium/Hafnium-Based Metal–Organic Frameworks. *ACS Catalysis*, 7(2), pp.997–1014. doi:https://doi.org/10.1021/acscatal.6b02923.
- [25] Hu, Z., Wang, Y. and Zhao, D. (2021). The chemistry and applications of hafnium and cerium(IV) metal–organic frameworks. *Chemical Society Reviews*, 50(7), pp.4629–4683. doi:https://doi.org/10.1039/d0cs00920b.
- [26] Aparicio, S. (2007). Computational Study on the Properties and Structure of Methyl Lactate. *Journal of Physical Chemistry A*, 111(21), pp.4671–4683. doi:https://doi.org/10.1021/jp070841t.
- [27] Saadatkah, N., Carillo Garcia, A., Ackermann, S., Leclerc, P., Latifi, M., Samih, S., Patience, G.S. and Chaouki, J. (2019). Experimental methods in chemical engineering: Thermogravimetric analysis—TGA. *The Canadian Journal of Chemical Engineering*, 98(1), pp.34–43. doi:https://doi.org/10.1002/cjce.23673.
- [28] Khan, H., Yerramilli, A.S., D'Oliveira, A., Alford, T.L., Boffito, D.C. and Patience, G.S. (2020). Experimental methods in chemical engineering: X-ray diffraction spectroscopy—XRD. *The Canadian Journal of Chemical Engineering*, 98(6), pp.1255–1266. doi:https://doi.org/10.1002/cjce.23747.
- [29] Feng, X., Himanshu Sekhar Jena, Chidharth Krishnaraj, Daniel Arenas Esteban, Leus, K., Wang, G., Sun, J., Rüschler, M., Timoshenko, J., Beatriz Roldán Cuenya, Bals, S. and Der, V. (2021a). Creation of Exclusive Artificial Cluster Defects by Selective Metal Removal in the (Zn, Zr) Mixed-Metal UiO-66. *Journal of the American Chemical Society*, 143(51), pp.21511–21518. doi:https://doi.org/10.1021/jacs.1c05357.
- [30] Leyden, D.E. and Murthy, R.S. (1988). Diffuse reflectance Fourier transform IR spectroscopy. *TrAC Trends in Analytical Chemistry*, 7(5), pp.164–169. doi:https://doi.org/10.1016/0165-9936(88)85044-1.
- [31] Bartle, K.D. and Myers, P. (2002). History of gas chromatography. *TrAC Trends in Analytical Chemistry*, 21(9-10), pp.547–557. doi:https://doi.org/10.1016/s0165-9936(02)00806-3.
- [32] Lu, T., Fu, X., Zhou, L., Su, Y., Yang, X., Han, L., Wang, J. and Song, C.-Y. (2017). Promotion Effect of Sn on Au/Sn-USY Catalysts for One-Pot Conversion of Glycerol to Methyl Lactate. *ACS Catalysis*, 7(10), pp.7274–7284. doi:https://doi.org/10.1021/acscatal.7b02254.
- [33] Zhang, B., Zhang, J., Liu, C., Sang, X., Peng, L., Ma, X., Wu, T., Han, B. and Yang, G. (2015). Solvent determines the formation and properties of metal–organic frameworks. *RSC Advances*, 5(47), pp.37691–37696. doi:https://doi.org/10.1039/c5ra02440d.
- [34] Valenzano, L., Civalieri, B., Chavan, S., Bordiga, S., Nilsen, M.H., Jakobsen, S., Lillerud, K.P. and Lamberti, C. (2011). Disclosing the Complex Structure of UiO-66 Metal Organic Framework: A Synergic Combination of Experiment and Theory. *Chemistry of Materials*, 23(7), pp.1700–1718. doi:https://doi.org/10.1021/cm1022882.

Interpretable Supply Chain Optimisation for Inventory Management Problems with Genetic Decision Trees

Jay Geller and Lucia Zhan

Department of Chemical Engineering, Imperial College London, U.K.

Abstract

As machine learning becomes more interlinked with data-driven decision-making processes within critical industries like supply chain management, the transparency and explainability of such models will grow in importance. Decision making guided by black-box optimisation methods has proved to be powerful overall, but high risks are associated with such practices. If stakeholders in charge of decision making are not able to interpret complex machine learning model results, this can negatively impact the confidence that such stakeholders have in acting on the insights given by such models. Decision Trees (DTs) have been known to be highly interpretable in the field. Adding more complexity to DT models can result in boosts in performance, but with the downside of sacrificing interpretability. As a solution, this paper presents the use of Genetic Decision Trees (GDTs, DTs optimised with a genetic algorithm (GA)) as an interpretable machine learning method for solving the Inventory Management (IM) Problem. Different parameters of the GA were investigated, along with different population initialisation techniques and maximum tree depths. The approach was also benchmarked against the most popular IM heuristics. GDTs were shown to outperform all common heuristics whilst maintaining interpretability, and they were also able to pinpoint features of the problem with real-world relevance. Additional, use-case-specific insights were also uncovered, highlighting the promising nature of the proposed approach.

1 Introduction

A supply chain is a network of businesses which work together to procure, manufacture, store and distribute a product or service to fulfill a customer demand (Ganeshan & Harrison, 1995) (Garcia & You, 2015). It comes naturally then that cooperation and information sharing between businesses in the supply chain leads to significantly higher profits overall (Gavimani, 2005). Over the past few decades, the information sharing revolution has produced exponential amounts of data (Tiwari, et al., 2018). This surge of data opened the door for computational methods such as machine learning (ML) to overtake traditional logistics used for supply chain optimisation, which suffered from their limited ability to forecast demand, and their inability to adapt to real-time data (Makkar, et al., 2020).

Optimising supply chains is crucial for businesses who want to stay competitive. Minimising inefficiencies increases profits by increasing order fulfillment and customer satisfaction. To this end, neural networks (NNs) have been the predominantly used ML method (Toorajipour, et al., 2021). NNs can extract intricate patterns from large datasets due to their inherent complexity. However, this complexity represents a double-edged sword. Whilst NNs and other black-box optimisation methods have demonstrated tremendous success in achieving state-of-the-art performance across various fields and real-world applications, their complexity also limits their interpretability (Zhang, et al., 2021). This trade-off is illustrated in Figure 1 for different ML techniques.

Interpretability of supply chain decisions is crucial for several reasons. Understanding the rationale behind the decisions made by optimisation models increases stakeholders' trust and confidence to act upon the insights given by the model. If inefficiencies arise, interpretability can help with identifying the root causes for issues and hold appropriate parties accountable. Certain industries might also be subject to regulations, in which case

interpretability of decisions is critical to avoid non-compliance. Interpretable models have also been shown to be able to reveal important patterns that more complex techniques overlooked (Caruana, et al., 2015).

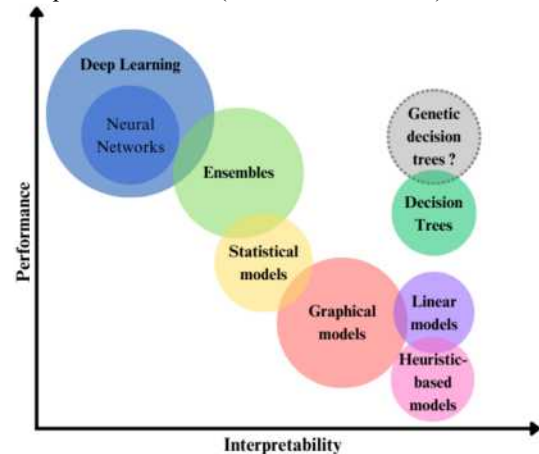


Figure 1: Model performance vs. model interpretability for different machine learning techniques. We expect our GDTs to perform better than individual decision trees and heuristic methods. This study does not investigate how GDTs perform against other methods. This figure was partially adapted from (Yang, et al., 2021).

Within the field of ML, there is a growing body of literature criticising the undervalued importance of interpretability over performance (Baryannis, et al., 2019). Governments across the globe are also developing new legislation around the need for increased transparency in the field of ML. The US with their “Blueprint for an AI bill of rights” (The White House, 2022), the EU’s “AI Act” (European Commission, 2021), and the UK’s “A pro innovation approach to AI regulation” (Secretary of State for Science, Innovation and Technology, 2023).

The aim of our study is to contribute to this growing, but still underdeveloped body of literature around interpretable methods in machine learning by promoting the use of Decision Trees (DTs) optimised with a genetic algorithm as an interpretable method for supply chain

optimisation, specifically to tackle the Inventory Management Problem.

To our knowledge, whilst genetic algorithms and DTs have been used separately to tackle supply chain optimisation in the past, we are the first to combine their use in this field.

2 Background

2.1 Defining interpretability

Interpretability is widely recognised by the ML community as being one of the most important aspects for improvement in the future of AI, and this sentiment is shared by governments, companies, and the public alike (Carvalho, et al., 2019). Despite this, there is no set consensus for the definition of interpretability. In part because it is difficult to mathematically define interpretability. Different academics and groups will vary in how they characterise interpretable models. As Miller (2019) puts it: “ultimately, it is a human-agent interaction problem”. Therefore, in this paper we define interpretability as the ability for a model’s result to be predicted by a human; in other words, a human will be able to follow through the model’s decision-making process even if they do not understand the rationale behind it.

Blockeel, et al. (2023) provides an extended list of different forms of interpretability which helps understanding how different methods can be interpretable in one or more ways as described below:

- (1) Understanding the full model
- (2) Understanding aspects of the full model (such as feature importance)
- (3) Understanding of a single prediction
- (4) Understanding the decision-making process behind a prediction

Within the field of interpretability, the word explainability is sometimes used interchangeably. However, it is important to point out their distinction. Explainable models are interpretable by default, but the opposite is not necessarily true (Gilpin, et al., 2019). We make the argument that explainable models, such as individual decision trees, are those which are interpretable in all four of the above ways.

2.2 Decision trees

Decision trees are sequential models that break down complex decision-making processes into a series of simple tests. Drawing inspiration from nature, decision trees start from a root node where the tree branches out from each sequential test (decision nodes) progressively until reaching a conclusive leaf node. For a visual depiction, see Figure 4.

Early research on decision trees for decision-making was focused on improving their performance, as it was widely acknowledged that individual trees were not able to compete with more complex models such as neural networks (Blockeel, et al., 2023). However, recent studies

have shown that certain tree methods such as ensembles and differentiable trees are able to perform as well, if not better, than complex neural networks whilst also being more interpretable (Silva, et al., 2020) (Frosst & Hinton, 2017) (Lundberg, et al., 2020). As a result, it is now widely accepted that such tree methods can compete with neural networks. Even so, these methods have shown that some level of interpretability is traded for their enhanced performance. For example, ensemble tree methods are inherently more complex than an individual decision tree. They lose interpretability when the decisions made by various trees are combined, adding a black-box nature to their decision-making process, and hindering understanding of how different variables contribute to the model’s predictions.

It is important to appreciate the trade-off between interpretability and performance of different decision tree methods to understand how further research in the field can advance both aspects of the DT model. In the next section, we propose how genetic decision trees can solve the issue of advanced DT methods inherently losing interpretability, whilst still being able to achieve good predictive performance.

2.3 Genetic decision trees (GDTs)

Pioneered by Holland (1975), genetic algorithms mimic the evolution of living organisms via natural selection and sexual reproduction to solve complex ill-defined problems (Holland, 1992). Through an adaptive process, genetic algorithms take a population of possible solutions (chromosomes), select those which have the higher fitness (those which have proven more useful), and use the genetic operators of crossovers and mutations to evolve said population. This is done iteratively, with the best chromosomes preserved from one generation to the next (elitism).

The combined use of genetic algorithms and decision trees has already been realised by various papers and found to produce robust and scalable predictions (Carvalho & Freitas, 2004) (Bala, et al., 1995) (Fu, et al., 2003) (Vandewiele, et al., 2017). More relevantly, these studies have also found that such an approach can reduce the description complexity of the model, thus improving interpretability of the results.

In this study, these properties of GDTs are leveraged to transform a sequential decision-making problem into a static, data-driven one.

2.4 The Inventory Management Problem

A critical aspect of supply chain optimisation is successfully managing inventory levels to meet uncertain customer demand. The Inventory Management Problem (IMP) describes the challenge of balancing the trade-off between meeting that demand (order fulfillment) and incurring holding costs for excess inventory. To do this, there are three main questions that modelling attempts should aim to answer. First, the frequency in which current inventory levels should be determined. Second, when

should a replenishment order be placed, and lastly how large should this order be.

In order to develop an optimal IM policy for each sequential stage of the supply chain, various constraints must be taken into account. These include supplier constraints (e.g. minimum order quantities, maximum production capacity) and internal constraints (e.g. storage capacity).

The costs associated with the IMP involve replenishment, holding, and shortage costs. The last of which refers to the costs incurred when a short-term shortage of inventory leads to backlog or loss of orders which result in lost profit on sales (Silver, 1981).

3 Methods

A schematic overview of the optimisation approach is presented in Figure 2 as a helpful reference to the reader. The individual steps within the approach will be explained in more detail in the upcoming sections.

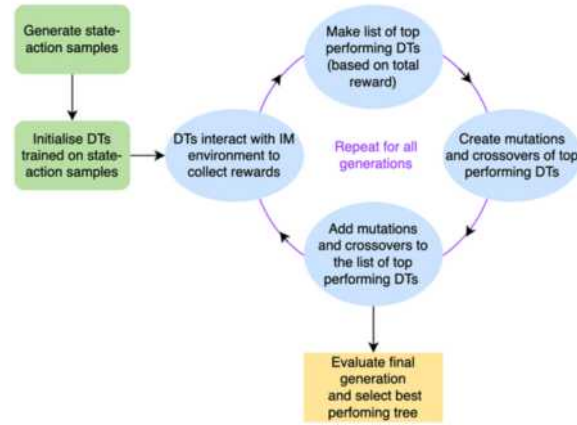


Figure 2: Genetic Decision Tree algorithmic framework.

3.1 Inventory Management environment configuration

To optimise a multi-stage IMP, the inventory management environment provided by Hubbs et.al. (2020) in their Operations Research gym (OR gym) is employed. OR gym is a Python library that is used to develop, test, and compare reinforcement learning techniques.

The IM environment's supply chain structure is illustrated in Figure 3 with the customer at the bottom of the chain. A retailer at stage 0 fulfills customer demand, sourcing products from the supplier at the stage directly above, which may also have another supplier above them. This hierarchy continues until reaching the initial supplier that processes raw materials.

Interacting with this environment involves providing an action (reorder quantity) as input. After that action is taken, the environment evaluates the supply chain and returns the resulting state and a reward. The state contains the resulting inventory levels and actions that were taken to get there, as far back as the value of the maximum lead time. For example, if the maximum lead time is 3 periods, the state will contain the current inventory levels, and actions from the past three time periods. The reward takes

into account costs and revenues from all stages, and thus represents the overall profit achieved at a given state. This sequence of taking an action and then observing the results is repeated for the specified number of time periods (i.e., until the end of the episode). The aim of solving the IMP is to maximise the total reward at the end of the episode.

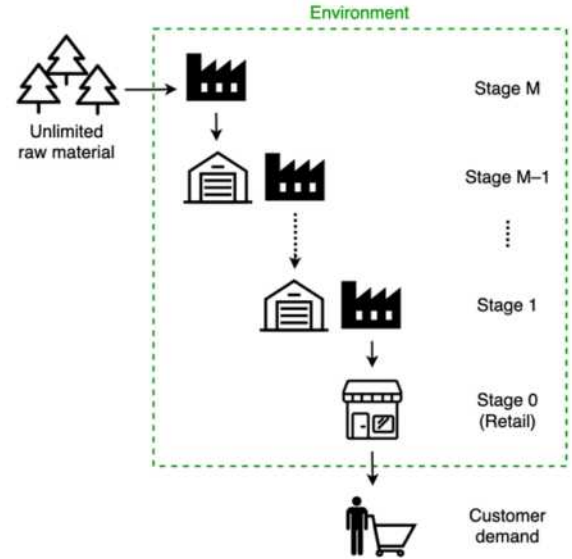


Figure 3: Inventory management supply chain set-up in the OR-gym environment. Partially adapted from (Hubbs, et al., 2020).

For the sake of simplicity, most environment parameters were left in their default state - however, in certain instances that was undesirable. The main reason for this is the tractability of the problem. Given the default parameters, the action space would be comprised of 744,471 actions, which was deemed too computationally demanding a problem, as there's scarce literature available to gauge how powerful the applied method would be *a priori*. Thus, the production capacities and initial inventories were sized down to reduce the size of the action space and state space. As a state is described by prior inventory levels and actions as far back as the biggest lead time, lead times were all deterministically set to 1 to further reduce the size of the state space (for a more detailed description of how a state is described, see (Hubbs, et al., 2020)). The rest of the parameters were left in their default state: an episode consisted of 30 episodes, the fluctuating customer demand was drawn from a Poisson distribution, the supply chain under consideration had four stages, and backlogging was allowed. Lastly, replenishment, backlog and holding costs per unit increased from stage 3 to 0, as more refined products cost more to produce and store. For more details on environment parameters, see the Supplementary Information.

3.2 Genetic Decision Tree Regressor (GDTR) class

The most suitable model for this study is a decision tree regressor. We opted to write a custom genetic decision tree regressor (GDTR) class which allows for the use of

crossover and mutation operators. The various methods involved in the GDTR class are described below.

3.2.1 Fit and predict methods

Binary splits were used, that is, the data is partitioned into a left and right dataset based on how the value of the split feature compares to the split threshold. To decide on the best split feature and threshold value, the Sum of Squared Error (SSE) was calculated for each possible split, and the minimum was selected. The formulation for SSE is shown in Eq. 1, where R denotes the part of the dataset going into the right partition, and L denotes that going into the left. Variables \bar{y}_R and \bar{y}_L denote the means of the two partitions.

$$SSE = \sum_{i \in R} (y_i - \bar{y}_R)^2 + \sum_{i \in L} (y_i - \bar{y}_L)^2 \quad (1)$$

Once the data is split into two sets at a given node, those two parts are split further recursively using the same technique until one of the stopping conditions is met. Two stopping conditions were defined: either the maximum depth needs to be reached, or the dataset needs to be left with only one datapoint (as then there is no point in splitting further). It should be noted that the data was standardised before fitting.

The ‘predict’ method is essentially a reverse of the ‘fit’ method, recursively checking how the datapoint to be predicted compares to the threshold value until a leaf node is reached. Then, the value of that leaf node is returned. If at any point the method encounters a criterion which cannot be fulfilled (e.g., $x[2] > 4$ followed by $x[2] < -3$) because a criterion from a node higher up has been modified by a genetic operator, it will default to the right-hand partition.

3.2.2 Crossover creation

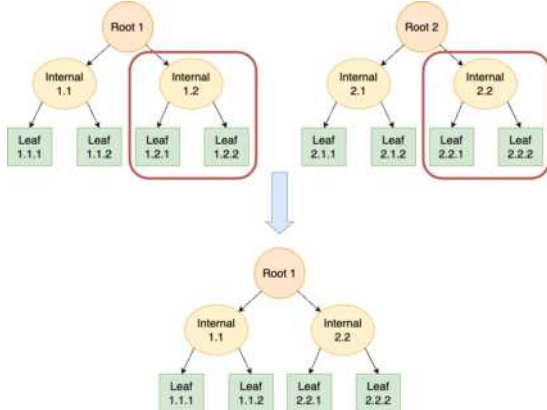


Figure 4: Visual representation of a decision tree crossover operation.

Crossovers are one of the two main genetic operators used to evolve a population. In the context of decision trees, this means exchanging subtrees of two parent trees. For the purposes of this study, the parent trees were selected randomly. Then, one internal node was selected from each parent tree randomly to act as the root node of the subtree to be exchanged. For a visual depiction of the

process, see Figure 4. It is to be noted that this operation might lead offspring trees to have a different depth than that of the parent trees, which may be helpful in avoiding suboptimal tree structures due to lack of sufficient exploration.

3.2.3 Mutation creation

Mutations are also genetic operators and, when discussing decision trees, they most often mean modifying one single node’s splitting criterion. In this paper the splitting criterion is modified by randomly selecting the split feature, and then perturbing the original threshold by a random percentage between -50% and +50%. The node to be mutated is also selected randomly.

3.3 Genetic algorithm framework

3.3.1 Generating initial training data

Decision trees (DTs) are classed as supervised learning algorithms, meaning that they require a training dataset to be trained on. Such a dataset was generated in three different ways within this study.

- (1) State-action pairs are generated randomly, leading to the initial tree population following purely random policies
- (2) Half of the state-action pairs are generated randomly, and the other half is generated using the base-stock policy heuristic (for more on this see Section 3.4)
- (3) State action pairs are generated fully using the base-stock policy heuristic

For the latter two approaches, each tree initialised using a heuristic was given 200 [state – base-stock action] data pairs to be trained on. It should be noted that stochasticity was also present in the latter two approaches, as the 200 states for which the heuristic was applied were selected randomly.

3.3.2 Fitting the initial DT population

To determine the number of trees in the population, the heuristic outlined in Eq. 2 was followed, where n_s is the number of variables describing a state (6 in this study) and n_a is the number of possible actions (60 in this study). Thus, the initial population was determined to consist of 1320 trees.

$$N_{initial\ trees} = (n_s + n_a) \times 20 \quad (2)$$

Each of these trees were then fit to data generated as described in the previous section, resulting in 1320 different policies.

3.3.3 Genetic algorithm loop

After the initial tree population has been created, a four-step sequence is carried out for 10 generations, after which the final best performing trees are selected:

1. The fitness of each DT's policy is evaluated by having it interact with the IM environment for 10 episodes, reviewing inventory levels at every time period. The end-of-episode rewards are averaged for the 10 episodes, resulting in a final "total average reward" for each of the trees.
2. The top performing DTs are selected to be kept, and the rest are discarded.
3. Crossovers are generated between the trees that were deemed to be the top performers – it is to be noted here that only one of the potential crossovers was generated, and that both parent trees were retained in their original states as well. Next, mutations of the top performing trees were generated and, similarly to crossovers, the original tree was also retained.
4. Newly generated crossovers and mutants were added to the list of top performing trees (so that there were 1320 trees in the population again), and the fitness of each of those trees was evaluated.

As part of the study, multiple setups of this GA were investigated by changing the percentage of trees being retained from the old generation and the percentage of trees which are results of crossovers and mutations. The effects of different maximum initial tree depths were also studied, for all GA strategies.

For a comprehensive tabulated view of all considered setups in this study, see Table 1.

Table 1: Experimental setup; all considered max depth and GA strategy combinations investigated in this study.

Maximum initial depths	Genetic strategy: percentage of trees from the previous generation which are:		
	Retained	Crossovers	Mutations
3, 5, 10, 13	50	30	20
3, 5, 10, 13	40	40	20
3, 5, 10, 13	30	40	30
3, 5, 10, 13	20	50	30
3, 5, 10, 13	10	50	40

Finally, the features of the 5 best performing models were investigated for each initialisation type, max depth, and GA strategy which yields 300 models in total. The investigation included an examination of both the overall occurrences of individual features and the occurrences of each feature when utilised as the initial (root) splitting feature.

3.4 Benchmarks

Heuristic methods are widely used for benchmarking IMPs (Jackson, et al., 2020). In this study we use five of the most popular ones to benchmark our GDTR method. These heuristics are described below:

- *Base-stock policy*: inventory levels are reviewed at each period, and the reorder quantity is placed to bring the inventory levels to a predefined base-stock level
- *(R, S) policy*: same as the base-stock policy, except inventory levels are only reviewed every period R (and the base-stock level is denoted by S)
- *(R, s, S) policy*: inventory levels are reviewed every period R, and if they are found to be below the minimum level s, a reorder is placed to bring the inventory levels to S
- *(r, Q) policy*: inventory levels are continuously reviewed, and if they are found to be below the minimum level r, a reorder of size Q is placed
- *(s, S) policy*: same as the (R, s, S) policy, except inventory levels are reviewed continuously. Also known as the 'Min-max' policy

The heuristics were run in an environment whose configuration was identical to that used for evaluating the GDTR. SciPy's minimise method (Pedregosa, et al., 2011) was used to find the optimal parameters for the heuristic policies, by reformulating them as black-box problems taking the policies' parameters as inputs. Additionally, a random agent taking a random action at each step was also tested.

4 Results

4.1 Benchmarking

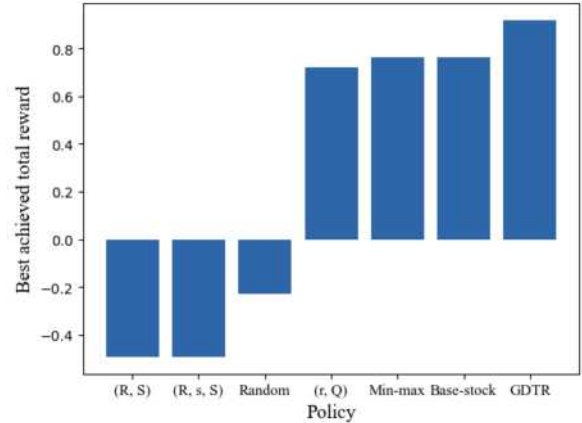


Figure 5: Benchmark test results for 5 different heuristics and a random policy against the Genetic Decision Tree Regressor (GDTR).

Benchmark tests reveal that the GDTR policy performs better than any of the other heuristic approaches outlined in Section 3.4. The highest achieved reward is shown for each policy in Figure 5. Even in the best-case scenario, the random, (R, S), and (R, s, S) policies were not able to break even in costs. In comparison, the (r, Q), min-max and base-stock policies performed much better, yet they were still outperformed by the GDTR. The best GDTR policy outperforms the worst-performing benchmark heuristic, (R, S), by 287%, and it performs better than the base-stock policy by 20%.

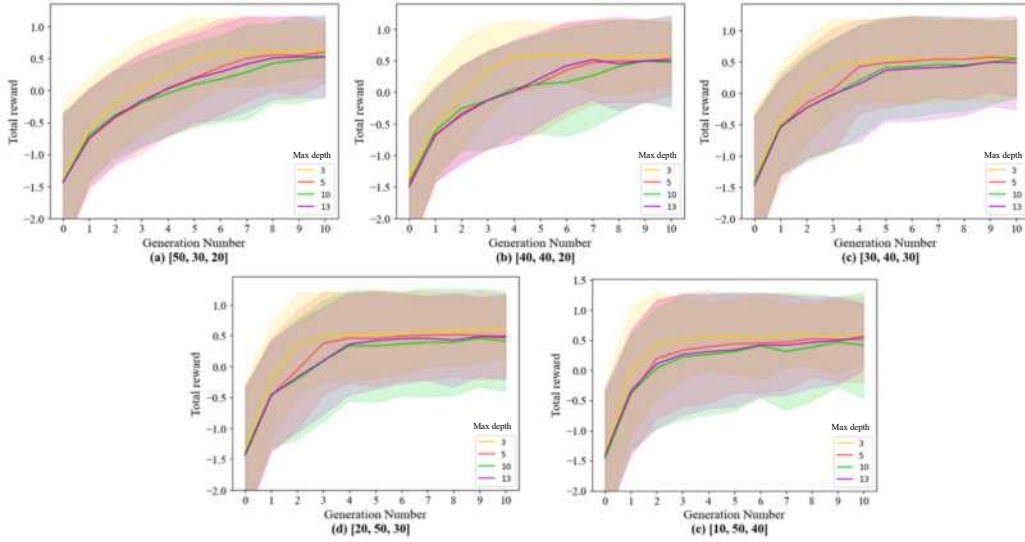


Figure 6: Genetic decision tree regressor reward functions using randomly initialized trees for evolution strategies (a)-(e) and different starting max-depth trees. The different strategies are identified by the number of trees which are [retained, crossover-ed, mutated].

4.2 Investigating the optimal combination of initialisation and GA strategies

Figures 6, 7 and 8 visualise how different combinations of genetic strategies, initial max depths of trees and initialisation methods yield different mean total rewards (solid lines) and standard deviations (shaded regions, $\pm 1\sigma$) across generations. The ‘mean of total rewards’ refers here to the mean of the rewards achieved by the populations of trees at each generation – it will be referred to as the “reward” in this section for simplicity.

Figure 6 displays how using a random initialisation method affects the performance of subsequent generations of trees for five different genetic strategies and four different max depth of trees. The initial rewards obtained for all strategies and depths were similar in value (-1.5), and notably this initial reward is negative. This is followed by a steep increase in reward from generation 0 and 2, after which most reward functions reach a positive value

(breaking even). The steepness of initial gradients is observed to increase with increasing proportion of crossovers and mutations across strategies (a) to (e). Generally, all strategies from generation 8 onwards seem to converge to reward values between 0.0 and 0.5.

For the vast majority of generations, the reward function for max depth 3 performs the best consistently for all strategies and improves at a faster rate than the rest of the depths. From strategy (a) to (e), the shaded regions representing the standard deviation of the rewards steadily increase. Note that no yellow regions are seen at the lower end of the figures, corresponding to trees with initial max depths of 3 performing better than other depths. On the other hand, it seems that for trees with initial max depths of 10, overall improvement is the slowest and with the highest standard deviation (as shown by the green shaded areas).

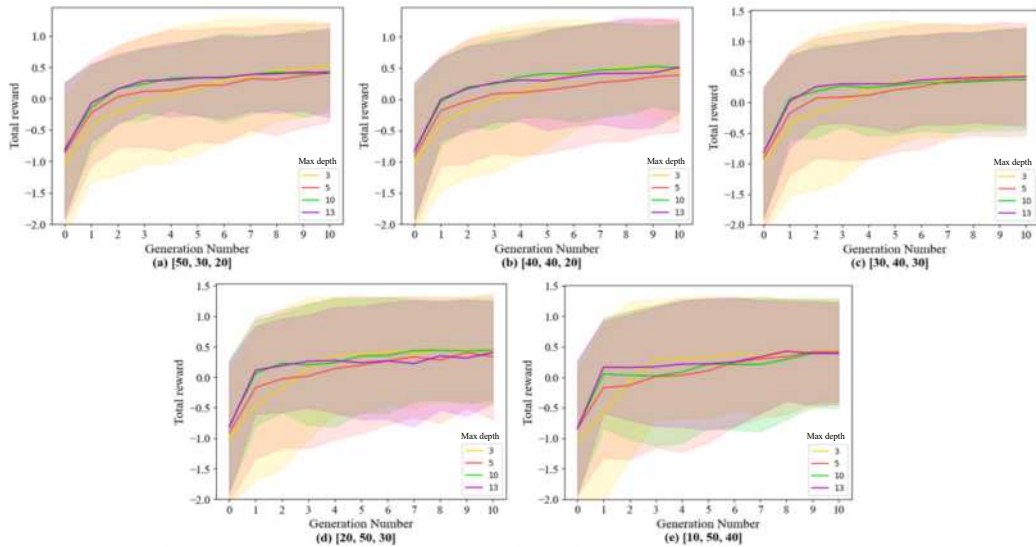


Figure 7: Genetic decision tree regressor reward functions using half random-half heuristic initialized trees for evolution strategies (a)-(e) and different starting max-depth trees. The different strategies are identified by the number of trees which are [retained, crossover-ed, mutated].

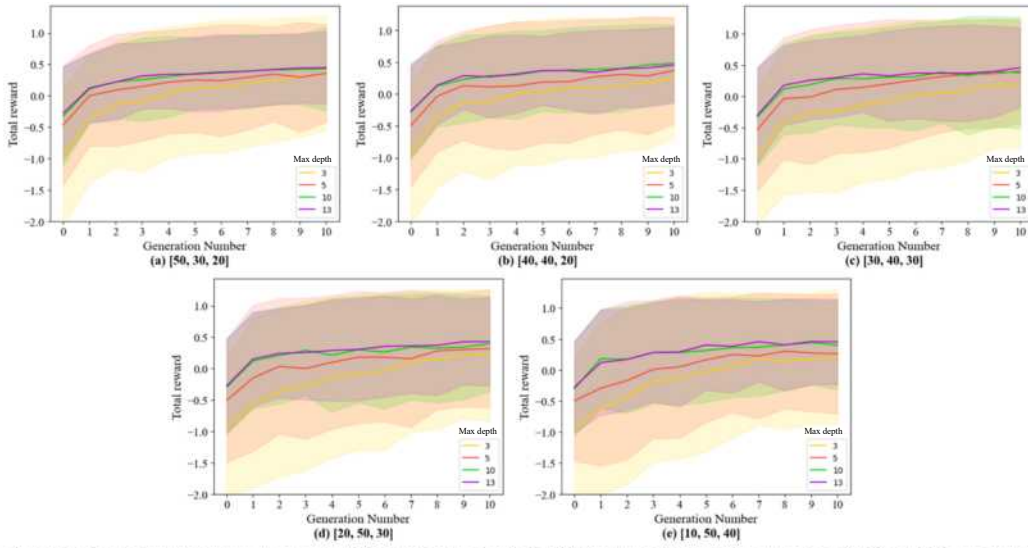


Figure 8: Genetic decision tree regressor reward functions using heuristic initialized trees for evolution strategies (a)-(e) and different starting max-depth trees. The different strategies are identified by the number of trees which are [retained, crossover-ed, mutated].

Figure 7 investigates the same parameters as Figure 6, but for a half-heuristic, half-random initialisation. The general upwards trend across generations is preserved, however, the initial starting point is generally higher (approx. -1, -0.5) than it is for fully random initialisation. Additionally, there is a steep initial gradient of improvement for all GA strategies, not just for the ones where less of the old population is retained.

The trend of lower max depths performing better seems to start reversing, however, the reward functions for the different depths run too closely together to tell exactly. Further, the standard deviations for lower max depth trees (3, 5) have increased when compared to the random initialisation, especially for GA strategies where more of the older generation is retained (strategies (a)-(c)). This may be seen best by observing the increased presence of yellow and red regions in the bottom half of the figures.

Figure 8 shows the results for a fully heuristic initialisation. The general upwards trend of mean rewards across generations for all GA strategies and max depths is preserved with this initialisation type as well. The steeper initial gradient of the reward function is also present, albeit the curve is less steep, as the starting reward is better than that of the previous two initialisation types (approx. -0.25, -0.9). With a fully heuristic initialisation, trees with lower max depths (3, 5) generally perform worse, with max depth 3 visibly performing the worst for strategies (a)-(d). This trend is particularly prominent in earlier generations. Moreover, the standard deviation in achieved rewards increases as max depth decreases, for all GA strategies—the prominence of the yellow and red regions in the bottom regions of the figures is the most noticeable here amongst all initialisation types.

Finally, the different GA strategies seem to be more similar to each other than they are at the previous initialisation types; employing a fully heuristic initialisation decreased the importance of the GA strategy but increased that of the max depth.

4.3 Assessing interpretability of top performing trees

Figure 9's blue bars show the number of occurrences for each of the features used for splits in the top performing trees. The first three features (prior inventory levels) were used much more than the last three (prior actions). There is also a decrease in how much the features were used across the inventory levels, and across actions.

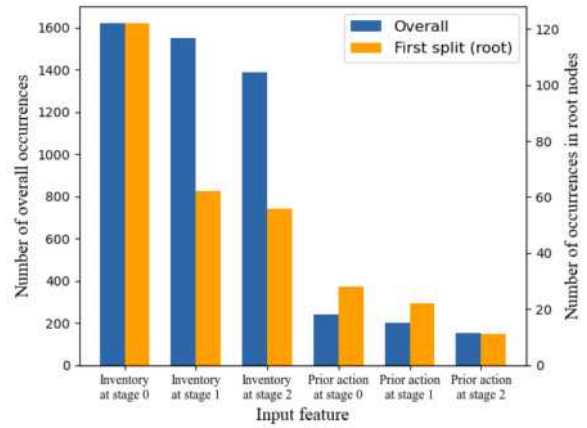


Figure 9: Feature importance for features 0-5 (left to right). Blue: feature occurrence at all root and decision nodes of the top performing trees. Orange: feature occurrence at the root node of the top performing trees. Features represent the splitting criteria at each root or decision node in the decision tree.

The split features at the root nodes were also examined, with the results presented in Figure 9's orange bars. A trend like the one described above was found, that is, the features corresponding to the inventory levels were much more frequently used as a first split than those corresponding to prior actions. However, in this instance the usage of feature 0 was outstandingly high, approximately double the usage of the second most used first split feature (120 vs. 60 occurrences). This trend is

further illustrated in Figure 10, which depicts one of the best performing trees.

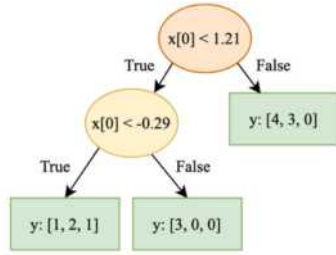


Figure 10: Example output of the approach: an interpretable decision tree of max depth 3. $x[i]$, where i is an integer between 0 and 5, refers to the splitting feature at a specific decision node. For example, $x[0]$ is the standardised current inventory at stage 0 of the supply chain. The variable y refers to the action the tree has outputted at a leaf node, where the values inside the brackets denote the reorder quantity at [stage 2, stage 1, stage 0]. Note that, at stage 3 we have unlimited raw materials, so there is no need for reordering at that stage.

5 Discussion

5.1 Benchmarking

Our genetic decision tree regressor performs better than 6 different heuristics and sequentially improves the performance of an initial population of trees to achieve higher rewards generation after generation. This is in line with existing literature around the ability of genetic algorithms to improve the performance of decision tree models, and it also confirms our expectation that GDTs perform better than heuristic-based models and individual DTs as illustrated in Figure 1.

As mentioned in Section 2.4, the frequency in which inventory levels should be reviewed is an important consideration. For the GDT, this was done at every time period. Regarding the benchmarks, neither of the two worst performing heuristic policies were allowed to do so, suggesting that having regular inventory reviews is advantageous in this setup.

5.2 Investigating the optimal combination of initialisation and GA strategies

Figures 6-8 show that there is a clear trend of improvement for all combinations of strategies as trees evolve from one generation to the next. From this, we can determine that our GDT method was able to improve the performance of an initial population of trees even when such trees were generated using a heuristic, which further supports the validity of our benchmarking results.

The trends seen in Figures 6-8 seem to be of similar shape at a glance. Deeper analysis into the effects of the initialisation method, max depth of trees and GA strategy reveal that they affect the starting rewards, improvement rates, standard deviations, final rewards, and convergence behaviour.

5.2.1 Effect of initialisation method and max depth of trees

The initialisation strategy had a profound influence on the performance of the GDT regressor, and it likely affects how different depth trees will evolve from earlier to later

generations. We see this in Figures 6-8, with the same GA strategies producing similar, but different trends for different initialisation methods.

A fully heuristic approach gives reward functions which increase approximately parallel to each other and seem to strongly suggest that the best depth trees to use are of max depth 10 or 13. On the other hand, initialisation methods with a fully random or half random approach tend to show the reward functions intersecting liberally. Although there is still some consistency as to which functions perform better overall, for different strategies there is no clear “winner”.

Certain depth trees might perform better than others as shorter trees might suffer from lacking the complexity to fit data accurately (low max-depth, heuristic initialisation), whilst higher depth trees are able to fit to more complex data but run the risk of overfitting and thus performing worse (high max-depth, random initialisation). The higher rates of intersection between different depths’ reward functions for the half-heuristic, half-random initialisation seem to illustrate this trade-off.

Using a fully heuristic initialisation also gives higher initial rewards, as expected. However, the reward functions do converge towards mean rewards which are interestingly slightly lower than those achieved by the half-random and fully random approaches. This might suggest that the fully heuristic method is not encouraging enough exploration, leading most of the trees following the algorithm to converge to locally optimal solutions. This, however, does not mean that the fully heuristic initialisation method did not produce top performing individual trees.

Further, the standard deviation of shallower trees increases and their performance decreases as more trees follow heuristically initialised policies. This is potentially due to deeper trees being able to capitalise better on the additional insights provided by the heuristics. This trend is especially prominent in earlier generations, as there was less opportunity to select fitter trees and close the gap between the performance of shallow and deep trees.

5.2.2 Effect of GA strategy

As the percentage of retained trees decreases from strategies (a) to (e), higher initial improvement rates are witnessed when there is randomness in the initialisation, yet the average rewards achieved across strategies at generation 10 are similar. This suggests that whilst increasing the rates of crossovers and mutations might not directly increase the rewards achieved at later generations, it does benefit the evolution of the initial population of trees by increasing the amount of initial exploration the algorithm is allowed. Thus, it can be said that under circumstances where there are limitations on budget, time or computing power, utilising higher rates of genetic operators can reduce the resources needed to reach a converging result. However, this also leads to an increase in the standard deviation of results, which might be undesirable. On the other hand, for the fully heuristic initialisation the effect of the GA strategy is less

pronounced as starting rewards are already higher, as discussed above.

For GA strategies with lower retention percentages, shallower trees have a higher standard deviation when there is some heuristic element present in the initialisation. This points to shorter trees being potentially more sensitive to crossovers and mutations giving them a more profound change than deeper trees would experience. To visualise this, one can think of a tree of depth 3 and another of depth 13 both undergoing a mutation at one of the decision nodes adjacent to one of the leaf nodes. In the grand scheme of things, the tree with depth 3 would be affected more than the tree with depth 13.

5.3 Assessing the interpretability of top performing trees

Features corresponding to inventory levels were found to be better predictors for optimising reorder quantities than features corresponding to prior actions. This was true even with non-zero lead times. Given a specified (albeit stochastic) customer distribution, this would make intuitive sense, and shows the model's ability to uncover real-world insights even when the environment setup presents a rather specific case-study.

The decreasing importance of features as one moves further away from the stage closest to the customer may be explained by their influence on how much of the end customer's order is getting fulfilled. The more an order is fulfilled, the less money is lost to backlogging. What makes this especially important at stage 0 is that replenishment, backlogging and holding costs are higher the closer one is to the end customer (more refined products cost more to store and produce), which means even smaller mismanagements lead to higher losses. Additionally, specifying suboptimal reorder quantities closest to the end customer could create a ripple effect, not only affecting that stage, but constraining the maximum achievable profit of all preceding stages. The exceptionally high occurrence of feature 0 (the inventory level at the stage closest to the customer) as the first, most influential split feature of trees further emphasises this point.

By visualising our best performing trees (Figure 10), we can clearly follow through the decision-making process behind individual predictions. This being that whenever the inventory levels at stage 0 are above a certain threshold, the model predicts that there is no need to reorder at that stage. It does, however, predict the need for reordering at higher stages, probably in anticipation of future demand.

Since the model's decision-making process can be easily navigated by a human, the method was not only proven to yield good rewards but was also found to be highly interpretable.

5.4 Limitations and further research

One of the most relevant limitations of this work is the reliance on a theoretical framework (OR-gym) for inventory management. After all, for this method to be

leveraged in real-world applications, additional research needs to be conducted to substantiate the practical utility of genetic decision trees. This is understandably a challenge as the real world also introduces much more uncertainty, but it is necessary to expand the confines of this study for the method to be more widely applicable to fields outside of supply chain optimisation.

Time and computational constraints limited the number of generations, initialisation methods, GA strategies and max depth trees that were possible to investigate. As a proof-of-concept study, this work was able to show that genetic decision trees indeed show promise in optimising IMPs, but further studies should expand on the parameters investigated and look at exploring higher generations and tree depths, as well as more test cases for different GA strategies. This is important in order to gain higher levels of confidence that an optimum has been reached, because in our study we had reward functions which ran fairly close to each other. Consequently, it was difficult to guarantee that an optimum had been found given our limited parameter space. Thus, we further encourage the use of new parameter investigations, for example, introducing different genetic operators.

A further study on benchmarking GDTs against more advanced methods would be crucial to gauge the gap between the proposed method and those at the forefront of the ML field. This includes benchmarking studies against black-box optimisation methods (namely deep neural networks), and other popular methods such as ensembles. This would aid stakeholders in making well-informed decisions when choosing the appropriate optimisation method for their specific application based on the trade-off they're willing to make between interpretability and performance.

Open sourcing our GDTR class can also invite additional collaboration to increase the functionality of our approach. For example, through the creation of more sophisticated interpretation tools to gain further insights into features, splits and decision patterns.

6 Conclusion

This proof-of-concept study investigated a novel approach to optimising an Inventory Management Problem using genetic decision trees.

Benchmark tests against the most relevant heuristics for an IMP revealed that there is promise in using genetic algorithms for improving the performance of tree models whilst maintaining interpretable results.

Parameter tests varying genetic strategies, max depth of trees and initialisation methods were not able to produce clear answers as to which combination of strategies would produce the best performing trees. The stochasticity of the problem also contributes to this problem. So, we have outlined further investigations to be undertaken that could improve the knowledge around which parameters would produce the best predictions.

Nevertheless, final analysis of the top performing trees at the end of the genetic algorithm loop shows that

this approach is indeed interpretable and has the potential to highlight features with real-world relevance. Referring to Section 2.1, the model is interpretable in all four ways described by Blockeel et al. (2023), so there is reasonable ground to claim that this novel approach is not only interpretable, but also inherently explainable. These results affirm the idea that a GDT approach is a promising candidate in the intersection of interpretable machine learning and inventory management optimisation.

7 Acknowledgements

We would like to thank Ilya Sandoval and Niki Kotecha for their invaluable feedback and support.

8 Supplementary information

Access the code for this paper through the following link: <https://github.com/jb-gell/Interpretable-SC-Opt-GDT>. An extended catalogue of tabulated data of results is also available in the supplementary information document.

9 References

- Ganeshan, R. & Harrison, T. P., 1995. *An Introduction to Supply Chain Management*, Department of Management Sciences and Information Systems, 303 Beam Business Building, Penn State University, University Park, Pennsylvania: s.n.
- Garcia, D. J. & You, F., 2015. *Supply chain design and optimization: Challenges and opportunities*, s.l.: Computers and Chemical Engineering.
- Gavirneni, S., 2005. *Information Centric Optimization of Inventories in Capacitated Supply Chains: Three Illustrative Examples*, Boston, MA: Springer.
- Tiwari, S., Wee, H. M. & Daryanto, Y., 2018. Big data analytics in supply chain management between 2010 and 2016: Insights to industries. *Computer and Industrial Engineering*, Volume 115, pp. 319-330.
- Makkar, S., Devi, G. & Solanki, V., 2020. Applications of Machine Learning Techniques in Supply Chain Optimization. *ICICCT 2019 – System Reliability, Quality Control, Safety, Maintenance and Management*.
- Toorajipour, R. et al., 2021. Artificial intelligence in supply chain management: A systematic literature review. *Journal of Business Research*, Volume 122.
- Zhang, Y., Tino, P., Leonardis, A. & Tang, K., 2021. A Survey on Neural Network Interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5), pp. 726-742.
- Caruana, R. et al., 2015. *Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission*. s.l., s.n.
- Yang, G., Ye, Q. & Xia, J., 2021. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Computer Science and Engineering*.
- Baryannis, G., Samir, D. & Antoniou, G., 2019. Predicting supply chain risks using machine learning: The trade-off between performance and interpretability. *Future Generation Computer Systems*, Volume 101.
- T. W. H., 2022. *Blueprint for an AI bill of rights*. s.l.:<https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf>.
- E. C., 2021. *The AI Act*. s.l.:s.n.
- S. o. S. f. S. I. a. T., 2023. *A pro-innovation approach to AI regulation*. s.l.:s.n.
- Carvalho, D. V., Pereira, E. M. & Cardoso, J. S., 2019. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics*, 8(832).
- Miller, T., 2019. Explanation in artificial intelligence: Insights from the social sciences.. *Artificial Intelligence*.
- Blockeel, H. et al., 2023. Decision trees: from efficient prediction to responsible AI. *Frontiers in Artificial Intelligence*, Volume 6.
- Gilpin, L. H. et al., 2019. Explaining Explanations: An Overview of Interpretability of Machine Learning.
- Silva, A. et al., 2020. Optimization Methods for Interpretable Differentiable Decision Trees in Reinforcement Learning. *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics 2020*, pp. 1855-1865.
- Frosst, N. & Hinton, G., 2017. Distilling a Neural Network Into a Soft Decision Tree. *Comprehensibility and Explanation in AI and ML*.
- Lundberg, S., Erion, G. & Chen, H., 2020. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature Machine Intelligence*, 2(1).
- Holland, J. H., 1975. *Adaptation in Natural and Artificial Systems*. Ann Arbor, MI: University of Michigan Press.
- Holland, J. H., 1992. Genetic Algorithms. *Scientific American*, 267(1), pp. 66-73.
- Carvalho, D. R. & Freitas, A. A., 2004. A hybrid decision tree/genetic algorithm method for data mining. *Information Sciences*, 163(1-3), pp. 13-35.
- Bala, J., Huang, J., Vafaie, H. & DeJong, K., 1995. *Hybrid Learning Using Genetic Algorithms and Decision Trees for Pattern Classification*. s.l., s.n.
- Fu, Z. et al., 2003. A Genetic Algorithm-Based Approach for Building Accurate Decision Trees. *INFORMS Journal on Computing*, 15(1), pp. 3-22.
- Vandewiele, G. et al., 2017. A Genetic Algorithm for Interpretable Model Extraction from Decision Tree Ensembles. *Trends and Application in Knowledge Discovery and Data Mining*.
- Silver, E. A., 1981. Operations Research in Inventory Management: A Review and Critique. *Operations Research*, 29(4), pp. 628-645.
- Hubbs, C. D. et al., 2020. OR-Gym: A Reinforcement Learning Library for Operations Research Problems. *Artificial Intelligence*.
- Jackson, I., Tolujevs, J. & Kegenbekov, Z., 2020. Review of Inventory Control Models: A Classification Based on Methods of Obtaining Optimal Control Parameters. *Transport and Telecommunication*, 21(3), pp. 191-202.
- Pedregosa, F. et al., 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, Volume 12, pp. 2825-2830.

In-Situ FTIR Spectroscopy of Interactions Between High-Pressure CO₂ and Porous Liquids

Feyintoluwa Mojeyinoluwa Delano and Victoria Ebubechukwu Nwokolo
Department of Chemical Engineering, Imperial College London, U.K.

1. Abstract

Type III porous liquids, a combination of porous solids and ionic liquids, have the potential to be deployed as a carbon capture storage (CCS) technology. The aim of this paper is to systematically investigate the CO₂ sorption process within type III porous liquids with a focus on swelling, CO₂ sorption capacity and the nature of CO₂-porous liquid interactions. In addition, the influence of pressure and ionic solvents on the sorption process is elucidated in this paper. Three type III porous liquids namely, 12.5 % ZIF-8 [BMIM][NTf₂], 12.5 % ZIF-8 [P₆₆₆₁₄][NTf₂] and 12.5 % ZIF-8 [BPy][NTf₂] were examined. In-situ ATR-FTIR spectroscopy was employed to probe the CO₂-porous liquid interactions, measure the swelling and CO₂ sorption capacity under diverse conditions. Samples of porous liquids were exposed to varying CO₂ pressures (5 – 40 bar) for 20 minutes, after which the cell was depressurized as CO₂ was released. It was found that CO₂ sorption and swelling increase as pressure increases and decrease as the steric hinderance of the porous liquid cations increases. Upon the release of CO₂ pressure, analysis of the spectra produced indicates a residual absorbance at the spectral band corresponding with the anti-symmetric stretching mode (ν_3) of CO₂. This observation is evidence that the CO₂-porous liquid interactions are not completely reversible. Irreversible interactions between CO₂ and type III porous liquids may affect the ease of regenerating the porous liquid and adversely impact its deployment as a CCS technology. To engineer porous liquids that efficiently capture CO₂, understanding the interactions that exist between CO₂ and porous liquids is vital to lay a foundation for its implementation in industry.

2. Introduction

Carbon dioxide (CO₂) has been identified as the greatest contributor to global warming (1). With CO₂ emissions on the rise due to industrial processes and energy production, the Paris Agreement was pivotal in rallying global effort to combat climate change (2). Although the penetration of renewable and nuclear energy into the world's energy mix has increased over the past few decades (3), conventional fossil fuels are still being used. While the world transitions to low-carbon energy, carbon capture storage (CCS) systems are needed to capture the CO₂ being released from burning these carbon intensive fuels. Supercritical CO₂ (scCO₂) has the potential to be a green solvent with applications in extracting natural products, counter current separation of liquid natural products, impregnation with supercritical fluids and supercritical drying, cleaning and degreasing (4). To achieve the goal of limiting global temperature increase to 1.5 °C by 2050 (5), the need to sequester CO₂ cannot be overemphasized. The most common CCS technology is chemical absorption using amine-based solvents. However, there is a high energy cost incurred in regenerating the solvent which negatively impacts the efficiency of the process (6). To improve the economic viability and environmental sustainability of the carbon sequestration process, novel CCS technologies need to be developed via adsorption using porous substances such as activated carbon, metal organic frameworks (MOFs), zeolites, porous liquids etc (7, 8).

Porous liquids (PLs) have shown potential to overcome the limitations currently facing other novel CCS technologies such as flowability. The main attraction of PLs is the possibility of process simplification due to their ability to be pumped into circulation and the potential for increased energy efficiency (9). PLs possess the mobility of liquids and the features of microporous solids. These PLs differ from conventional liquids in their permanent and accessible cavities (10).

PLs can currently be classified into four types (types I, II, III and IV) according to the method of porosity generation (11, 12). Among these four types of PLs, Type III PLs (T3PLs) are more widely discussed and studied due to their ease of preparation and consequently used for the purpose of this research (13). T3PLs consist of sterically hindered solvents and pore generators and present efficient sorption capabilities due to its permanent cavities. T3PLs can be produced using either ionic or non-ionic liquids as the sterically hindered solvent and a wide variety of porous frameworks (known for strong adsorption capabilities) may be used as pore generators such as zeolites, MOFs, porous organic cages (POCs) etc. It is easier to prepare T3PLs via well-established methods of production including surface hydrophobization or size-excluded dispersions since T3PLs are a suspension and not a neat liquid (14).

MOFs, one of the commonly utilised porous substances for CCS technologies, are a class of

crystalline material composed of a central metal ion surrounded by organic ligands (15). Their high porosity and large surface area make them a suitable porous framework for T3PL synthesis. The customizable nature of MOFs amongst other properties, captures the interest for a variety of applications such as catalysis, gas storage, drug delivery, gas vapor separation, water treatment, and CO₂ capture (16). Zeolite imidazolate frameworks (ZIFs) are a type of MOFs which derive their structure from tetrahedral units where every metal ion (M) attaches to four organic imidazolate (Im) linkers [M--Im--M] (Figure 1). ZIFs have a topology similar to zeolites and are sought after for the combined advantages of MOFs and zeolites (17). ZIF-8 has proven to be effective in separating CO₂ from a mixture of gases via adsorption and was consequently selected for this study (17, 18).

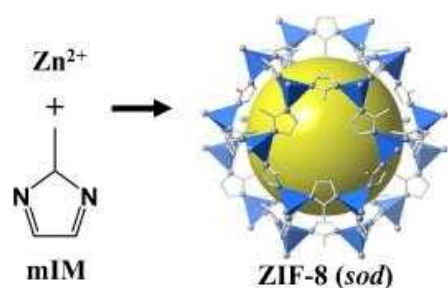


Figure 1: Crystal structure of ZIF-8: Zn (polyhedral), N (sphere), and C (line) (15)

T3PLs have been shown to present greater control and tunability over desired properties making T3PLs a good candidate for CO₂ sorption (16). The solvent used in producing T3PLs is vital as it must be sufficiently bulky to prevent invasion of the porous structure, leaving it empty for CO₂ sorption (19, 20). In selecting a porous framework and solvent combination, the characterisation of the porous framework was evaluated in order to ensure the size of the solvent molecules were suitable for the pore size of the porous framework to prevent high accessibility of the solvent molecules to the manufactured cavities (21). This allows for further tailoring of PLs in selective gas separation (22, 23).

The interactions between CO₂ and T3PLs are not limited to physical interactions. The possibility of chemical interactions between CO₂ and the PLs may hinder the ease of regeneration and CO₂ sorption capacity of PLs. Thus, further investigation into the CO₂-PLs interactions is required. In this paper, the interactions between high-pressure CO₂ and PLs were systematically studied and analysed using ATR-FTIR spectroscopy. Three ionic liquid solvents, [BMIM][NTf₂], [BPy][NTf₂] and [P₆₆₆₁₄][NTf₂] (known for their impressive CO₂

absorption properties) and the porous framework, ZIF-8 (MOF) formed the basis of this study.

FTIR spectroscopy, a vibrational spectroscopic technique, has broad applications in fields ranging from biopharmaceuticals (24), petrochemicals (25), material science (26) and forensics (27). The non-destructive and rapid qualities of the FTIR spectroscopy led to its popularity. The Attenuated Total Reflectance (ATR) technique has lower penetration depth making it suitable to analyse high absorbing samples (28, 29). The changes in the modes of vibration of CO₂ and PLs allows for the use of in-situ ATR-FTIR spectroscopy to gain insight into their behaviours (30).

The objective of this paper is to deepen the understanding of the sorption of CO₂ in PLs with the aid of in-situ ATR-FTIR spectroscopy. To achieve this, the impact of pressure and functional groups in PLs on the swelling of the PL sample, CO₂ sorption capacity of the PL and nature of CO₂-PL interactions will be investigated.

3. Methodology

3.1 Materials

The T3PLs used in this study were synthesised from the following ionic liquids (ILs) as solvents.

Ionic Liquids

1. 1-Butyl-3-methylimidazolium bis(trifluoromethylsulfonyl)imide [BMIM][NTf₂]
2. N-Butylpyridinium bis(trifluoromethanesulfonyl)imide [BPy][NTf₂]
3. Trihexyltetradecylphosphonium bis(trifluoromethylsulfonyl)imide [P₆₆₆₁₄][NTf₂]

The addition of the ZIF-8 porous framework to the ILs above yielded the following T3PLs.

1. 12.5 % ZIF-8 [BMIM][NTf₂]
2. 12.5 % ZIF-8 [P₆₆₆₁₄][NTf₂]
3. 12.5 % ZIF-8 [BPy][NTf₂]

3.2 ATR-FTIR Spectroscopy

The ATR-FTIR spectrometer, BRUKER EQUINOX 55, was used in conjunction with the computer software 'OPUS' to obtain the absorbance and wavenumbers for the IR spectra. The ATR-FTIR spectrometer utilizes an IR light source which was then passed through a golden gate ATR accessory located in the sample chamber of EQUINOX 55, a detector and a diamond placed on the ATR accessory. A Teflon O-ring was placed above the crystal and secured in place to seal the high-pressure cell and prevent leakage. CO₂ pressure within the cell was monitored and controlled using a pressure gauge and six pressure valves. Specac

Heated Golden Gate Controller connected to the sample chamber allowed for temperature control. For all measurements of PLs in this investigation, 64 co-addition scans and a spectral resolution of 4 cm^{-1} were used.

3.3 Experimental Procedure

The PL samples were pipetted onto the diamond crystal surface. At 20°C , the samples were subject to varying CO_2 pressures for 20 minutes (at which point equilibrium was reached) after which the CO_2 pressure was released. Measurements were taken in 5-minute intervals. CO_2 pressure was varied with 5 bar intervals, from 5 bar to 40 bar. Between runs, the crystal was cleaned with acetone to remove any remnants that may obscure readings. To achieve robust results, each experiment configuration for all PL samples was conducted three times. The repeats allowed for the precision and reliability of experimental results by determining the mean and standard deviation of the results. The absence of a spectral band in the 2388 cm^{-1} region in the spectrum confirms that gaseous CO_2 does not interact with the IR radiation and obscure the results.

3.4 Quantitative Analysis

3.4.1 Swelling Phenomenon

Swelling phenomenon is the increase in volume observed with the introduction of CO_2 into the sample. The higher the amount of CO_2 sorbed the greater the increase in volumetric and sorption-induced strain. Swelling was calculated using the ratio of absorbance at a characteristic band of the sample before and after CO_2 is sorbed into the high-pressure cell.

$$S = \frac{A_0}{A} - 1 \quad (\text{Equation 1})$$

where S is the swelling extent, A_0 and A are the absorbance of a spectral band before and after CO_2 sorption into the sample (31).

3.4.2 CO_2 Sorption

The concentration of CO_2 sorbed into the sample can be calculated using the Beer-Lambert Law which states a linear relationship exists between the absorbance and the concentration, molar absorption coefficient and optical path length of a sample (32).

$$A = \epsilon cl \quad (\text{Equation 2})$$

where A is the absorbance, ϵ is the molar absorptivity, c is the concentration, and l is the pathlength (31). The concentration of CO_2 in the sample was calculated using the absorbance at 2339 cm^{-1} which corresponds with the anti-symmetric stretching mode (ν_3) of CO_2 . The pathlength, l , was based on the effective depth of IR penetration in the sample. Unlike the depth of

penetration, it factors in IR absorption in transmission (31). The effective depth, d_e , was calculated using the p -polarized effective thickness, $d_{e,||}$, and s -polarized effective thickness, $d_{e,\perp}$ (33).

$$d_e = d_{e,||} + d_{e,\perp} \quad (\text{Equation 3})$$

$$d_{e,||} = \frac{n_1^2 n_2 \cos \theta (2n_1^2 \sin^2 \theta - n_2^2) \lambda}{\pi (n_1^2 \sin^2 \theta - n_2^2)^{\frac{1}{2}} (n_1^2 - n_2^2) [(n_1^2 - n_2^2) \sin^2 \theta - n_2^2]} \quad (\text{Equation 4})$$

$$d_{e,\perp} = \frac{\lambda n_1^2 n_2 \cos \theta}{\pi (n_1^2 - n_2^2) (n_1^2 \sin^2 \theta - n_2^2)^{\frac{1}{2}}} \quad (\text{Equation 5})$$

Where λ is the wavelength of the incident beam, θ is the incident angle and n_1 and n_2 are the refractive index of the ATR-FTIR crystal (diamond, $n_1 = 2.419$) (34) and the PL sample (Section 4.1), respectively.

Based on the assumption of one reflection within the crystal, the pathlength is equal to the effective depth, d_e . The molar absorptivity of the ν_3 absorbance band for CO_2 was estimated as $1500\text{ dm}^3\text{mol}^{-1}\text{cm}^{-1}$ at 20°C (35).

3.5 Assumptions and origin of errors

The refractive index of the porous liquid samples was assumed to be the same as the literature values of their corresponding ionic liquids. These assumptions could lead to errors and would require further verification. Additionally, the refractive index was assumed to remain constant even with CO_2 uptake based on literature (36). Baseline correction was also performed, however as the baseline was corrected to Absorbance = 0 (± 0.01), this error is considerably negligible.

3.6 Reproducibility of results

The reliability of data was ensured by repeating each experiment three times. The overall experimental error of all data was within 5% indicating reliable and reproducible results.

4. Results And Discussion

4.1 Angle of Incidence and Refractive Index

Distilled water was selected as the sample used to calculate the angle of incidence for the diamond ATR-crystal. The absorbance of the spectral band (1643 cm^{-1}) in the bending region (ν_2) of water was obtained from the resulting spectrum. The molar absorptivity of the O-H band ($\epsilon_{1643} = 21.8\text{ dm}^3\text{mol}^{-1}\text{cm}^{-1}$) and concentration of water ($c = 55.34\text{ mol dm}^{-3}$) alongside the absorbance of the band was used to calculate the effective depth using equation 2. The effective depth of water was found to be $1.32\text{ }\mu\text{m}$. The refractive index of water

is 1.333 (34). The angle of incidence for the diamond crystal was found to be 46.85° after applying equations 3, 4 & 5. Variations in the refractive index of the PLs as it is subjected to pressures of CO_2 is found to have a negligible impact on the effective depth of the sample (36). As stated, the refractive indices of the PL samples were assumed to be the same as that of the corresponding ILs. The refractive indices of [BPy][NTf₂], [P₆₆₆₁₄][NTf₂] and [BMIM][NTf₂] are 1.444 (37), 1.45164 (38), 1.4282 (39), respectively.

4.2 CO_2 Sorption Capacity

4.2.1. Dependence on Pressure

Absorbance at the band corresponding to the anti-symmetric stretching mode (ν_3) of CO_2 at 2339 cm^{-1} increases with rising pressure (Figure 2) across all samples (Figure 3). These results indicate that CO_2 sorption increases with intensifying CO_2 pressure. The resulting spectra (Figure 2) is expected as the concentration of CO_2 sorbed is directly proportional to the absorbance measured $A = \epsilon cl$ (Equation 2). At increased pressures, there is an increased amount CO_2 molecules the PL is exposed to which in turn increases their likelihood of being sorbed. Furthermore, CO_2 molecules are able to gain entry to more pores as the pressure increases (40) yielding the results in Figure 2.

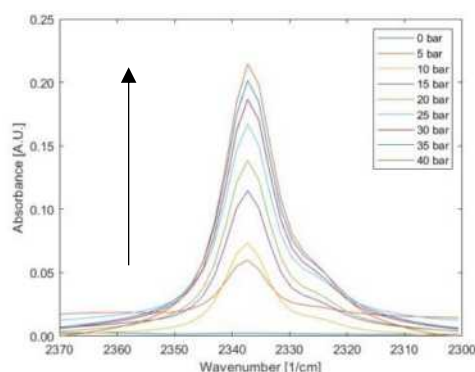


Figure 2: Increasing absorbance seen at the absorbance band 2339 cm^{-1} with increasing pressure observed in the spectra of 12.5% ZIF-8 [P₆₆₆₁₄][NTf₂]

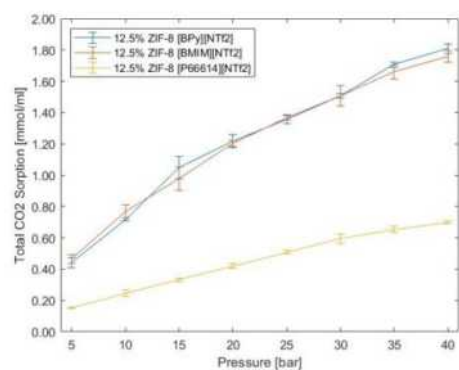


Figure 3: CO_2 sorption capacity [mmol ml^{-1}] for various Type III porous liquids across the pressure range

4.2.2. Dependence on Functional Groups Present in PLs

The PL samples have a common porous framework (ZIF-8) and anion ([NTf₂]). Consequently, the variation in CO_2 sorption capacity across the PL samples (Figure 3) is attributed to the difference in functional groups present in the ionic liquid i.e.e., the cation. In Figure 4 and Figure 5, the PLs with the cations [BMIM] and [BPy] show significantly larger absorbance i.e.e., CO_2 sorption capacity in comparison to the [P₆₆₆₁₄] cation.

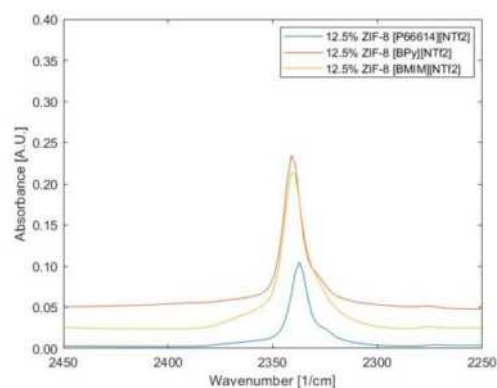


Figure 4: Absorbance band at 2339 cm^{-1} for the Type III porous liquids samples at 15 bar

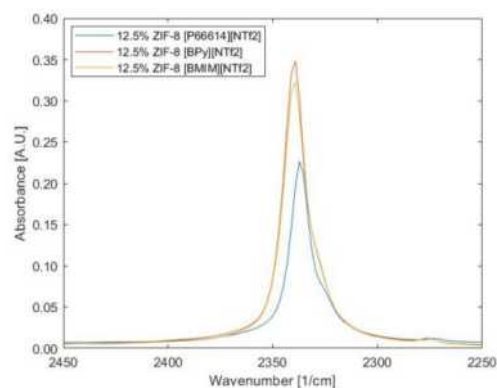


Figure 5: Absorbance band at 2339 cm^{-1} for the Type III porous liquids samples at 40 bar

Both [BPy] and [BMIM] are heterocyclic compounds with Nitrogen atom(s) present in the rings, as shown in Figure 6. These structural similarities of [BPy] and [BMIM] may suggest the similar and greater CO_2 sorption (Figures 3-6) is due to the presence of an interaction that is promoted by the presence of certain functional groups within the cation e.g., the N present in the rings. As ILs are being investigated as a CO_2 absorbent (31, 41), investigations into the absorption of CO_2 in the ILs used in synthesizing the PLs can provide some further insight.

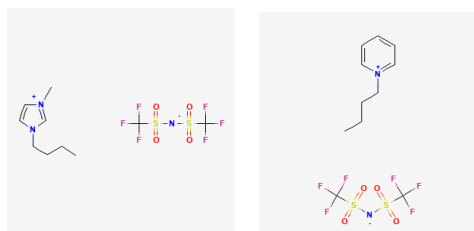


Figure 6: Structure of Neat Solvents [BPy][NTf₂](left) and [BMIM][NTf₂](right) (42, 43)

In ILs, anions were found to have a much greater impact on CO₂ absorption compared to cations (44). Similarly, in alkylimidazolium-based ILs, anions had a greater influence on CO₂ absorption compared to cations (45). Blanchard et al found anions varied CO₂ absorption by up to 25 % whilst the cation had a small effect of only up to 1.5 % at 40 bars and 40 °C (46). However, it is difficult to confirm this is consistent across all ionic liquids. In the pairing of [N-bupy][BF₄] and [C₈mim][BF₄], cations had a much greater contribution (of up to 20 % at 40 bar and 40°C) compared to anions (46). From the results obtained in this study (Figure 3), cations are shown to alter CO₂ sorption by up to 160% (at 40 bar) insinuating its contribution to CO₂ sorption within PLs are far more prominent and cannot be overlooked.

The potential formation of a C2-C bond between the cation and CO₂ (carbamate) was investigated (41). Furthermore, the formation of N-CO₂ complexes was investigated as both [BPy] and [BMIM] contain stacking binding sites for CO₂ with basic and polar N atoms (47-49). The potential of Lewis acid-base interactions (50) occurring between CO₂ and the electron-donating functional groups present in the PL's cation were also investigated. The resulting spectra of the PLs under high-pressure CO₂ were examined for the presence of characteristic spectral bands of carbamates such as 1568 cm⁻¹ (COO —) (51) and N-complexes. No characteristic spectral bands of carbamates and N-complexes are observed. Additionally, the release in degeneracy expected at the ν_2 mode (667 cm⁻¹) characteristic of CO₂ (this has been shown to be evidence of Lewis acid-base interactions (50)) was not as prominent as expected. These findings are reasonable because such interactions might be too weak to be detected using ATR-FTIR spectroscopy. It is therefore difficult to confirm interactions exist between the cation and CO₂. However, the presence of such interactions would imply that absorption via cation interactions contribute significantly to the uptake of CO₂ by T3PLs. As discussed earlier, the anion is the main determining factor for absorption capacity in ILs, so these cation-PL interactions are less prominent in ionic liquids. It is possible that in PLs, these cation-PL interactions are enhanced by the presence of a porous framework (ZIF-8).

Another plausible consideration is that the difference in CO₂ sorption capacity is due to the steric hindrance of the cations. As the [P₆₆₆₁₄] cation contains four lengthy alkyl chains (Figure 7), the steric hindrance of [P₆₆₆₁₄] reduces the interaction between CO₂ and the PL sample, by hindering accessibility to the porous framework (ZIF-8), reducing the sample's uptake of CO₂.

It is likely that the reduced steric hindrance of the [BPy] and [BMIM] cations are responsible for the increased interaction between CO₂ and ZIF-8 within PL sample resulting in an increased uptake of CO₂. This consideration will imply that the primary method of CO₂ uptake in T3PLs is via adsorption.

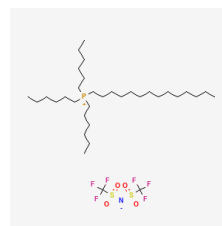


Figure 7: Structure of Neat Solvent [P₆₆₆₁₄][NTf₂] (52)

As PLs show enhanced CO₂ sorption capacity in comparison to their respective ILs, there is confirmation that adsorption plays a key role in CO₂ sorption in T3PLs. However, the question remains if absorption (via interactions with ionic liquid ions) plays an additional role, further enhancing the PL's efficiency as a sorbent.

4.3 Swelling of Porous Liquids

4.3.1. Dependence on Pressure

Swelling in porous materials has been attributed to the penetration of liquid or gas molecules into the pores of the material, resulting in volumetric strain (53). This is a well-known phenomenon in gas sorption. In polymers, swelling was observed using ATR-FTIR spectroscopy and the degree of swelling was found to increase with increasing pressure (54). The increased absorbance observed in the ν_3 mode of CO₂ with elevated pressures occurred simultaneously with a decline in the absorbance of polymer bands. This correlation is due to the reduction in the number of polymeric molecules present per unit volume (55), due to the increase in volume caused by swelling.

To compute swelling in PLs, the sulfonyl (S=O) stretching band at 1346 cm⁻¹ present in [NTf₂] was selected as a reference band. Figure 8 illustrates the decrease in absorbance of the reference band in

12.5 % ZIF-8 [P₆₆₆₁₄][NTf₂] with increasing pressure.

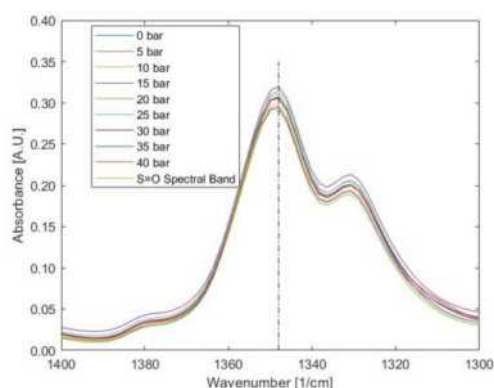


Figure 8: Swelling observed at the S=O Spectral band 1346 cm⁻¹, characteristic of the porous liquid anion [NTf₂] as observed in the spectra of 12.5% ZIF-8 [P₆₆₆₁₄][NTf₂]

This result proves PLs exhibit swelling (56) and pressure has similar impact on the swelling of PLs and polymers. A clear linear trend of swelling increasing with increasing CO₂ pressure is highlighted in Figure 9.

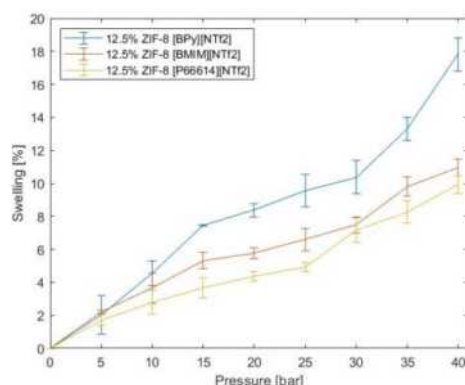


Figure 9: Swelling across pressure range for Type III porous liquids

4.3.2. Dependence on Functional Groups Present in PLs

Similar to the observation made in comparing the CO₂ sorption capacity of the PLs (Section 4.2.2), the variation in swelling across the samples under the same conditions are attributed to the functional groups present in the cation. 12.5 % ZIF-8 [BPy][NTf₂] and 12.5 % ZIF-8 [BMIM][NTf₂] show a higher degree of swelling compared to 12.5 % ZIF-8 [P₆₆₆₁₄][NTf₂] (Figure 9). Table 1 shows how swelling per mole sorbed compares for the PLs. The similar value of approximately between 5.0 – 7.0x suggest that [BMIM] and [BPy] behave rather similarly and exhibit similar swelling potential(%swelling per mol of CO₂ sorbed). Although 12.5 % ZIF-8 [P₆₆₆₁₄][NTf₂] shows lower

swelling for a given pressure, it is seen to exhibit greater swelling potential of about 12.0x.

Table 1: %Swelling per mmol of CO₂ sorbed across pressure range for Type III porous liquids

Pressure / bar	%Swelling per mole adsorbed		
	12.5% ZIF-8 [BMIM][NTf ₂]	12.5% ZIF-8 [BPy][NTf ₂]	12.5% ZIF-8 [P ₆₆₆₁₄][NTf ₂]
5	4.8	4.6	11.2
10	4.8	6.4	11.3
15	5.4	7.1	11.0
20	4.8	6.9	10.5
25	4.8	7.0	9.7
30	5.0	6.9	12.0
35	6.0	7.8	12.7
40	6.2	9.8	14.2
Mean	5.2	7.0	11.6

This is likely due to steric hinderance of the molecules. [P₆₆₆₁₄] is bulky and thus the [P₆₆₆₁₄] based PL is likely to have less efficient packing compared to the PLs containing [BPy] and [BMIM]. The sorption of the same amount of CO₂ will cause the 12.5 % ZIF-8 [P₆₆₆₁₄][NTf₂] to undergo a greater degree of swelling compared to PLs containing [BPy] and [BMIM] that have more efficient packing. The swelling potential (%swelling mole sorbed⁻¹) is a more accurate way to compare swelling across PLs with varying sorption capacities. PLs with bulky cations have less efficient packing and the sorption of CO₂ by the PL will cause it to swell more.

4.4 Nature of interactions between CO₂ and PLs

4.4.1. Dependence on Pressure

The potential of cation-PL interactions occurring between CO₂ and the electron-donating functional groups present in PLs suggest that swelling and sorption processes may not be completely reversible (50, 57). The degree of swelling and CO₂ sorption within the polymer before and after the CO₂ pressure was released were analysed to gain insight into the nature of these CO₂-PL interactions. Upon releasing the pressure within the cell and performing baseline correction on the resulting IR spectra, a band at 2339 cm⁻¹ (ν_3 mode of CO₂) remained (Figure 10). The presence of this band validates the hypothesis that the PLs-CO₂ interactions are not purely physical but rather a mix of reversible and irreversible interactions. The absence of a spectral band in the 2388 cm⁻¹ region in the spectrum demonstrates that gaseous CO₂ has not integrated with the IR radiation and the CO₂ detected by the spectrometer is situated within the PL sample. It can be inferred that the CO₂ remaining in the sample after releasing CO₂ pressure from the cell is a measure of the CO₂ irreversibly sorbed by the PL.

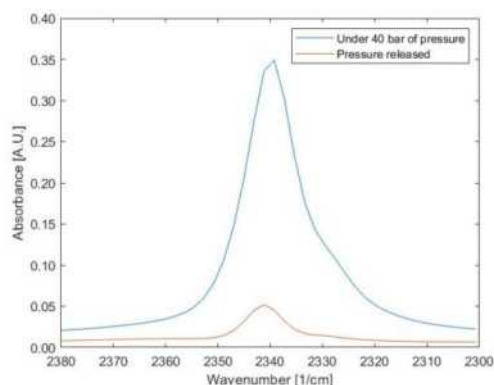


Figure 10: The spectral band of 12.5% ZIF-8 [BPy][NTf₂] at 2339 cm⁻¹ under 40 bar of CO₂ pressure and after pressure was released

Using the Beer-Lambert law, the concentration of CO₂ sorbed before and after releasing the CO₂ pressure was computed. Interestingly, at low pressure of CO₂, these irreversible contributions dominated the sorption of CO₂ (57). As pressure increased, reversible sorption known as physisorption overtook these interactions as the dominant mechanism. We can see that the interactions are clearly a mix of irreversible (chemical) interactions and physisorption whereby the dominant interaction is dependent on the pressure of CO₂. For each PL under probe, the amount of CO₂ irreversibly sorbed by the sample remained constant across the pressure range indicating its pressure independence (Figure 10). This suggests that these irreversible interactions are independent of pressure.

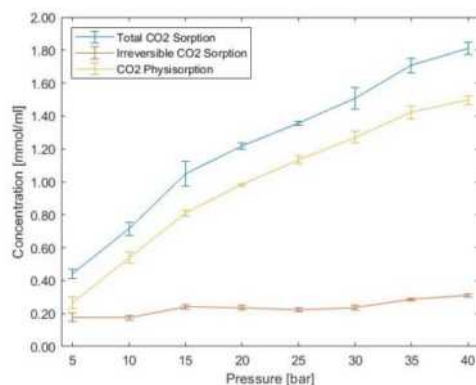


Figure 11: Variation of total CO₂ sorption and Irreversible sorption of 12.5% ZIF-8 [BPy][NTf₂] with pressure

CO₂ can be physically and chemically trapped by adsorbents (57, 58). As alluded to in section 4.4.2, the sorption process of CO₂ in T3PLs is unclear. As these irreversible interactions are independent of pressure (Figure 10), further investigation is required in order to confirm this irreversible sorption is as a result of CO₂ absorption via chemical interactions with IL ions or via chemical and irreversible adsorption (chemisorption).

4.4.2. Dependence on Functional Groups Present in PLs

Investigating the dependence of irreversible sorption on cation functional groups can provide insight on the nature of these interactions i.e., chemisorption or absorption.

The [P₆₆₆₁₄] based cation shows a lower irreversible sorption contribution (Figure 11). As the porous liquids have the same anion of the same concentration, if the irreversible sorption was due to chemical reactions with [NTf₂] we would expect similar levels of irreversible sorption across all PL samples similar to the behaviour of ionic liquids (as the difference in absorption capacity in ILs is primarily dependent on anions) (41, 44, 45). The difference in the amount of CO₂ irreversibly sorbed can be narrowed further to the chemical interactions between CO₂ and the functional group of the cation as discussed in section 4.2.2.

The PL with the [P₆₆₆₁₄] cation exhibits the lowest level of irreversible sorption compared to the other samples (Figure 12). The carbon atom in CO₂ is positively charged and strongly absorbed on basic and polar groups such as the N-atoms present in [BMIM] and [BPy] (47, 48). Additionally, [BPy] and [BMIM] exhibit resonance and are more stable than [P₆₆₆₁₄]. Thus, the interactions between CO₂ and the PLs containing [BMIM] and [BPy] are more stable, and a larger amount of CO₂ has been irreversibly absorbed in the sample after CO₂ pressure is released compared to [P₆₆₆₁₄] which does not have a resonant structure. The amount of CO₂ irreversibly sorbed by [BMIM] and [BPy] are roughly the same. The binding energy between CO₂ and [BMIM] is greater than that of [BPy]. [BPy] has 3 stacking binding sites for CO₂ while [BMIM] has 2 (49). It is unclear if the binding energy between the cation and CO₂ dominates over the number of stacking binding sites. Similar to the results discussed in section 4.2.2, Although the IR spectra of the sample after the release of CO₂ pressure shows no evidence of these interactions, the results shown in Figure 12 is further evidence of the possible presence of these interactions as absorption requires energy to be reverted.

The steric hinderance of the cations could also potentially be the influencing factor for the difference in irreversible CO₂ sorption as irreversible chemical induced adsorption also exists (chemisorption). The steric hinderance of the [P₆₆₆₁₄] cation hinders accessibility of CO₂ molecules to the porous framework, which in turn reduces the interaction between CO₂ and the MOF (ZIF-8) within the PL sample reducing the uptake of CO₂ via chemisorption (irreversible adsorption). Hence, it remains unclear if these irreversible

interactions are as a result of chemisorption or absorption or both.

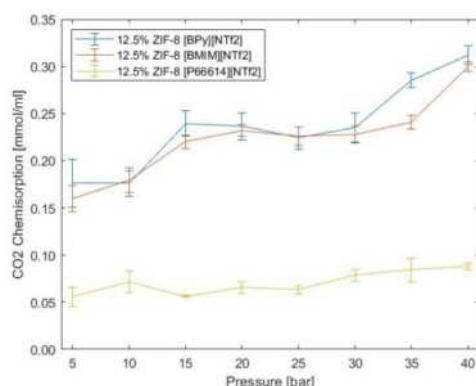


Figure 12: Irreversible sorption across pressure range for Type III porous liquids

5. Conclusion

The influence of pressure and functional groups on the interactions between PLs and CO₂, swelling and CO₂ sorption capacity has been probed using ATR-FTIR spectroscopy. The spectroscopic data obtained from the experiments provided useful insights into the behaviour of this novel technology.

CO₂ sorption by PLs increases as pressure increases. PLs were found to behave differently to their corresponding IL as it is found that cations are consequential to the PL's CO₂ sorption capacity. This is seen as a change in cations altered the CO₂ sorption capacity significantly. 12.5% ZIF-8 [BPy][Ntf₂] is seen to possess the highest CO₂ sorption capacity 12.5%, then ZIF-8 [BMIM][Ntf₂] with 12.5% ZIF-8 [P₆₆₆₁₄][Ntf₂] having the lowest concentration of CO₂ sorbed.

The degree of swelling of PLs was found to increase with increasing pressure of CO₂. PLs with bulky functional groups were found to undergo more swelling compared to PLs with less bulky functional

groups. The T3PLs exhibited swelling in the following order 12.5% ZIF-8 [P₆₆₆₁₄][Ntf₂] > 12.5% ZIF-8 [BPy][Ntf₂] > 12.5% ZIF-8 [BMIM][Ntf₂]. The process of CO₂ sorption was found to not be completely reversible, proving the process of CO₂ sorption is a mix of physisorption and some irreversible sorption. The nature of these irreversible interactions is yet to be confirmed but is likely due to absorption, chemisorption, or both. Further investigation is required using other investigative equipment such as FTIR imaging, Scanning Electron Microscopy (SEM) etc. which can better reveal the location of the irreversibly sorbed CO₂.

The degree of irreversible sorption compared to the degree of reversible physisorption observed in T3PLs at high pressures, is evidence of their applicability to CCS processes. T3PL will not only demand much less energy for regeneration but also the pressure independence of the irreversible sorption experienced suggests that the energy required for regeneration will remain relatively constant regardless of its application.

The findings of this report have laid the groundwork for further investigation in the field of PLs such as research into the amount of heat energy required to fully regenerate the PLs after CO₂ sorption i.e reversing the effects of chemisorption and/or absorption. This will provide insight into the potential costs that may be incurred when deploying T3PLs as a CCS technology in industry. Furthermore, investigations into the swelling extent of T3PLs provides crucial information on the behaviour of the PLs under pressure, allowing for considerations into the design and implementation of PLs into operating CCS processes. Lastly, investigations into the influence of cations i.e functional groups, provides useful insight on the tunability of T3PLs, allowing for bespoke and tailored applications where control of CO₂ sorption capacity and swelling is required.

6. References

1. European Commission. Causes of Climate Change. https://climate.ec.europa.eu/climate-change/causes-climate-change_en. [Accessed November 19 2023].
2. United Nations Climate Change. The Paris Agreement. <https://unfccc.int/process-and-meetings/the-paris-agreement#:~:text=The%20Paris%20Agreement%20is%20a,force%20on%204%20November%202016>. [Accessed November 17 2023].
3. Ritchie H, Rosado P. Energy Mix <https://ourworldindata.org/energy-mix>. [Accessed October 16 2023].
4. Kaziunas A, Schlake R. Why is CO₂ a Green Solvent? Green Chemistry: The Nexus Blog. Weblog. <https://communities.acs.org/t5/GCI-Nexus-Blog/Why-is-CO2-a-Green-Solvent/ba-p/16071>
5. United Nations. Net Zero Coalition. <https://www.un.org/en/climatechange/net-zero-coalition>. [Accessed November 14 2023].
6. Yu C-H, Huang C-H, Tan C-S. A Review of CO₂ Capture by Absorption and Adsorption. Aerosol and Air Quality Research. 2012; 12(5): 745-769.
7. Raganati F, Miccio F, Ammendola P. Adsorption of Carbon Dioxide for Post-Combustion Capture: A Review. Energy & Fuels. 2021; 35(16): 12845-12868.

8. Gao X, Yang S, Hu L, Cai S, Wu L, Kawi S. Carbonaceous Materials as Adsorbents for CO₂ Capture: Synthesis and Modification. *Carbon Capture Science & Technology*. 2022; 3: 100039.
9. Wang D, Ying Y, Xin Y, Li P, Yang Z, Zheng Y. Porous Liquids Open New Horizons: Synthesis, Applications, and Prospects. *Accounts of Materials Research*. 2023; 4(10): 854-866.
10. Egleston BD, Mroz A, Jelfs KE, Greenaway RL. Porous Liquids – The Future is Looking Emptier. *Chemical Science*. 2022; 13(18): 5042-5054.
11. O'Reilly N, Giri N, James SL. Porous Liquids. *Chemistry – A European Journal*. 2007; 13(11): 3020-3025.
12. Bennett TD, Coudert FX, James SL, Cooper AI. The Changing State of Porous Materials. *Nature Materials*. 2021; 20(9): 1179-1187.
13. Cahir J, Tsang MY, Lai B, Hughes D, Alam MA, Jacquemin J, et al. Type 3 Porous Liquids Based on Non-Ionic Liquid Phases – A Broad and Tailorable Platform of Selective, Fluid Gas Sorbents. *Chemical Science*. 2020; 11(8): 2077-2084.
14. Fulvio PF, Dai S. Porous Liquids: The Next Frontier. *Chem*. 2020; 6(12): 3263-3287.
15. Lee Y-R, Jang M-S, Cho H-Y, Kwon H-J, Kim S, Ahn W-S. ZIF-8: A comparison of synthesis methods. *Chemical Engineering Journal*. 2015; 271: 276-280.
16. Sakamaki Y, Tsuji M, Heidrick Z, Watson O, Durchman J, Salmon C, et al. Preparation and Applications of Metal-Organic Frameworks (MOFs): A Laboratory Activity and Demonstration for High School and/or Undergraduate Students. *Journal of Chemical Education*. 2020; 97(4): 1109-1116.
17. Bergaoui M, Khalfaoui M, Awadallah-F A, Al-Muhtaseb S. A review of the features and applications of ZIF-8 and its derivatives for separating CO₂ and isomers of C₃- and C₄- hydrocarbons. *Journal of Natural Gas Science and Engineering*. 2021; 96: 104289.
18. Gong X, Wang Y, Kuang T. ZIF-8-Based Membranes for Carbon Dioxide Capture and Separation. *ACS Sustainable Chemistry & Engineering*. 2017; 5(12): 11204-11214.
19. Li Y. Research Progress of Porous Liquids. *ChemistrySelect*. 2020; 5(43): 13664-13672.
20. Li P, Schott JA, Zhang J, Mahurin SM, Sheng Y, Qiao Z-A, et al. Electrostatic-Assisted Liquefaction of Porous Carbons. *Angewandte Chemie International Edition*. 2017; 56(47): 14958-14962.
21. Liu S, Liu J, Hou X, Xu T, Tong J, Zhang J, et al. Porous Liquid: A Stable ZIF-8 Colloid in Ionic Liquid with Permanent Porosity. *Langmuir*. 2018; 34(12): 3654-3660.
22. Torralba-Calleja E, Skinner J, Gutiérrez-Tauste D. CO₂ Capture in Ionic Liquids: A Review of Solubilities and Experimental Methods. *Journal of Chemistry*. 2013; 2013: 473584.
23. Bavykina A, Cadiau A, Gascon J. Porous Liquids Based on Porous Cages, Metal Organic Frameworks and Metal Organic Polyhedra. *Coordination Chemistry Reviews*. 2019; 386: 85-95.
24. Tiernan H, Byrne B, Kazarian SG. ATR-FTIR Spectroscopy and Spectroscopic Imaging for the Analysis of Biopharmaceuticals. *Spectrochim Acta A Mol Biomol Spectrosc*. 2020; 241: 118636.
25. Shalygin AS, Milovanov ES, Kovalev EP, Yakushkin SS, Kazarian SG, Martynov ON. In Situ FTIR Spectroscopic Imaging of Asphaltene Deposition from Crude Oil under n-Heptane and Acetone Flows. *Petroleum Chemistry*. 2022; 62(9): 1087-1095.
26. Kowalczyk D, Pitucha M. Application of FTIR Method for the Assessment of Immobilization of Active Substances in the Matrix of Biomedical Materials. *Materials (Basel)*. 2019; 12(18).
27. Mistek-Morabito E, Lednev IK. FT-IR Spectroscopy for Identification of Biological Stains for Forensic Purposes. *Spectroscopy (Santa Monica)*. 2018; 33: 8-19.
28. Fadlilmoula A, Pinho D, Carvalho VH, Catarino SO, Minas G. Fourier Transform Infrared (FTIR) Spectroscopy to Analyse Human Blood over the Last 20 Years: A Review towards Lab-on-a-Chip Devices. *Micromachines (Basel)*. 2022; 13(2).
29. Grdadolnik J. ATR-FTIR Spectroscopy: Its Advantages and Limitations. *Acta Chimica Slovenica*. 2002; 49: 631-642.
30. Schott JA, Do-Thanh C-L, Shan W, Puskar NG, Dai S, Mahurin SM. FTIR Investigation of the Interfacial Properties and Mechanisms of CO₂ Sorption in Porous Ionic Liquids. *Green Chemical Engineering*. 2021; 2(4): 392-401.
31. Sakellarios NI, Kazarian SG. In Situ IR Spectroscopic Study of the CO₂-Induced Swelling of Ionic Liquid Media. *American Chemical Society*; 2005. p. 89-101.
32. Swinehart DF. The Beer-Lambert Law. *Journal of Chemical Education*. 1962; 39(7): 333.
33. PIKE Technologies. ATR - Theory and Applications. n.d. <https://mmrc.caltech.edu/FTIR/Literature/ATR/Intro%20to%20ATR.pdf>.
34. Douglas College Physics Department, OpenStax. Douglas College Physics 1207. Pressbooks; n.d. <https://pressbooks.bccampus.ca/introductorygeneralphysics2phys1207/chapter/25-3-the-law-of-refraction/>.

35. Kieke ML, Schoppelrei JW, Brill TB. Spectroscopy of Hydrothermal Reactions. 1. The CO₂–H₂O System and Kinetics of Urea Decomposition in an FTIR Spectroscopy Flow Reactor Cell Operable to 725 K and 335 bar. *The Journal of Physical Chemistry*. 1996; 100(18): 7455-7462.
36. Flichy NMB, Kazarian SG, Lawrence CJ, Briscoe BJ. An ATR–IR Study of Poly (Dimethylsiloxane) under High-Pressure Carbon Dioxide: Simultaneous Measurement of Sorption and Swelling. *The Journal of Physical Chemistry B*. 2002; 106(4): 754-759.
37. 1-butylpyridinium bis(trifluoromethylsulfonyl)imide. n.d.
https://www.chemicalbook.com/ChemicalProductProperty_EN_CB92469772.htm
38. Mendivelso-Pérez DL, Farooq MQ, Santra K, Anderson JL, Petrich JW, Smith EA. Diffusional Dynamics of Tetraalkylphosphonium Ionic Liquid Films Measured by Fluorescence Correlation Spectroscopy. *The Journal of Physical Chemistry B*. 2019; 123(23): 4943-4949.
39. Solvionic. 1-butyl-3-methylimidazolium bis(trifluoromethanesulfonyl)imide. [online] n.d.
https://en.solvionic.com/files/solvionic/fiches/Product_Im0408c_R1.pdf?PHPSESSID=1cb0d5b24fa4b69e29dc89a803b4194a. [Accessed 19 October 2023].
40. Bahadur J, Melnichenko Y, He L, Contescu C, Gallego N, Carmichael J. SANS Investigations of CO₂ Adsorption in Microporous Carbon. *Carbon*. 2015; 95: 535-544.
41. Simon NM, Zanatta M, Neumann J, Girard A-L, Marin G, Stassen H, Dupont J. Cation–Anion–CO₂ Interactions in Imidazolium-Based Ionic Liquid Sorbents. *ChemPhysChem*. 2018; 19(21): 2879-2884.
42. PubChem Compound Summary for CID 12184310, 1-Butylpyridinium Bis(trifluoromethanesulfonyl)imide. 2023. <https://pubchem.ncbi.nlm.nih.gov/compound/12184310>
43. PubChem Substance Record for SID 329763082, 1-Butyl-3-methylimidazolium bis(trifluoromethylsulfonyl)imide, >=98%, Source: Sigma-Aldrich. 2023.
<https://pubchem.ncbi.nlm.nih.gov/substance/329763082>
44. Anthony JL, Anderson JL, Maginn EJ, Brennecke JF. Anion Effects on Gas Solubility in Ionic Liquids. *The Journal of Physical Chemistry B*. 2005; 109(13): 6366-6374.
45. Cadena C, Anthony JL, Shah JK, Morrow TI, Brennecke JF, Maginn EJ. Why Is CO₂ So Soluble in Imidazolium-Based Ionic Liquids? *Journal of the American Chemical Society*. 2004; 126(16): 5300-5308.
46. Blanchard LA, Gu Z, Brennecke JF. High-Pressure Phase Behavior of Ionic Liquid/CO₂ Systems. *The Journal of Physical Chemistry B*. 2001; 105(12): 2437-2444.
47. Sumida K, Rogow DL, Mason JA, McDonald TM, Bloch ED, Herm ZR, et al. Carbon Dioxide Capture in Metal–Organic Frameworks. *Chemical Reviews*. 2012; 112(2): 724-781.
48. Petrovic B, Gorbounov M, Masoudi Soltani S. Impact of Surface Functional Groups and Their Introduction Methods on the Mechanisms of CO₂ Adsorption on Porous Carbonaceous Adsorbents. *Carbon Capture Science & Technology*. 2022; 3: 100045.
49. Lee HM, Youn IS, Saleh M, Lee JW, Kim KS. Interactions of CO₂ with Various Functional Molecules. *Physical Chemistry Chemical Physics*. 2015; 17(16): 10925-10933.
50. Kazarian SG, Vincent MF, Bright FV, Liotta CL, Eckert CA. Specific Intermolecular Interaction of Carbon Dioxide with Polymers. *Journal of the American Chemical Society*. 1996; 118(7): 1729-1736.
51. Sun C, Dutta PK. Infrared Spectroscopic Study of Reaction of Carbon Dioxide with Aqueous Monoethanolamine Solutions. *Industrial & Engineering Chemistry Research*. 2016; 55(22): 6276-6283.
52. Heller R, Zoback M. Adsorption of Methane and Carbon Dioxide on Gas Shale and Pure Mineral Samples. *Journal of Unconventional Oil and Gas Resources*. 2014; 8: 14-24.
53. Ewing AV, Gabrienko AA, Semikolenov SV, Dubkov KA, Kazarian SG. How Do Intermolecular Interactions Affect Swelling of Polyketones with a Differing Number of Carbonyl Groups? An In Situ ATR-FTIR Spectroscopic Study of CO₂ Sorption in Polymers. *The Journal of Physical Chemistry C*. 2015; 119(1): 431-440.
54. Hasell T, Armstrong JA, Jelfs KE, Tay FH, Thomas KM, Kazarian SG, Cooper AI. High-Pressure Carbon Dioxide Uptake for Porous Organic Cages: Comparison of Spectroscopic and Manometric Measurement Techniques. *Chemical Communications*. 2013; 49(82): 9410-9412.
55. Dong K, Zhai Z, Jia B. Swelling Characteristics and Interaction Mechanism of High-Rank Coal during CO₂ Injection: A Molecular Simulation Study. *ACS Omega*. 2022; 7(8): 6911-6923.
56. Kazarian SG. Polymer Processing with Supercritical Fluids. *Polymer Science - Series C*. 2000; 42: 78-101.
57. Gunathilake C, Manchanda AS, Ghimire P, Kruk M, Jaroniec M. Amine-Modified Silica Nanotubes and Nanospheres: Synthesis and CO₂ Sorption Properties. *Environmental Science: Nano*. 2016; 3(4): 806-817.
58. Díaz-Herrera PR, Ramírez-Moreno MJ, Pfeiffer H. The Effects of High-Pressure on the Chemisorption Process of CO₂ on Lithium Oxosilicate (Li₈SiO₆). *Chemical Engineering Journal*. 2015; 264: 10-15.

High-Flux Ethanol-Water Separation via Mildly Reduced Graphene Oxide Membranes

Haochuan Chen

Department of Chemical Engineering, Imperial College London, U.K.

Abstract Ethanol-water mixtures of azeotropic concentrations cannot be separated by conventional distillation. Pervaporation with graphene oxide (GO) membrane can selectively allow water molecules to permeate while rejecting ethanol. Pristine GO does not allow for high selectivity and high permeance simultaneously, consequently structural modifications to it are desired. GO was synthesised with a modified Hummers' method, and mildly reduced graphene oxide (mrGO) was synthesised with an eco-friendly reduction procedure from GO. GO and mrGO were deposited onto alumina hollow fibre (HF) substrate via a pressure assisted filtration method. Membrane thickness is controlled by doubling coating times (10 s, 20 s, 40 s) at a constant 0.0125 mg/mL GO or mrGO concentration. The mildly reduced nature of mrGO in contrast with GO is confirmed via X-ray photoelectron spectroscopy (XPS), Raman spectroscopy, and contact angle analysis. Pervaporation tests were carried out at 75 °C and were fed a feed concentration of 95 wt.% ethanol. The mrGO membranes above the 20-second coating time exhibited slightly reduced flux but disproportionately greater selectivity towards water permeation. mrGO outperforms GO in separation factor 11-fold (133 vs 11.5) in the 20-s samples and 15-fold (250 vs 16.1) in the 40-s samples. Long-term (8 hours) test with mrGO 20 s confirmed that near-absolute ethanol can be produced.

1 Introduction

Ethanol, organically, is made with yeast through fermentation. Ethanol is also manufactured from petrochemicals via the hydration of ethylene. Regardless of process, however, the end products are usually a diluted aqueous solution of ethanol. Purification of ethanol by distillation is the usual next step. However, distillation is limited as ethanol forms an azeotrope with water at 95.6% purity. To be used as a biofuel, a near-anhydrous level of purity may be desirable. Currently, industrial processes involve azeotropic distillation, extractive distillation, and solvent extraction as means to dry ethanol [13]. In the laboratory, drying with calcium oxide or with elemental magnesium are common practices. Water adsorption onto molecular sieves is a practice seen at scales both small in laboratories and large in industry. Membrane based technologies, and most notably pervaporation (PV), are attracting the attention of research. PV achieves separation by selectively allowing vapours of one molecule to permeate through a membrane over another. PV as a mean of separation is of great interest due to its capacity to operate in a continuous fashion and its ease to be scaled. Moreover, a well-tailored membrane excels in selectivity and in energy efficiency. Graphene oxide, among other materials, are one of such membranes capable of this feat.

2 Background

2.1 Advantages of Graphene Oxide

In 2012, Nair et al. found that GO membranes are capable at rejecting the smallest of atoms in the gaseous phase (helium) but is permeable to water

vapour [16]. This property had attracted many researchers to investigate into its probable application in pervaporation as a mean to dehydrate organic solvents [3][4][5][6][8][15][19]. Most attempts at pervaporation with unmodified GO membranes yielded rather low selectivity, or if higher selectivity is desired, the membranes would be too thick to allow for a swift permeation [5]. This is most often due to undesirable packing and surface functional groups [9]. A high-performance membrane for the dehydration of ethanol is expected to grant high permeance and high separation factor. A higher permeance in an industrial setting is a boost to the rate of production, while a high separation factor is desirable as less ethanol seeps into the permeate side of the membrane, requiring re-distillation to re-acquire the dilute ethanol.

2.2 GO Modification Methods

Dan Hua et al. demonstrated the performance of GO frameworks (GOF) [5]. GOF are fabricated by inserting aldehyde functional group appendages in between two layers of GO as shown in Figure 2.2.1:

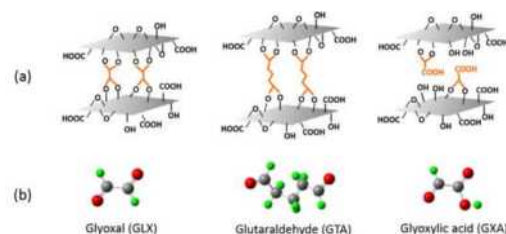


Fig 2.2.1 (a) Graphene oxide framework structures as modified by a specialised aldehyde and (b) GLX, GTA, and GYA aldehydes used in the interlayer standoffs. Green: hydrogen; grey: carbon; red: oxygen. Figure by Dan Hua et al. [5].

The resultant GOF membranes not only shows GO-aldehyde covalent bonds, but it also displays adjustable microstructural properties. GLX, GTA, and GYA-modified GO all exhibited improved separation factor and long-term membrane stability at the cost of some flux. Fresh, unmodified GO membrane can achieve a separation factor of around 28 (as calculated from a feed ethanol wt% of 85% to a permeate concentration of 13% as per Equation 3). The separation factor decreases to 10.5 (35 wt.% ethanol in the permeate) over a 170-h testing period. In comparison, the GO-GTA membranes showed a consistent <10 wt.% of ethanol in the permeate over the same 170-h testing period, indicative of a separation factor of over 50.

Roberto et al. [3] demonstrated the performance of cross-linked poly(vinyl alcohol) and GO (PVA/GO) as a mixed-matrix membrane (MMM), where they showed the inclusion of GO in the matrix increases the water partial permeance (Equation 2 demonstrates the formula for partial permeance) whilst the ethanol partial permeance stays consistent. These membranes, with 2% and 1% GO loading, showed a separation factor of 65.9 and 263 with a permeance of 0.185 and 0.137 kg m⁻² h⁻¹ respectively. The membranes with some of the best separation factor in the MMM category are the cross-linked sodium alginate beta zeolites, demonstrating a separation factor of 1600 [1] and 1334 [2]. The zeolite membranes, however, displayed relatively low permeance at 0.130 and 0.138 kg m⁻² h⁻¹ respectively.

2.3 Investigated Material and Chosen Conditions

Mildly reduced graphene oxide (mrGO) is chosen as the main test subject for this work. mrGO had already seen study in many other water-based applications such as in desalination [21] and in water purification [12]. The synthesis of mrGO from GO requires no organic solvents, no high temperature, and no specialised additives, and thus can be deemed eco-friendly. Environmental friendliness in the synthesis adds on top the already efficient nature of pervaporation, which is what makes mrGO desirable to be researched.

Alumina hollow fibre (HF) is chosen as the substrate for membrane settlement as it boasts a high specific surface area, possesses high stability over time, is heat resistant, and is compatible with many organic compounds. The porous granular geometry of alumina HF allows for an unimpeded permeation of vapours, which is desirable to maximise flux once the membranes are applied.

Most previous experiments focused on feed concentrations of either 85 wt.% or 90 wt.% ethanol. This work, however, used 95% ethanol to best

simulate scenarios of dehydrating ethanol from an azeotropic concentration (95.6%); this also serves to stress-test the system.

This work also focused on fabricating and testing relatively thin membranes (~50 – 100 nm) as an attempt to better manifest the selectivity disparity between GO and mrGO. Thinner membranes should also allow for a higher permeance. In most membranes, a higher permeance is usually linked to a lower selectivity. It would be of the utmost interest in this work to see whether selectivity can be maintained at a respectable level by material selection despite the high permeance.

3 Methods

3.1 Chemicals

Graphite powder (99% carbon basis), potassium permanganate, 1-methyl-2-pyrrolidionone (NMP, 99%) as ceramic suspension carrier, azeotropic sulphuric acid (98%), hydrogen peroxide (30%), ammonia (~30%), hydrochloric acid (~37%), absolute ethanol, alumina powder (99.9% metal basis) obtained from Alpha Aesar, poly(methylmethacrylate) (PMMA) as the ceramic binder, dispersant Arlacel P135 supplied by Croda, dichloromethane (DCM, stabilised with 0.2% ethanol), Araldite Rapid epoxy, Araldite 2014-2 epoxy. Chemicals were not further purified before use; chemicals may need to be diluted with deionised water before use.

3.2 GO Synthesis and Purification

GO was synthesised per a modified Hummers method [7]. Chiefly, sulphuric acid (450 mL) was added to graphite powder (10g) at a solution temperature below 10 °C and was stirred for 90 min. 1.5 g of potassium permanganate (KMnO₄) was then added to the solution and was stirred for a further 90 min. Then, a larger amount (30g) of KMnO₄ was added and was stirred again for 1 hour below 10 °C. The expected colour change would be from a purplish black into a dark shade of green, indicating the reduction of manganese from a +7 oxidation state to +6, and thus a correspondent partial oxidation of carbon in the form of appended functional groups. The solution was then heated to 40 °C to expedite the redox reaction and then was stirred for 1h to allow the reaction to finish. Deionised water (450 mL) was added to the solution dropwise at a temperature below 50 °C. It is important to note that heat control is crucial for this step as to not allow for an explosive thermal runaway. The solution should then turn brown from particulates of manganese dioxide, which was indicative of a further reduction of manganese to a +4 oxidation state. The solution was heated to 95°C

for 30 min, following with the slow addition of a hydrogen peroxide solution at 10 wt.% and volumed at 300 mL. The solution was then stirred for a further 30 min. The expected colour change was to a light yellow.

Following the synthesis of GO comes purification. The synthesised GO was filtered and washed with hydrochloric acid (10 wt.%; 5,000 mL) five times. Metal impurities were checked with inductive coupled plasma optical emission spectrometry; wash repeats should terminate only when GO samples show no metal impurities present. GO filtrate cake was dried over phosphorous pentoxide as a desiccant at 40 °C for 24 h under vacuum. The GO powder was resuspended in acetone (5,000 mL) to be filtered and washed for a further five times and was desiccated again at 40 °C for 24 h under vacuum. A standard solution of 0.0125 mg/mL of GO in ultrapure water is prepared for experimentation.

3.3 Synthesis of mrGO [11]

The synthesis of mildly reduced graphene oxide (mrGO) follows the synthesis of GO. The synthesised pure GO is suspended homogeneously in water by sonication. The dispersed GO solution (0.1 wt.%) was loaded into a round-bottom (500 mL) flask, maintained at 100 °C, and stirred overnight under a flowing nitrogen atmosphere. This is the stage where the GO is converted into mrGO, which is then rinsed on filter several times (exact number not mentioned in source, 5 if in doubt) with ultrapure water. A standard solution of 0.0125 mg/mL of mrGO in ultrapure water is prepared for experimentation.

3.4 Preparation of Alumina Hollow Fibre (HF)

150 g of alumina powder is suspended in 180g of NMP in a ceramic jar via 3 g of dispersant Arlacel P135. The ceramic jar is then loaded onto a planetary ball miller at 283 rpm for 48 hours. PMMA is then added to the mixture, which is then mixed again for 48 hours. For the bubbles to be evacuated before spinning, the suspension was degassed under vacuum for a minimum of 4 h. Degassing shouldn't terminate until bubbling comes to a near full stop, which can take as long as 8 h depending on environmental conditions and natural variations between samples. The degassed suspension was then transferred to a stainless-steel syringe. The Alumina HFs were prepared with a combined phase-inversion/sintering process, where sintering was conducted at 1450 °C as to improve mechanical performance.

3.5 Forming Membranes on HF substrate

HF strands of 4 cm length were fastened onto a vacuum adapter with Araldite Rapid epoxy for fixing and PTFE tape for sealing. Araldite Rapid was advertised to cure in 5 min on packaging. It is found imperially that it would be best to allow around 30 min for curing for the best results. Curing can be expedited to 5-10 min in an oven. Araldite Rapid was also found to have excellent compatibility with GO and mrGO suspensions. For a good seal, it was ensured that all PTFE sealings were fully covered with epoxy resin. The other opening end of the HF strand not connected in the direction of the vacuum adapter was blocked off with a droplet of epoxy. It would be advisable to wait around 2-3 minutes for the epoxy to partially cure and gain some viscosity before applying the droplet, as uncured epoxy may seep inside the fibre, thus losing testable length. Practically, this meant sealing the vacuum adapter end first before sealing the exposed end. Membranes were loaded onto the substrate through a process akin to vacuum filtration. The thicknesses of membranes were controlled by filtration times and suspension concentration. GO and mrGO suspensions were prepared at a 0.0125 mg/mL concentration, and loading was done over 10 s, 20 s, and 40 s of vacuum filtration. After loading, the HFs were pulled out of their suspensions and were left to dry in air for a further 60 s with the vacuum pump running. HF fibres should now be covered in a brown-grey sediment of various shades, with the longer deposition exhibiting a darker shade. It is of paramount importance to not tamper with the membranes or to subject them to dusty air before curing them in a vacuum oven at 40 °C for at least 3 hours. Uncured GO and mrGO are prone to dislodging from the substrate with the smallest of perturbations on the surface.

3.6 Ethanol-Water PV Testing

Araldite Rapid is incompatible with the ethanol-rich test environment, thus the epoxy needed swapping. This was done by first plucking off the HF from the original vacuum connector to remove epoxy from both ends. New epoxy Araldite 2014-2 hardener was applied in a similar fashion, though it would take around 12-18 hours for Araldite 2014-2 to cure. It would be useful to use keep an epoxy sample on the side as to ease checking the state of curing. It needs to be noted that Araldite 2014-2 cannot be used directly during the step of forming membranes in section 3.5, where Araldite 2014-2 would cause the fine GO and mrGO particulates to clump and settle out of suspension. Once the epoxy cures, the vacuum connector was connected to a liquid nitrogen cold trap and then to a vacuum pump. The outside of the membrane was submerged in a warm (75 °C) bath of ethanol feed (250 mL, 95 wt.%). Once the vacuuming commences, ice (of water or ethanol) crystals were expected to deposit

in the cold trap and were to be collected. Experiments can be terminated as soon as at least 0.5 g of deposit is collected.

Tests usually go on for 30 min to 2 h, depending on the visible amount of permeate in the cold trap, which was monitored continuously for the first 10 minutes and checked regularly every 30 minutes for any anomalies. Collected permeate samples are weighed and membrane rods have their length and thickness measured with a digital calliper.

The total permeate flux (J) is calculated as follows:

$$J = \frac{Q}{At} \quad (1)$$

Where Q is the mass of the permeate in kg, A is the membrane area in m^2 , and t is the operating time in hours. Partial flux J_i was defined as the flux of component i . It was calculated by multiplying the weight fraction y_i in the permeate by the total permeate flux J .

$$J_i = y_i J \quad (2)$$

Ethanol concentration was determined by two means: gas chromatography (GC, Shimadzu SH-Rxi-5ms) and density tests.

In gas chromatography, 0.300 mL of permeate was added with 0.500 mL of DCM in a 1.5 mL centrifuge tube. The tubes were then shaken to expedite extraction. For phases to separate, the centrifuge tubes were spun in a centrifuge at 8.5k rpm for 5 min, and the denser DCM phase was collected for GC analysis against a calibration line of 2%, 5%, 10%, 15%, 20%, 30%, and 40% ethanol in water (wt.%).

For concentrations high enough that doesn't allow for phase separation when attempting extraction with DCM, density tests would be opted instead. The density values are converted to ethanol concentration with a reference table [20]. A generous error of ± 5 wt.% was given. While density testing is less precise, blindly using GC on high-ethanol-concentration samples carried the risk of excessive water content, which may damage the Shimadzu SH-Rxi-5ms column used in the GC instrument. It should also be noted that for those concentrations that uses the density method would usually imply a low-selectivity (and often leaking) membrane, thus having the exact measurement wouldn't be significant in the first place. In practice, density testing mainly applies to the thin GO 10 s samples and mrGO 10 s samples.

With the permeate ethanol concentration, the separation factor α is calculated according to (3):

$$\alpha = \frac{y_{water}/y_{ethanol}}{x_{water}/x_{ethanol}} \quad (3)$$

Where y and x are the mass fractions of the components in the permeate and feed, respectively.

All conditions for each membrane thickness were tested at least twice with their results averaged as evidence of repeatability and for accuracy. The mrGO membrane with 20-s coating time was further tested for 8 hours to ensure stability and to examine separation performance as the ethanol in the feed bottle approaches complete dryness. While longer stability tests could be desirable, the materials used in the current method seems limited. Notably, the Araldite 2014-2 epoxy, when subject to a warm ethanol environment for prolonged periods of time (>12 h), can soften, which may compromise seals.

3.7 Membrane Material Characterisation

The settlement geometry of GO and mrGO particles were determined with a high-resolution field emission gun scanning electron microscope (FEG-SEM, LEO Gemini 1525). Membrane thickness can be directly measured on a close-up zoom setting.

X-ray photoelectron spectroscopy (XPS) was used to determine the elemental proportion of carbon and oxygen in GO and mrGO membranes, with mrGO a larger peak representing carbon and a smaller peak representing oxygen can be expected.

Contact angles (CA) of ultrapure water on GO and mrGO membrane samples were measured with three repeats on the Ramé-hart Model 590 Advanced Automated Goniometer. The average and standard deviation of the values were determined. CA aim at characterising membrane surface hydrophilicity. A more obtuse (higher) contact angle would indicate a more hydrophobic surface, and vice versa.

Raman spectroscopy was used to detect membrane defects and bonding nature in GO and mrGO samples. For this test, instrument SENTERRA II is operated with 2.5 mW and 532 nm laser and 10 s integration time. It also provided further observations on the degree-of-oxidation differences between GO and mrGO (mildly reduced GO). This was done by measuring the intensity ratio of the D peak to the G peak (ID/IG). The G peak represents a sp^2 hybridised carbon atom in a single sheet of graphene (2 dimensional). Meanwhile, the D peak indicates chemical functionalisation (oxidation in GO) of the carbon atoms.

4 Results and Discussion

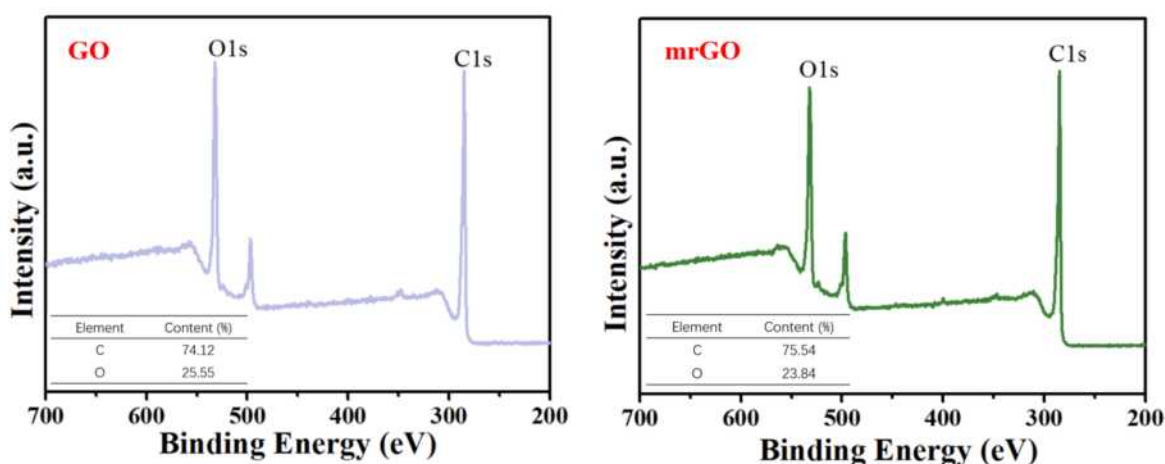


Fig 4.1.1 XPS analysis of GO (left) and mrGO (right).

4.1 XPS Analysis on Elemental Composition

The XPS is a suitable analysis tool in that it is the most useful for analysing the superficial surfaces. By the photo electric effect, it identifies the elemental makeup on the membrane surface with little interference from the aluminium in the substrate that the membrane settles on

Figure 4.1.1 shows oxygen peaks at ~540 eV and carbon peaks at ~280 eV. The area under each peak post integration shows that GO composes of 74.12% carbon and 25.55% oxygen. Meanwhile, mrGO composes of 75.54% oxygen and 23.84% carbon. The slightly higher carbon content and the slightly lower oxygen content in mrGO (mildly reduced GO) than GO is indicative of a mild reduction.

4.2 Raman Spectroscopy

Raman spectroscopy is a fitting analysis tool to monitor oxidation levels in the mrGO in contrast with unmodified GO as the spectroscopy is sensitive to discern geometric structure and bonding within the 2D lattice.

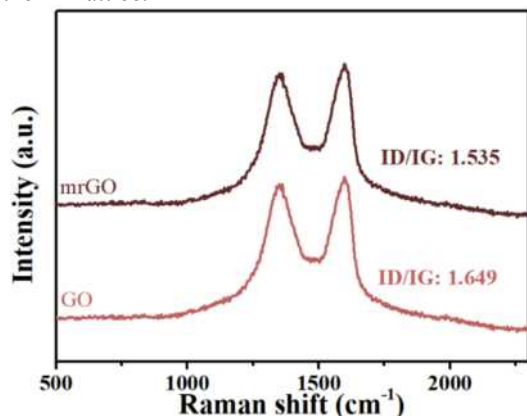


Fig 4.2.1 Raman spectroscopy of GO (bottom) and mrGO (top).

Raman spectra in Figure 4.2.1 shows that both the G peak and D peak are higher in mrGO than GO. The higher G peak indicates a heightened graphitic (sp^2 – hybridised) domain. Meanwhile, the higher D peak can possibly be attributed to defects in the form of misplaced carbon atoms. It is also worth noting that the ID/IG ratio shows values of 1.649 and 1.535 in GO and mrGO respectively. The lower ID/IG value of mrGO can be the evidence for net chemical reduction in mrGO coming from GO.

4.3 SEM Imagery and Membrane Morphologies

Figure 4.3.1 demonstrates the surface and cross-sectional profiles of GO and mrGO coated HF membranes. In the cross-sectional images, the granular portions below represent the fine deposits of alumina that constitute the fibre substrate, while the tightly packed layers on top would be the formed membranes. The thickness of 10-s, 20-s, and 40-s coatings times of GO membranes are 45.27 nm, 67.39 nm, and 68.02 nm respectively. The thickness of 10-s, 20-s, and 40-s coating times of mrGO membranes are 57.63 nm, 75.63 nm, and 102.0 nm respectively.

It is noteworthy that the membranes on the 10-s GO and mrGO samples developed wrinkles that chiefly follows the granular substrate structures underneath. On the SEM images, the wrinkles are seen as light-coloured streaks. These wrinkles get milder as membrane thickness increases. Wrinkles, especially in the 10-s samples, can be the main contributing factor to their low selectivity. The wrinkles, when under vacuum pressure, can become pinch points against the granular alumina substrate and tear voids in the membrane, thus allowing much of the permeates through without it going through any membrane material at all. This effect is the most evident in the pervaporation results in section 4.5.

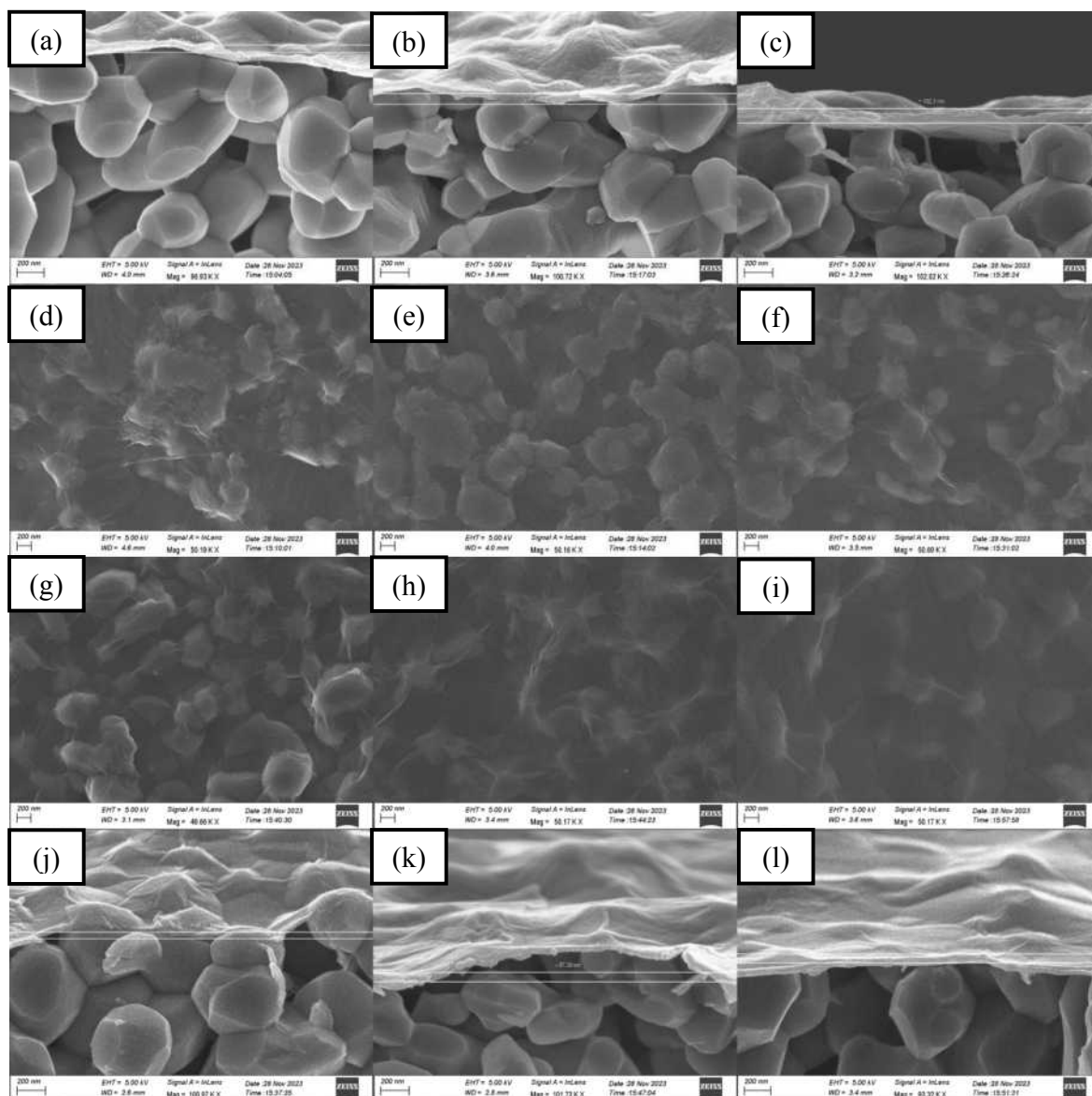


Fig 4.3.1 SEM images of GO and mrGO. Cross-sectional views: (a) GO 10 s; (b) GO 20 s; (c) GO 40 s; (j) mrGO 10 s; (k) mrGO 20 s; (l) mrGO 40s. Surface views: (d) GO 10 s; (e) GO 20 s; (f) GO 40 s; (g) mrGO 10 s; (h) mrGO 20 s; (i) mrGO 30 s.

4.4 Contact Angle

Contact angle examines the hydrophobicity of a material's surface. Functionalisation is the factor that influences the hydrophobicity on modified GO membrane surfaces to the greatest extent. Some functional groups in GO, notably hydroxyl, carboxyl, carboxylic acid, and epoxide rings can form hydrogen bonds with water molecules. Hydrogen bonds attaches the water droplet strongly to a membrane surface, and such downward pulling force decreases the contact angle. The contact angle measurements of GO surface with ultrapure water were measured to be 67.91° , 64.8° , and 64.89° , with an average angle of 65.87° and a standard deviation $\sigma = 1.45^\circ$. On the surface of mrGO, the measurements are 72.66° , 73.93° , and 73.93° , with

an average of 73.25° and a standard deviation $\sigma = 1.77^\circ$. The larger contact angles in mrGO indicates an enhanced hydrophobicity, which signifies the removal of some hydrogen-bonding functional groups in the mild-reduction process.

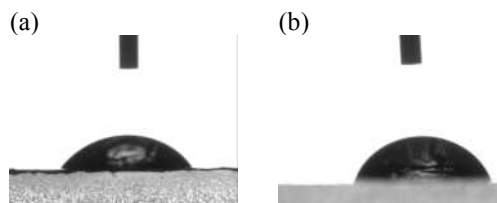


Fig. 4.4.1 Contact angle of (a) GO and (b) mrGO.

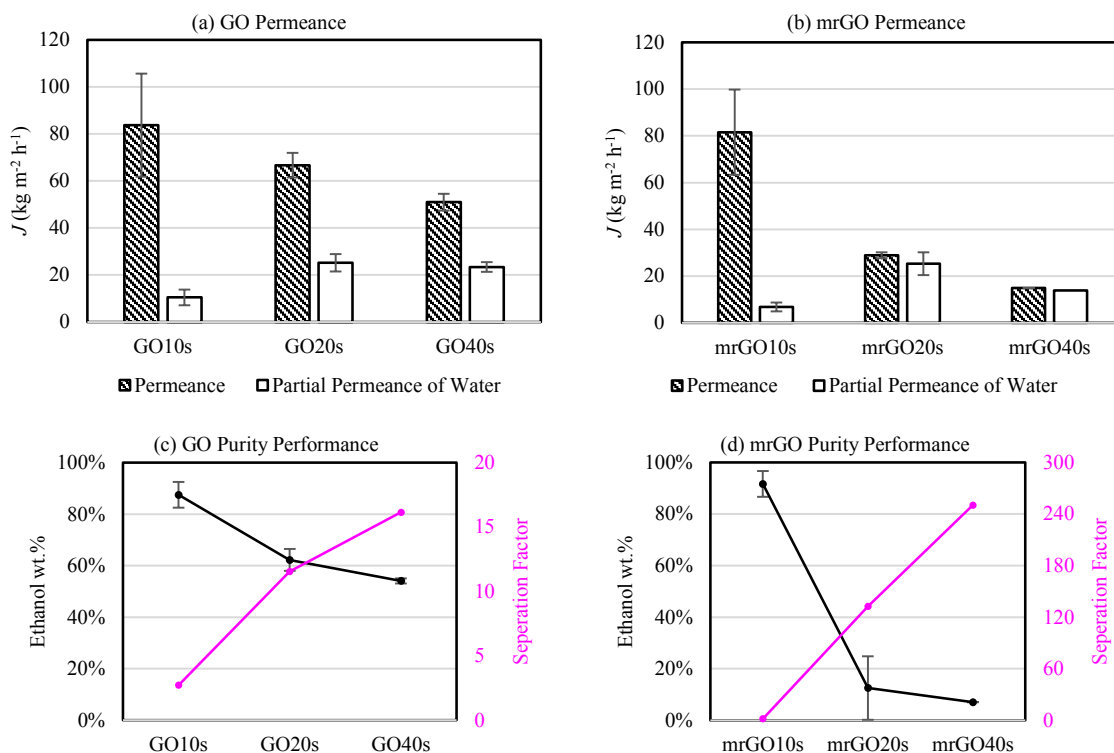


Fig. 4.5.1 (a), (b): GO & mrGO permeance; (c), (d): GO and mrGO purity. Error for (b), (d) mrGO40s are too small to be shown properly.

4.5 PV testing

Generally, thicker membranes exhibited lower flux as seen in Figure 4.5.1 (a) and (b), but higher selectivity towards water in (c) and (d). As the membrane gets thicker, there was more distance any permeate would need to push through, hence the lower flux. Thicker membranes, on the other hand, allowed more 2-D layers for the selective nature of the membrane to take effect, which leads to the higher selectivity.

Membranes with 10-s coating times all showed next to negligible separation (see section 4.3 and Figure 4.3.1). Between the 20-s samples, the mrGO sample showed a lower overall permeance of $28.9 \text{ kg m}^{-2} \text{h}^{-1}$ compared to GO's $66.7 \text{ kg m}^{-2} \text{h}^{-1}$. The partial permeance of water, as calculated in Equation 2, are similar at $25.3 \text{ kg m}^{-2} \text{h}^{-1}$ for mrGO and $25.2 \text{ kg m}^{-2} \text{h}^{-1}$ for GO. This shows that mrGO has greater rejection towards ethanol, where it can be theorised that the removal of some functional groups in mrGO reduced the size of defects, making it harder for the larger ethanol molecules to fit through. The permeate ethanol percentage in the 20-s samples read 12.5% for mrGO against 62.2% for GO, a five-fold reduction.

Among the 40-s samples, the mrGO membrane showed both a lower total permeance (14.9 vs $51.0 \text{ kg m}^{-2} \text{h}^{-1}$ in GO) and a lower partial water permeance (13.9 vs $23.4 \text{ kg m}^{-2} \text{h}^{-1}$ in GO). It can be

theorised that the intricacies in the mrGO membrane that arises from the functional group removal is now affecting the water molecule to some extent. While it should still be of a similar difficulty for the water molecules to pass through any single void section in between layers, there was more entwinement as produced by this thicker membrane, which leads to the lower partial water permeance. On the flip side, the 40-s mrGO sample demonstrates much improved selectivity towards water (7.06% vs 54.1% in GO) attributable to the same entwinements that makes the larger ethanol molecules disproportionately harder to navigate through.

Overall, the mrGO 20 s and 40 s membranes both demonstrated fluxes that can be compared favourably to some other membranes as seen in Figure 4.5.2. In the meantime, a respectable selectivity is maintained.

Another notable fact about the purity measurements of the 20-s samples is that their measurements have a wide error range (standard deviation). This means that in experiments their results are quite varied. It can be hypothesised that there might be thickness variations on the membrane. It is improbable that the HF substrate is perfectly homogeneous such that an equal vacuum is applied to the entire substrate surface in the coating step. Variations at the scale of the 20-s membrane can happen from any natural, random, and uncontrollable perturbations in the system. These

effects are the most pronounced when the membrane is thin and when the coating time is short, as much of the purity inconsistencies disappear in the thicker 40-s membranes. In the 40-s membranes, more time allows the thickness of the membrane to equilibrate: the areas that are initially more thickly coated attracts material with a weaker vacuum, during which the thinner parts can catch up in thickness.

The 8-h long term stability test on mrGO 20 s returned a permeance of $25.8 \text{ kg m}^{-2} \text{ h}^{-1}$ and an

ethanol concentration of 58.0% in the permeate. Additionally, the ending ethanol density was measured to be 779 kg m^{-3} (nearly 100%) in comparison with 797 kg m^{-3} at the start of the experiment (95% ethanol). This provides evidence that the mrGO membrane is operable at a near-anhydrous concentration, i.e. dehydrating ethanol approaching an anhydrous state. However, the high ethanol concentration in the filtrate can serve as preliminary evidence to a weak longevity in the mrGO membranes.

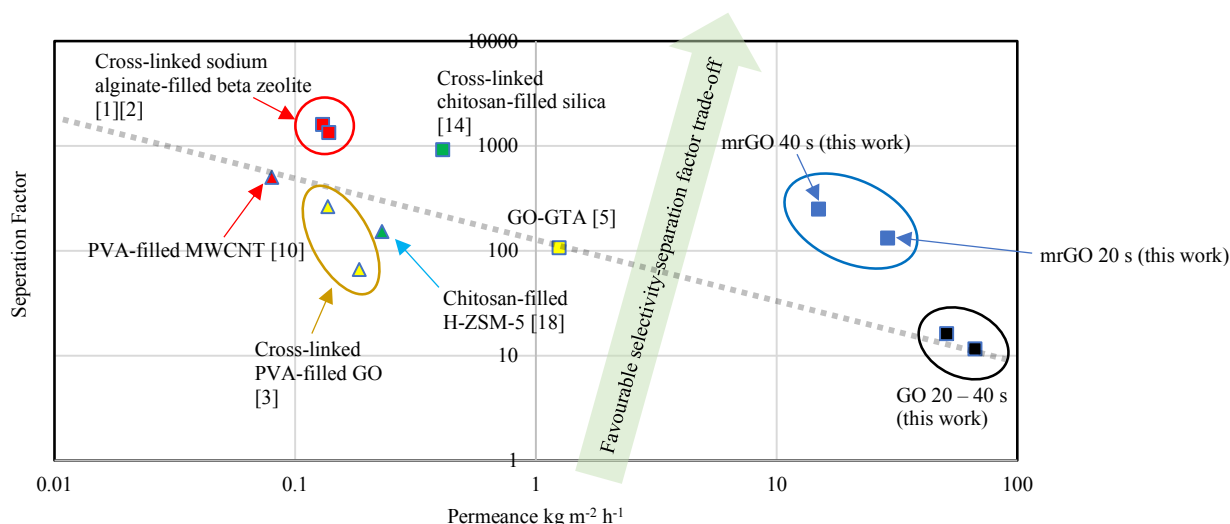


Fig. 4.5.2 Permeance and separation factor scatter graph of various membranes.

5 Conclusions

In conclusion, it is confirmed that the mrGO membranes are successfully fabricated. The reduced nature of mrGO in contrast with unmodified GO was confirmed with XPS, where the results showed an increase in carbon ratio from 74.12% to 75.54%. Raman spectroscopy showed a decrease in ID/IG from 1.649 to 1.535, which is also symbolic of a successful reduction procedure. The contact angle test, where the mrGO showed an increasing contact angle from 65.87° to 73.25° , which indicates an enhanced hydrophobicity leading from the removal of oxygen-related functional groups in the reduction procedure. In pervaporation tests, the mrGO 20 s and mrGO 40 s had both outperformed their GO counterparts in a heightened separation factor. Additionally, the mrGO 40 s membrane exhibited permeance that is two orders of magnitude higher than conventional GO membranes whilst still displaying a respectable separation factor of 250, this can be an improvement from current membranes as most high-selectivity membranes suffer from poor permeance and vice versa. The long-term test confirms that dehydration of ethanol until absolute (100%) is feasible.

Moving forward, it would be interesting to see how if the mrGO membranes can sustain the high selectivity and permeance over the very long term ($>100 \text{ h}$). For example, a feed of a much larger volume could be used as to intentionally not allow the water content to dry out. The Araldite 2014-2 isn't suitable for this experiment, however. Another more ethanol-resistant epoxy resin or means of membrane installation to the apparatus would need to be found.

6 Acknowledgements

Haochuan Chen would like to express his utmost gratitude to Mengjiao Zhai for her professional guidance and support throughout this work.

7 References

- [1] Adoor, S.G., Manjeshwar, L.S., Bhat, S.D. and Aminabhavi, T.M. (2008). Aluminum-rich zeolite beta incorporated sodium alginate mixed matrix membranes for pervaporation dehydration and esterification of ethanol and acetic acid. *Journal of Membrane Science*, 318(1-2), pp.233–246. doi:https://doi.org/10.1016/j.memsci.2008.02.043.

- [2] Bhat, S.D. and Aminabhavi, T.M. (2009). Pervaporation-aided dehydration and esterification of acetic acid with ethanol using 4A zeolite-filled cross-linked sodium alginate-mixed matrix membranes. *Journal of Applied Polymer Science*, 113(1), pp.157–168. doi:<https://doi.org/10.1002/app.29545>.
- [3] Castro-Muñoz, R., Buera-González, J., Iglesia, Ó. de la, Galiano, F., Fila, V., Malankowska, M., Rubio, C., Figoli, A., Téllez, C. and Coronas, J. (2019). Towards the dehydration of ethanol using pervaporation cross-linked poly(vinyl alcohol)/graphene oxide membranes. *Journal of Membrane Science*, [online] 582, pp.423–434. doi:<https://doi.org/10.1016/j.memsci.2019.03.076>.
- [4] Chen, X., Liu, G., Zhang, H. and Fan, Y. (2015). Fabrication of graphene oxide composite membranes and their application for pervaporation dehydration of butanol. *Chinese Journal of Chemical Engineering*, 23(7), pp.1102–1109. doi:<https://doi.org/10.1016/j.cjche.2015.04.018>.
- [5] Hua, D., Rai, R.K., Zhang, Y. and Chung, T.-S. (2017). Aldehyde functionalized graphene oxide frameworks as robust membrane materials for pervaporative alcohol dehydration. *Chemical Engineering Science*, 161, pp.341–349. doi:<https://doi.org/10.1016/j.ces.2016.12.061>.
- [6] Huang, H., Song, Z., Wei, N., Shi, L., Mao, Y., Ying, Y., Sun, L., Xu, Z. and Peng, X. (2013). Ultrafast viscous water flow through nanostrand-channelled graphene oxide membranes. *Nature Communications*, 4(1). doi:<https://doi.org/10.1038/ncomms3979>.
- [7] Hummers, W.S. and Offeman, R.E. (1958). Preparation of Graphitic Oxide. *Journal of the American Chemical Society*, 80(6), pp.1339–1339. doi:<https://doi.org/10.1021/ja01539a017>.
- [8] Hung, W.-S., An, Q.-F., De Guzman, M., Lin, H.-Y., Huang, S.-H., Liu, W.-R., Hu, C.-C., Lee, K.-R. and Lai, J.-Y. (2014a). Pressure-assisted self-assembly technique for fabricating composite membranes consisting of highly ordered selective laminate layers of amphiphilic graphene oxide. *Carbon*, 68, pp.670–677. doi:<https://doi.org/10.1016/j.carbon.2013.11.048>.
- [9] Hung, W.-S., Tsou, C.-H., De Guzman, M., An, Q.-F., Liu, Y.-L., Zhang, Y.-M., Hu, C.-C., Lee, K.-R. and Lai, J.-Y. (2014b). Cross-Linking with Diamine Monomers To Prepare Composite Graphene Oxide-Framework Membranes with Varying d-Spacing. *Chemistry of Materials*, 26(9), pp.2983–2990. doi:<https://doi.org/10.1021/cm5007873>.
- [10] Jae Hyun Choi, Jonggeon Jegal, Woo Nyon Kim and Han Suk Choi (2008). Incorporation of multiwalled carbon nanotubes into poly(vinyl alcohol) membranes for use in the pervaporation of water/ethanol mixtures. *Journal of Applied Polymer Science*, 111(5), pp.2186–2193. doi:<https://doi.org/10.1002/app.29222>.
- [11] Kim, H.W., Ross, M.B., Kornienko, N., Zhang, L., Guo, J., Yang, P. and McCloskey, B.D. (2018). Efficient hydrogen peroxide generation using reduced graphene oxide-based oxygen reduction electrocatalysts. *Nature Catalysis*, 1(4), pp.282–290. doi:<https://doi.org/10.1038/s41929-018-0044-2>.
- [12] Kumar, S., Garg, A. and Chowdhuri, A. (2022). Mildly Reduced Graphene Oxide Membranes for Water Purification Applications. *Nano Express*. doi:<https://doi.org/10.1088/2632-959x/aca7d6>.
- [13] Lee, Y.H., Chen, C.H., Umi Fazara, M.A. and Irfan Hatim, M.D.M. (2021). Production of fuel grade anhydrous ethanol: a review. *IOP Conference Series: Earth and Environmental Science*, 765(1), p.012016. doi:<https://doi.org/10.1088/1755-1315/765/1/012016>.
- [14] Liu, Y.-L., Hsu, C.-Y., Su, Y.-H. and Lai, J.-Y. (2004). Chitosan–Silica Complex Membranes from Sulfonic Acid Functionalized Silica Nanoparticles for Pervaporation Dehydration of Ethanol–Water Solutions. *Biomacromolecules*, 6(1), pp.368–373. doi:<https://doi.org/10.1021/bm049531w>.
- [15] Lou, Y., Liu, G., Liu, S., Shen, J. and Jin, W. (2014). A facile way to prepare ceramic-supported graphene oxide composite membrane via silane-graft modification. *Applied Surface Science*, [online] 307, pp.631–637. doi:<https://doi.org/10.1016/j.apsusc.2014.04.088>.
- [16] Nair, R.R., Wu, H.A., Jayaram, P.N., Grigorieva, I.V. and Geim, A.K. (2012). Unimpeded Permeation of Water Through Helium-Leak-Tight Graphene-Based Membranes. *Science*, [online] 335(6067), pp.442–444. doi:<https://doi.org/10.1126/science.1211694>.
- [17] Princeton.edu. (2022). nglos324 - carbon. [online] Available at: <https://www.princeton.edu/~maelabs/mae324/glos324/carbon.htm#:~:text=Carbon%20is%20in%20gro up%20IV>.
- [18] Sun, H., Lu, L., Chen, X. and Jiang, Z. (2008). Pervaporation dehydration of aqueous ethanol solution using H-ZSM-5 filled chitosan

membranes. *Separation and Purification Technology*, 58(3), pp.429–436.
doi:<https://doi.org/10.1016/j.seppur.2007.09.012>.

[19] Tang, Y., Widjojo, N., Gui Min Shi, Chung, T.-S., Weber, M. and Maletzko, C. (2012). Development of flat-sheet membranes for C1–C4 alcohols dehydration via pervaporation from sulfonated polyphenylsulfone (sPPSU). 415-416, pp.686–695.
doi:<https://doi.org/10.1016/j.memsci.2012.05.056>.

[20] www.engineeringtoolbox.com. (n.d.). Density of Ethanol Water Mixtures. [online] Available at: https://www.engineeringtoolbox.com/ethanol-water-mixture-density-d_2162.html [Accessed 14 Dec. 2023].

[21] Yuan, S., Li, Y., Qiu, R., Xia, Y., Selomulya, C. and Zhang, X. (2022). Minimising non-selective defects in ultrathin reduced graphene oxide membranes with graphene quantum dots for enhanced water and NaCl separation. *Chinese Journal of Chemical Engineering*, [online] 41, pp.278–285.
doi:<https://doi.org/10.1016/j.cjche.2021.12.009>.

Assessing Pyrolytic Carbon Derived from Methane Pyrolysis as an Anode Material

Alexander Tandon and Alexander Vaughan
Department of Chemical Engineering, Imperial College London, U.K.

Received December 13, 2023

Abstract

As the world becomes increasingly electrified, methane pyrolysis emerges as a promising avenue for the production of turquoise hydrogen, offering a viable resource to power the transition toward a net-zero future. This study investigates the suitability of the pyrolytic carbon (PyC) coproduct of this process as a lithium-ion battery anode. Three samples - low, medium and high PyC – each corresponding to different durations of the methane pyrolysis reaction – were used. Several performance analysis parameters were deployed to assess the anode performance of each PyC sample: namely specific discharge capacity, Coulombic Efficiency (CE), Electrochemical Impedance Spectroscopy (EIS) and Equivalent Circuit Modelling (ECM). This revealed the high PyC sample exhibited the most promising results with the highest CE (98.4%) and lowest overall impedance (most notably $R_{SEI} = 1.608 \Omega$; $CPE_D = 0.293 \Omega$). When comparing these parameters with corresponding values for graphite in literature, it was determined that the high PyC presented as a competitive alternative to conventional graphite anodes. The results were attributed to the unique crystalline nature of the high PyC structure through its high graphitisation with the presence of few amorphous structures. Thus, as each sample was increasingly pyrolysed, they displayed increasing graphitisation, which led to superior electrochemical and ionic properties.

1. Introduction

The global energy market is moving toward a zero-emission system whilst simultaneously meeting increasing energy demands. The Hydrogen Council labelling hydrogen as “the missing piece of the clean energy puzzle”^[1], has intensified the search for sustainable energy solutions. From this, the production of hydrogen from methane pyrolysis (MP), a product colloquially known as ‘Turquoise Hydrogen’^[2], has emerged as a promising solution for completing this intricate puzzle.

1.1 Comparison of Different Approaches to Hydrogen Synthesis and Cost Analysis

The World Energy Council reports that 96% of hydrogen produced is derived from fossil fuels^[3]. This predominantly involves the traditional processes of Steam-Methane Reforming and Coal Gasification, designated as grey and brown, respectively, contributing to 797 million tons of CO₂ per year^[3]. Operating these processes with integrated Carbon Capture and Storage apparatuses gives blue hydrogen. The utilisation of electrolysis for hydrogen synthesis, has a palette of colours assigned to its production, with green being the preeminent hue, occurring when the electricity in this process is sourced from renewable means.

While green and blue hydrogen appear as viable emission cutting alternatives, both harbour intrinsic challenges. Green hydrogen’s notable elevated cost diminishes its practicality. Similarly, blue hydrogen “is not clean, not a low-carbon source of energy and not a solution to the global climate crisis” as criticised by the Institute for Energy Economics and Financial Analysis^[4] since it fails to

eliminate greenhouse gas emissions and only captures 90% of CO₂ emitted without incurring additional costs.

Turquoise hydrogen holds promise as a compelling alternative and potentially transformative replacement to conventional methods. Forbes anticipates significant market growth, predicting an expansion from the current \$15 million USD to \$144 million by 2030^[5]. Diab et al.’s study emphasises turquoise hydrogen is a “game changer”, demonstrating through a LCA that MP offers a clear and sustainable path forward, removing up to 5.22 kgCO₂eq/kg for each kilogram produced^[6].

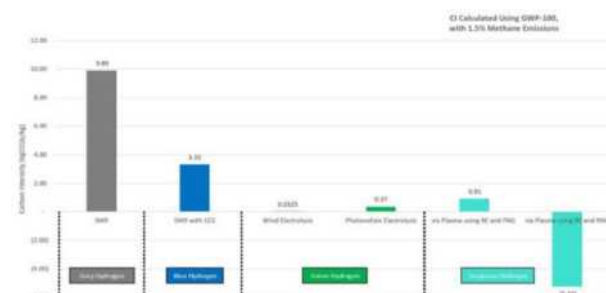


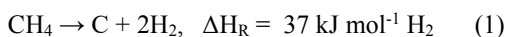
Figure 1. Comparing hydrogen types, turquoise hydrogen excels in removing CO₂ when powered by the same source as green hydrogen [6].

1.3 Fundamentals of Methane Pyrolysis

Pyrolysis denotes molecular decomposition in the presence of heat. In MP, natural gas serves as a favourable feedstock owing to its abundance, high energy density, availability, cost-effectiveness and notably its high methane content of over 85%^[7].

This methane undergoes pyrolysis to yield ‘turquoise’ hydrogen and a pyrolytic carbon (PyC) by-product. The

endothermic nature of this reaction adheres to the following chemical equation^[8].



Considering this, the produced carbon emerges as a keystone of interest, showing promise as an anode material in lithium-ion batteries (LIBs), potentially mirroring characteristics of graphite.

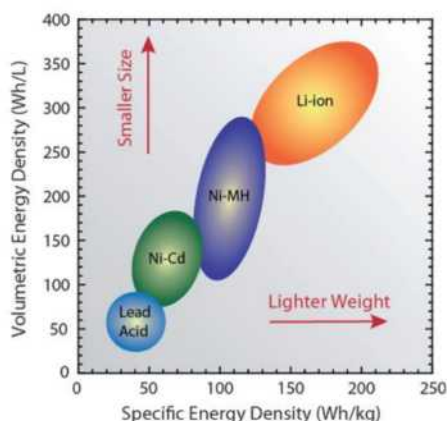
The primary endeavour of this paper is to assess the performance of samples of varying pyrolytic extents as potential replacements and compare these samples to graphite through literature. In simulating carbonaceous anodes, three distinct samples of PyC were produced, each indicative of an increasing MP reaction duration. Subsequently these were labelled low PyC, medium (med) PyC and high PyC denoting the shortest to longest reaction times.

2. Background

2.1 Rechargeable Batteries

The rechargeable battery market is currently valued at \$90bn USD in 2020 and is poised to almost double by 2030, reaching \$150bn^[9]. This growth is fuelled by demand from sectors like portable electronic devices, electric vehicles and renewable energy storage. While early developments included Nickel-Cadmium (NiCd) and Nickel-Metal (NiMH) batteries, LIBs have replaced these models due to their higher energy densities, low self-discharge rates (0.5-3% vs 10-20% per month^[10, 11]), enhanced storage potential, lightweight properties, and environmental benefits. Higher energy densities and low self-discharge rates contribute to increased battery performance by enabling the storage of more energy per unit of mass and maintaining stored energy for longer periods, respectively.

Figure 2. A 2009 study by Landi et al. compared the energy densities of rechargeable batteries^[12].



2.2 Lithium-Ion Batteries

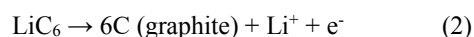
Global demand for LIBs is set to grow markedly to 4.7 TWh in 2030 from 700GWh in 2022^[13].

Over decades of research on LIBs, the structural development has favoured layered oxides Li_xMO_2 predominantly lithium cobalt oxide (LiCoO_2) and lithium manganese cobalt oxide. Of particular significance to this

study is the anode material with early LIBs employing Li-metals and Li-alloys. However, safety concerns prompted their discontinuation as such materials led to dendrite formation - metallic growth networks developing at the anode-electrode interface during charging - posing risks of short circuits and overheating^[14, 15].

In 1985, research by Akira Yoshino with Asahi Kasei led to the development of a LIB fitted with LiCoO_2 and a carbonaceous material. Today, graphite is the anode of choice in LIBs due to its optimal structure and low chemical potential of 0.1 eV, boosting energy output and efficiency^[16].

During discharge, lithium is oxidised from a 0 to +1 oxidation state in the anode^[17].



These Li^+ ions migrate through the electrolyte medium until they reach the cathode where they intercalate into the layered crystal structure of lithium cobalt oxide, reducing cobalt from a +4 to +3 oxidation state. This is depicted by reaction (3)^[17].



When reactions (2) and (3) are run in reverse, the charging cycle is initiated where Li^+ ions deintercalate the cobalt oxide structure and are incorporated into the graphite network of the anode.

2.3 Graphitic Anode

LIBs require up to 30 times more carbon to lithium during synthesis. Anodes tend to age quicker, reported by Sarkar et al. The increased electrolyte in reactions lead to the formation and growth of a Solid Electrolyte Interphase (SEI) layer, which is a primary ageing mechanism^[18]. As the world shifts from traditional energy sources, the escalating demand for lithium-ion batteries (LIBs) is set to cause graphite demand to concurrently soar. A surge in MP would add another dimension. In this scenario, incorporating PyCs into LIBs appears to be a compelling path forward. Graphite in anodes is characterised by graphene sheets intercalated with lithium. This gives a high theoretical capacity of 372 mAh/g. The high crystalline structure is key in influencing battery performance and is advantageous in anodes with Igarashi et al. reporting robust cycle stability and a capacity retention potentially exceeding 200 cycles in such systems^[19].

2.4 Sources of Resistance in LIBs

Coulombic efficiency (CE) refers to the ratio of the discharge capacity (following a full charge) to the charging capacity, as represented by Equation 1. As such, the measure is correlated with battery life, and is utilised in the calculation of capacity retention^[20]. In Li-ion batteries, the most significant contributors to a diminished coulombic efficiency relate to the development of either the SEI or the cathode electrolyte interphase, in addition to shuttle reactions in polysulfide batteries^[21, 22].

$$CE = \frac{\text{charge released during discharge}}{\text{charge previously stored}} \quad \text{Equation 1}$$

The internal resistance activity of focus for this paper is the growth of the SEI, the limitation of which has been a topic of focus in recent literature^[23, 24]. As An et al. posits, the “entirety of the anode surface” should have coverage provided by SEI growth, to minimise Li-ion depletion within the electrolyte medium^[25, p.54]. The first charge-discharge cycle of an Li-ion battery is generally associated with a 10% lithium loss in the process of synthesising this SEI^[26]. Nonetheless, further thickening of the SEI over the battery life results in a gradual loss of Li-ions and solvent, further increasing the internal resistance. Thus, its management and optimisation continues to be an area of focus in the field of electrochemistry^[21, 27]. Zhang et al. highlight the rate of SEI formation following this initial charge-discharge cycle is related to several properties, notably: particle size, basal/edge ratio, pore size, degree of crystallinity and surface chemical composition^[28, p. 35139].

2.5 Crystallinity and Anode Performance

In the context of pyrolytic carbon, the degree of crystallinity of the active anode material of a Li-ion battery is of particular interest. Where high PyC is produced at high pyrolysis temperatures, an increasingly ordered, graphitic-like structure with high crystallinity is developed. Conversely, low PyC is related to a lower degree of crystallinity, and thus a lesser extent of graphitisation^[29]. In the context of factors affecting SEI formation, the likelihood of solid lithium metal dendritic formation is highly related to the crystallinity of the anode material. Crystalline anode materials such as graphite readily provide nucleation sites for dendrites to form due to their propensity for non-uniform lithium deposition at the anode-electrolyte interface^[30]. As such, Guo et al. posit PyC anodes are related to higher coulombic efficiencies, higher discharge capacities and thus improved cycling stability when compared to graphite-based anodes^[31]. The increase in electrolyte retention and lithium-ion diffusivity invoked by the less ordered structure and reduced SEI development of PyC (when compared to graphite), render this alternative potentially competitive where cycling stability is critical^[32].

2.6 Electrical Impedance Spectroscopy (EIS)

Electrochemical Impedance Spectroscopy (EIS) is a commonly used technique which involves the application of an AC signal across a minute amplitude to an electrochemical system and observing its current response over a large frequency range^[33]. In turn, this provides insight into the nature of resistance, capacitance, and inductance within the system. Data acquired from EIS regimes may be fitted to an Equivalent Circuit Model (ECM) using computational software over a wide range of simulated frequencies. In this way, individual contributors to resistance within the battery assembly (e.g. electrodes, electrolyte etc.) may be described by circuit elements (i.e. resistors, capacitors, Constant Phase Elements {CPE}

etc.)^[34]. The quality of fit may be assessed through scrutiny of the error pertaining to each model parameter, in addition to the chi-squared value of the model.

The most common ECM employed for LIBs is displayed in Figure 3^[35]. One contributor to impedance, the bulk resistance (R_b), refers to the sum of the resistance derived from the electrodes ($R_{\text{current collector}}$), separator ($R_{\text{separator}}$), and electrolyte ($R_{\text{electrolyte}}$). As the bulk materials do not significantly affect measured capacitance or inductance, a resistor is sufficient in modelling their effect on impedance^[35]. The electrolyte-electrode interface where the SEI is formed is related to a more complex impedance effect, where the reversibility of SEI formation can influence the nature of both the system’s resistance and capacitance, in addition to imparting a phase shift on the response^[36]. Thus, both the real (R_{SEI}) and complex (CPE_{SEI}) influences on impedance about the interfacial layer are included in the model. The resistance associated with charge transfer interactions (R_{ct}) relates to electron migration between phases, whilst the diffusion CPE (CPE_D) accounts for the real and complex resistance imparted from the mass transport of Li-ions by diffusion, which is most strongly associated with the chemical characteristics of the electrode-electrolyte interface^[37]. The resistance and capacitance of the double layer is related to the ECM by the model parameter CPE_{dl} . Figure 4 illustrates how these individual contributors to impedance can be attributed to different frequency ranges in a Nyquist plot, one of the products of EIS.

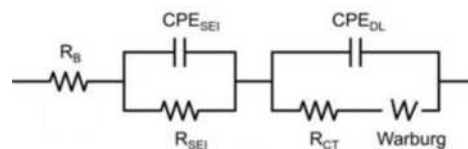


Figure 3. Typical ECM representing a Li-ion battery assembly^[35].

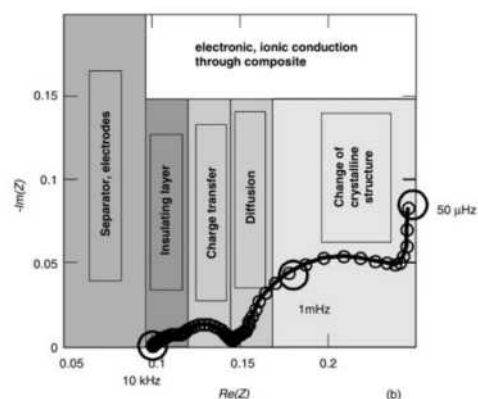


Figure 4. Example of a Nyquist plot produced in EIS analysis, with annotations illustrating the contribution of each element/process to impedance across the frequency range^[33].

3. Methodology

3.1 Anode Synthesis and Casting

Samples of high, med and low PyC co-product synthesised in the thermal decomposition of methane were received from ExxonMobil. These samples were ground using a planetary ball mill, reducing the median particle

sizes to 11.3, 10.4, and 13.0 microns for low, med, and high PyC samples, respectively (see Supplementary Materials Figure S1 for particle size distribution analysis plot). Three anode material formulations were prepared using a 9:1 weight ratio of each carbon sample received to polyvinylidene fluoride (PVDF). PVDF is the main binding agent of the synthesised anode, which was selected due to its preference as an industry standard and high thermal and electrochemical stability^[38].

N-methyl pyrrolidone (NMP) was introduced dropwise to each dry mixture over a magnetic stirrer hotplate until a viscous slurry-like consistency was achieved. Stirring was continued for a period of two hours until the PVDF binder was thoroughly dispersed. This slurry was then applied and evenly distributed across a copper foil-covered mounting slide using a coverslip to ensure even anode coating thickness (~0.1 mm thick) (see Figure 5 for details of slide preparation).

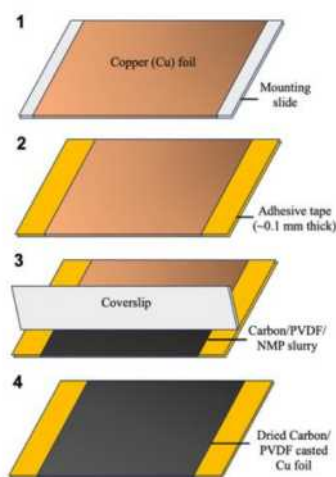


Figure 5. Schematic of slide preparation prior to punching. Carbon/PVDF/NMP slurry is applied to copper foil to the thickness of a single layer of tape, then dried.

The addition of sufficient NMP solvent to produce a slurry with viscosity conducive to application is critical. However, as highlighted by Li et al., the minimisation of trace solvent present in the final anode product has a favourable effect on the coulombic efficiency of the final LIB^[39]. Thus, the prepared slide was dried at room temperature within a fumehood overnight prior to punching of the coated foil into 19 mm discs. These discs were further dried in a vacuum oven prior to coin cell battery assembly to minimise the presence of trace NMP.

As a result of issues with achieving sufficient wettability in the slurry produced from the low PyC sample, a smaller electrode disc size was selected (11 mm). As such, the specific charge/discharge capacity will be a more robust comparative tool to assess anode performance relative to the charge/discharge capacity, to account for the variation in active anode mass.

3.2 Coin Cell Battery Assembly

The assembly of the coin cell battery involved a series of materials in addition to the anode synthesised above. The scope and aims of these research activities exclusively

involves the synthesis of the anode electrode, whilst the lithium cobalt oxide (LiCoO_2) cathode and the separator (polypropylene) were commercially sourced from MTI Corporation.

The coin cell battery was assembled by positioning the prepared anode disc, a 19 mm polypropylene (PP) separator, and a 15 mm cathode disc (comprised of LiCoO_2) within a coin cell casing. A spacer was positioned over the assembled anode/separator/cathode apparatus and tightened using a spring and peg. A 1.0 M lithium hexafluorophosphate (LiPF_6) in a 50/50 (v/v) ethylene carbonate-diethyl carbonate (EC/DEC) solution was selected as the most suitable electrolyte based on existing literature, and was applied prior to, and following, the positioning of each electrode disc and the PP separator at a quantity of five drops, as illustrated in Figure 6^[40]. This process was repeated for all three carbon samples being investigated.

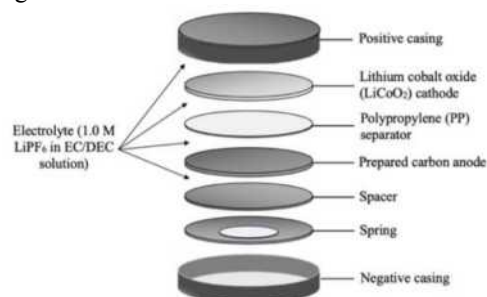


Figure 6. Visual representation of coin battery assembly, with five drops of electrolyte being introduced between each electrode and separator.

The cathode material was selectively chosen over alternatives to prevent the introduction of further internal resistance within the battery through external means, such as shuttle reactions (e.g. in LFP/graphite and NMC811/graphite assemblies)^[22, 41, 42].

3.3 Charge and Discharge Cycling

Electrochemical testing of the four produced coin cell batteries was undertaken by charge cycling using a Keithley 2461 Source Measure Unit (SMU). For graphite/ LiCoO_2 assemblies, literature suggests a nominal voltage of 3.7 V and a standard operating range of 2.75 – 4.2 V is standard^[43]. This informed the selected charge and discharge cycle regimes for all three assembled batteries, as presented in Table A. The charging cycle was conducted in a two-stage process, with the first a constant current phase, where an estimated 20% of the batteries' capacity rating (0.2 C) was used. The second phase of charging involved 10% of this current being applied (this occurred once the battery reached the maximum operating voltage). The discharging regime followed a similar pattern with a current equivalent to 20% of the battery capacity being applied, and a target voltage equivalent to the minimum of its operating range.

The Keithley 2461 SMU produced charge/discharge profiles for each charge/discharge cycle. These were utilised to synthesise corresponding voltage/capacity profiles and calculate the coulombic efficiency and final

specific capacities of each battery tested, as presented in the Results section of this report.

Table A. Input specifications for Keithley 2461 SMU for charge/discharge cycling of assembled batteries.

	Charge Regime	Discharge Regime
Target Voltage	4.3 V	2.8 V
Initial Current	0.3 mA	
Target Current	0.03 mA	
Constant Current		0.3 mA

3.4 Electrical Impedance Spectroscopy (EIS)

A Solartron Instruments 1280B Electrochemical Measurement Unit (EMU) was utilised to undertake impedance spectroscopy for each coin cell battery tested. Electrochemical Impedance Spectroscopy (EIS) was undertaken with an AC amplitude of 10 mV across a frequency range of 0.1 Hz to 20 000 Hz, or 20 kHz.

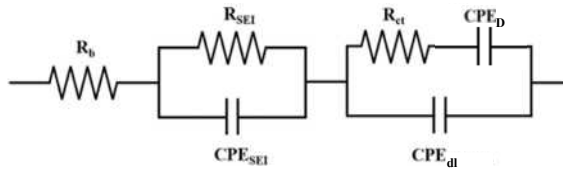


Figure 7. Schematic of equivalent electrical circuit model employed; adapted from Choi et al.^[44].

An ECM which reflected the features of the battery assembly presented in this paper was developed to fit the EIS results produced. Figure 7 presents the model, which includes three resistors and three common phase elements (CPEs), as adapted from Choi et al.^[44]. Although some literature suggest the use of a Warburg parameter in place of CPE_D, the non-ideal surface of the anode synthesised in this paper calls for a CPE, which produces a more robust model at low frequencies^[45].

4. Results

4.1 Capacity and Coulombic Efficiency

The resulting galvanostatic charge/discharge profiles (presented in Figures S2 to S11 in Supplementary Materials) were then utilised to produce several metrics. Initially, the theoretical capacity profile corresponding to each galvanostatic charge/discharge profile was synthesised. As illustrated in Equation 2, the sum of the capacity at each time interval over the length of the profile (from time 0 to time n) is equal to the theoretical charge/discharge capacity ($Q_{charge/discharge}$) of the battery.

$$Q_{charge/discharge} [mAh] = \sum_{i=0}^n I \times \Delta t_i \quad \text{Equation 2}$$

From the calculated theoretical capacity profile for each charge/discharge cycle, an equivalent specific capacity profile could then be synthesised which accounted for the different quantities of active material present within the anodes utilised. Firstly, six discs of copper foil of equivalent sizing to those utilised in anode casting (i.e. 19 mm for high and med PyC and 11 mm for low PyC) were

prepared and weighed. The average mass of these discs, in addition to the masses of each casted anode following drying procedures are presented in Table S1 in Supplementary Materials. Equation 3 presents the conversion of theoretical charge capacity (Q_{charge}) to specific theoretical charge capacity (\dot{Q}_{charge}) using the active mass of the casted anode, calculated by subtracting the mean mass of punched copper discs ($\bar{m}_{Cu disc}$) from the dried anode mass (m_{anode}). Determination of the specific theoretical discharge capacity ($\dot{Q}_{discharge}$) also follows Equation 3 where \dot{Q}_{charge} is substituted by $\dot{Q}_{discharge}$.

$$\dot{Q}_{charge} \left[\frac{mAh}{g} \right] = \frac{Q_{charge}}{m_{anode} - \bar{m}_{Cu disc}} \quad \text{Equation 3}$$

Calculating the coulombic efficiency (CE) from the specific theoretical charge (\dot{Q}_{charge}) and discharge ($\dot{Q}_{discharge}$) capacities then follows Equation 4.

$$CE (\%) = \frac{\dot{Q}_{discharge}}{\dot{Q}_{charge}} \times 100\% \quad \text{Equation 4}$$

The charge/discharge capacities are presented in Supplementary Material Figure S12. Figure 8 presents the specific theoretical charge/discharge capacities calculated from low, med and high PyC respectively. Note that the med PyC anode possessed both the highest charge and discharge capacities, in addition to the highest specific charge and discharge capacities (247 mAh/g and 232 mAh/g, respectively). Moreover, the low PyC battery assembly exhibited the lowest specific discharge capacity, despite possessing a specific charge capacity (242 mAh/g) similar to that of the med PyC sample. This greatly exceeded the specific charge capacity of the high PyC sample (195 mAh/g).

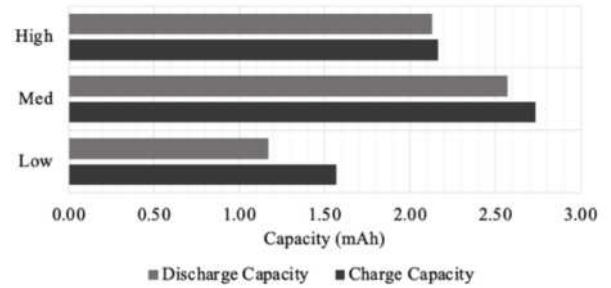


Figure 8. Comparison of specific theoretical charge and discharge capacity (mAh/g) of low, med and high PyC anodes.

The coulombic efficiencies from the first and second full charge cycle were calculated for each of the low, med and high PyC anode batteries, as presented in Figure 9. Note the first cycle coulombic efficiency for low PyC could not be calculated due to data loss from an equipment operational error. Nonetheless, it can be observed the highest coulombic efficiency in the second charge-discharge cycle was produced from the high PyC anode (98.4%). Conversely, the low PyC anode was associated with the lowest coulombic efficiency at the second charge-discharge cycle (74.5%).

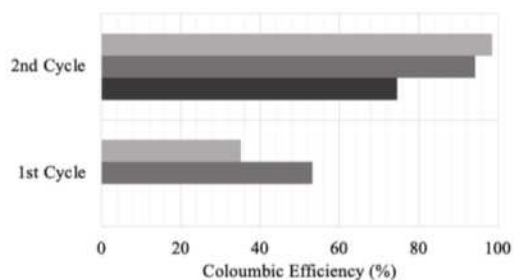


Figure 9. Comparison of coulombic efficiency (%) of low, med and high PyC anodes.

Discharge curves were then produced for each of the anode types, by plotting voltage against the specific discharge capacity of each anode. Figure 10 illustrates the discharge curves for the batteries produced from low, med and high PyC anodes. The discharge curves of all three batteries using different types of carbon anode exhibit fairly linear reductions in potential (voltage) over time. This is reflective of the second-cycle discharging regime profiles presented in Supplementary Materials (Figures S9 to S11), which all produced similar patterns of discharge.

4.2 Electrochemical Impedance Spectroscopy (EIS)

Presented in Figure 11, the Nyquist plot output from EIS conducted illustrates the extent of variation in the nature of measured impedance between each PyC sample. Where the x-axis describes the real contribution to impedance (Z'), the y-axis is associated with its imaginary component^[46]. Moreover, another feature of note is the frequency distribution of the curve. The right-most section of each curve refers to the impedance measurements observed in the low frequency region, providing insight into the factors affecting mass transfer or diffusion rates within the system^[33]. Conversely, bulk and SEI resistance affect the nature of the curve in the high frequency region, illustrated in the left-most section and the curve's proximity to the origin. Individual Nyquist plots which illustrate the ECM fitting of each EIS data set are provided in the Supplementary Materials (Figures S13-S15).

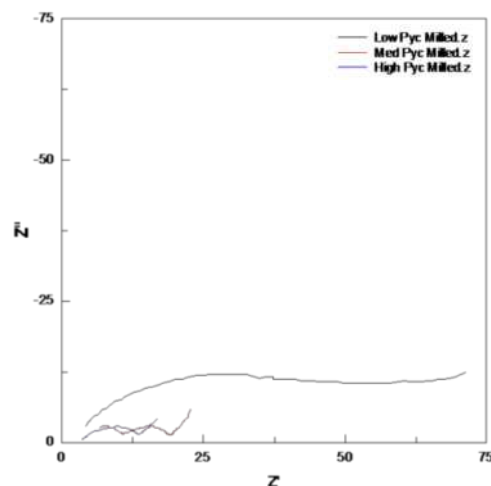


Figure 11. Nyquist plots of the negative imaginary impedance (Z'' ; y axis) against the real impedance (Z' ; x axis) at each excitation frequency from 1- 20 000 Hz for low, med, and high PyC anodes.

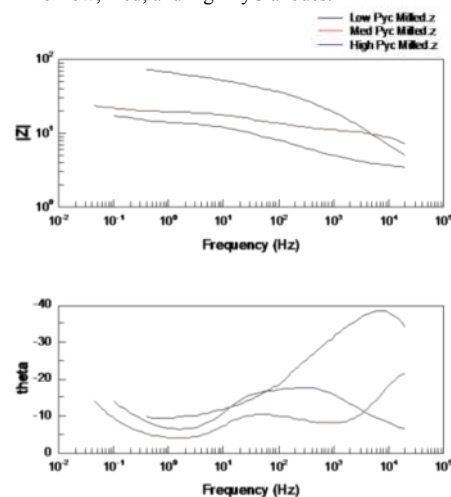


Figure 12. Bode plots for low, med, and high PyC anodes; illustrating the logarithm of the impedance (upper plot) and the phase shift (θ) (lower plot) against the logarithm of the frequency applied to the system.

Figure 12 illustrates the Bode plot outputs of the EIS conducted, with the upper plot presenting the logarithm of the measured impedance $|Z|$, and the lower plot the phase

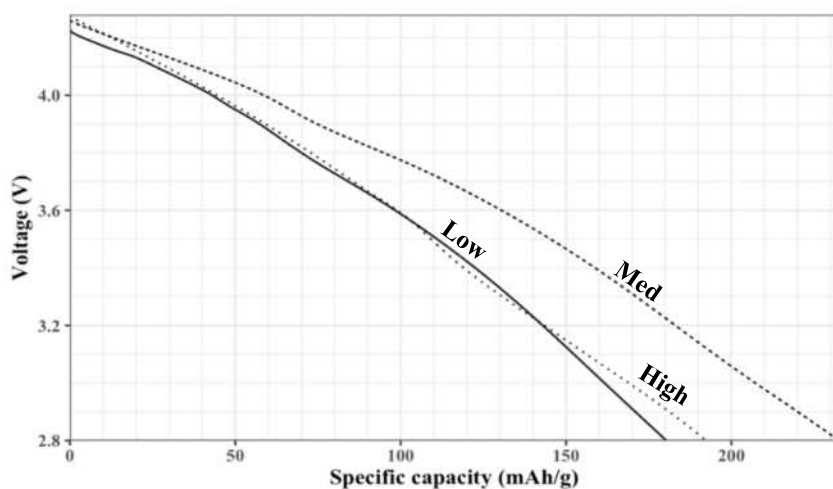


Figure 10. Voltage (V) plotted against specific discharge capacity (mAh/g) for high (dotted line), medium (dashed line) and low (solid line) PyC. The highest specific discharge capacity is observed in medium PyC.

shift (θ) against the logarithm of frequencies at which measurements took place. The variation in impedance across frequencies is visualised with ease through the upper plot, which illustrates the low relative impedance of the battery developed from the high PyC sample, compared to both the med and low PyC samples.

From the Bode plot in Figure 12, it can also be noted that at high frequencies, the phase shift of both low and med PyC batteries was elevated compared to the high PyC battery. The curvature of each of the phase shift plots is highly varied between samples across the frequency range. Similarly, in Figure 11, the curvature of each Nyquist plot is varied quite significantly, potentially suggesting the model parameters resulting from fitting to the ECM may vary between pyrolytic sample.

Table B presents the output values for the resistor elements of the ECM each EIS output was fitted to. Considering the chi-squared values of fit were all below the significance level ($p = 0.05$) (presented alongside error values in the Supplementary Materials – Tables S3 and S4), it may be validated that the tabulated ECM elements vary significantly between the different PyC samples used in anode synthesis.

The highest R_b (resistance associated with the bulk assembly – electrodes, electrolyte and separator) was observed in the med PyC (3.418 Ω), with the lowest resistance observed in the low PyC battery (2.365 Ω). Nonetheless, the R_{SEI} was significantly elevated for the low PyC sample (16.43 Ω) compared to the high PyC sample, which exhibited the lowest resistance for this parameter (1.612 Ω). The R_{ct} of high and med PyC were very similar, at 7.255 Ω and 8.84 Ω , respectively. Moreover, the CPE_D was also similar between the med and high PyC, at 0.342 Ω and 0.293 Ω , respectively. This is reflected in the similarities in the curvature of the med and high Nyquist plots at low frequencies (observed in the right-most parts of their curves in Figure 12). Note that due to the failure of the low PyC battery during the low frequency regime in EIS testing, the R_{ct} CPE_D could not be obtained. Nonetheless, the general scale and trajectory of the low PyC Nyquist plot suggests a high R_{ct} and CPE_D relative to low and med PyC samples could be expected. During data cleaning, any abnormal data points associated with battery failure were removed from the EIS data produced.

Table B. Output values for model resistance parameters following fitting of EIS results to ECM. *Note a result for R_{ct} and CPE_D could not be obtained for low PyC due to battery failure.

	Low PyC	Med PyC	High PyC
R_b (Ω)	2.365	3.418	3.129
R_{SEI} (Ω)	17.09	8.687	1.608
R_{ct} (Ω)	N/A*	7.255	8.927
CPE_D (Ω)	N/A*	0.342	0.293

5. Discussion

5.1 Charge and Discharge Capacities

The specific charge/discharge capacities provided essential insights into the battery storage capabilities and operational time, representing the amount of electric charge stored per unit mass. As per Figure 8, med PyC displayed the highest specific discharge capacity of 247 mAh/g, closely followed by the low PyC and subsequently the high PyC. These trends were incongruent with what was expected and that of literature. The high PyC experienced higher crystallisation and graphitisation and is structured as captured by Honorato et al. in Figure 13^[47].

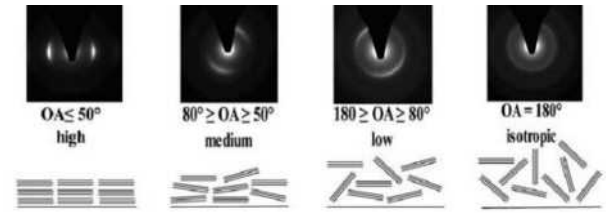


Figure 13. Three samples of PyC made from varying reaction temperatures producing a similar low, med and high PyC, displaying greater crystallisation for High samples.

In light of the superior structural order of high PyC, the capacity in this sample was expected to be the greatest. Higher crystallinity facilitates efficient electron and ion movement giving enhanced conductivity, stability producing higher anode capacities. Hence, the discrepancy in capacity can be attributed to the greatest particle size of high PyC of 13.0 microns which is seen in Supplementary Materials Figure S1. Analysing this graph shows an outlier in the high PyC sample compared to the low and med samples being 11.3 and 10.4 microns, respectively. The milling process was conducted for 10 minutes. For future experiments, prioritising uniform particle size over consistent milling duration would be prudent to mitigate potential discrepancies. However, it is important to note that the graphitisation of med and high PyC experience similar degrees of graphitisation. Comparing the charging capacity of the med PyC to that of graphite indicates a competitive sample. Studies conducted by Laziz et al. provided charging capacities of a raw and concentrated graphite of 145 mAh/g and 291 mAh/g, respectively^[48]. Comparing the med PyC sample of 247 mAh/g to the experimental value of 291 mAh/g (considering that high PyC should experience higher charging capacities), demonstrates the competitiveness of Med and potentially high PyC samples to that of graphite.

5.2 Coulombic Efficiencies

From this, the CE was computed as represented in Figure 10. As expected, the high PyC displayed the highest CE of 98.4%, closely followed by 94.1% of med PyC and low PyC having a 74.5% CE. This outcome aligns with expectations and is attributable to the enhanced graphitization of high and med PyC, fostering improved pathways and channels for the improved mobility and de/intercalation of Li^+ ions to efficiently move between electrodes during charging.

Comparing the CE of the PyC samples to that of graphite found in literature sheds a positive light on the performance of high PyC. Laziz et al. calculated CE values of 98.3% and 99.7%, values that are similar to that of med and high PyC, demonstrating the potentially competitive nature of the high and med PyC samples^[48].

In analysing the CE, only second cycle values were considered. This is due to the formation of the SEI on the anode during the first cycle. This passivating layer protects the anode from the electrolyte. This SEI layer prevents unwanted reactions, selectively allowing the passage of Li^+ ions. As such, this regulates ion transport during charge and discharge cycles, contributing to a greater CE. Conversely, it is important to note that this is only beneficial to an extent. Due to excessive cycling, the accumulation of solid products may occur which in turn can hinder Li^+ transport.

5.3 Charge and Discharge Profiles

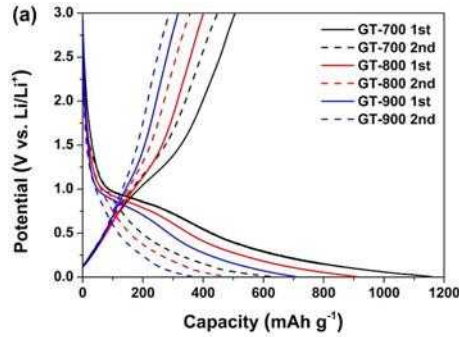


Figure 15. The charge/discharge profiles computed by Han et al. on varying PyC temperatures (GT-700 = low PyC)^[49].

Figure 15 illustrates the charge/discharge profiles of low, med and high PyC samples (distinguished by varying reaction temperatures). A typical representation of such plots entail a steep initial drop in voltage, followed by a stable plateau and a sustained decrease in voltage. The plateau in the graph represents the nominal voltage, which is the average or standard voltage in the anode. In contrast, the discharge profiles for all three PyC samples from this study in Figure 10 deviated from this trend, displaying a linear-line slope, devoid of a plateau. The observed variation is likely due to the C-rate, representing the current as a fraction to the battery capacity. The quoted discharge profile operated at 0.1C, whereas this experiment was executed at a discharge of 0.2C, a much more moderate rate. For subsequent experiments, slowing discharge to 0.1C might be preferred, allowing for more data points over time and potentially revealing a nominal voltage, given the discharge at 10% capacity per hour.

5.4 Trends in Electrochemical Impedance Spectroscopy

The trends observed in the combined Nyquist plot (Figure 11) appear to indicate the most significant overall impedance was observed in the anode developed from low PyC. This is supported by the ECM model parameters obtained when the EIS data was fitted to the ECM employed in this paper. As noted in Table B, the R_{SEI} was heightened for low PyC, and although the R_{ct} and CPE_D could not be obtained due to battery failure, the general

trajectory of the low PyC plot in Figure 11 indicates these values would likely significantly exceed that of the med or high PyC. This trend was further supported in the upper Bode plot presented in Figure 12, which indicated the highest impedance at all frequencies below ~5 kilohertz was associated with the low PyC sample. Furthermore, the lower Bode plot in Figure 13 illustrated the high phase shift observed in the low PyC sample. This suggests issues pertaining to the adaptability of the battery and its response rate to phenomena (e.g. changes in load) could be expected.

Table C. Output values for graphite anode model resistance parameters following fitting of EIS results to ECM, adapted from Paul [50].

	Graphite
$R_b (\Omega)$	1.39
$R_{\text{SEI}} (\Omega)$	12.70
$R_{\text{ct}} (\Omega)$	5.77
$\text{CPE}_D (\Omega)$	1.25

In an analogous study using an equivalent ECM model, Paul identified the resistances and common phase element statistics for a graphite anode, as presented in Table C^[50]. Of note is the significantly reduced R_{SEI} achieved in the high PyC sample compared to the graphite anode, at 1.608 Ω and 12.7 Ω , respectively. The CPE_D was also reduced significantly in the high PyC sample, suggesting an increasingly ideal and uniform charge distribution, whilst the R_b and R_{ct} remain slightly elevated in the high PyC assembly compared to that of graphite.

In a 2020 study, Rezaei et al. identified similar trends in EIS results as were observed in this study, where low pyrolysis temperature was associated with higher overall impedance than higher pyrolysis temperatures^[51]. As elucidated in Figure 16, the magnitude of the semi-circular behaviour of the 900 °C impedance spectrum is indicative of an elevated charge transfer resistance (R_{ct}), compared to the 1000 °C and 1100 °C pyrolysed carbon assemblies^[52]. It is worth noting that the highest pyrolysis temperature employed in this paper remains relatively low (1100 °C), thus further investigation into the effect of varying pyrolysis regimes by also measuring the impedance of higher pyrolytic temperatures samples could be beneficial to further deciphering trends in impedance.

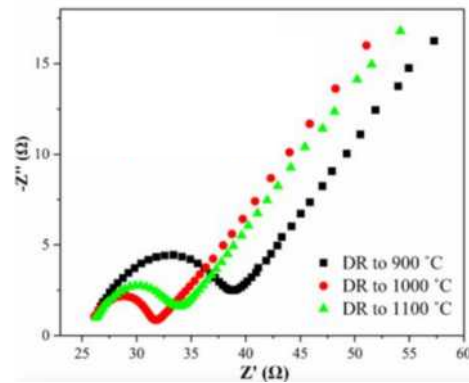


Figure 16. Nyquist plots of Li-ion batteries assembled with anodes of carbon produced at varying pyrolytic temperatures^[51].

5.5 Crystallinity and SEI Formation

Mao et al. identified anode material crystallinity as a significant factor in SEI formation and resistance. FTIR spectra was employed to characterise the chemical composition of anode SEI formed following battery cycling^[53]. The thinnest SEI layers and lowest active lithium loss were observed in the two anodes formed from disordered carbon powders presenting with a greater degree of amorphous structures (as measured by Raman spectra). A trend can be discerned where SEI thickness is partially modulated by crystallinity, where high crystallinity anodes possess generally thicker SEI layers. It should be noted that SEI formation is also modulated by several other anode material characteristics, such as porosity and particle surface area^[54]. However, due to the milling procedure undertaken in the methodology, these factors have been significantly mitigated by reducing the particle size to equivalent normal distributions between samples and through pore collapse, respectively^[55].

In the context of the ECM results of this study, the increased presence of amorphous structures in the low PyC sample were a contributing factor to its elevated R_{SEI} (17.09 Ω). Conversely, the optimised presence of amorphous structures occurred in the relatively ordered and graphitised high PyC sample, as demonstrated by its minimised R_{SEI} (1.608 Ω). Although the development of a thin SEI layer can be beneficial to the reduction of active lithium depletion in the electrolyte, if sufficient SEI coverage is not achieved, a capacity reduction effect is observed, in turn reducing anode performance^[56]. Furthermore, the literature value of R_{SEI} for the graphite anode assembly (12.70 Ω) indicates the complete absence of amorphous structures may also have a negative effect on SEI resistance^[50]. Where an entirely graphitised anode structure encourages SEI formation, a thick SEI layer may form which can contribute to this observed increase in R_{SEI} . Thus, high PyC anodes present as an optimised alternative for ensuring adequate SEI formation occurs without unnecessarily contributing to the total system resistance and reducing battery capacity.

6. Conclusion

From our testing regimes, it can be concluded that the high PyC outperformed low and med PyC anodes based on its superior coulombic efficiency (98.4%) and competitive specific discharge capacity. When this specific discharge capacity was compared to equivalent literature values for graphite, the potential efficacy of high PyC as an anode material was further reinforced. Furthermore, EIS testing indicated that compared to low and med samples, high PyC presents itself with a minimised impedance spectrum, which was further supported when viewing its parameters for the ECM employed compared to literature reference values of graphite. In particular, the significantly reduced R_{SEI} (1.608 Ω) and CPE_D (0.293 Ω) relative to the low (R_{SEI} of 17.09 Ω) and med (R_{SEI} of 8.687 Ω , CPE_D of 0.342 Ω) PyC samples highlighted the efficacy of high PyC as an anode material. The higher degree of crystallinity in the high PyC sample may have a positive effect on reducing

some resistances such as that of SEI, where the formation of a thinner, more stable SEI than that of graphite also has a positive effect on reducing impedance. Nonetheless, further testing would be required for a holistic assessment of the performance of PyC as an anode material. For example, the implementation of a lower current (i.e. 0.1 C instead of 0.2 C rate) in the charge-discharge cycling could provide further insights into the trends observed, by providing a more comprehensive data inventory and higher resolution discharge curves.

7. References

- [1] Hydrogen Council, "Why hydrogen", n.d. [Online] Available: <https://hydrogencouncil.com/en/why-hydrogen/> (Accessed: November 28, 2023).
- [2] J. M. Arcos and D. M. Santos, "The hydrogen color spectrum: Techno-economic analysis of the available technologies for Hydrogen production", *Gases*, vol. 3, no. 1, pp. 25–46, 2023. doi:10.3390/gases3010002
- [3] World Energy Council, "New Hydrogen Economy – Hope or Hype", World Energy Council, 2019. [Online] Available: <https://www.worldenergy.org/assets/downloads/WEInnovation-Insights-Brief-New-Hydrogen-Economy-Hype-or-Hope.pdf>. (Accessed: November 7, 2022).
- [4] IEEFA, "Blue hydrogen: Not clean, not low carbon, not a solution", IEEFA, 2023. [Online] Available: <https://ieefa.org/articles/blue-hydrogen-not-clean-not-low-carbon-not-solution#:~:text=Blue%20hydrogen%20is%20not%20clean,to%20he%20global%20climate%20crisis.&text=The%20model%20used%20by%20the,produce%20hydrogen%20with%20fossil%20fuels> (Accessed: November 18, 2023).
- [5] D. Bhamhani, "Turquoise hydrogen producers could capture flourishing graphite market", *Forbes*, [2023]. [Online] Available: <https://www.forbes.com/sites/dipkabhamhani/2023/11/01/turquoise-hydrogen-producers-could-capture-flourishing-graphite-market/> (Accessed: November 18, 2023).
- [6] J. Diab, L. Fulcheri, V. Hessel, V. Rohani, and M. Frenklach, "Why turquoise hydrogen will be a game changer for the Energy Transition", *International Journal of Hydrogen Energy*, vol. 47, no. 61, pp. 25831–25848, 2022. doi:10.1016/j.ijhydene.2022.05.299
- [7] M. Mezni, "LPG Recovery Unit Optimization", *Insat*, p. 15, 2015. doi:10.13140/RG.2.2.14475.90402
- [8] S. Schneider, S. Bajohr, F. Graf, and T. Kolb, "State of the art of hydrogen production via pyrolysis of natural gas," *ChemBioEng Reviews*, vol. 7, no. 5, pp. 150–158, 2020. doi:10.1002/cben.202000014
- [9] Allied Market Research, "Rechargeable batteries market size, share and growth analysis - 2030", n.d. [Online] Available: <https://www.alliedmarketresearch.com/rechargeable-batteries-market-A09294> (Accessed: December 1, 2023).
- [10] K. Kang et al., "Abnormal self-discharge in lithium-ion batteries", *ECS Meeting Abstracts*, vol. MA2018-01, no. 3, pp. 294–294, 2018. doi:10.1149/ma2018-01/3/294
- [11] D. G. Vutetakis, "Applications – transportation | aviation: Battery", *Encyclopedia of Electrochemical Power Sources*, pp. 174–185, 2009. doi:10.1016/b978-044452745-5.00370-1
- [12] B. J. Landi, M. J. Ganter, C. D. Cress, R. A. DiLeo, and R. P. Raffaele, "Carbon nanotubes for lithium ion batteries", *Energy & Environmental Science*, vol. 2, no. 6, p. 638, 2009. doi:10.1039/b904116h
- [13] J. Fleischmann et al., "Battery 2030: Resilient, sustainable, and Circular," McKinsey & Company, <https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/battery-2030-resilient-sustainable-and-circular> (accessed Dec. 7, 2023).
- [14] K. J. Siczek, "Negative electrode (anode) materials", *Next-Generation Batteries with Sulfur Cathodes*, pp. 117–131, 2019. doi:10.1016/b978-0-12-816392-4.00008-6.
- [15] T. Vegge et al., "Computational design of catalysts, electrolytes, and materials for Energy Storage," *New and Future Developments in Catalysis*, pp. 499–521, 2013. doi:10.1016/b978-0-444-53880-2.00023-5.
- [16] R. M. Salgado et al., "The latest trends in Electric Vehicles Batteries", *Molecules*, vol. 26, no. 11, p. 3188, 2021. doi:10.3390/molecules26113188.

- [17] H. D. Abruña, Y. Kiya, and J. C. Henderson, "Batteries and electrochemical capacitors," *Physics Today*, vol. 61, no. 12, pp. 43–47, 2008. doi:10.1063/1.3047681.
- [18] A. Sarkar, I. C. Nlebedim, and P. Shrotriya, "Performance degradation due to anodic failure mechanisms in lithium-ion batteries", *Journal of Power Sources*, vol. 502, p. 229145, 2021. doi:10.1016/j.jpowsour.2020.229145
- [19] D. Igarashi *et al.*, "Effect of crystallinity of synthetic graphite on electrochemical potassium intercalation into graphite", *Electrochemistry*, vol. 89, no. 5, pp. 433–438, 2021. doi:10.5796/electrochemistry.21-00062
- [20] W. H. Zhu, Y. Zhu, Z. Davis, and B. J. Tatarchuk, "Energy efficiency and capacity retention of Ni–MH batteries for storage applications", *Applied Energy*, vol. 106, pp. 307–313, 2012. doi:10.1016/j.apenergy.2012.12.025
- [21] A. Wang, S. Kadam, H. Li, S. Shi, and Y. Qi, "Review on modelling of the anode solid electrolyte interphase (SEI) for lithium-ion batteries", *NPJ Computational Materials*, vol. 4, no. 1, p. 15, 2018. doi:10.1038/s41524-018-0064-0
- [22] Y. Huang *et al.*, "Recent advances and strategies toward polysulfides shuttle inhibition for high-performance Li–S batteries", *Advanced Science*, vol. 9, no. 12, p. 2106004, 2022. doi:10.1002/adv.202106004
- [23] Z. Li, J. Liu, Y. Qin, and T. Gao, "Enhancing the charging performance of lithium-ion batteries by reducing SEI and charge transfer resistances", *ACS Applied Materials & Interfaces*, vol. 14, no. 29, pp. 33004–33012, 2022. doi:10.1021/acsami.2c04319
- [24] S. Solchenbach *et al.*, "Monitoring SEI formation on graphite electrodes in lithium-ion cells by impedance spectroscopy", *Journal of The Electrochemical Society*, vol. 168, no. 11, p. 110503, 2021. doi:10.1149/1945-7111/ac3158
- [25] S. J. An *et al.*, "The state of understanding of the lithium-ion-battery graphite solid electrolyte interphase (SEI) and its relationship to formation cycling", *Carbon*, vol. 105, p. 54, 2016. doi:10.1016/j.carbon.2016.04.008
- [26] A. Patil *et al.*, "Issue and challenges facing rechargeable thin film lithium batteries", *Materials Research Bulletin*, vol. 43, no. 8–9, pp. 1913–1942, 2007. doi:10.1021/acsenergylett.0c02336
- [27] D. L. Wood, J. Li, and S. J. An, "Formation challenges of lithium-ion battery manufacturing", *Joule*, vol. 3, no. 12, pp. 2884–2888, 2019. doi:10.1016/j.joule.2019.11.002
- [28] Z. Zhang *et al.*, "Operando electrochemical atomic force microscopy of solid–electrolyte interphase formation on graphite anodes: the evolution of SEI morphology and mechanical properties", *ACS Applied Materials & Interfaces*, vol. 12, no. 31, p. 35139, 2020. doi:10.1021/acsami.0c11190
- [29] N. Tsubouchi, C. Xu, and Y. Ohtsuka, "Carbon crystallization during high-temperature pyrolysis of coals and the enhancement by calcium", *Energy & Fuels*, vol. 17, no. 5, pp. 1119–1125, 2003. doi:10.1021/ef020265u
- [30] S. Di *et al.*, "A crystalline carbon nitride–based separator for high-performance lithium metal batteries", *Proceedings of the National Academy of Sciences*, vol. 120, no. 33, p. e2302375120, 2023. doi:10.1073/pnas.2302375120
- [31] W. Guo *et al.*, "Microstructure-controlled amorphous carbon anode via pre-oxidation engineering for superior potassium-ion storage", *Journal of Colloid and Interface Science*, vol. 623, pp. 1075–1084, 2022. doi:10.1016/j.jcis.2022.05.073
- [32] K. Chang *et al.*, "Graphene-like MoS₂/amorphous carbon composites with high capacity and excellent stability as anode materials for lithium-ion batteries," *Journal of Materials Chemistry*, vol.21, no.17, pp.6251–6257, 2011. doi:10.1039/C1JM10174A
- [33] Y. Barsukov and J. R. Macdonald, "Electrochemical impedance spectroscopy", *Characterization of Materials*, vol. 2, pp. 898–913, 2012. doi:10.1002/0471266965
- [34] Z. Lukács and T. Kristóf, "A generalized model of the equivalent circuits in the electrochemical impedance spectroscopy", *Electrochimica Acta*, vol. 363, p. 137199, 2020. doi:10.1016/j.powtec.2023.119252
- [35] B. R. Chen *et al.*, "A mathematical approach to survey electrochemical impedance spectroscopy for aging in lithium-ion batteries", *Frontiers in Energy Research*, vol. 11, p. 167, 2023. doi:10.3389/fenrg.2023.1132876
- [36] B. Y. Chang, "The effective capacitance of a constant phase element with resistors in series", *Journal of Electrochemical Science and Technology*, vol. 13, no. 4, pp. 479–485, 2022. doi:10.2139/ssrn.4079679
- [37] M. Janssen and J. Bisquert, "Locating the frequency of turnover in thin-film diffusion impedance", *The Journal of Physical Chemistry*, vol. 125, no. 28, pp. 15737–15741, 2021. doi:10.1021/acs.jpcc.1c04572
- [38] J. Y. Eom and L. Cao, "Effect of anode binders on low-temperature performance of automotive lithium-ion batteries," *Journal of Power Sources*, vol. 441, p. 227178, 2019. doi:10.1016/j.jpowsour.2019.227178
- [39] J. Li, C. Daniel, S. J. An & D. Wood, "Evaluation residual moisture in lithium-ion battery electrodes and its effect on electrode performance. MRS advances", *Research Snapshots from MRS*, vol.1, no.15, pp.1029–1035, 2016. doi:10.1557/adv.2016.6
- [40] H. Lundgren, M. Behm, and G. Lindbergh, "Electrochemical characterization and temperature dependency of mass-transport properties of LiPF₆ in EC: DEC", *Journal of The Electrochemical Society*, vol. 162, no. 3, p. A413, 2014. doi:10.1149/2.0641503jes
- [41] Y. Lyu *et al.*, "An overview on the advances of LiCoO₂ cathodes for lithium-ion batteries", *Advanced Energy Materials*, vol. 11, no. 2, p. 2000982, 2020. doi:10.1002/aenm.202000982
- [42] T. Boulanger *et al.*, "Investigation of redox shuttle generation in LFP/graphite and NMC811/graphite cells", *Journal of The Electrochemical Society*, vol. 169, no. 4, p. 040518, 2022. doi:10.1149/1945-7111/ac62c6
- [43] S. Saxena, C. Hendricks, and M. Pecht, "Cycle life testing and modeling of graphite/LiCoO₂ cells under different state of charge ranges", *Journal of Power Sources*, vol. 327, pp. 394–400, 2016. doi:10.1016/j.jpowsour.2016.07.057
- [44] W. Choi, H. C. Shin, J. M. Kim, J. Y. Choi, and W. S. Yoon, "Modelling and applications of electrochemical impedance spectroscopy (EIS) for lithium-ion batteries", *Journal of Electrochemical Science and Technology*, vol. 11, no. 1, pp. 1–13, 2020. doi:10.33961/jecst.2019.00528
- [45] J. Q. Liu *et al.*, "Tertiary butyl hydroquinone as a novel additive for SEI film formation in lithium-ion batteries", *RSC Advances*, vol. 6, no. 49, pp. 42885–42891, 2016. doi:10.1039/C6RA04839K
- [46] N. Meddings *et al.*, "Application of electrochemical impedance spectroscopy to commercial Li-ion cells: A review", *Journal of Power Sources*, vol. 480, p. 228742, 2020. doi:10.1016/j.jpowsour.2020.228742
- [47] E. López-Honorato, P. J. Meadows, and P. Xiao, "Fluidized bed chemical vapor deposition of pyrolytic carbon – I. effect of deposition conditions on microstructure", *Carbon*, vol. 47, no. 2, pp. 396–410, 2009. doi:10.1016/j.carbon.2008.10.023
- [48] N. A. Laziz *et al.*, "Li- and na-ion storage performance of natural graphite via simple flotation process", *Journal of Electrochemical Science and Technology*, vol. 9, no. 4, pp. 320–329, 2018. doi:10.33961/jecst.2018.9.4.320
- [49] S. W. Han, D. W. Jung, J. H. Jeong, and E. S. Oh, "Effect of pyrolysis temperature on carbon obtained from green tea biomass for superior lithium ion battery anodes", *Chemical Engineering Journal*, vol. 254, pp. 597–604, 2014. doi:10.1016/j.cej.2014.06.021
- [50] M. Paul, "Synthesis and Evaluation of Battery Electrodes Using Pyrolytic Carbon from Methane Pyrolysis", pp. 27–28, 2023.
- [51] B. Rezaei, J. Y. Pan, C. Gundlach, and S. S. Keller, "Highly structured 3D pyrolytic carbon electrodes derived from additive manufacturing technology", *Materials & Design*, vol. 193, 2020. doi:10.1016/j.matdes.2020.108834
- [52] S. Rodrigues, N. S. A. K. Munichandraiah, and A. K. Shukla, "AC impedance and state-of-charge analysis of a sealed lithium-ion rechargeable battery", *Journal of Solid State Electrochemistry*, vol. 3, pp. 397–405, 1999. doi:10.1007/s1000800050173
- [53] C. Mao *et al.*, "Selecting the best graphite for long-life, high-energy Li-ion batteries", *Journal of The Electrochemical Society*, vol. 165, no. 9, pp. A1837, 2018. doi:10.1149/2.1111809jes
- [54] J. B. Cook *et al.*, "Tuning porosity and surface area in mesoporous silicon for application in Li-ion battery electrodes", *ACS Applied Materials & Interfaces*, vol. 9, no. 22, pp. 19063–19073, 2017. doi:10.1021/acsami.6b16447
- [55] R. Yuan *et al.*, "Structural transformation of porous and disordered carbon during ball-milling," *Chemical Engineering Journal*, vol. 454, p. 140418, 2023. doi:10.1016/j.cej.2022.140418
- [56] T. Yoshida *et al.*, "Degradation mechanism and life prediction of lithium-ion batteries", *Journal of The Electrochemical Society*, vol. 153, no. 3, pp. A576, 1999. doi:10.1149/1.2162467

Using Agent-Based Modelling to Investigate Effects of the Socioeconomic Climate on the UK Power Sector

Le Bouillier, Noah and Berry, Joseph

Abstract

The global commitment to combat climate change has been fortified by individual nations' pledges to reduce carbon emissions. The UK's ambitious target to achieve carbon neutrality by 2050 stands as a testament to this endeavour. Integral to this vision is the success of renewable energy sources, among which wind power has been identified as a key player in the UK's energy matrix. However, the intricacies of establishing a robust wind energy infrastructure are vast, encompassing not just technological challenges, but also economic, social, and political dimensions. This project used an agent-based model to investigate how varying certain prices associated with different technologies would affect the future landscape of the UK power section. Techno-economic data was sourced to accurately model an array of technology options, with an LCOE (levelized cost of energy) agent in place to make investment decisions to fulfil demand. The findings of this project indicated an expected dependence on wind, both onshore and offshore, and solar generation to shift to a carbon-neutral power sector. Capital costs of wind were found to be very influential, with small increases in cost resulting in a significant drop in wind investment. This provided a useful insight into the future of wind power.

1. Introduction

The energy landscape of the United Kingdom is currently undergoing an unprecedented change, propelled by the ever-changing socio-economic climate of the world. This is greatly affecting the development of the nation alongside technological advancements shown by the share of renewable energy increasing from 2% to 43% over the last 30 years (**National Grid, 2023**). Currently, the UK aims to reach a net-zero carbon footprint by 2050 in its aim to reach a sustainable and resilient future. This means that it is critical to understand the relationships between stakeholders and their objectives within the power industry. This study explores this complexity by using an integrated agent-based model, ModUlar energy system Simulation Environment (MUSE), to reveal the intricate connection between the socio-economic environment and the UK power industry.

The socio-economic climate, defined by societal behaviour, economic trends and governmental policies, greatly affects the path of the power sector. Agent-based modelling (ABM) provides

more detailed insight into this by using agents to replicate the behaviour of individual investors in a dynamic environment where standard models can often oversimplify these influences. In their interactions with one another; agents, representing various stakeholders, ranging from governing bodies to individual producers, respond to simulated objectives. This gives ABMs the ability to depict the power industry's non-linear, adaptive reaction to outside stimuli more precisely.

Pledging to reach net-zero carbon emissions by 2050 (**Gov.uk, 2019**) and a net-zero power industry by 2035 (**Gov.uk, 2021**) has dramatically accelerated the energy transition in the UK. As of 2022, the carbon dioxide emissions from the power sector were 85 million tonnes, a decrease of more than 50% since 2000 (**Tiseo, 2023**). This means that in the next 13 years, a similar number of emissions need to be reduced as the previous 22. Considering that the emissions have decreased by less than 2% over the last 2 years (**Tiseo, 2023**), the enormity of this task cannot be overlooked. As shown in Figure 1, approximately 43GW of the UK capacity is generated by fossil fuels. By 2035 this all must be

phased out or counterbalanced with carbon-negative fuel sources such as sustainable biomass

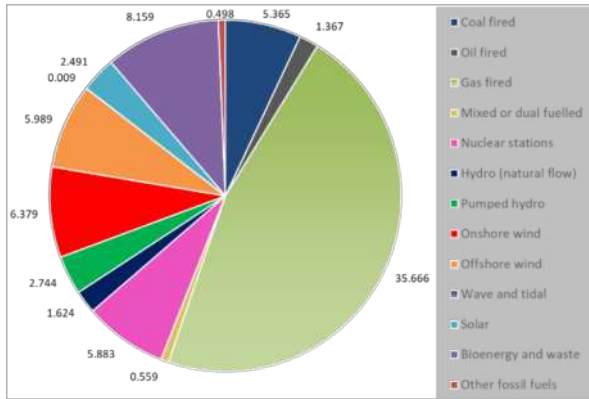


Figure 1: Pie Chart representing 2022 capacity share per fuel. (Department for Energy Security and Net Zero, 2023)

combustion fitted with carbon capture.

This project aims to investigate the effects different socio-economic circumstances, such as supply chain problems, could have on the stability, resilience and sustainability of the UK power sector. Through the integration of agent behaviours and real-world data, the model aims to reproduce the investment decisions within the sector. This entails how energy providers diversify their portfolios of energy generation and how governmental interventions and reforms affect the distribution of resulting capacity.

2. Background

The modelling of UK energy sectors is by no means a new way to forecast the feasibility of government targets with optimisation programmes such as the UK TIMES model (UKTM) (Broad 2017) being used by the government to aid in the clean growth strategy of 2017. The UKTM itself was a successor to UK MARKAL (Dodds et al., 2003) which was developed to give insights for the Energy White Paper 2003 and continued to be developed and used until 2012.

ModUlar energy system Simulation Environment (MUSE) is an agent-based simulation model which is a more novel approach to this field. This model has been used in looking at energy investments in the

residential sector of the UK (Sachs, et al., 2019) as well as across the board in other countries. MUSE is unique in the way that has an agent-based approach, representing possible decisions made by real investors with the ability to be adjusted depending on each investor type. These small motivations by individuals could have large effects on the overall system and by using them. The objective of MUSE is to reach an economic equilibrium by finding a price-quantity tension between supply and demand. This leads to unique solutions to otherwise near-impossible tasks counterbalancing different individuals' aims as well as cost reduction. The ability to individually characterise each technology aids in this approach as all technologies have unique challenges and restrictions.

3. Methods

Utilising a market-clearing algorithm, MUSE depends on multiple technical and economic parameters. The majority of this data was obtained from either the UK TIMES model or government reports. Different scenarios were investigated using MUSE setting various objectives as well as introducing carbon budgets, all with the overarching aim of being used for investment decisions by energy companies. A few parameters, therefore, were set with the aim to be non-restrictive on the rest of the model. This was done mainly because, as previously mentioned, this paper aimed to show an insight to investors in the industry so restrictive scenarios would not be as useful.

3.1 Technology Selection

A technology is defined as a technology that services a service demand. The selection of technologies was divided into two distinct categories: existing and future. The existing technologies were simple to obtain as they were all power-producing technologies that are used to provide electricity in the UK. Future technologies were taken from the UK TIMES model.

3.2 Techno-economic parameters:

As mentioned, MUSE uses individual data for each technology to determine the uptake of each according to their capacity throughout time. A selection of these parameters is provided here with an explanation as to what data was used where necessary.

3.2.1 Technical Parameters

Existing Capacity

The existing capacity for each technology was taken from Plant capacity: United Kingdom (DUKES 5.7) (Department for Energy Security and Net Zero, 2023). This provides a base on which for MUSE to build future decisions.

Decommissioning profile

The decommissioning profile of existing technologies was carried out linearly through the existing capacity input file. This decommissioning profile was based on the technical life of the technologies, as taken from the UK TIMES model.

Utilisation factor

The Utilisation factor is the achievable output from the installed capacity of each technology. This is used as most technologies have downtime for maintenance and repairs as well as certain renewable power sources only running given certain conditions. For most technologies, this was taken to be 0.9. For technologies affected by environmental factors i.e. solar and wind, utilisation factors were calculated through an average of availability factors sourced from the UK TIMES model.

Carbon Emissions Factor

The Emission factor was used based on the mass of carbon dioxide released per petajoule of electricity produced. For each technology this was found using data taken from Greenhouse gas reporting: conversion factors 2023 (Department for Energy

Security and Net Zero, 2023). For combustion technologies fitted with carbon capture and storage, this number was reduced by 90%. Biomass combustion has a net zero lifetime emissions factor as in the UK it is produced sustainably and therefore, when fitted with CCS, it has a negative emission factor.

Electricity Demand projection

The demand for UK electricity was found by using data taken from the UK TIMES model. This data was required in ten-year intervals from the base year of 2020 up to 2050 for MUSE to accurately project the necessary total capacity.

Growth constraints

Total Capacity, Max capacity growth and addition. Growth constraints were tailored to simulate realistic growth without restricting capacity addition such that demand could not be filled. However, some technologies with specific policy and/or physical restrictions in growth available in literature were constrained to preserve realism.

Efficiency

The efficiency of each technology was defined as the percentage of usable output electricity of the total energy in the fuel. For renewables, this was taken to be one as there is no fuel associated with each process.

3.2.2 Economic parameters

Capital and Operational Expenditures

The Capital Expenditure (CapEx) for new technologies were taken from the UK TIMES model. For existing technologies, this was decided to be slightly higher than their respective new technologies to prevent investment in existing, implemented technologies. Fixed and variable parameters for Operational Expenditure (OpEx) were similarly taken from the UK TIMES model.

Fuel Price

Fuel prices were taken from government documentation (DESNZ, 2023). Projections were carried out through moving averages where necessary and specific prices were altered to explore scenarios such as the price of gas.

Carbon Price

Carbon price was controlled through the projections input file, pricing the CO2f commodity accordingly. The UK government currently sets carbon price at £40/kWh thus the initial run was carried out on this basis.

3.3 Agent Characterisation

The population does not have a huge say on what occurs in the power industry, only slightly governing through the ability to advocate for policy reforms and introduction. Therefore, it was decided that for the power sector, Agents would have to be representative of government and energy providers. Stakeholders in the industry itself generally have the same goal of minimising the costs associated with electricity production, thus a levelized cost agent was introduced to represent the industry. The government's only ability to interact with the industry is through mandated policies surrounding decreasing emissions. However, decision making on investment is carried out principally on cost, therefore the Agent objective was defined by LCOE (Levelised Cost Of Energy). The equation to calculate LCOE is:

$$LCOE = \frac{\sum Total Lifetime Costs}{Total Lifetime Energy Produced}$$

Where:

$$\begin{aligned} \sum Costs over Lifetime \\ &= CapEx + Fixed OpEx \\ &+ Variable OpEx + Fuel Cost \\ &+ Carbon Emission Cost \end{aligned}$$

And:

$$\begin{aligned} Total Lifetime Energy Produced \\ &= Technical Life * Capacity \\ &* Utilisation Factor \end{aligned}$$

3.4 Scenario Development

Three scenarios were developed based on economic incentives alongside UK policy. These scenarios decided which agents would be prioritised which was further described in the agent characterisation section.

3.4.1 Scenario 1: Varying Carbon pricing

Purpose of this scenario was to explore the effects of changing carbon pricing on the decision-making within investments into future power-generation technologies.

3.4.2 Scenario 2: Return of gas prices to pre-2022 levels

This scenario investigates the possible return of gas pricing to pre-2022 rates and the subsequent changes in investment with regards to renewable technologies over non-renewable gas-based power generation.

3.4.3 Scenario 3: Effect of wind implementation price

Scenario 3 was divided into two stages, the first being investigating how increasing the CapEx of offshore wind affects its feasibility. The second stage studied the capacity distribution of the sector if wind was no longer feasible until certain points in time.

3.4.3.1 Stage 1

MUSE uses an input parameter named cap_par to determine the CapEx of a technology with the relationship: $CapEx = cap_par * Capacity$. Adjusting cap_par would have the same proportional effect on CapEx meaning that it would be possible to determine a percentage increase in CapEx that would make it infeasible.

3.4.3.2 Stage 2

At this point, it was necessary to investigate the effects of no new. This was accomplished by establishing the price of wind as a commodity at

£1,000,000M /PJ until the year in which new wind would be commissioned again. The pricing then reverted to zero five years later. This approach aligns with MUSE's assumption of an instantaneous building process, despite the actual transition from commissioning to operational status taking approximately five years. Importantly, this strategy to curtail wind growth did not impact already operational wind capacities, as their capacity profiles were updated every 5 years, incorporating considerations for decommissioning.

4. Results & Discussion

4.1 Scenario 1

4.1.1 Business-as-usual

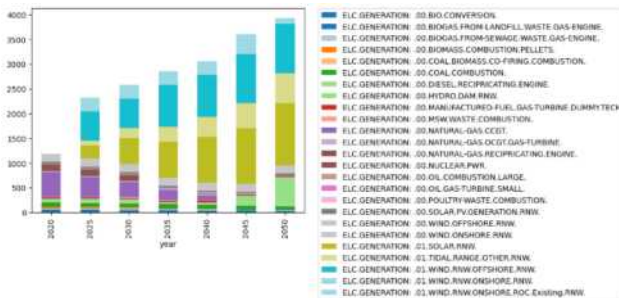


Figure 2: Capacity distribution in PJ over time with a carbon price of £40 /kt

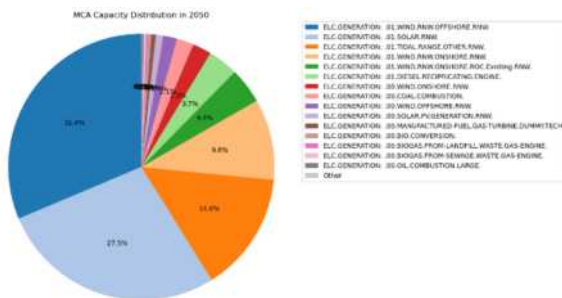


Figure 3: Capacity distribution in PJ over time with a carbon price of £40 /kt

For scenario 1, Carbon pricing was initially set to the current rate of £40 /kWh, resulting in the following capacity distribution. Investment into both onshore and offshore wind is evident from 2025 onwards resulting in a capacity share dominated by both wind technologies and a combination of solar and

tidal power. This is a positive result in line with Government policy and suggests that current Carbon pricing provides a sufficient incentive to invest towards greener and thus more sustainable alternatives to traditional carbon-based generation. This also doubles as a base case for this project as it implies a business-as-usual approach, wherein current policy and constraints are in place.

4.1.2 Increasing carbon price

Following this, the effects of increasing carbon price were investigated by raising projections linearly from current rates (£40 /kWh) in 2020, to

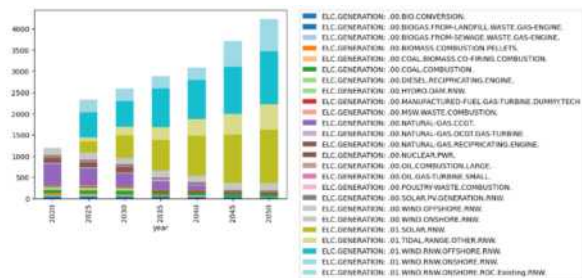


Figure 4: Capacity distribution in PJ over time with a carbon price increasing linearly to £180 /kt

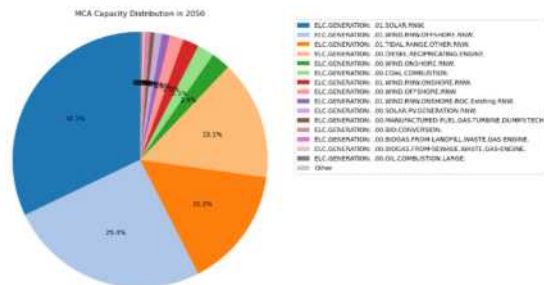


Figure 5: Capacity distribution in 2050 with a carbon price increasing linearly to £180 /kt

£180 /kWh by 2050. The results of this demonstrate an almost identical trend, with investment into wind and solar at the same rates. This suggests that current carbon prices are sufficient in achieving net-zero and raising the prices would be negligible. However, further investigation into the time frame in which prices are altered could yield different results. For example, introducing the higher rate by

2035 may influence the rate of uptake of clean energy.

4.1.3 Zero carbon price

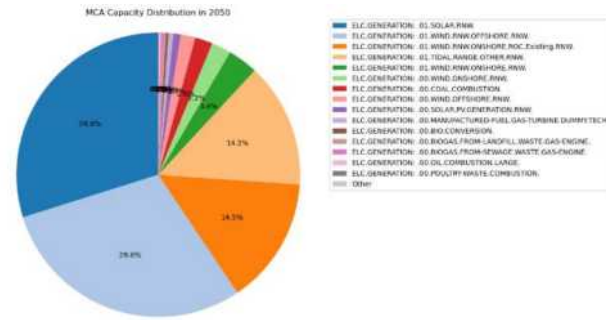


Figure 2: Capacity distribution in 2050 with a carbon price of £0 /kt

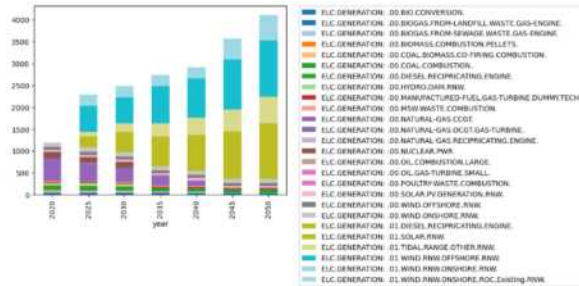


Figure 3: Capacity distribution in PJ over time with a carbon price of £0 /kt

Finally, the effect of carbon price was investigated through setting to £0 /kWh. This results in a capacity distribution in 2050 made up of around 19% combustion based. This presents an infeasible scenario as policy dictates a decrease in emissions and this would not align with aforementioned carbon targets. The unsurprising conclusion presents the importance of a carbon price when approaching policy and investment decisions.

4.2 Scenario 2

4.2.1 Gas price

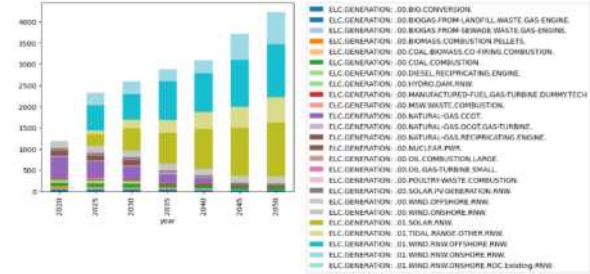


Figure 4: Capacity distribution in PJ over time with a gas price remaining at £13.3M /PJ

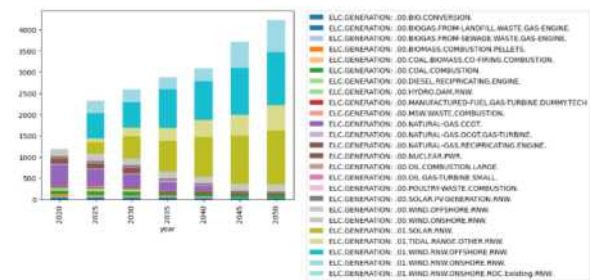


Figure 5: Capacity distribution in PJ over time with gas prices reverting back to pre-2022 rates

To investigate the effects of gas prices, two potential routes were explored. Due to the 'Russian War' gas prices have shot up to levels as high as £13.3M /PJ, due to current reliance on natural gas generation this greatly affected the UK household gas prices. Initially the effect of rates remaining at this high level was explored and then compared to the possible outcome of rates returning to pre-2022 levels. As evident in Figures 8 and 9, the outcome is identical. This suggests that existing cost implications of natural gas generation are sufficient in limiting investment and encouraging clean energy. Another contextual implication of this is that 'energy security' can be increased and prices can be insulated from future socio-political factors.

4.3 Scenario 3

4.3.1 Stage 1:

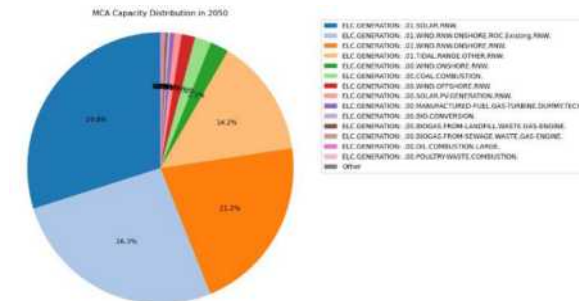


Figure 6: Capacity distribution in 2050 with offshore wind cap_par at £31.16M /PJ

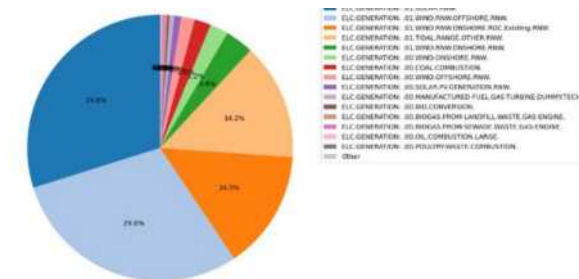


Figure 11: Capacity distribution in 2050 with offshore wind cap_par at £31.155M /PJ

In this stage, offshore wind projects' capital cost parameter (cap_par) was continuously adjusted until it ceased to appear in the capacity distribution. According to Figures 10 and 11, this critical point was found to be £31.16 million per petajoule (£31.16M /PJ). It is evident that offshore wind projects are observable at £31.155M /PJ but absent at £31.16M /PJ. Interestingly, this was just a 3.1% increase from the initial cap_par value of £30.23M /PJ, highlighting the fact that offshore wind projects are quickly getting closer to becoming economically infeasible.

4.3.2 Stage 2:

Five Situations were developed at this point: Business as usual, no new wind to be commissioned until 2025 and 2035, no new wind and no new wind or solar being commissioned for the foreseeable future.

4.3.2.1 Business as usual:

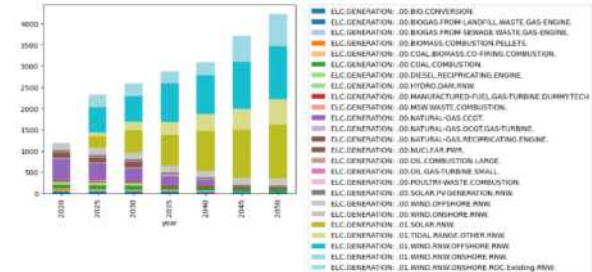


Figure 12: Capacity distribution in PJ over time

Figure 12 depicts wind energy's crucial significance in the context of sustainable energy, with a contribution of approximately 50% of total capacity forecasted by 2045. Notably, offshore wind is expected to account for around two-thirds of this significant share, highlighting its importance in the expanding landscape of renewable energy sources.

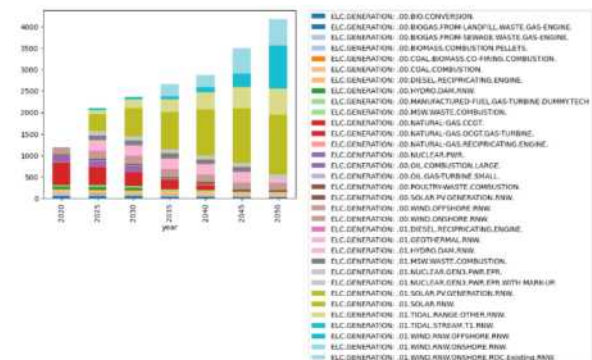


Figure 7: Capacity distribution over time with no new wind commissioned until 2025

4.3.2.2 No new wind commissioned until 2025:

As seen in Figure 13, hydro dams will make up for the capacity shortfall caused by wind in 2025 and 2030. However, starting from 2030, wind emerges as the most advantageous option, exhibiting the largest year-on-year growth in each incremental step up to 2050. Nuclear sources and the burning of MSW can meet additional capacity requirements in 2025, albeit in small amounts. Though technically possible, this scenario highlights systemic flaws, especially as there is less chance that the

government will commission several hydro dams due to popular opposition stemming from land inundation concerns.

4.3.2.3 No Wind commissioned until 2035:

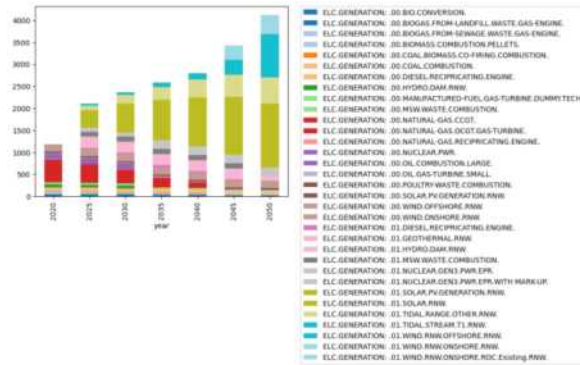


Figure 8: Capacity distribution in PJ over time with no new wind commissioned until 2035

In the absence of new wind operations until 2040, the ensuing portfolio shown in Figure 14 remarkably mirrors that of 2030. This similarity results from a relatively little rise in energy consumption between 2030 and 2040, or slightly more than one-third of the growth that was noted between 2020 and 2030. As such, alternative technologies are not overly burdened by the increased capacity gap arising from the lack of wind. Nevertheless, solar energy poses a potential obstacle because it will account for 45% of energy production by 2050 and is not feasible without improvements in energy storage technologies due to its dependence on daylight to function.

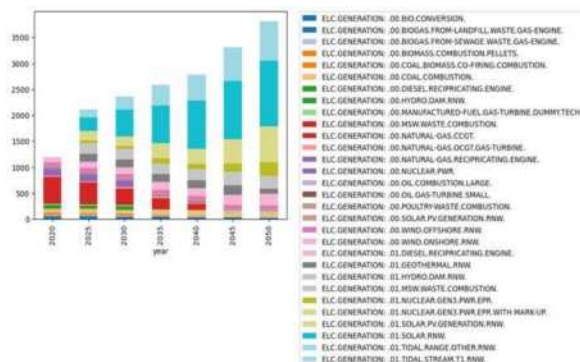


Figure 9: Capacity Distribution in PJ over time with no new wind commissioned

4.3.2.4 No new wind commissioned:

This capacity portfolio's feasibility is further called into question as demonstrated by the continued use of combustion technologies shown in Figure 15, which will account for about 25% of capacity by 2035 and 8% by 2050. Notably, tidal electricity generation assumes a substantial role in mitigating the wind deficit, underscoring its potential as a significant contributor in the extended course of energy generation.

4.3.2.5 No wind or solar:

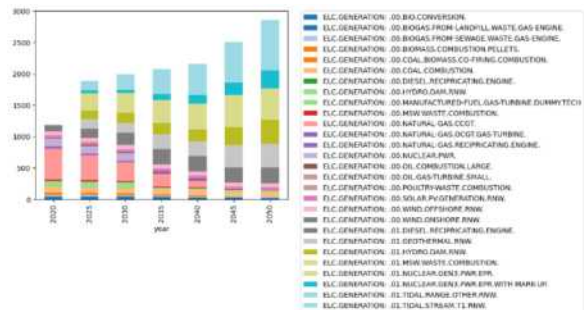


Figure 10: Capacity distribution in PJ over time with no new wind or solar commissioned

This hypothetical situation was purposefully created with the knowledge that it is not feasible, yet it can be used as a learning tool to determine which electricity sources are worth investing in to save costs. Because alternative renewables are continuously operational regardless of weather, they have the potential to outperform wind and solar in terms of economic feasibility, provided they receive sufficient financing and development. As anticipated from previous results, tidal power exhibits the highest capacity, closely followed by nuclear generation, with geothermal energy playing a notably substantive role in this prospective energy landscape.

5. Conclusion

Significant ideas can be drawn from the results of this project. Looking at the variation of carbon pricing, having a price for carbon emissions is a necessity, however increasing it further may not be

needed to incentivise energy producers any further to focus more heavily on renewable sources. Increasing it may have an adverse effect on the economy due to the 'tax' effectively being paid by the population's energy bills and since all businesses use electricity, this could have a much larger effect on the cost of living.

Reverting gas prices to what they were before 2022 somewhat surprisingly did not influence the future power landscape. This is a welcoming sign that renewable sources can provide the UK with energy security if there is another unforeseeable occurrence that increases fuel prices. This in turn will give the possibility to have set electricity costs with no major fluctuations.

Scenario three displayed that the future power landscape relies heavily on offshore wind. Unfortunately, the CapEx of offshore wind cannot grow by more than 3% compared to other technologies before it becomes unfeasible.

In addition to these findings, it is critical to recognize that the current supply chain concerns (**Hodgson, 2023**) may increase this thin margin of infeasibility. Furthermore, the recent COP28 resolution to triple the proportion of renewable energy by 2030 (**Mcfarlane & Twidale, 2023**) adds another layer of complexity, potentially putting further strain on the global wind energy supply chain. As a result, the increased demand for renewable energy sources may contribute to an increase in CapEx in the offshore wind sector.

6. Outlook

This project is part of a growing body of work that uses ABM to evaluate complicated systems, thus it doesn't work on its own. It adds to the larger conversation on the role of ABM in realizing and reducing difficulties related to large-scale socio-technical systems by applying this methodology to the UK power industry. The hope is that in conducting this project, the results will act as a model for similar studies in other areas struggling

with the difficult path to a sustainable energy transition.

Looking into the future of this model, integrating it into an entire UK energy sector model with the already produced residential sector as well as new transportation and industrial sectors could give a holistic view of how the whole system reacts to various changes in the socioeconomic climate. This in turn could provide valuable insight into the intricacies and possible flaws in the sector.

The addition of other agents into this model to represent different stakeholders and their objectives could give further sensitivity analysis of the power sector. For example, even though the general population doesn't control the sector, people do prefer certain technologies over others and companies do want to have a positive perception by the public leading them to possibly listen.

Finally, using this model with different output parameters such as price or carbon dioxide emissions to show feasibility rather than just capacity could give a better understanding and detail into which scenarios are more favourable than others. This was an aim of this project but unfortunately, due to time constraints, it was not a possibility.

7. Acknowledgements

The authors would like to finish by acknowledging Martin J Stringer for his assistance in making this project a possibility.

References

Broad, O., 2017. *UK TIMES*. [Online]
Available at: <https://www.ucl.ac.uk/energy-models/models/uk-times>
[Accessed 20 10 2023].

Department for Energy Security and Net Zero, 2023. *Greenhouse gas reporting: conversion factors 2023*. [Online]
Available at:
<https://www.gov.uk/government/publications/greenhouse-gas-reporting-conversion-factors-2023>
[Accessed 15 10 2023].

Department for Energy Security and Net Zero, 2023. *Plant capacity: United Kingdom (DUKES 5.7)*. [Online]
Available at:
https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Fassets.publishing.service.gov.uk%2Fgovernment%2Fuploads%2Fsystem%2Fuploads%2Fattachment_data%2Ffile%2F1174123%2FDUKES_5.7.xlsx&wdOrigin=BROWSELINK
[Accessed 20 10 2023].

DESNZ, 2023. *Energy Prices*. s.l.: Department for Energy Security and Net Zero.

Dodds et al., 2003. *UK MARKAL*. [Online]
Available at: <https://www.ucl.ac.uk/energy-models/models/uk-markal>
[Accessed 20 10 2023].

Gov.uk, 2019. *UK becomes first major economy to pass net zero emissions law*. [Online]
Available at:
<https://www.gov.uk/government/news/uk-becomes-first-major-economy-to-pass-net-zero-emissions-law>
[Accessed 20 10 2023].

Gov.uk, 2021. *Plans unveiled to decarbonise UK power system by 2035*. [Online]
Available at:
<https://www.gov.uk/government/news/plans-unveiled-to-decarbonise-uk-power-system-by-2035>
[Accessed 20 10 2023].

Hodgson, C., 2023. *Wind sector faces supply chain crunch this decade, industry body warns*. [Online]
Available at:
<https://www.ft.com/content/f324be0d-191e-4943-97fd-51a8d46e286c>
[Accessed 12 10 2023].

National Grid, 2023. *National Grid Energy Explained*. [Online]
Available at:
<https://www.nationalgrid.com/stories/energy-explained/how-much-uks-energy-renewable>
[Accessed 15 10 2023].

Sachs, J., Meng, Y., Giarola, S. & Hawkes, A., 2019. *An agent-based model for energy investment decisions in the residential sector,,* s.l.: Science Direct.

Tiseo, I., 2023. *Power sector emissions in the United Kingdom from 2000 to 2022, by source*. [Online]
Available at:
<https://www.statista.com/statistics/1405567/power-sector-emissions-united-kingdom-by-source/#:~:text=The%20United%20Kingdom%27s%20power%20sector%20emissions%20totalled%2085,gas%20accounted%20for%20almost%20three-quarters%20of%20this%20total.>
[Accessed 20 10 2023].

Direct Visualisation of Surfactant Flooding in Micromodels

Oluwanifemi Emmanuel Adejumobi and Tazz Bennett-Gant

Imperial College London, Department of Chemical Engineering, London, UK

Abstract – Hypothesis: Crude oil trapped in porous rocks can be emulsified and extracted via surfactant flooding based enhanced oil recovery (EOR). In-situ emulsification is reported to increase the efficacy of the EOR process, but the mechanism behind this phenomenon is not yet fully understood. This study aimed to provide a better comprehension of these mechanisms and their effect on oil recovery through the use of glass microfluidic models (transparent pore networks which allow the direct visualization of fluid flow) and microscopy imaging, a capability beyond the reach of conventional imaging techniques like rock core-flood visualisation using CT scans.

Experiments: This research work employed both homogeneous and heterogeneous microfluidic models to simulate different types of underground oil reservoir pore networks. Two surfactants, ALFOTERRA L-145-10s 90 (Surfactant 1) and Sodium dodecyl-benzenesulfonate (Surfactant 2), were used in optimally formulated brine solutions to flood oil-saturated micromodels. Throughout the experiments, pore-scale snapshots were captured to visually document and analyse how these surfactants facilitated oil emulsification and oil recovery.

Findings: The study confirmed that both surfactants, through different mechanisms (W/O emulsions for ALFOTERRA L-145-10s 90 and O/W emulsions for Sodium dodecyl-benzenesulfonate), contributed significantly to oil recovery. It was also observed that micromodel geometry influenced emulsion size in the case of Sodium dodecyl-benzenesulfonate. Additionally, both surfactants demonstrated enhanced oil recovery effectiveness across both geometrical setups. These findings underscore the importance of surfactant selection and micromodel geometry in optimizing oil recovery processes. The direct visualization offered by micromodels provides a unique insight into the interaction between surfactants and trapped oil, thereby informing more efficient and economically viable recovery strategies.

Keywords – Enhanced oil recovery, surfactant flooding, microfluidics, interfacial tension, emulsions, pore-scale visualization, microscopy

1. Introduction and Background

The global demand for energy and fuels relies significantly on crude oil, constituting 34.21% of the total primary energy supply in 2017 [1]. Crude oil is extracted from geological oil reservoirs located across the globe and is industrially performed through recovery methods of increasing complexity; denominated as primary, secondary and tertiary methods.

Primary recovery methods utilize natural drives such as reservoir pressure and gravity, complemented by artificial lift from pumps, to extract oil to the surface at production wells. Secondary recovery involves fluid injection,

such as waterfloods, to raise and maintain reservoir pressure and also displace oil towards production wells. Tertiary recovery methods are used after secondary recoveries and include Enhanced Oil Recovery (EOR).

Despite these considerable extraction efforts, primary and secondary recovery methods applied in sequence only achieve oil recoveries of around 35% [2]. With two-thirds of the original reservoir oil still unobtained, the primary challenge of extraction owes to strong capillary forces which trap oil in the porous reservoir sediment. These pore-scale capillary effects can be described by the capillary number:

$$Ca = \frac{v\mu}{\gamma} \quad (1)$$

This is the ratio of viscous forces to capillary forces during an immiscible displacement, where v is the interstitial velocity of the injected fluid, μ is the injected fluid viscosity and γ is the interfacial tension (IFT) between the two immiscible fluids. Higher viscous forces relative to capillary forces lead to greater deformation and displacement of the trapped oleic phase within the pores, increasing capillary number and thus oil recovery.

While it's theoretically feasible to increase the interstitial velocity of the injected fluid to achieve this, practically, such an increase is limited, especially after secondary recoveries. From an engineering and economic standpoint, it's often not viable to increase the viscous forces by multiple orders of magnitude. Conversely, a more feasible and impactful approach involves reducing the IFT between the immiscible phases. By decreasing the IFT, usually by 2-3 orders of magnitude, a significant change in the capillary number can be achieved, thereby enhancing oil recovery more efficiently and economically. [3].

Tertiary oil recovery methods, specifically chemically enhanced oil recovery (cEOR) by the utilisation of surfactants, have emerged as a key solution to recovering the immobilised oil following initial extraction methods. Utilising the interfacial tension (IFT) altering properties of the surfactant molecules, the IFT between the oil and the injected fluid can be considerably decreased during surfactant flood injection. Consequently, capillary forces are reduced, increasing the systems capillary number. Correspondingly trapped oil ganglia within the pores are more easily liberated by the viscous forces of the flood injection, increasing oil recovery. As such, surfactant-based flooding is recognized for its cost-effectiveness in achieving substantially high ultimate oil recoveries of 50%-70% [3].

Additionally, reservoir rock wettability alteration is recognized as another crucial mechanism to liberating oil as modifying oil-

wet reservoirs into water-wet conditions during a surfactant flood drastically reduces resistance to flow of the wetted oil thus increasing recovery [4].

However, despite these advantages, surfactant flooding faces significant drawbacks in the industry. Primarily, surfactants are expensive, which introduces a higher financial risk when implementing these methods. This cost factor necessitates a deeper understanding of the process to ensure efficient application and optimization of resources. To address this, direct visualization becomes imperative. Other imaging techniques such as X-ray, can't directly observe the phases, and their properties are often inferred from other parameters like the CT number. In contrast, micromodels offer unique insights by allowing direct observation of these intricate processes.

Aside from the widely recognised oil recovery mechanisms mentioned already, existing surfactant flood studies have reported the presence of a macroemulsion in the extracted oil product with droplets as small as 10 μm in diameter [3]. Micromodel flooding experiments have demonstrated that these O/W and W/O emulsions occur in-situ of the porous structure and demonstrate how they increase sweep efficiency and local oil recoveries within the micromodel by 14-30% compared to waterflooding [4].

Even more intriguing is the possibility of microemulsions forming in-situ of porous media [5] - emulsions which are thermodynamically stable in comparison to macroemulsions, and do not coalesce after long periods of time. Micromodels enable the direct visualisation of the mechanisms which facilitate the formation of these microemulsions. This provides valuable insights into the behaviour of surfactants within the reservoir and bridges the gap in understanding the complex dynamics at play.

This project, recognizing the great economic interest in surfactant flooding, aimed to directly observe, document, and quantify these processes' contribution to enhanced oil recovery. By utilizing microfluidic

approaches, the study provided direct visualization of pore-scale behaviours and displacement processes.

Central to this research was the construction of a viable lab-scale experimental system, designed to observe and confirm whether macro-emulsions and micro-emulsions form in-situ or arise through alternative mechanisms.

The project also aimed to explore the impact of micromodel geometry on oil recovery. Homogeneous micromodels (uniform pore structures) and heterogeneous micromodels (mimicking complex, non-uniform rock structures) were employed to understand how different pore structures influence the formation and behaviour of emulsions, and other mechanisms of oil recovery.

Additionally, the study evaluated the effects of two different surfactants on oil liberation mechanisms and overall oil recovery. Key aspects of this investigation included the emulsion size and types formed, and how these properties varied with different micromodel geometries.

By comparing and contrasting these outcomes across the range of micromodel geometries and surfactant types, the research aimed to provide deeper insights into the nuanced interplay of surfactant properties, micromodel geometries, and their collective impact on enhancing oil recovery.

2. Methodology

2.1. Experimental Methods

2.1.1. Materials

Carbon dioxide (housed in an Industrial-grade compressed gas cylinder, $\geq 99\%$) was purchased from BOC Ltd., Woking, United Kingdom.

Mineral Oil ('light') and Decane ($\geq 95\%$) were purchased from Sigma Aldrich, United Kingdom and used without further treatment.

ALFOTERRA L-145-10s 90 (Surfactant 1) (90% active) was provided by Sasol Ltd., United Kingdom.

Sodium chloride (NaCl, $\geq 99.5\%$), Sodium dodecyl-benzenesulfonate (Surfactant 2) (SDBS, Technical-grade) with co-solvent Isobutanol ($\geq 99.5\%$), Lissamine Green B powder dye (60%) and Sudan II powder dye (90%) were purchased from Sigma Aldrich, United Kingdom.

2.1.2. Sample Preparation

Two brine solutions were prepared by the addition of designated amounts of NaCl and Lissamine Green B dye to deionised water to make two dark-blue solutions at optimal salinities for their respective surfactant (3.7wt% NaCl and 3.5wt% NaCl for SDBS and L-145-10s 90 respectively).

SDBS solution was prepared by adding specific amounts of SDBS powder, isobutanol and Lissamine Green B dye to deionised water to make a dark-blue 3wt% SDBS, 5wt% isobutanol solution.

L-145-10s 90 powder and Lissamine Green B dye were added to deionised water to make a dark-blue 1wt% L-145-10s 90 solution. Sudan II dye was added to decane (used as purchased, without further treatment) to make a dark-orange decane sample.

All samples and solutions were mildly agitated upon preparation. The blue and orange dyes enabled easy differentiation between oleic and aqueous phases in images.

2.1.3. Experimental Setup

The experimental set-up consisted of two pumps (Teledyne ISCO 100D and Teledyne ISCO 260D, ISCO Inc., Nebraska, United States) containing mineral oil and decane respectively.

A micromodel/microfluidic chip (NJ1 (homogeneous) or EOR PR 20 2 (heterogeneous)) was housed in a chip holder (Fluidic Connect 4515) (all from Microunit BV, Enschede, Netherlands) and placed on the motorised stage (MAC 600, Ludl Electronic Products, Ltd., New York, United States) of an inverted light optical telescope (Axio Observer A.1m, Carl Zeiss AG, Oberkochen, Germany).

Connected to the microscope was a high-speed colour CCD camera (AxioCam HSc CCD, Carl Zeiss AG, Oberkochen, Germany) with which digital images and video recordings were taken for visual data collection. A stage micrometres glass slide (R1L3S1P, Thorlabs, New Jersey, United States) was used for image resolution calculations.

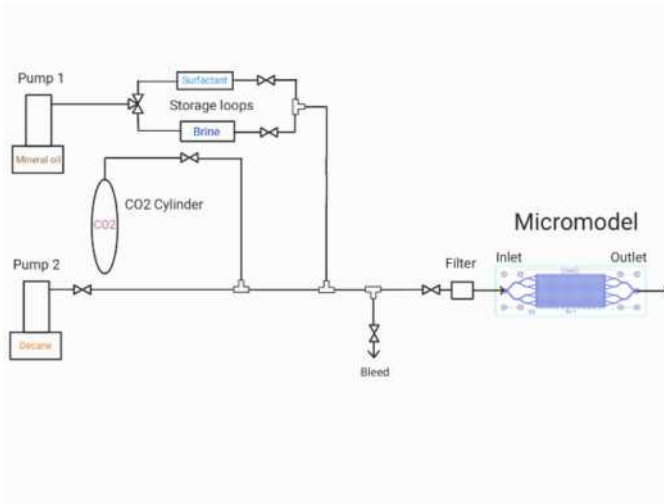


Figure 1: Schematic of microfluidic experimental set-up.

2.1.4. Experimental Procedure

Two surfactants and two micromodel geometries (homogeneous and heterogeneous) were investigated which gave rise to four experimental combinations. For each surfactant/micromodel combination, two styles of observation were conducted; a fixed camera set-up and a ‘free-roam’ camera set-up. The former enabled a quantitative analysis while the latter enabled a more qualitative, observational analysis.

A video recording was started, Pump 100D was filled with 20ml of mineral oil, Pump 260D was filled with 15ml of decane and the brine and surfactant coils were loaded. The pressurised CO₂ cylinder was opened, set to 4 bar and CO₂ was injected for 5 minutes to saturate the micromodel.

Brine solution was then injected at 0.05ml/min for 10 minutes after entry to dissolve CO₂ into aqueous phase. This was followed by decane injection at 0.01ml/min for 10 minutes after entry to fill the micromodel with an ample amount of oleic phase.

For the fixed camera experiments, an elementary area of the micromodel is selected (based on an RVEA analysis – **Section 2.2**) and an initial image of this area is taken.

Brine solution was injected into the micromodel for a second time. For the fixed-camera experiments, this injection rate was 0.005ml/min for 15 minutes after entry and a second image was taken after this time had elapsed. For the ‘free-roam’ experiments, injection rate was 0.001ml/min for 15 minutes. Surfactant solution was then injected into the micromodel at the same respective flowrate, also for 15 minutes.

For the fixed-camera experiments, images were taken in 30 second intervals until 15 minutes had elapsed. For the free-roam experiments, all zones of the micromodel were explored thoroughly during the surfactant flood to enable qualitative observations e.g. surfactant clearance mechanisms. The video recording was then stopped.

2.2. Analytical Methods

Ilastik, a machine-learning based image analysis tool, was used for image segmentation to convert raw microscope snapshots into two-colour segmented images.

MATLAB was utilised for porosity analysis, representative volume element analysis (RVEA), microscope image sharpening, oil recovery analysis and emulsion/droplet size distribution analysis.

3. Results and Discussion

3.1. Micromodel Characterisation

3.1.1. Normalised Representative Elementary Volume (REV) Analysis

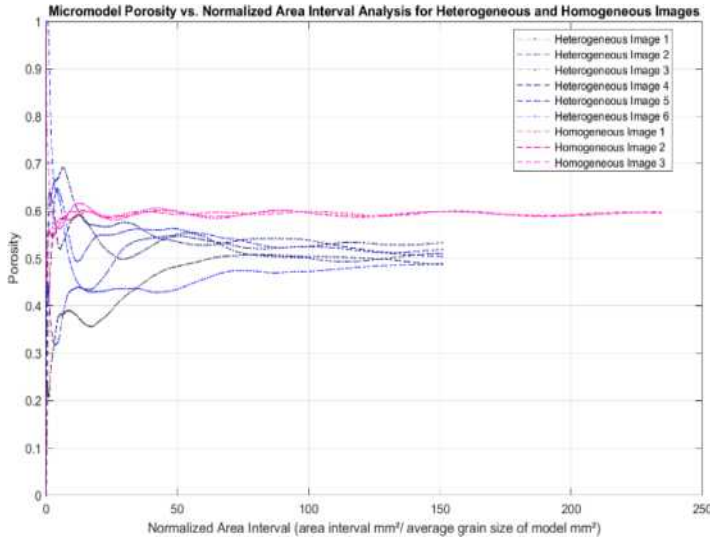


Figure 2: Representative Area Element Analysis for porosity of micromodel. Mean solid grain size for each respective geometry: Heterogeneous = 0.07757mm^2 and Homogeneous = 0.0499mm^2

The REV analysis was normalised across micromodel geometry to account for differences in solid grain size distribution in the heterogeneous micromodel compared to the homogeneous micromodel. In **Figure 2**, it is clear that the homogeneous plots equilibrate faster than the heterogeneous plots. This is because all homogeneous grains were identical in size whereas the heterogeneous grains were more randomly distributed.

These minimum equilibrated area values for each geometry were then multiplied by the respective average grain sizes to obtain the minimum elementary micromodel area which was sufficiently representative in porosity of the whole micromodel. This minimum representative micromodel area was determined to be approximately 11.6mm^2 and 7.5mm^2 for the heterogeneous and homogeneous micromodels respectively. As such, an area of 11.6mm^2 was employed to satisfy both geometries.

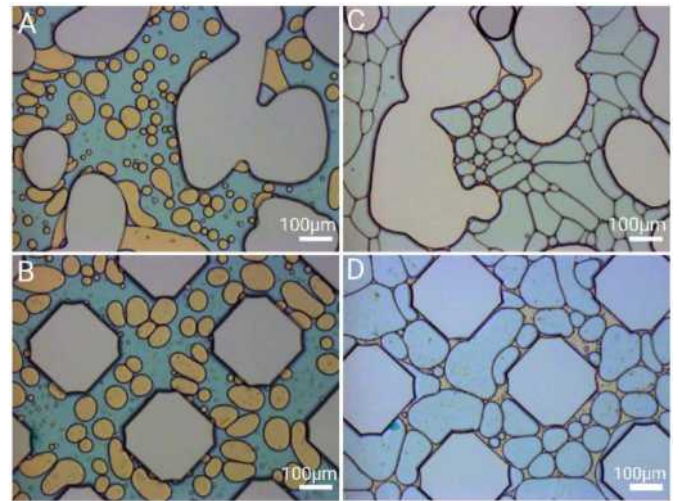


Figure 3: Photos taken during micromodel surfactant floods. Depicts emulsion types and sizes formed by surfactants 1 and 2 in different micromodel geometries: (A) is surfactant 2 in the heterogeneous micromodel, (B) is surfactant 2 in the homogeneous micromodel, (C) is surfactant 1 in the heterogeneous micromodel, (D) is surfactant 1 in the homogeneous micromodel. The scale bar is $100\mu\text{m}$.

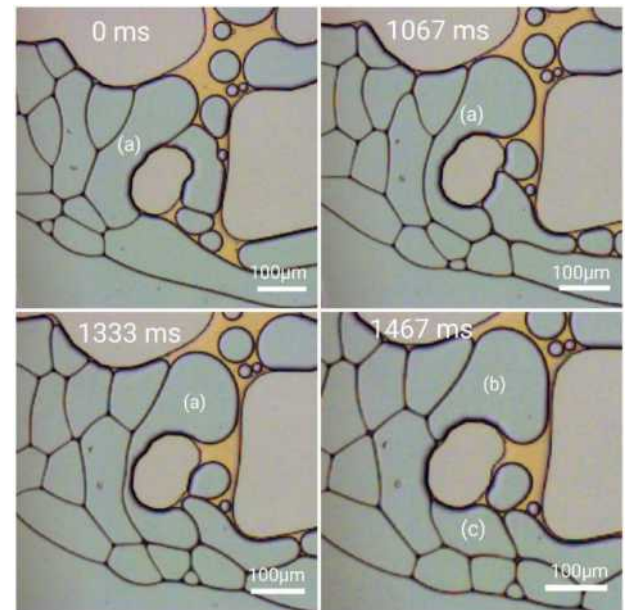


Figure 4: Photos taken during Surfactant 1 flood of heterogeneous micromodel. Depicts the splitting of droplet (a) into smaller blobs (b) and (c). Also depicts the displacement of bulk oil phase from pore. The scale bar is $100\mu\text{m}$.

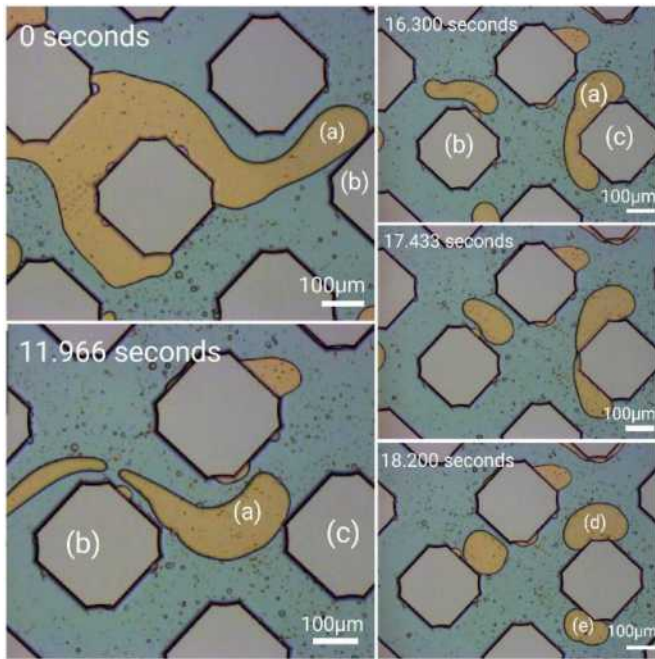


Figure 5: Timed snapshots depicting detachment of oil ganglion (a) from main oil ganglion body and further separation into smaller O/W emulsions (d) and (e) in the presence of Surfactant 2. The scale bar is 100µm.

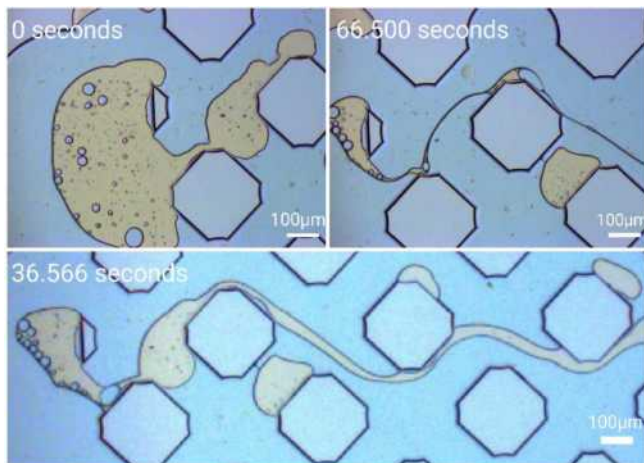


Figure 6: Timed snapshots depicting oil phase liberation via filament stretching, in the presence of Surfactant 1. The scale bar is 100µm.

3.2. Qualitative Analyses

3.2.1. Oil Recovery Mechanisms

As displayed in **Figure 3**, L-145-10s 90 (Surfactant 1) and SDBS (Surfactant 2) utilised different oil liberation mechanisms. Experiments involving Surfactant 1 were characterised mainly by two mechanisms. The first was the formation of water-in-oil (W/O) emulsions (**Figure 3 – C, D**) which displaced oil-occupied spaces, propagating oleic phase

along the main flow path. A magnified image of this mechanism is displayed in **Figure 4**. The second was the ‘stretching’ of oil ganglia into thin streaks which were pulled along pore throats in the micromodel **Figure 6**. In the homogeneous geometry, these observed streaks rarely broke into smaller droplets or segments; they are simply ‘snaked’ through the entire length of the micromodel. This phenomenon can be attributed in part to the ‘wetting’ effect of the surfactant which caused the micromodel inner wall surface to transition from ‘oil-wet’ to ‘water-wet’. This low-friction environment yielded an increased mobility of the oil streaks thus enabling easier flow.

This effect was also due to the lowering of interfacial tension (IFT) by the surfactant molecules resulting in a more thermodynamically stable system with a more fluid and flexible oil-water interface. The ‘stretching’ of these oil filaments generated new interfacial area which typically requires a large input of work energy;

$$W = \delta_i \times \Delta A \quad (2)$$

Where W is work energy, δ_i is the work per unit area required to form an interface and ΔA is the change in interface area. However, since this stretching of the oil ganglia was accompanied by adsorption of surfactant at the oil/water interface, the resulting interfacial tension was much lower therefore the required work energy per unit area was also lower. This yields a process more thermodynamically feasible than if it were to proceed in the absence of surfactant.

Surfactant 2 also initially conveyed oil in observed streaks as displayed in **Figure 5**, but these were quickly segmented into smaller oil ganglia or ‘snapped off’ at pore throats into oil-in-water (O/W) emulsion droplets (**Figure 3 – A, B**). In the heterogeneous micromodel, larger oil droplets often blocked pore throats in frequented paths thus forcing flow in a new unswept path, increasing sweep efficiency. These droplets and segments rarely coalesced as they flowed downstream implying they

were stabilised by SDBS surfactant. This is in agreement with observations made by Zhao et al. [4] who also employed SDBS surfactant in an EOR microfluidic set-up and reported similar results.

3.2.2. Emulsion Types

As stated in **Section 3.2.1.**, Surfactant 1 exhibited W/O emulsions while Surfactant 2 was characterised by O/W emulsions. The complex nature of surfactant, oil and water systems raises difficulties in predicting emulsion properties and behaviours. However, there exists some general principles to help predict/explain the occurrence of emulsion types in particular surfactant systems.

Bancroft [7] and Clowes [8] first recognised that for two-phase emulsion systems, the interface would curve to maximise the tension of the inner surface relative to the outer surface. Bancroft's Rule states - "The external phase of an emulsion system will be the phase in which the surfactant is the most stable." [9]

Winsor divided emulsion types into different classes; O/W emulsions being Type I forming at higher salinities whilst W/O emulsions were classed as Type II, forming at lower salinities [9].

Based on these principles and given that Surfactant 1 - being a sulfate-based surfactant - is less water-soluble than Surfactant 2 (sulfonate-based) [10][11], it can be reasoned as to why Surfactant 1 exhibited W/O emulsions while Surfactant 2 exhibited O/W emulsions.

Uncertainty exists, however, due to the fact that there are many more variables which govern emulsion system properties and behaviour such as temperature and oleic phase properties, but the principles introduced by Bancroft, Clowes and Winsor provide an excellent starting point to understanding and predicting emulsion formation.

3.2.3. Emulsion Size

From visual inspection of **Figure 3**, it is apparent that the generated emulsions not only differed in type but in size also.

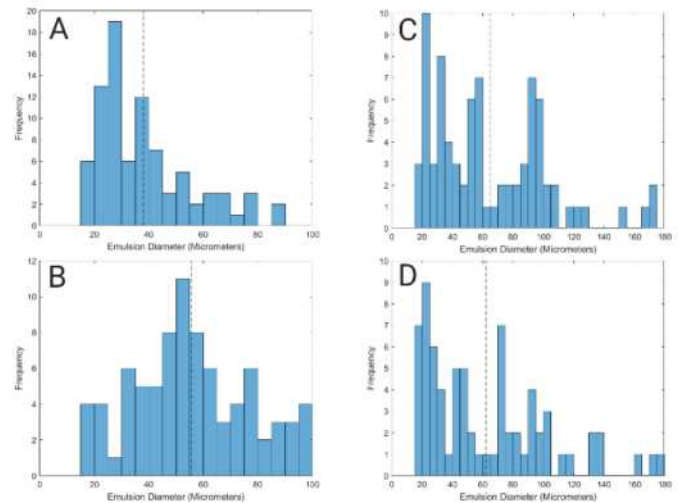


Figure 7: Emulsion droplet diameter distributions. Mean diameter values marked by black dotted line.

As depicted by the left two graphs in **Figure 7**, Surfactant 2 typically exhibited a normal emulsion size distribution in both micromodel geometries. However this distribution was significantly shifted to the left when employed in the heterogeneous micromodel along with a 32% decrease in average emulsion droplet size.

The majority of Surfactant 2 O/W emulsions were observed to form in-situ of the micromodel via 'snapping action' largely mediated by pore throat diameters. And given the heterogeneous micromodel possessed larger solid grains and smaller pore throats, smaller emulsion droplets were expected to form thus explaining the left-shifted distribution and decreased average emulsion droplet diameter.

Contrarily, as shown in the right two graphs in **Figure 7**, Surfactant 1 exhibited a more erratic emulsion size distribution seemingly independent of micromodel geometry. Heterogeneous and homogeneous distributions were almost identical, with similar average emulsion diameters also (4% difference compared to 32% difference for Surfactant 2).

This apparent independence of emulsion size on micromodel geometry was because a

smaller percentage of Surfactant 1 (W/O) emulsions were formed in-situ of the micromodel compared to Surfactant 2 (O/W) emulsions. The majority of these W/O emulsions were observed to form upstream prior to micromodel entry (unmediated by pore throat sizes) explaining why the emulsion size distributions for Surfactant 1 were less dependent on micromodel geometry.

3.2.4. In-situ micro-emulsification

Microemulsions are typically categorised as a special sub-class of emulsions. They are classed as ranging from 5-200nm in size, only forming in specific system compositions and most importantly, are thermodynamically stable [9].

In relation to enhanced oil recovery (EOR), due to the technical inability thus far to monitor flow in porous media, the nature of the emulsification process and its importance to EOR is largely unknown. However, the advent of microfluidics has enabled a means of direct visualisation of flow systems in porous media.

Tiny green emulsion droplets were observed in the micromodel during experiments where Surfactant 2 was employed. These were observed to form within the micromodel and did not coalesce over the course of the surfactant flooding period implying that an in-situ microemulsification of surfactant, oil and water phases had occurred.

These microemulsions were only observed with Surfactant 2 which may owe to the ultralow interfacial tension (IFT) conditions required for microemulsification. IFT relates to surfactant bulk and surface amounts of surfactant species 'i' through the following equation;

$$d\delta = -\Gamma_i RT \times d(\ln C_i) \quad (3)$$

Where δ is interfacial tension, Γ_i is surface excess of surfactant 'i' and C_i is the concentration of surfactant 'i' in the bulk solution [9]. It is worth noting that the maximum value of Γ_i is also proportional to solubility.

As such, it can be said that generally, the more soluble a surfactant, the higher its C_i and Γ_i which constitutes a lower δ (IFT), therefore the more likely the system is to undergo spontaneous microemulsification.

Since Surfactant 2 has been established to be more water-soluble than Surfactant 1 (Section 3.2.2.), it resolves at a lower IFT therefore its system is more likely to initiate microemulsification, explaining the occurrence of green microemulsions in Surfactant 2 systems and the absence in Surfactant 1 emulsion systems.

3.3. Quantitative Analyses

3.3.1 Oil Recovery Analysis

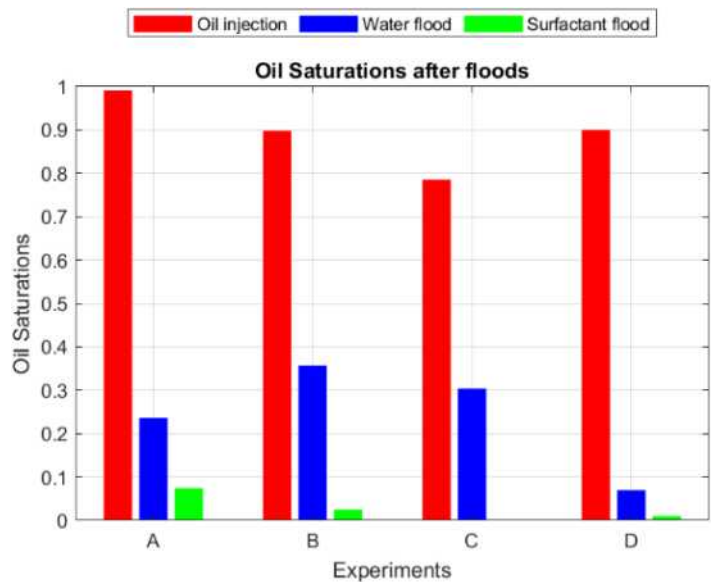


Figure 8: Oil saturations following initial oil injection, water flood and surfactant flood for experiments A, B, C and D.

(A) is surfactant 2 in the heterogeneous micromodel, (B) is surfactant 2 in the homogeneous micromodel, (C) is surfactant 1 in the heterogeneous micromodel, (D) is surfactant 1 in the homogeneous micromodel.

Analysing the bar chart presented in **Figure 8**; If surfactant type is fixed (Chart A is compared with B and Chart C is compared with D), it can be said that on average, the homogeneous micromodel achieved about 5% more oil recovery over the brine and surfactant flooding sequence than the heterogeneous micromodel. This is because, as discussed in **Section 3.2.3.**, the homogeneous micromodel had a larger average pore throat diameter than the heterogeneous micromodel. Therefore,

there existed a greater total area for O/W and W/O emulsions to propagate oil flow through the micromodel, thus yielding an increased overall oil recovery.

If micromodel geometry is fixed (Chart A is compared with C and Chart B is compared with D), one can note that, on average, Surfactant 1 achieved around 4.5% more oil recovery than Surfactant 2. This owed to Surfactant 1's oil recovery mechanisms (W/O emulsions and oil streaks) being more effective at mobilising and conveying the oil through the micromodel than Surfactant 2 (O/W emulsions). As such, this increased oil liberation led to an increased oil recovery.

4. Conclusions and Recommendations

In this research work, a novel, functional microfluidic set-up was successfully developed to investigate effects of surfactant types and micromodel geometry on oil recovery. This microfluidic approach also enabled the direct visualisation of flow phenomena and emulsification processes as a result of surfactant flooding.

Surfactant 1 was characterised by long oil streaks and W/O emulsions whilst Surfactant 2 utilised O/W emulsions to mobilise and liberate the oleic phase. Both surfactants facilitated these mechanisms by reducing IFT and shifting micromodel wettability.

Emulsion size was found to be micromodel geometry dependent for Surfactant 2 only; with the average W/O emulsion diameter being 32% smaller in the heterogeneous model compared to the homogeneous. Surfactant 1 O/W emulsion size was independent of micromodel geometry; exhibiting only a 4% difference in diameter across micromodel geometries. This invariance in geometry may prove advantageous on at the industrial scale as Surfactant 1 can be employed in a wider range of reservoir types.

Both surfactants improved overall oil recovery, but Surfactant 1 proved more effective than Surfactant 2; achieving 4.5% greater oil recovery. However, regarding industrial scale application, with Surfactant 1

being less commercially available than Surfactant 2, further economic considerations are required to determine whether the additional cost of employing Surfactant 1 instead of Surfactant 2 is justified by the marginal improvement in oil recovery.

Additional reservoir-scale research is also required to assess the validity of the results for full-scale field applications. This is to account for potential differences in experimental conditions, increased complexity of 3-dimensional porous rock media as opposed to glass micromodels alongside other general difficulties in scale-up.

Considering the limitations of this work, the accuracy and reliability of quantitative oil recovery analysis could be improved by implementing automation of the image segmentation workflow and ensuring complete separation of oleic and aqueous phases prior to micromodel entry by implementing a dual micromodel inlet. These improvements would enable a more reliable investigation into in-situ microemulsions and provide more rigorous evidence of their existence and importance to enhanced oil recovery as a whole.

The microfluidic set-up developed in this research work can also be adapted for a plethora of other disciplines and applications e.g. catalytic bed reactors [12], water-fuel separations and wastewater treatment [13] where visualisation of multi-phase flow behaviour would prove useful.

This research work could also be extended, using the current microfluidic set-up to engage in sensitivity studies e.g. investigating the effect of varying flowrates, brine salinities or surfactant concentrations.

5. Acknowledgements

We thank Dr Ronny Pini and Andrea Rovelli for their sustained support and guidance in the completion of this research project.

6. References

[1] Wu XF, Chen GQ. Global overview of crude oil use: From source to sink through inter-regional

- trade. Energy Policy. 2019;128. doi:10.1016/j.enpol.2019.01.022
- [2] Lake L, Johns RT, Rossen WR, Pope GA. Fundamentals of Enhanced Oil Recovery. 2014; doi:10.2118/9781613993286
- [3] Muggeridge A, Cockin A, Webb K, Frampton H, Collins I, Moulds T, et al. Recovery rates, enhanced oil recovery and technological limits. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2014;372(2006):20120320. doi:10.1098/rsta.2012.0320
- [4] Zhao X, Feng Y, Liao G, Liu W. Visualizing in-situ emulsification in porous media during surfactant flooding: A microfluidic study. Journal of Colloid and Interface Science. 2020;578. doi:10.1016/j.jcis.2020.06.019
- [5] Zhao X-Z, Liao G-Z, Gong L-Y, Luan H-X, Chen Q-S, Liu W-D, et al. New insights into the mechanism of surfactant enhanced oil recovery: Micellar solubilization and in-situ emulsification. Petroleum Science. 2022;19(2). doi:10.1016/j.petsci.2021.11.014
- [6] Winsor, P. A. Trans. Faraday Soc. 1948, 44, 376.
- [7] Bancroft, W. D. J. Phys. Chem. 1913, 17, 501
- [8] Clowes, G. H. A. J. Phys. Chem. 1916, 20, 407
- [9] Myers D. Surfactant science and technology. Hoboken: John Wiley & Sons; 2006.
- [10] Iglaier S, Wu Y, Shuler P, Tang Y, Goddard WA. New surfactant classes for enhanced oil recovery and their tertiary oil recovery potential. Journal of Petroleum Science and Engineering. 2010;71(1–2). doi:10.1016/j.petrol.2009.12.009
- [11] Negin C, Ali S, Xie Q. Most common surfactants employed in chemical enhanced oil recovery. Petroleum. 2017 Jun;3(2):197–211. doi:10.1016/j.petlm.2016.11.007
- [12] Trambouze P. Countercurrent two-phase flow fixed bed catalytic reactors. Chemical Engineering Science. 1990;45(8). doi:10.1016/0009-2509(90)80105-n
- [13] Santana HS, Silva JL, Aghel B, Ortega-Casanova J. Review on microfluidic device applications for fluids separation and water treatment processes. SN Applied Sciences. 2020 Feb 12;2(3). doi:10.1007/s42452-020-2176-7

Ultrathin Graphene Oxide Based Membranes with Tailored Graphitic Domain for Organic Solvent Nanofiltration

Junxin Yu and Yingqing Zhu

Department of Chemical Engineering, Imperial College London, U.K.

Abstract The reduced graphene oxide (mrGO) membranes attracted intense interests in the field of organic solvent nanofiltration (OSN) as a reliable alternative to graphene oxide (GO) membranes suffering from instability and swelling issue. The separation capability of graphene oxide-based membranes can be evaluated by permeance and size exclusion ability. The permeance is controlled by the thickness of layer-stacked GO, pore density, the chemistry of GO, and the physical properties of the solvent; and the selectivity largely relies on the interlayer d-spacing of membranes. Four GO-based membranes were discussed in this research: graphene oxide (GO), mildly reduced graphene oxide (mrGO), porous graphene oxide (PGO) and mildly reduced porous graphene oxide (mrPGO). The mild reduction approach proposed in this study allows for preferential control of the graphitic domain in mrGO membranes and fabricating ultrathin membranes with the thickness of 18-50 nm. The mrGO_9h_10s exhibited the most significant and feasible selectivity of 93.6% with a methanol permeance rate of 3.98 LMHbar⁻¹. Although the GO membrane showed higher permeance compared with mrGO, it was not structurally stable in the solvent. Both PGO and mrPGO membranes were not taken to further research since the oversized pores lost the selectivity even though a feasible permeance rate was presented.

Keywords Graphene oxide, Mildly reduced graphene oxide, Mildly reduced porous graphene oxide, Hollow fibre membranes, Organic Solvent Nanofiltration

1. Introduction

Two-dimensional (2D) materials are rapidly being investigated as a promising platform for the development of gas and liquid separation technologies because of their unique atomic thickness, micrometre lateral dimensions and physicochemical features [1]. Graphene is a single-atom-thick flat 2D carbon material consisting of a monolayer of carbon atoms in a honeycomb array that exhibits electrical and thermal conductivities. As the earliest practical discovered and common 2D material, graphene-based materials tailored with different structural and physiochemical structures have been explored extensively and show the potential for assembling high-performance membranes [2].

Organic solvents were widely used in the chemical and pharmaceutical industries. Separation between products and residual reactants, as well as catalysts, is widespread and of great importance. As a result, membrane separation technology that can work in various organic solvent environments was expected and more research into multipurpose membranes that act stable in various organic solvent nanofiltration (OSN) is necessary [3].

The membrane technology is widely used for industrial separation and purification processes because it is energy-saving, cost-saving and environmentally friendly [2]. Nanofiltration (NF) membrane has various industrial applications such as, desalination and wastewater pre-treatment because it is mesoporous (pore size in a range of 1-10 nm) with higher performance, selectivity and

lower pressure requirement compared to Reverse Osmosis (RO) [4].

Graphene Oxide (GO), the oxidative form of graphene, is highly rated due to its unique permeation pattern, large surface area, and high chemical tolerance [5]. GO is expected to be a feasible choice and has great potential in separation since GO could be easily and rapidly mass-produced from affordable raw materials graphene by chemical oxidation and ultrasonic exfoliation [6]. The nanochannels formed between GO nanosheet layers allow molecule separation, which confirms the extensive application of GO in nanofiltration.

The ceramic hollow fibre (HF) supports are widely used in the industrial separation process because of their advantages of high-packing density, chemical, thermal and structural stability, and strong and rigid properties [7,8]. The ultrathin GO membranes were coated on the HF support by vacuum filtration method. The hydrogen bonding formed by oxygen functional groups on GO nanosheets and hydroxyl groups on the surface of HF stabilised the GO/alumina HF membranes. Because of this strong interfacial interactions, GO/alumina HF membrane could withstand harsh conditions so that could be applied in multipurpose of the chemical and pharmaceutical industries [6, 8].

GO is a single layer of carbon monoatomic on the basal planes and sides, formed in a honeycomb with oxide groups (carbonyl, epoxy, and carboxyl). Its single-atom thickness property means it can be stacked easily. The presence of these oxygen functional groups in the GO makes GO nanosheets very hydrophilic and becomes the reason for membrane swelling in wet conditions. The swelling

effect of GO could be reflected by the interlayer d-spacing. That is, the dry GO has around 0.8 nm interlayer d-spacing while after hydration the d-spacing has enlarged to around 6 nm [9]. The interlayer d-spacing could be efficiently controlled by crosslinking GO nanosheets with multivalent ions or organic crosslinkers so that their stability could be improved [2].

To further improve the stability of GO, reduced GO (rGO) that removes the in-plane oxygen functional groups of GO can be prepared. However, the fully reduced GO membranes are not stable since only a few oxide groups are mainly responsible for GO flakes stacking and show low permeance [3]. Hence, the performance of GO-based membranes could be optimised by investigating the mildly reduced graphene oxide mrGO membranes.

The feasibility and performance of mrGO membranes for OSN were studied and explored in this research. With the comparison of other GO-based membranes (GO, PGO and mrPGO), four characterisation techniques are introduced to evaluate the effect of mild thermal reduction degree on membranes nanostructures and interlayer d-spacing for OSN.

To further analyse the sieving properties of ultrathin graphene oxide-based HF membranes with tailored graphitic domain for OSN, the dye Erythrosin B (EB) was used for the rejection test.

2. Methods

2.1 Materials

The chemical materials used for this experiment were referenced to Wu, Moghadam and Li (2022) [10]. The graphite powder (99% carbon basis, 325 mesh), potassium permanganate (KMnO_4 , $\geq 99.3\%$), and 1-Methyl-2-pyrrolidinone (NMP, 99.5%, anhydrous) were obtained from Sigma Aldrich. VWR supplied hydrogen peroxide (H_2O_2 , 30%), hydrochloric acid (HCl, 37%), and ethanol (absolute, VLSI). Sulfuric acid (H_2SO_4 , 98%) and ammonia solution (NH_4OH , 28–30%) were supplied by Supelco. The alumina hollow fibre was prepared by alumina oxide powder (alpha-phase, 99.9% metals basis) which was supplied by Alpha Aesar. Glass fibre filters were obtained from Whatman (US reference).

2.2 GO synthesis

According to the previously reported method [10,11], GO was synthesised by the well-known modified Hummer's method and then dispersed in water by sonication, which ultimately resulted in stable GO dispersions with 2 mg/ml concentrations.

Sulphuric acid (390 mL) was added to graphite powder (10 g) in a two-wall glass reactor, and the solution was agitated for 10 minutes while the temperature was kept at 5 °C. Then, 50g of potassium permanganate (KMnO_4) was added to the mixture and stirred for 12 h at 35 °C. The 500 mL of deionised water (DI water) was added to the solution dropwise, and the temperature was kept below 5 °C and stirred for an hour. A 10% hydrogen peroxide (H_2O_2) solution was slowly added to the solution until the colour turned into golden yellow followed by stirring for an hour. Hydrochloric acid (HCl) aqueous solution (10 wt%) was used to purify the synthesised GO. The obtained GO cake was dried under a vacuum at room temperature for more than 3 days. The GO powder was redispersed in acetone and washed in bath sonication for 10 minutes followed by vacuum filtration pass through. To obtain GO powder, the cake layer was dried at room temperature for 3 days. The GO_{xs} referred to the coating time x seconds of GO dispersions.

2.3. Porous Graphene Oxide (PGO) synthesis

The mild chemical etching method of PGO nanosheet synthesis was conducted as previously reported [10, 11]. Before synthesis, a 400 ml GO (2mg/ml) dispersion was prepared by using bath sonication for 30 minutes. NH_4OH and H_2O_2 were added into GO dispersion in the volume ratio of $\text{NH}_4\text{OH}:\text{H}_2\text{O}_2:\text{GO} = 20:1:1$ and the dispersion was stirred in a double-walled glass reactor at 50 °C for 5h. Depending on the etching time, the pore size and porosity of PGO nanosheets could be tailored. The PGO dispersion was cooled in an ice bath and centrifuged at 12000 rpm for 1 h. In order to remove the residual chemical agents NH_4OH and H_2O_2 and purify the PGO dispersion, a dialysis tubing membrane was used to dialyse redispersed PGO flakes against deionised water (DI water) for several days. PGO_{xh} referred to the GO nanosheets etched by x hours.

2.4 mrGO and mrPGO synthesis

According to Kim et al. (2018) [12], the synthesized and purified GO was dispersed homogeneously in water using a sonication bath filled with oil to maintain a uniform temperature. The dispersed GO solution was added to a round-bottom flask and stirred under flowing nitrogen that can sweep off the vapour and avoid the rebonding of functional groups and GO nanosheets. The prepared solution was maintained at 100 °C and stirred for 3h, 5h, 7h and 9h to reduce the GO dispersions. The mrGO_{xh_ys} referred to as the GO nanosheets were mildly reduced by x hours and coated by y seconds. By increasing the thermal reduction time, the interlayer d-spacing was decreased. Analogously, mrPGO was synthesised from the PGO dispersions. And

mrPGO_xh_ys referred to PGO nanosheets were mildly reduced by x hours and coated by y seconds.

2.5 Membrane fabrication and preparation

The preparation of alumina hollow fibres was referred to Tan, Liu and Li (2001) [8, 13]. Before, performance tests, alumina hollow fibres were assembled on the stainless-steel holder by using epoxy resin to seal.

The preparation of GO/alumina HF membrane used the vacuum filtration method. In order to get a uniform coating and ultrathin membranes, the GO-based dispersions were diluted to low concentrations of 0.1 mg/ml and 0.025 mg/ml before vacuum filtering. The aqueous GO-based dispersions were filtered through hollow fibre, and then the synthesized GO/mrGO/mrPGO membranes were deposited on hollow fibres. After filtration, the membranes were dried under a vacuum condition at 40 °C for 2 to 3 hours before conducting the permeance testing of OSN.

The GO-based membranes are formed by stacking the flakes layer by layer, with the aid of vacuum filtration. The thickness of membranes could be controlled by adjusting the concentrations of dispersions and the coating time.

2.6 Membranes Performance Test

2.6.1. Solvent Permeance Test

The permeance for pure solvents Methanol, Hexane and water of membranes were evaluated using a dead-end filtration cell. The schematic diagram of the permeance test could be demonstrated in Figure 1. The nitrogen gas was injected into the cells that were partially filled with pure organic solvents or dye solution to adjust the pressure applied in the cells by a pressure regulator. In this experiment, the pressure applied from 6 to 10 bar depended on the permeance behaviour. The permeate through the membranes was collected by a container and sealed with a parafilm in order to avoid the evaporation of solvents during nanofiltration. The programmed balances were used to record the weight changes of permeate in a time interval of every minute until the permeance reached a steady state for a certain time. The permeance J of each membrane was calculated with Equation 1.

$$J = \frac{\Delta M}{\rho \times A \times \Delta t \times P} \quad (1)$$

Where J is the permeance in LMHBar⁻¹, ΔM is the changed mass of permeate in g, A is the effective surface area in m², t is the testing time in h, ρ is the density of each pure solvent in g/L, and P pressure in bar.

2.6.2. Rejection Test

The EB dye with a concentration of 20 mg/L was dissolved in methanol. The rejection test was carried out in dead-end cells with a similar setup as shown in Figure 1.

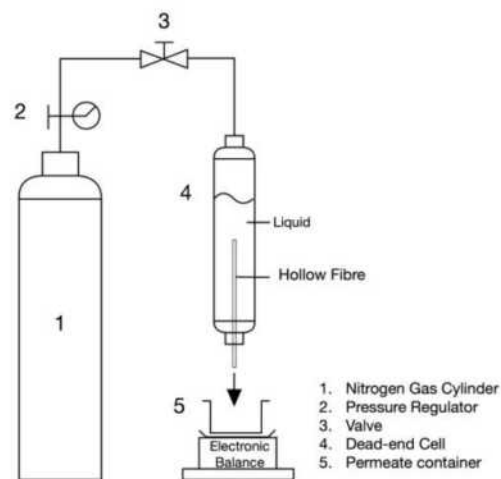


Figure 1. The scheme of permeance test set up for organic solvent nanofiltration and dye rejection test [8].

Took around 1-2 ml of permeate from the collector and feed solutions (EB dye), and stored them in separate tubes for detecting the rejection. The concentration of feed and permeate was measured by optical absorption spectroscopy UV-vis absorption. Then the rejections R of each membrane for EB dye solution could be calculated from the equation X.

$$R = \left(1 - \frac{C_p}{C_f}\right) \times 100\% \quad (2)$$

Where R is the rejection rate, C_p is the permeate concentration and C_f is the feed solution concentration in mg/L.

2.7 Membrane Characterisation Techniques

The morphology of the GO-based membranes and dimensions of flakes were observed by the Scanning Electronic Microscope (SEM). X-ray Diffraction (XRD) was applied to detect interlayer d-spacing of membranes by adjusting the 2θ contact angle in a range from 5-20 °. The interlayer d-spacing of each membrane was calculated by Bragg's equation [14] as shown in Equation 3.

$$\lambda = 2d \cdot \sin(\theta) \quad (3)$$

Where λ is the X-ray wavelength, d is the distance between adjacent nanosheets and layers, θ is the diffraction angle.

Dynamic Light Scattering (DLS) measured the particle size distribution and gave the average

particle size. The changes in membranes' particle compositions and functional groups were detected by X-ray Photoelectron Spectroscopy (XPS).

3. Results and discussions

3.1 Morphology features and microstructures of GO-based hollow fibre membranes

3.1.1 Scanning Electron Microscopy (SEM)

Four types of modified-GO membranes were chosen for further analysis in this research, including

primitive graphene oxide (GO), porous graphene (PGO), mildly reduced graphene oxide (mrGO) and mildly reduced porous graphene oxide (mrPGO). Figure 2 showed the outer surface and cross-sectional images of (a,b) GO, (c,d) mrPGO_7h, and (e,f) mrGO_9h nanosheets coated upon the alumina hollow fibre substrates using scanning electron microscopy (SEM) method under the same coating time and concentrations. With the same magnification (10,000x), the morphology of GO-based membranes were observed. The porosity behaviour of PGO could be visualised from Figure 2(c) since this more mottled display exhibited an increased pore density compared with primitive GO

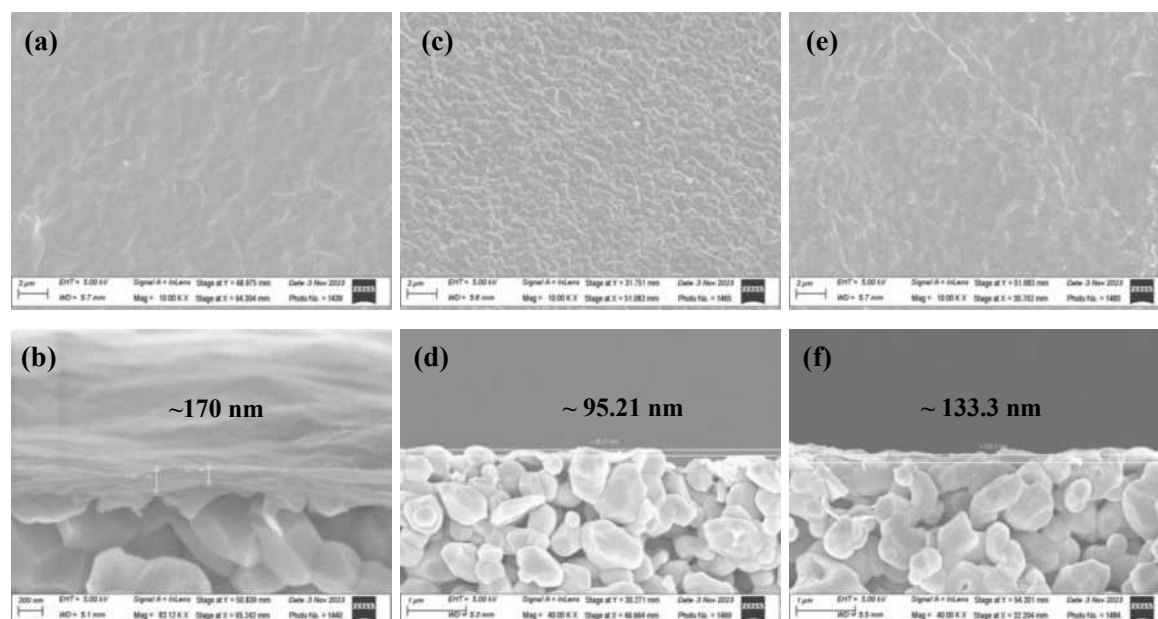


Figure 2. Outer surface and cross-sectional images of (a,b) GO, (c,d) mrPGO_7h and (e,f) mrGO_9h under same coating time and concentrations of 15s and 0.1mg/ml respectively.

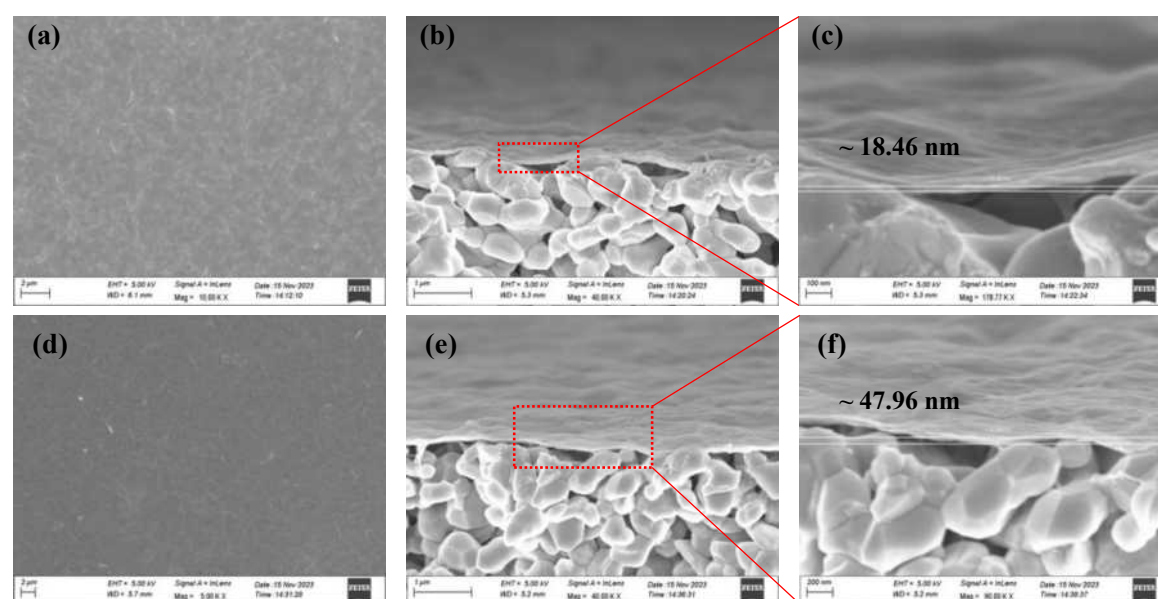


Figure 3. Outer surface and cross-sectional images of mrGO hollow fibre membranes fabricated by (a-c) 5s and (d-f) 10s coating time in concentration of 0.025mg/ml.

in Figure 2(a). Likewise, the same comparison result could be obtained from Figure 2 (c) and (e). In terms of the cross-sectional surface of GO, mrPGO and mrGO layers in Figure 2 (b), (d) and (f) respectively, the laminar structure of coating layers was observed, and the thicknesses were detected: an average thickness value of approximately 170 nm in base GO nanosheets, followed by 133.3 nm of thickness of mrGO nanosheets and the thinnest was mrPGO nanosheets with about 95.2 nm.

The morphology of mrGO hollow fibre membranes with varying coating time of 5s and 10s and 0.025 mg/ml concentration was observed, as shown in Figure 3. The effectiveness of coating was represented by the different colour gradations in Figure 3 (a) and (d), where the darker the surface image, the more stacking of mrGO layers. Moreover, Figure 3 (c) exhibited a thickness of 18.46 nm with 5s coating and as expected, thicker layers of mrGO were observed in Figure 3 (f) with a thickness of 47.96nm.

Thickness measurements of GO modifications in this research were set for two initiatives: Firstly, thickness acted as the key factor in deciding the permeation performance and the scientific detection provided strong evidence to explain the behind story; besides, the conditions which influenced the thickness of coating layers were also taking into consideration. Therefore, we can conclude that the mildly reduced GO and PGO would have smaller thickness since the oxygen functional groups on the basal plane were removed and reducing the coating time would also help with thinner the coating layers.

3.1.2 X-Ray Diffraction Spectra Analysis (XRD)

Selectivity of membranes largely relied on the interlayer d-spacing which varied for different

based-GO nanosheets compositions and functional groups. The laminar structure of GO membranes prepared by the vacuum-filtration method could be confirmed by using XRD. From XRD spectra analysis in Figure 4(a), the peak of the GO membrane was found at $2\theta = 9.56$ degrees. By applying Bragg's equation [14], with the known wavelength of x-ray beam of 0.154 nm, the d-spacing value was calculatable and analogously, the d-spacing values of mrPGO, mrGO_3h, mrGO_7h and mrGO_9h with the peak at $2\theta = 11.53, 11.58, 12.12, 12.40$ degrees were determined in Figure 4(b). From the results calculated, GO had the largest interlayer spacing followed by mrPGO and mrGO alternatives. Furthermore, it is noticeable that with longer thermal reduction time, the 2θ values corresponding to the peaks shifted to the right under the same coating concentration and time condition, which contributed to smaller values of interlayer distance. The explanation would be with a higher percentage of reduction reaction, more oxygen functional groups between GO layers were removed, therefore, confirming the inverse proportion between reduction and interlayer d-spacing, and explaining the rationality for the application of mildly reduced GO.

3.1.3 Dynamic Light Scattering Analysis (DLS)

Dynamic Light Scattering analysis [15] was prepared for studying the diffusion behaviour of molecules in solution whilst a summarized particle size distribution of all four types of modified-GO membranes could be observed from Figure 5. Three measurements were taken for each membrane detection, and the average distribution shape was compared in terms of the particle size. Figure 5

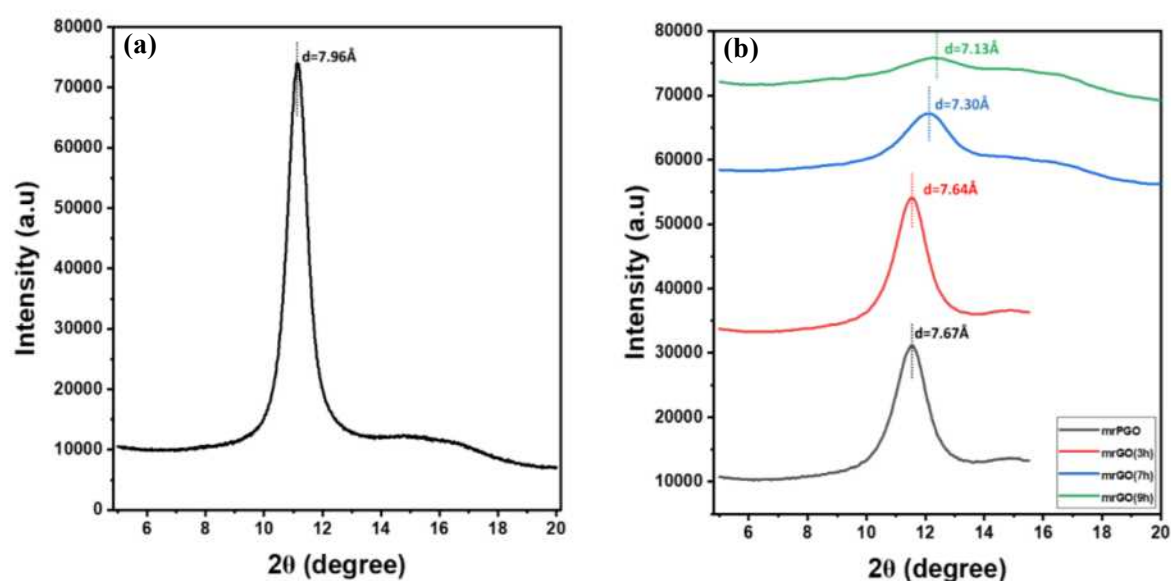


Figure 4. X-ray diffraction spectra of (a) GO and (b) mrPGO and mrGO reduced for 3h, 7h and 9h respectively.

illustrated that the size distribution of all nanosheets used for the membranes preparation were nearly uniform and within an acceptable deviation range. It then could be assumed that the effect of particle size on the permeation performance could be neglected.

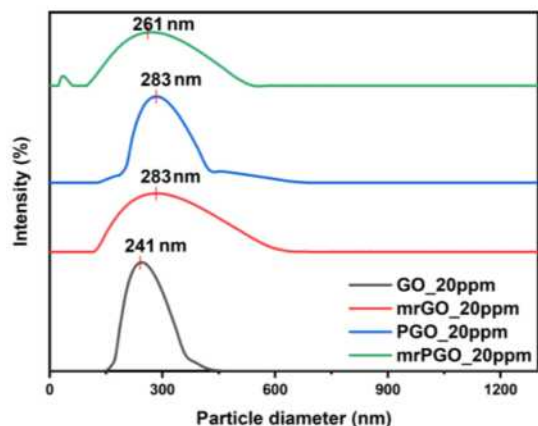


Figure 5. Particle size distributions of GO, mrGO, PGO and mrPGO in concentration of 20 ppm using detecting equipment Litesizer DLS.

3.1.4 X-ray Photoelectron Spectroscopy (XPS)

The reduction process of the GO membrane involved chemical reaction which removed oxygen functional groups, and X-ray Photoelectron Spectroscopy was applied to identify the elemental composition and the surface chemistry. From Figure 6(a) of the GO membrane, two significant peaks of C-O which is dominant and C-C bonds were observed with the highest centred at $E_{\text{binding}} \approx 287$ eV and followed by 285 eV. The abundance of C-O bonds specified that the dominant bonds were related to oxygen functional groups. Looking into mrPGO_7h in Figure 6(b), the binding energy corresponding to the highest peak was $E_{\text{binding}} \approx 284.5$ eV followed by the second high peak at $E_{\text{binding}} \approx 287$ eV. However, compared with Figure 6(a) of GO, the dominant C-O bonds were replaced by C-C bonds in mrPGO_7h, indicating the changing of compositions and the internal structures. Likewise, Figure 6(c) and (d) showed identical behaviour with mrPGO_7h since the mild reduction reaction was implemented. Similar binding energy data of the

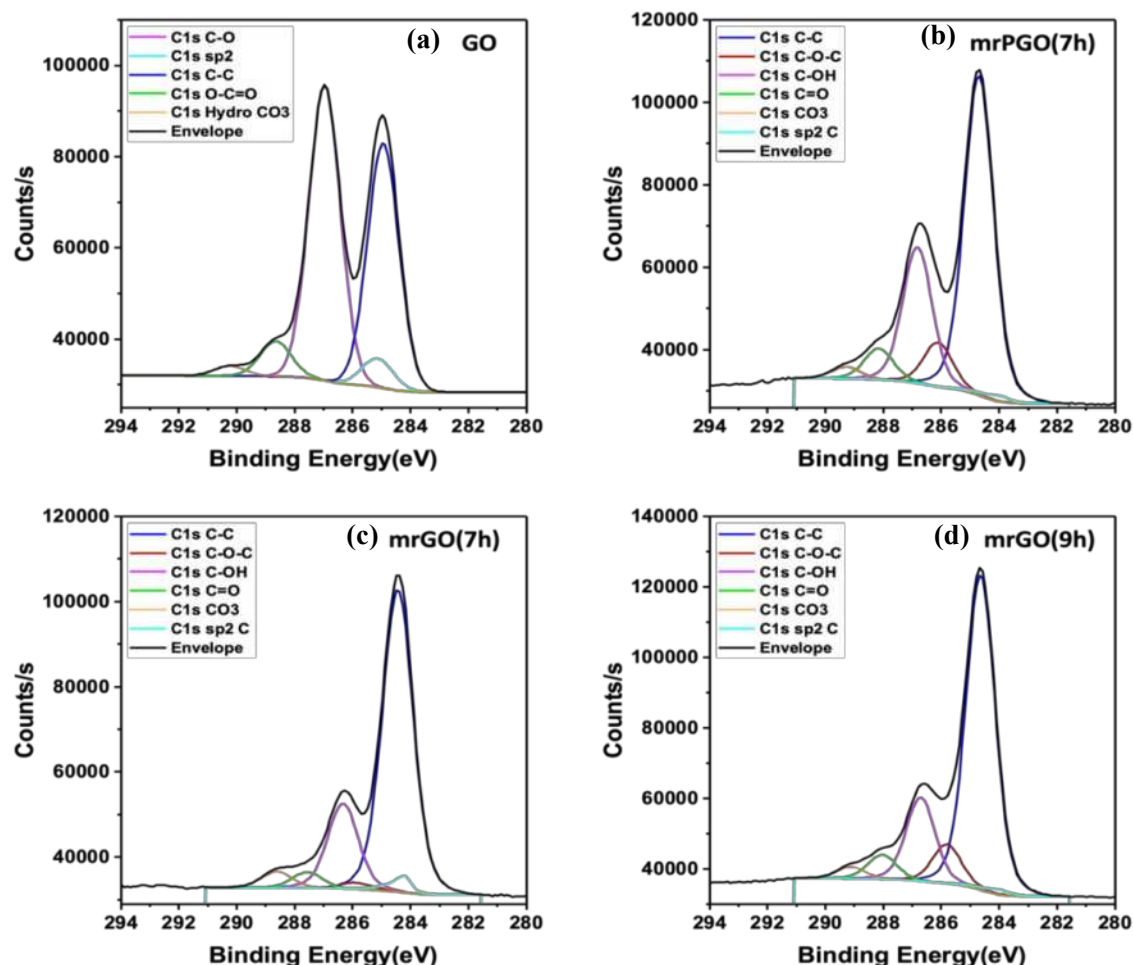


Figure 6. XPS spectra of (a) GO, (b) mrPGO_7h, (c) mrGO_7h and (d) mrGO_9h

peaks were observed as $E_{\text{binding}} \approx 284$ eV and 286.5 eV for the highest peak and second peak respectively. Another finding from the graphs was the increased C-C peak intensity with the longer thermal reduction time, and oxygen-related functional groups including O-C=O, C-O, C-OH and C=O showed a decreasing trend. A larger degree of reduction proved the enhanced predominance of C-O bonds.

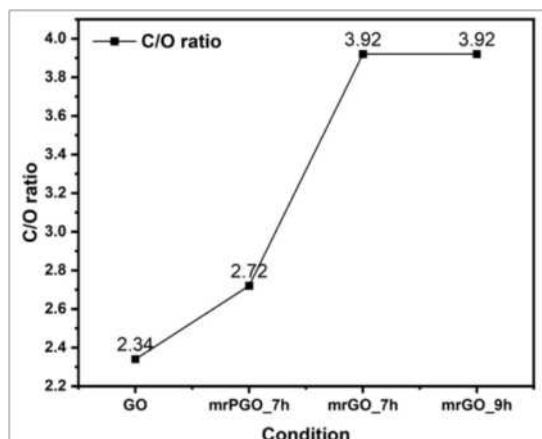


Figure 7. Calculated carbon-to-oxygen ratio of membranes (GO, mrPGO_7h, mrGO_7h and mrGO_9h) from XPS analysis.

Further analysis of the carbon-to-oxygen (C/O) ratio with respect to various material conditions was presented in Figure 7. With the higher percentage of reduction, the carbon-to-oxygen ratio data increased. mrGO groups exhibited the highest value which proved the stability nature because reduction lessened the undesirable interactions between membrane and solvent. Besides, although mrPGO_7h had a higher carbon-to-oxygen ratio than that of GO which indicated the partial reduction happened, it is still around 1.44 times lower than the value of mrGO. Porosity characterization of mrPGO made it still maintain some functional groups on the pore edges that maintained the stability of the mrPGO dispersions and its processibility into membranes. Lastly, the effect of reduction time on the C/O ratio could be neglected as no obvious differences could be found.

3.2 Solvent permeance and dye rejection test

3.2.1 Solvent permeance test

Based on the understanding of the membrane structures using characterisation techniques, the pure solvent permeance test was carried out with GO-based membranes in 0.1 mg/ml dispersion concentrations and 15 s coating time. Pure solvents were categorised into two types: organic solvents including methanol and hexane, as well as inorganic water solvents for comparison. The permeance testing results were summarized in Figure 8, the pure

solvent permeance rates of each membrane were measured and averaged after 4-5 tests. In general, four membranes all witnessed the fastest permeance in the hexane environment followed by methanol and water. The mrPGO_7h_15s HF membrane had the highest permeance of organic solvents among four membranes with 8.72 LMHBar⁻¹ in hexane, 4.76 LMHBar⁻¹ in methanol and 4.20 LMHBar⁻¹ in water. In terms of water permeance, it was enhanced by porosity as the pores created more shortcuts for molecules to pass through and shorten the required permeance length. Besides, mrPGO_7h_15s HF membrane had around two-fold faster than PGO_5h_15s because the mild reduction removed the oxygen functional groups in PGO nanosheets and the interlayer d-spacing decreased so that shorter distance for molecules to travel and showed higher permeance. Additionally, no data was recorded for water permeance by using GO_15s and mrGO_9h_15s membrane, therefore, the concept of making ultrathin membranes has arisen, which was achieved by lower dispersion concentrations and less coating time, and used to further analyse the solvents permeation performance.

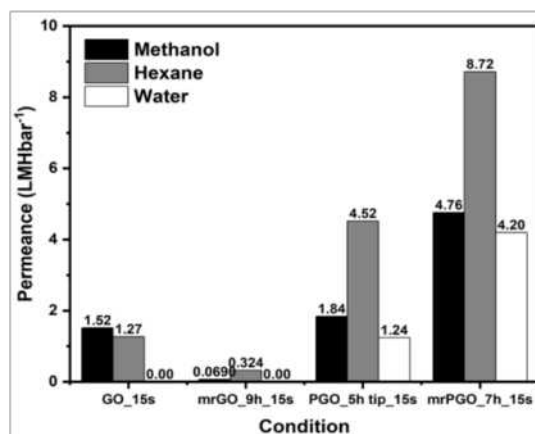


Figure 8. Organic solvent nanofiltration test results in GO-based (GO, mrGO_9h, PGO_5h and mrPGO_7h) membranes with 0.1 mg/ml concentrations of dispersions and 15s coating time.

As aforementioned, GO and mrGO HF membranes with 0.025 mg/ml dispersion concentrations and 5s, and 10s coating time were glued and tested. From Figure 9 and 10, it can be observed that with lower dispersion concentrations and less coating time, the permeability of organic solvent nanofiltration was enhanced. Although GO membranes showed more than a five-fold faster permeance rate than mrGO membranes, they were not considered for further research because they exhibited unstable and irreproducible behaviour during the experiment. The abundance existence of oxygen functional groups generated undesirable

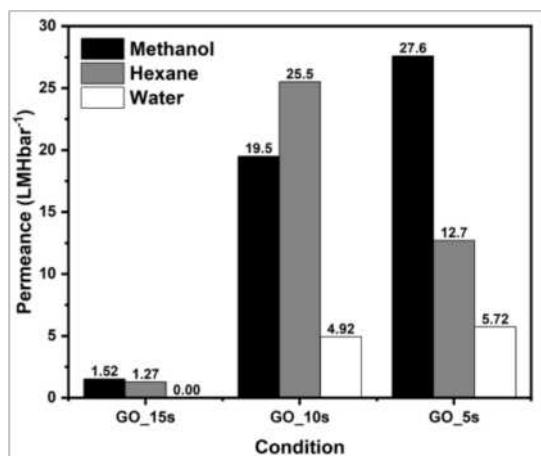


Figure 9. Permeance tests in OSN for GO membranes - 0.1mg/ml GO dispersion with 15s coating time; 0.025mg/ml GO dispersion with 10s and 5s coating time respectively.

interactions with surrounding solvent molecules and turned the GO membrane into unstable properties if it was used in the long term.

From Figure 10, it can be illustrated that the permeance rate in OSN showed an increased trend with a higher reduction degree of GO nanosheets. This trend could be explained by the decreasing percentage of oxygen functional groups that contained in mrGO membranes with higher reduction degree. The water permeance rate of mrGO_9h membranes increased to 2.00 LMHBar⁻¹ and 2.75 LMHBar⁻¹ for 5s and 10s coating time respectively. The presence of water permeates confirmed that the shorter permeation length was experienced with reduced coating time. This coincided with the cross-sectional of membranes visualised by the SEM measurements.

Furthermore, to investigate the fluid mechanisms that happened during the permeance test, several solvent parameters were correlated such as viscosity, polarity [18,19], and surface tension. Among these three solvents, hexane had the smallest viscosity and relative polarity leading to less bonding formation and interactions between membranes and solvents. Thus, membranes would be more stable in hexane solvent so that higher permeation. The values of solvent surface tension could also explain the higher permeance rate in hexane. The lowest surface tension of hexane solvent at 20 °C was 17.9 mN·m⁻¹ [20], compared with 22.5 mN·m⁻¹ of methanol [21] and 72.9 mN·m⁻¹ of water [22]. The liquid surface tension was also an important factor for predicting the membrane-solvent interactions. Lower surface tension leads to hydrophilicity in water solvent (similar behaviour could be predicted in other solvents) and wetting conditions could be easily achieved in hexane solvent.

Figure 11 was an example of plotting permeance as a function of the reciprocal of viscosity, whereas

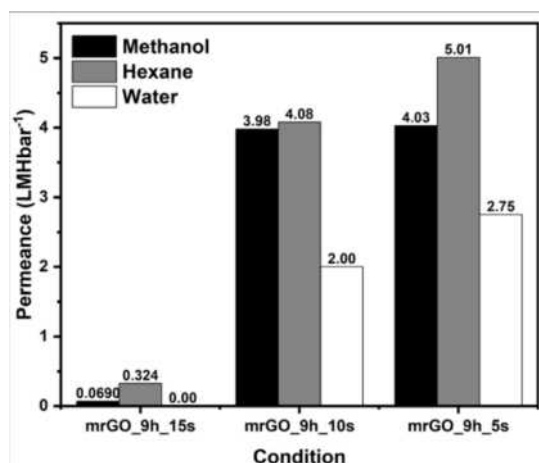


Figure 10. Permeance tests in OSN for mrGO_9h membranes - 0.1mg/ml GO dispersion with 15s coating time; 0.025mg/ml GO dispersion with 10s and 5s coating time respectively.

the viscosity of hexane, methanol and water were 0.3 mPa s, 0.544 mPa s and 0.89 mPa s respectively [16,17]. The linear best-fit line indicated that lower-viscosity solvent would have a better permeability with 5.01 LMHbar⁻¹ of hexane, 4.03 LMHbar⁻¹ of methanol and 2.75 LMHbar⁻¹ of water in mrGO_9h_5s membrane. Thus, permeance was determined to be inversely proportional to the solvent viscosity, which indicated the viscous nature of the flowing solvent in the membrane.

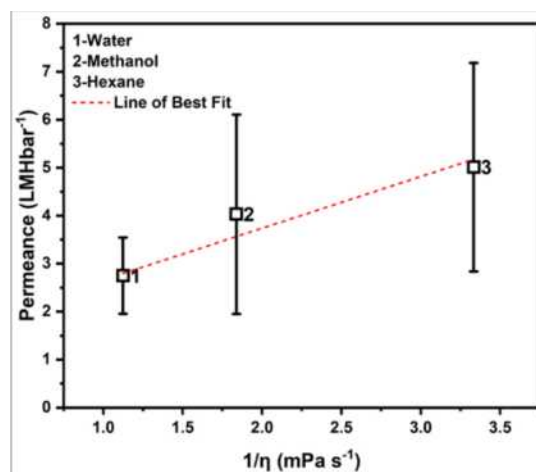


Figure 11. Permeance rate of mrGO_9h_5s HF membranes in OSN as a function of the inverse viscosity [16, 17].

3.2.2 Dye rejection test

The rejection of EB dye was tested for all developed membranes. From Figure 12, the GO_5s membrane gave the maximum methanol permeance of 27.6 LMHBar⁻¹ with dye rejection of 81.1%. Although the GO membrane showed high permeability and dye selectivity, its performance was unstable and not reproducible. The existence of oxygen functional groups on GO nanosheets made the GO very hydrophilic and the interlayer d-spacing was

enlarged under wet conditions compared to dry GO. The swelling effect exerted irreversibly on compressible membranes and resulted in an unstable state in OSN. Therefore, GO_5s membrane data was not discussed in this experiment.

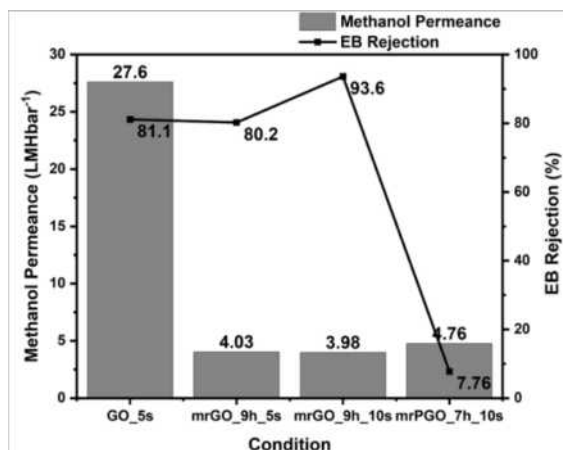


Figure 12. Methanol permeance rate and dye rejection for 0.025mg/ml dispersions including GO_5s, mrGO_9h_5s, mrGO_9h_10s and 0.1 mg/ml mrPGO_7h_10s

Compared to GO_5s HF membrane, mrGO_9h_10s improved the dye selectivity to 93.6% because mrGO contained less oxygen functional groups and smaller interlayer d-spacing. Under the same mrGO_9h dispersion concentration, mrGO_9h_5s gave only 0.05 LMHBar⁻¹ higher permeance rate than that of mrGO_9h_10s whereas the latter showed a much higher dye rejection. Although mrPGO_7h_10s membrane permeated methanol slightly faster than mrGO, its porosity allowed EB dye molecules to pass through the pores and lost its selectivity which reduced to only 7.76%. This could be explained that mrPGO nanosheets were synthesized from PGO nanosheets, and the presence of large pores in the nanosheets benefited the flowing but greatly weakened the rejection capability. Especially in ultrathin PGO/mrPGO membrane, the large pores became dominant and poor dye rejection result was foreseeable. Therefore, the ultrathin membrane mrPGO_7h with lower dispersion concentration (0.025 mg/ml) and less coating time (5s) was not considered to develop for further study in this research.

4. Conclusions

In conclusion, mrGO membranes were developed to conduct the relevant organic solvent permeation and solute rejection performances. The nanofiltration permeation was decided by several parameters such as interlayer spacing, coating thickness, pore density, the chemistry of GO, and the physical properties of the solvent.

Besides, when designing the membrane thickness to provide diverse permeation

performances, it could be tailored by adjusting the dispersion concentration and coating time.

XRD, DLS and XPS analysis results demonstrated the relationship between interlayer d-spacing and the amount of oxygen functional groups. Selectivity was subject to the interlayer spacing, where the highest rejection percentage was achieved by mrGO which has the narrowest interlayer spacing. The trade-off between selectivity and permeability was applied in the optimization of GO-based membranes.

Such stable and high selectivity performance of mrGO hollow fibre membranes indicated that it would have the potential to contribute more to the field of organic solvent nanofiltration.

Acknowledgements

The authors gratefully acknowledge the research funding provided by EPSRC in the United Kingdom (SynHiSel project grant no: EP/V047078/1) and the consistent teaching and help offered by Dr. Farhad Moghadam throughout the research project.

References

- [1] Moghadam, F. and Park, H.B. (2019). 2D nanoporous materials: membrane platform for gas and liquid separations. *2D Materials*, 6(4), p.042002.
- [2] Moghadam, F., Ji Soo Roh, Jae Eun Shin and Ho Bum Park (2020a). *Toward Sustainable Chemical Processing With Graphene-Based Materials*. Elsevier eBooks, pp.195–229.
- [3] Kumar, S., Garg, A. and Chowdhuri, A. (2022). Mildly Reduced Graphene Oxide Membranes for Water Purification Applications. *Nano Express*.
- [4] Chuntanalerg, P., Bureekaew, S., Klaysom, C., Lau, W.-J. and Faungnawakij, K. (2019). Nanomaterial-incorporated nanofiltration membranes for organic solvent recovery. *Advanced Nanomaterials for Membrane Synthesis and its Applications*, pp.159–181.
- [5] Papageorgiou, D.G., Kinloch, I.A. and Young, R.J. (2017). Mechanical properties of graphene and graphene-based nanocomposites. *Progress in Materials Science*, [online] 90, pp.75–127.
- [6] Huang, K., Liu, G., Lou, Y., Dong, Z., Shen, J. and Jin, W. (2014). A Graphene Oxide Membrane with Highly Selective Molecular Separation of Aqueous Organic Solution. *Angewandte Chemie International Edition*, 53(27), pp.6929–6932.

- [7] Wang, Z., Ge, Q., Shao, J. and Yan, Y. (2009). High Performance Zeolite LTA Pervaporation Membranes on Ceramic Hollow Fibers by Dipcoating–Wiping Seed Deposition. *Journal of the American Chemical Society*, 131(20), pp.6910–6911.
- [8] Aba, N.F.D., Chong, J.Y., Wang, B., Mattevi, C. and Li, K. (2015). Graphene oxide membranes on ceramic hollow fibers – Microstructural stability and nanofiltration performance. *Journal of Membrane Science*, [online] 484, pp.87–94.
- [9] Zheng, S., Tu, Q., Urban, J.J., Li, S. and Mi, B. (2017). Swelling of Graphene Oxide Membranes in Aqueous Solution: Characterization of Interlayer Spacing and Insight into Water Transport Mechanisms. *ACS Nano*, 11(6), pp.6440–6450.
- [10] Wu, T., Moghadam, F. and Li, K. (2022). High-performance porous graphene oxide hollow fiber membranes with tailored pore sizes for water purification. *Journal of Membrane Science*, [online] 645, p.120216.
- [11] Moghadam, F., Lee, T.H., Park, I. and Park, H.B. (2020b). Thermally annealed polyimide-based mixed matrix membrane containing ZIF-67 decorated porous graphene oxide nanosheets with enhanced propylene/propane selectivity. *Journal of Membrane Science*, 603, p.118019.
- [12] Kim, H.W., Ross, M.B., Kornienko, N., Zhang, L., Guo, J., Yang, P. and McCloskey, B.D. (2018). Efficient hydrogen peroxide generation using reduced graphene oxide-based oxygen reduction electrocatalysts. *Nature Catalysis*, 1(4), pp.282–290.
- [13] Tan, X., Liu, S. and Li, K. (2001). Preparation and characterization of inorganic hollow fiber membranes. *Journal of Membrane Science*, 188(1), pp.87–95.
- [14] Khan, H., Yerramilli, A.S., D'Oliveira, A., Alford, T.L., Boffito, D.C. and Patience, G.S. (2020). Experimental methods in chemical engineering: X - ray diffraction spectroscopy — XRD. *The Canadian Journal of Chemical Engineering*, 98(6), pp.1255–1266.
- [15] Stetefeld, J., McKenna, S.A. and Patel, T.R. (2016). Dynamic light scattering: a practical guide and applications in biomedical sciences. *Biophysical Reviews*, 8(4), pp.409–427.
- [16] Buekenhoudt, A., Bisignano, F., De Luca, G., Vandezande, P., Wouters, M. and Verhulst, K. (2013). Unravelling the solvent flux behaviour of ceramic nanofiltration and ultrafiltration membranes. *Journal of Membrane Science*, 439, pp.36–47.
- [17] Karan, S., Jiang, Z. and Livingston, A.G. (2015). Sub-10 nm polyamide nanofilms with ultrafast solvent transport for molecular separation. *Science*, 348(6241), pp.1347–1351.
- [18] Liu, R., Girish Arabale, Kim, J.-S., Sun, K., Lee, Y., Ryu, C. and Lee, C. (2014). Graphene oxide membrane for liquid phase organic molecular separation. *Carbon*, 77, pp.933–938.
- [19] Linstrom, P.J. and Mallard, W.G. (2001). The NIST Chemistry WebBook: A Chemical Data Resource on the Internet. *Journal of Chemical & Engineering Data*, 46(5), pp.1059-1063.
- [20] Klein, T., Yan, S., Cui, J., Magee, J.W., Kroenlein, K., Rausch, M.H., Koller, T.M. and Fröba, A.P. (2019). Liquid Viscosity and Surface Tension of n-Hexane, n-Octane, n-Decane, and n-Hexadecane up to 573 K by Surface Light Scattering. *Journal of Chemical & Engineering Data*, 64(9), pp.4116–4131.
- [21] Adamson, A.W. and Gast, A.P. (1997). *Physical chemistry of surfaces*. New York: Wiley.
- [22] Pallas, N.R. and Harrison, Y. (1990). An automated drop shape apparatus and the surface tension of pure water. *Colloids and Surfaces*, 43(2), pp.169–194.

Operation and Modelling of a New Reactor for Solar Water Splitting

Konrad Reents and Alexander Kovacs

Department of Chemical Engineering, Imperial College London, U. K

Abstract

Green hydrogen is proposed as a potential solution to the climate change crisis by acting as a clean store of energy. Photoelectrochemical (PEC) water splitting is one of the most promising technologies for green hydrogen production. PEC reactors utilise sunlight to drive a hydrogen evolution reaction. The aim of this report is to optimise the flow rate, concentration and temperature of the borate buffer electrolyte in the reactor to reach the best performance in terms of maximising photocurrent density whilst not compromising long term stability. The photoanodes used are made from bismuth vanadate (BiVO_4). BiVO_4 is widely named by many sources as the leading material for photoanode use. The optimal conditions are found to be a flowrate of 2.38 ml s^{-1} with a 1 M borate buffer electrolyte at 50°C . However, these conditions only apply to short term use to the mechanical degradation effects of flow on the electrode enhanced by the higher temperature. For long term use, no flow or an extremely low flow rate would have to be used in order to reduce the effects of mechanical degradation as much as possible whilst still removing bubbles forming on the electrode surface which decrease performance. A lower concentration of electrolyte would have to be used as well such as 0.5 M in order to prevent precipitation onto the electrode. During this project it has become apparent that the main issue holding back the development of PEC water splitting reactors is the poor stability of the BiVO_4 electrode. There is an excess of literature solely focussing on using BiVO_4 whilst avoiding these issues and not considering expanding to investigate other materials.

Introduction

Climate change is a generational challenge which has become increasingly serious due to significant growth of extreme climate scenarios and consequences. The still growing consumption of fossil fuels to generate energy causes a significant proportion of greenhouse gas emissions. The worldwide target is to realise a renewable energy transition. Most members of the United Nations signed the Paris Agreement to keep the global temperature rise “well below two degrees Celsius”. This ambition is targeted in national actions to reach a net zero emission in Europe, Japan, and the UK.

However, a successful transition is also threatened by increasing energy demand caused by growing world population and living standards. According to IEA research, by 2050, the demand for energy will double (1) (2). Green hydrogen as a renewable energy carrier presents a potential solution (3). It is the most abundant element in the universe and boasts a high energy density per kilogram (4). Note that hydrogen provides only a high density when stored under high pressure or in a cryogenic atmosphere. Both conditions are costly and limit the efficient application of hydrogen. Moreover, existing internal combustion technologies can be adapted to use hydrogen as a fuel. The transport industry already uses hydrogen fuel cells for special means of transportation due to their short charging/refilling time in combination with emission-free operation, as for instance in warehouses (e.g. forklifts from Still or excavators from JCB and Liebherr) (3). Standard automobiles are still far

away from incorporating hydrogen fuel cells due to an extraordinary expensive refilling infrastructure.

Furthermore, internal combustion and fuel cells processes produce very limited greenhouse gas emissions. Note that lifecycle assessments show that isolated, the reactions produce no pollutants, but the production and transportation of hydrogen produces greenhouse gases (5). The growing hydrogen market is predicted to achieve 10 GW of low-carbon hydrogen production in the United Kingdom by 2030. This goal will be reached by making available up to £11 billion in private capital (6). Despite the enormous potential of hydrogen, there are still short-term barriers to overcome. Grey hydrogen (hydrogen produced by natural gas or coal) dominates the market due to its significantly lower cost than green hydrogen (hydrogen produced by renewable energy) (2) (7). Notably, hydrogen storage and transfer are indicated as a barrier due to its highly flammable nature and expensive storage conditions (high pressures need expensive tanks and cryogenic storage is realised with energy consuming cooling devices) (8).

Photoelectrochemical (PEC) water splitting is an exceptionally promising technology for Solar-to-Hydrogen (STH) conversion. The notion is to use sunlight to produce green hydrogen (4).

The focus of this project is to optimise the flow rate, temperature, and concentration of electrolyte regarding the best performance of a PEC reactor. Firstly, bubbles on the surface are a threat for efficient PEC devices (9) (10). In theory, a flowing electrolyte pushes the bubbles away. Secondly, sunlight heats the electrolyte up, and therefore the temperature effect on the reactor performance needs further investigation. Thirdly, the

concentration optimisation takes place because literature presents the beneficial effect of adding ionic activators to the electrolyte (9) (11). In this project, the key performance indicators (KPIs) for optimal reactor performance are photocurrent density (KPI1) and long-term stability (KPI2). In detail, the highest photocurrent density is normally the optimal reactor performance. Additionally, the goal is constant reactor performance over a long period of time.

Background

PEC Water Splitting

There are multiple approaches to perform a Solar-to-Hydrogen (STH) conversion (12). The solar energy is stored in energy-rich chemical bonds (13). Among all these technologies, the photoelectrochemical water splitting is the most promising (12). PEC combines all steps (energy capture, conversion, and storage) into a stand-alone reactor. The integration into a single device is the major advantage (7). Additionally, this approach provides simple product separation, high efficiency, and good durability (14). The simple design enables flexibility to adapt the device for each specific application (15). Monolithic PEC cells only have a single multi-layered material which oversees the oxidation and reduction reactions. The more common approach is to work with multiple electrodes and separate the cathode and the anode. The direct solution consists of a photoactive working electrode whilst the counter electrode only has catalytic function. The separation allows individual material development. The most effective research approach is to use two photoactive electrodes. In this way, one can optimise the easier half-reactions independently because of the spatial separation (16).

The first report about PEC is published in 1972. Honda and Fujishima tested PEC water splitting with a TiO_2 photoanode and a Pt counter electrode. This is the starting point for a growing research area. The PEC device consists of three essential parts, electrodes, electrolyte and membrane. At least one of the electrodes must be a light absorbing photoelectrode. The electrodes are immersed in an aqueous electrolyte (13). The preferred electrolyte solution is either acidic or alkaline cause of better ionic conductivity than pure water (17). PEC devices use semiconductor materials as photoelectrodes. In detail, one differentiates between negative (n-type) and positive (p-type) semiconductors. The n-type semiconductor works as a photoanode because of its excess of electrons. Conversely, the p-type semiconductors function as a photocathode by creating electron holes (18). In addition, the PEC water splitting process is divided into four fundamental steps: charge carrier generation, charge transfer, electrochemical reaction, and mass transport and storage (17).

In the charge carrier generation step, sunlight shines on the photoactive electrode and provides photoenergy to excite an electron. The reaction is

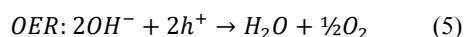
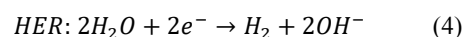
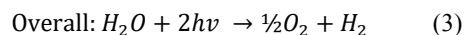
thermodynamically endothermic with a Gibb's free energy of $\Delta G = 272.15 \text{ kJ mol}^{-1}$ (18). Consequently, the provided energy must be higher than the band gap energy of 1.23 V. The electron jumps from the valance band to the conduction band. This way, electron – hole pairs are created. The necessary energy is in reality higher due to carrier separation, transport and catalysis losses (16). Equation 1 and 2 display the explained process.

$$E_{g,min} = 1.23V + \eta_{sep} + \eta_{trans} + \eta_{cat} \quad (1)$$

$$h\nu \rightarrow e^- + h^+ \quad (2)$$

Several reports show that semiconductor material undergo a subtle shift in thermodynamic potential under the influence of illumination (16) (19). The reason is the formation of quasi-fermi levels. In detail, when the semiconductor absorbs photons with more energy than the band gap, it produces photoexcited pairs of electrons in the conduction band and holes in the valence band. Additionally, it creates an equilibrium between photons and electrons in the conduction band, as well as between photons and holes in the valence band. This thermal equilibrium only occurs under stable conditions and leads to the quasi-fermi level of electrons and the quasi-fermi level of holes (20). The quasi-fermi level of electrons is higher than the level of the semiconductor, and the quasi-fermi level of electrons is lower than the level of the semiconductor during photoexcitation. When light shines, the water oxides when the quasi-fermi level of holes is greater than the redox potential of the oxygen electrode reaction (19). Conversely, in the dark the quasi-fermi levels are the level of the semiconductor cause an equilibrium between the electrons in the conduction band and the holes in the valence band exists (21). The charged carriers travel subsequently to their reaction site immediately to avoid recombination.

During the electrochemical reaction, the overall redox reaction assembles out of the hydrogen evolution reaction (HER) and the oxygen evolution reaction (OER) are established. The HER takes place at the photocathode and the OER at the photoanode (14). The reactions are illustrated by the equations 3 to 5.



These are the half reactions only for alkaline electrolyte. The half reactions for acidic electrolytes slightly change.

During the mass transport and storage step, the reaction products are removed from the reaction chamber and accumulated in storage tanks.

An impactful tool to review the PEC water splitting is the External Quantum Efficiency (EQE) shown in equation 6 (12).

$$EQE(\lambda) = \phi_{\text{gen}} \cdot \phi_{\text{sep}} \cdot \phi_{\text{trans}} \cdot \phi_{\text{cat}} \quad (6)$$

The indicator is the product of the light harvesting efficiency, the charge separation efficiency, the charge transport efficiency and the quantum efficiency. Every factor represents a single part of the fundamental steps.

Despite the promising future, the PEC water splitting technology has challenges to overcome (22). The major problems are related to efficiency or durability issues. Firstly, 10 % efficiency for a PEC device is required to commercialise the system (23). Secondly, the target is to reach lifetime of 10 years. Moreover, a material has not been found which enables to run the reactor for a longer period (24). Additionally, the technology is based on solar irradiation and therefore is only useable during the day. The sunlight dependency also leads to unpredictable performance.

Electrode Material

The electrodes used in the reactor testing are made from a piece of fluorine doped tin oxide (FTO) glass coated on one side with a flat layer of Tungsten trioxide (WO_3), then WO_3 nanoneedles on top and finally the nanoneedles are coated in a layer of bismuth vanadate (BiVO_4). The nanoneedle structure improves the rate of reaction by increasing the electrode surface area but decreases the mechanical stability when subject to flow. BiVO_4 is named by many papers as one of the most effective materials for photoanodes. The bandgap of BiVO_4 is roughly 2.4 eV which gives it a maximum theoretical current density of about 7.4 mA cm^{-2} (10). WO_3 is a good photo absorber and its bandgap lines up well with BiVO_4 as BiVO_4 absorbs well in the visible spectrum of light and WO_3 absorbs well in the UV spectrum of light. This allows for more efficient absorption of light.

Despite literature supporting BiVO_4 being one of the most effective materials for use in the photoanode, there is also literature showing that the bismuth easily dissolves into the electrolyte, particularly borate buffer, making it unstable for long term use (25). This can be seen in Appendix 11 with images of the electrodes after being used in a stability test. This is also an important issue as bismuth is listed as a scarce material.

One of the possible degradation reactions of the electrode is from BiVO_4 to V_2O_5 and $\text{Bi(III-IV)}_{\text{aq}}$. Then there are the possible dissolution reactions of V_2O_5 to V(V)_{aq} and $\text{Bi(III-IV)}_{\text{aq}}$ to Bi_{aq} and the possible precipitation reaction of $\text{Bi(III-IV)}_{\text{aq}}$ to $\text{Bi}_2\text{O}_{3-5}$ (26).

When it comes to development of larger scale PEC devices, the existing literature is mixed in opinions on the viability of use of BiVO_4 electrodes for larger reactors. The current materials in use are too expensive and unscalable to hit the required targets. The poor stability of BiVO_4 is also one of the major factors holding back scale up. Much more research is required into larger reactor sizes as almost all research done into PEC water splitting uses only one photo active electrode instead of both and electrodes of a very small size (10).

Methods

Experiments

Linear sweep voltammetry tests (LSVs) are performed by increasing the potential across the cell step wise 0.005 V every 0.10071 seconds from 0 to $0.8 V_{\text{REF}}$ over a period of 21.038 seconds and recording the output current. Each experiment setup has three tests with no incident light, three with light and one ‘chopped’ test with the shutter on the solar simulator periodically opening and closing every 2 seconds.

A reference test is used to compare the performance of each electrode, being a chopped LSV with a 2.38 ml s^{-1} flow rate and a 0.1 M borate buffer at room temperature.

Flow rate is varied by adjusting the power level on the pumps. Temperature is modified by placing the electrolyte reservoir on a hotplate and placing a temperature probe into the solution. Temperature tests are performed with flow to allow for continuous heating of the electrolyte to maintain the desired temperature. Concentration is altered by replacing the electrolyte in the reactor with electrolytes of different concentrations. Concentration tests are performed without flow to conserve boric acid supplies as large amounts of electrolyte would have had to been produced to use flow.

Chronoamperometry tests are carried out by having the reactor run with $0.5 V_{\text{REF}}$ potential and the solar simulator illuminating the anode and recording the output current. The system is given 20 minutes to stabilise at room temperature and the lowest flow rate.

For the flowrate test, pump speed is increased by one pump power level every ten minutes following the initial stabilisation period. The settings are not increased above power level seven as at this speed the tubing started to shake vigorously and large bubbles started to be created.

For the temperature test, the beaker containing the electrolyte is placed on a hot plate which is turned on at the end of the stabilisation period to heat the electrolyte. The temperature of the electrolyte is measured at the start and end of the stabilisation period with a temperature probe then every five minutes after the hot plate is turned on. The experiment is stopped when the electrolyte reaches around 62°C as it begins to produce a lot of vapour and the current readings starts to decrease.

For testing the effects of concentration, it is not possible to carry out a chronoamperometry test as concentration could not be precisely varied during operation.

Stability tests are performed by setting the desired test variable (flow speed, temperature or concentration) then applying a potential of $0.5 V_{\text{REF}}$ across the illuminated cell and allowing the reaction to proceed for at least 8 hours whilst the computer records the current output every 5 minutes. $0.5 V_{\text{REF}}$ reference is chosen as this gives a voltage of 1.23V against the right-hand electrode which is the potential required to split water.

For all types of tests, if new electrodes are used during the testing of a single parameter, results plotted as a ratio against the maximum current given in that electrodes reference test to normalise the results and remove the influence of electrode performance on our results.

Four dissolution experiments are performed to determine how the varied parameters affect degradation of the BiVO_4 electrode. Two tests are performed in the small reactor without flow and the other two are performed in the small reactor with flow. For the reactor without flow, one chamber is filled with 15 ml of electrolyte whereas for the reactor with flow, two chambers are supplied with electrolyte by a pump from two separate reservoirs containing 40 ml of electrolyte each.

Before the start of each test, a 1 ml sample of the electrolyte stock is taken with a syringe and placed into a container before adding the electrolyte to the reactor. A potential of $0.5 V_{\text{REF}}$ is applied across the illuminated reactor, and it is allowed to run for three hours. At the end of every hour, a 1 ml sample is taken from the electrolyte using a syringe and placed into a container for analysis. For the flow reactor, samples are taken from both reservoirs. At the end of the test, after the 1 ml sample is taken, the remaining electrolyte is collected and placed into a container.

The samples are analysed using an inductively coupled plasma (ICP) mass spectrometer to determine the concentration of bismuth. The charge required to produce the determined amount of bismuth can be calculated using equation 7, assuming that the oxidation state of bismuth changes from III^+ to V^+ .

$$Q = 2 \cdot F \cdot m \quad (7)$$

Q is the charge, F is Faraday's constant and m is the amount of bismuth in moles. The total charge produced by the reactor during the test equals the area under the graph of current density against time multiplied by the electrode area. These values determine how much of the charge produced during the tests are involved in hydrogen production and in electrode degradation.

Reactor Setup

Three distinct reactors are used to produce the results. The main difference is reactor size which varies photoactive areas and reaction chamber volumes. According to the PEC water splitting fundamentals, a reactor consists of electrodes, an electrolyte, and a membrane (13). BiVO_4 is the material used for the photoactive electrode (Figure 1: Photoanode FTO glass). It is a n-type semiconductor. The reference electrode is made of silver/silver chloride. Additionally, borate buffer is the used electrolyte at pH 9. These set-up parameters are constant regardless of the reactor. A few minor adjustments are made to the different reactors.

The large reactor, developed through the Hankin and Kafizas lab, is used to produce the majority of results.

The reactor consists of three compartment systems. The individual photoanode active area is 36 cm^2 , and the total photoactive area is 108 cm^2 . Only one compartment is used for each experiment. The single compartment is separated through a Nafion 115 membrane into an anode and a cathode reaction chamber. The maximum volume of each chamber is 85 ml. Furthermore, the non-photoactive nickel mesh serves as the counter electrode. The reactor composition is displayed in Figure 1. The rainbow-coloured arrow represents the simulated sunlight.

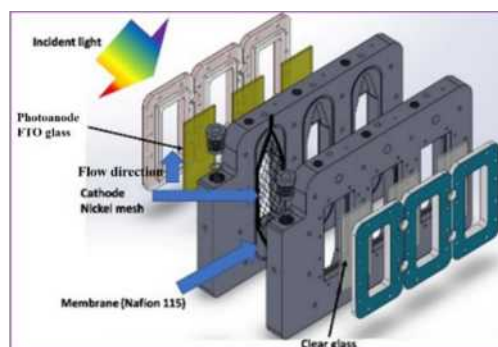


Figure 1: Composition of large reactor

In this project, smaller reactors are used to perform investigations addressing the dissolution of the BiVO_4 electrode. The small flow reactor is utilised to optimise the flow rate and temperature of the electrolyte and measure the dissolution. A perfluorosulfonic acid (PFSA) ionomer membrane divides the cell. Each of the two parts has a capacity of 1.5 ml. In addition, the cathode is made mainly out of platinum (99.9%), and the photoactive area comprises 1 cm^2 .

The flow cycle system is not needed to optimise the electrolyte's concentration. Therefore, the results for this parameter are created with another reactor. Exceptionally, the two experiments run with this reactor have a slightly changed set-up. It only includes one reaction chamber without a membrane. The maximum electrolyte volume is 15 ml and the photoactive area is 1 cm^2 . Moreover, the counter electrode is primarily composed of platinum.

The Sun2000 Solar Simulator replicates the sunlight for the experiments. Abet Technologies' sun simulator generates a cumulative light intensity of 2.5 suns using an XE lamp with an AM 1.5 G filter (Figure 2). Furthermore, it is possible to switch the shutter between open, close, and timed. This function enables to perform experiments with light changes. The lamp current must be 25 A to maintain the (nearly) identical spectrum during all the operations. Additionally, it's essential that the distance between the PEC reactor and the simulator always be 125 cm. A calibrated portable UV-Vis spectrometer (Stellarnet Black Comet Concave Grating) measures the spectrometer weekly to make sure everything is currently set up. Figure 2 shows the AM 1.5 and the solar simulator spectrum. The measured and illustrated spectral irradiance is from 271 nm to 900 nm.

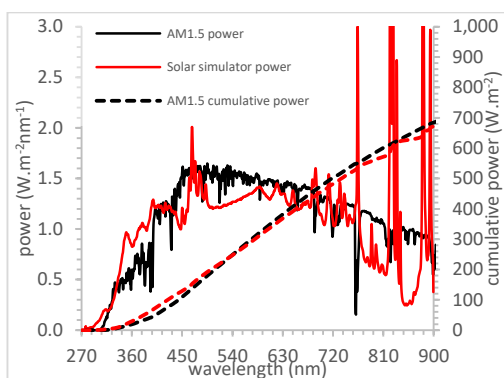


Figure 2: AM 1.5 G and solar simulator spectrum.

The aim of the project is to optimise the flow rate, the temperature, and the concentration of an electrolyte for a maximum PEC reactor performance. The reactor set-up is modified for testing each parameter. Only for the concentration observations is the set-up stationary. The electrolyte is manually filled into the reactor, which is then ready for experiments. For the flow rate and temperature measurements, the reactor is connected to the flow system. The electrolyte flows from a reservoir through the pump into the reactor. The outlet stream starts at the reactor and ends in the reservoir. This way, the fluid cycle is set up. This cycle exists for each of the two reaction chambers. The performance-addressing tests only include one reservoir for both cycles, so the solution is mixed. Especially for dissolution tests, another electrolyte reservoir is added to entirely separate the cycles. It is important that each reactor chamber is at least filled with its minimum electrolyte volume value. In this specific case, it's ensured because the inlet stream is at the bottom of the reactor chamber and the outlet is at the top. The flow direction is displayed in Figure 1. This way, the electrolyte only leaves the reactor chamber when it is completely filled up. This set-up is only partially changed for the experiments regarding the heated electrolyte. The reservoir (reservoirs) is (are) placed on a heat plate, which heats up the flowing electrolyte. Additionally, a thermometer inside the reservoir enables one to measure the current temperature and control the system for a specific temperature. By reason of the large discrepancy in the reaction chambers' volumes, the fluid cycle is slightly different for the reactors. The small reactor requires smaller tubes and a lower flow rate.

It's necessary to maintain all reactor and light set-up set points to perform the best possible optimisation. Only sensible results are appropriate for the assessment.

Electrolyte Preparation

Electrolyte is prepared by first measuring out the desired weights of boric acid and sodium hydroxide in weighing boats on a mass scale. These are then added slowly to a beaker of distilled water on a low heat with a magnetic stirrer. When the chemicals are dissolved, the stirrer is

removed, and the solution is added to a volumetric flask of the desired volume and topped up with more distilled water up to the level mark on the flask. The stirrer is then added back into the solution and a pH probe is used to test the pH of the solution. Sodium hydroxide is then slowly added to the solution to bring the pH up to 9.

To make 1 liter of a 1 M borate buffer solution, 61.83 g of boric acid and 10 g of sodium hydroxide are required. Necessary masses for the preparation of other concentrations and volumes are calculated from these values.

Results

Flowrate

Flowrates are tested from 2.38 to 17.24 ml s⁻¹. Each flow rate represents a pump power level. Beyond the 17.24 ml s⁻¹, the flowrate became too high which created a lot of vibration in the tubing and risked mechanical degradation. Tests are also performed with no flowrate (0 ml s⁻¹) to test the impact of having flow at all. Figure 3 shows that there is no clear correlation between the flowrate and current output. All flowrates have a similar ratio to the reference test. At a potential 1.23 V_{RHE} there is a point at which all the curves converge. Also, at this point, the test with no flowrate crosses the other speeds to become the highest performing speed.

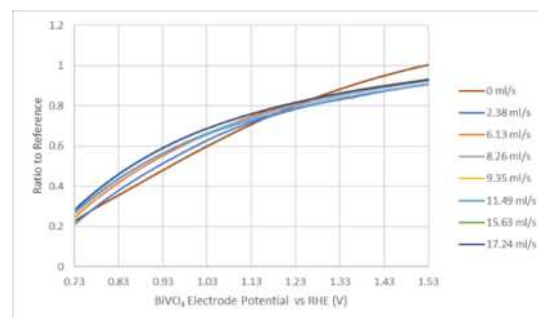


Figure 3: Graph of average light LSVs for varying flowrates with the ratio of current to the maximum reference test current plotted against potential.

The chronoamperometry test starts at 2.38 ml s⁻¹ and speed is increased every 10 minutes after the 20-minute stabilisation period up to the highest flow rate. During the stabilisation period, the current density decreases exponentially from about 0.4 to 0.16 mA cm⁻². The current then stays constant at around 0.14 mA cm⁻² from 1800 seconds onward. The peaks at around 1450 and 1800 seconds are most likely caused by disturbances in the system. The current density stays the same regardless of the flow rate. Figure 4 supports the findings of the LSV tests that there is no change in current with varying flowrate.

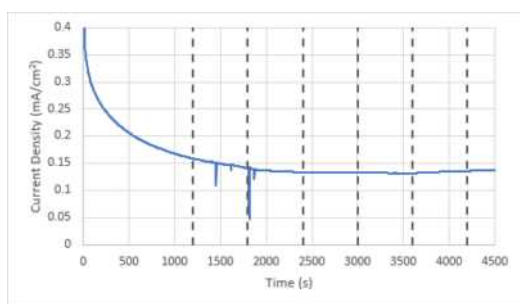


Figure 4: Graph of the flowrate chronoamperometry test with increasing flowrate at times denoted by the dotted lines with current density plotted against time.

Stability tests are performed with no flowrate and with the flowrate 2.38 ml s^{-1} . The 0 ml s^{-1} curve decreases rapidly to a minimum ratio of about 0.15 after about 15000 seconds. Afterwards, the ratio stays constant at this level. The 2.38 ml s^{-1} curve is still decreasing at the end of the test at 60000 seconds but reached a final value of about 0.06. Due to the slower rate of ratio decrease for the flow rate test, the period from about 8000 to 33000 seconds where flow has a higher ratio value than no flow. However, before and after this period, 0 ml s^{-1} has a higher ratio value.

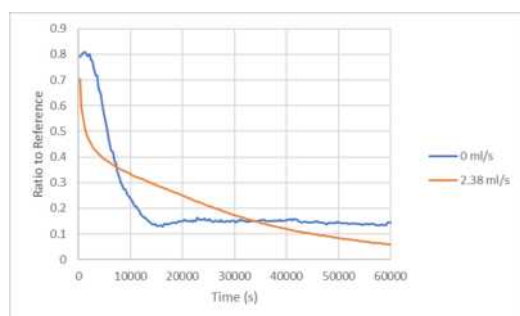


Figure 5: Graph of the stability tests for speeds 0 and 1 with the ratio of current to reference test maximum current plotted against time.

Temperature

The performance of the reactor is tested at a variety of temperatures. First, LSVs are performed for temperatures from room temperature (17°C) to 70°C . The results in Figure 6 show an increase in current density with increasing temperature up to 65°C where performance decreases at 70°C . Closer to $0.73 V_{\text{RHE}}$, values of current density for all temperatures are closer together, but as the voltage increases towards $1.53 V_{\text{RHE}}$, the current density values become further apart. The lowest performing temperature, 17°C , starts at a current density of about 0.03 mA cm^{-2} and ends at a value of about 0.23 mA cm^{-2} whereas the best performing temperature, 65°C starts at about 0.125 mA cm^{-2} and ends at about 0.48 mA cm^{-2} . Also, for the lower temperatures, the current density increases more linearly with potential but for higher temperatures, the rate of current density increase starts high and decreases with increasing potential.

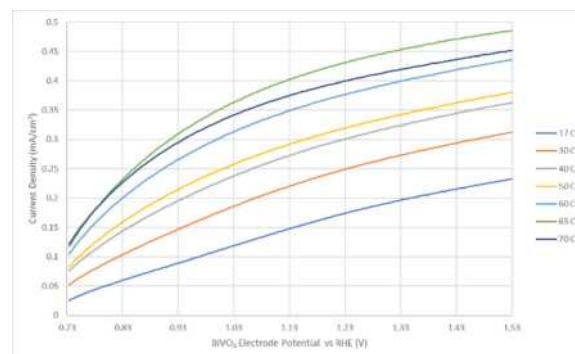


Figure 6: Graph of average light LSVs for varying temperatures with the ratio of current to the maximum reference test current plotted against potential.

The chronoamperometry test starts at room temperature then, after 20 minutes, the hot plate is turned on and the sample heats up. Figure 7 shows that current density increases with temperature from about 0.19 mA cm^{-2} at the end of the stabilisation period up to a maximum of about 0.28 mA cm^{-2} at about 55°C then begins to decrease. This supports the findings of the LSV tests.

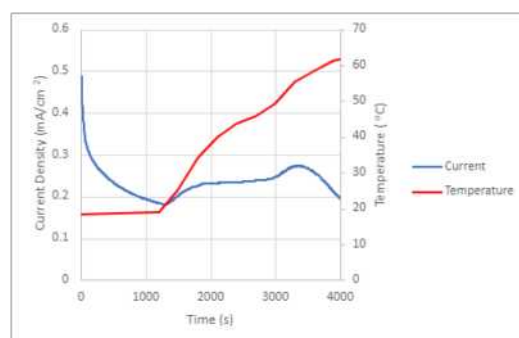


Figure 7: Graph of the temperature chronoamperometry test with temperature of the electrolyte and current density plotted against time.

Stability tests are performed at 50°C and at room temperature. 50°C is chosen as the high temperature to reduce evaporation of the electrolyte during a long experiment. Figure 8 shows that both tests decrease at a similar rate but the 50°C test decreases to a lower value which contradicts the findings of the LSV and chronoamperometry tests. The initial decrease is exponential up to about 5000 seconds but then the decline becomes more linear. The room temperature test starts at a ratio of 0.7 and decreases to a ratio of about 0.18. The 50°C test starts at a ratio of about 0.53 and decreases to a ratio of about 0.09.

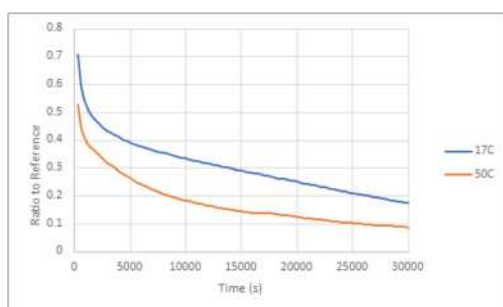


Figure 8: Graph of the stability tests for 50°C and 17°C with the ratio of current to reference test maximum current plotted against time.

Concentration

The concentrations of borate buffer are tested with 0.1, 0.5, and 1 M. Figure 9 shows a trend of increasing current output with increasing concentration. For all concentrations, the rate of ratio increase starts high and decreases with increasing potential. The 0.1 M test starts at a ratio of 0.4 and finishes at 0.9. The 1 M test starts at a ratio of 0.7 and finishes at 1.2. The difference in performance stays almost constant across all potentials.

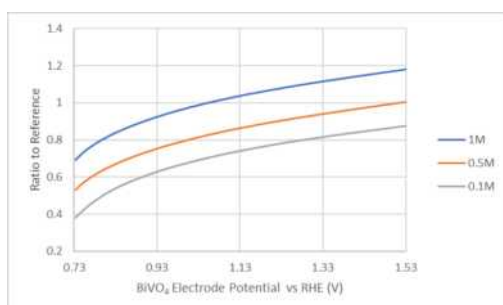


Figure 9: Graph of the light LSVs for ratio of current to the reference test maximum current against potential for varying concentration.

Concentration stability tests are performed with 0.1 and 1 M solutions. Results are normalised against their electrode's reference test and are plotted as shown in Figure 10. The 1 M test has a much higher ratio and takes much longer to decrease to the minimum current output. The 0.1 M curve decreases rapidly to a minimum ratio of about 0.15 after about 15,000 seconds. At the end of the test, the 1 M curve is still decreasing but reaches a final value of about 0.32 at 70,000 seconds.

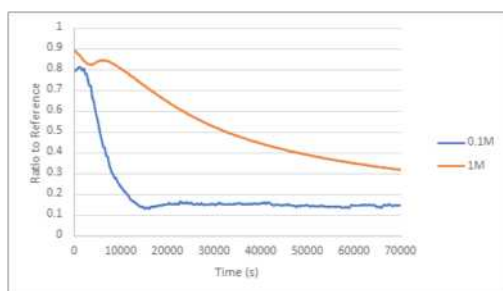


Figure 10: Graph of stability tests for 0.1M and 1M showing ratio of current output to maximum reference current against time.

Dissolution

The ICP machine was broken during the project timeline. Therefore, results are not analysed. The project explains how the results would have been used in the methodology and discussion.

Discussion

Flowrate

Bubbles are a positive sign of high a gaseous product generation rate. However, the introduced bubbles lead to a decrease in PEC reactor performance due several effects. The effective electrocatalytic area is smaller, for example (10). The introduction of flow rate solves the problem of bubbles on the photoelectrode's surface. This is visible regardless the speed level. Nevertheless, the results shown in Figure 3 show all flowrates having similar performance at specified voltages. With error bars add to the LSV data as shown in Appendix 13, one can see that they overlap which suggests that flowrate has no significant effect on reactor performance. There is no difference between no flow and flowing electrolyte at all (KPI1).

The results from the chronoamperometry show no change in the current output with increasing flowrate. This provides further evidence that flowrate has no effect on performance (KPI1). During a stabilisation phase the stationery test's performance drops significantly. After an initial minimum value, the system stabilises at a specific level. The constant ratio indicates little degradation of the photoelectrode. Figure 11 shows the surface of the photoelectrode after the run and validates the low degradation. The surface is very uniform. Only in one area does the surface looks thinner. The stationary test does not include any electrolyte motion therefore, it is solely photoelectrochemical degradation. For the stability test with speed one flowrate, the curve is more exponential and does not stabilise during the experiment. The instability indicates a high degradation of the photoelectrode. Furthermore, the degradation is visualised on the surface of the sample in Figure 11. Also, the degradation pattern is different after the flow rate experiment. The sample surface is not uniform. As well as photoelectrochemical degradation, there is evidence for mechanical degradation. The shear stress on the surface is the reason for the complete dissolution of BiVO_4 on the top part of the surface.

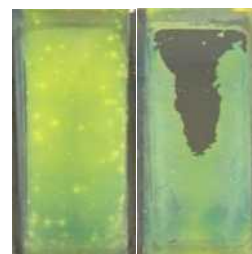


Figure 11: Photoelectrodes' surface after stability test for speed 0 and 1. left: sample GC64, speed 0; right: sample GC59, speed 1

In general, literature regarding the impacts of electrolyte flowrate is limited. The flowing electrolyte fulfils his task and pushes the bubbles from the photoelectrode's surface away. In addition, the flow enables to heat the electrolyte in a separated reservoir and to replenish the electrolyte during long term use. Nevertheless, all the advantages are neglectable next to immense degradation due to the additional mechanical stresses (KPI2). Consequently, the optimal flow rate is no flow rate at all to protect the photoelectrode's surface.

Temperature

The results of the temperature LSV and chronoamperometry tests shown in Figure 6 and Figure 7 show similar trends of reactor performance improving with increasing temperature up to a maximum between 50 and 60°C then decreasing at higher temperatures (KPI1). This drop off in performance at elevated temperatures is most likely due to increased degradation of the electrode. This would also explain the results of the stability tests in Figure 8 with the lower minimum value of the high temperature test caused by the electrode being degraded at higher rate than the room temperature test. Therefore, the results of the stability test at varying temperatures feature the same problem as the stability tests at varying flow rates. The current of these tests is never constant which indicate that the system never stabilises (KPI2). The reason for the instability is the degradation of the electrode additional intensified to the flow cycle system.

Nevertheless, at short-term experiments the increasing temperature has a beneficial effect on the performance of the reactor (KPI1). The improved performance with increasing temperature is most likely due to increased electrocatalytic activity which increases the rate of the reaction at the photoanode and reduces the overpotential of the cell. The fundamentals for this relation present the Arrhenius equations. The equation 8 shows the reaction rate regarding the factors: pre-exponential factor A, gas constant R activation energy E and temperature T. Higher temperatures also increase the conductivity of the membrane in the cell.

$$k = A \cdot e^{\frac{-E}{RT}} \quad (8)$$

Despite all the improving temperature effects the long-term runs present oppositional results. The degradation effect outweighs the improving effects and leads to unsatisfying stability results (KPI2). The tremendous mechanical degradation is set-up related and can be solved through different heating approaches. However, the Arrhenius equation is also the base for raised photo-corrosion reactions due to higher temperatures. The photoelectrochemical degradation is not set-up related. In a reactor set-up without mechanical degradation the optimal temperature can be found. The efficiency due to increased target reactions and stability loss due to photo-corrosion reactions need to be balanced. These analyses

illustrate that a temperature colder than a normal temperature, as tested in this project, does not improve the performance of the PEC reactor. Instead, colder temperature reduces the reaction rate and yield a lower performance.

In this project, the use of a hotplate in the temperature experiment hinders a constant temperature so the use of a water bath would improve the accuracy of our results. In addition, the heating process leads to loss of electrolyte due to vaporisation. The maximum used temperature is 70 °C.

According to Lamers, Marlene et. al., even temperature treatment up to 500 °C are a positive factor for the performance of PEC devices. Until this temperature the increase of the grain size leads to a better carrier mobility and diffusion length. Above 500 °C the dissolution of the vanadium dominates and dismisses the beneficial temperature effects. Lamers, Marlene et. al., present the solution to anneal the photoelectrodes in a vanadium-rich atmosphere. This approach limits the vanadium losses and prevents the decrease in current.

To sum it up, heated electrolyte increases the photocurrent density for short-term PEC reactor runs significantly (KPI1). The optimal temperature is in the range 50°C of the given device. On the basis of the immense degradation of the photoelectrode, the influence of temperature cannot be fairly evaluated for long-term runs (KPI2). The ICP results would have given a better understanding on the temperature impact on the dissolution. The degradation for heated runs adds the mechanical degradation and the boosted photoelectrochemical degradation together. By comparing these results to the flow rate dissolution results the exact effect of temperature related degradation is identifiable.

Concentration

It is predicted that higher concentrations of electrolyte improve the performance of the reactor due to the additional ionic activators to transfer charge for hydrogen evolution (9) (11). The results of the concentration LSV and stability tests shown in Figure 9 and Figure 10 show an increase in performance with increasing concentration as expected (KPI1). This is most likely due to the increased rate of the reaction at the photoanode at a higher concentration. Also, more concentrated solutions will have more ions in the electrolyte to carry charge, increasing conductivity and therefore the reactor performance.

The much longer time taken for the 1 M stability test to decrease in current output compared to the 0.1 M test may be due to the maintained higher concentration of ions at the electrode surface even after the formation of bubbles which causes the current output to decrease (KPI2). The effect of bubbles would be much greater on the 0.1 M solution with there only being a tenth of the number of ions.

The results obtained from the minimum effective concentration test presented in Appendix 14 show that

below 0.01 M electrolyte concentration, the current density drops to an insignificant value and the reactor effectively stops functioning properly (KPI1). The sudden drop in current density may be due to a change in the reaction mechanism caused by the very low number of ions in the electrolyte.

Unfortunately, samples of electrolyte at 1 M begin to crystallise when not in use which may cause issues in the system so use of a lower concentration like 0.1 M may be advisable for long term use (KPI2). Boric acid is moderately soluble in water, being able to form a 0.9 M solution at 25 °C and 3 M at 80 °C. These are subject to solution pH. This means that the 1 M optimal solution can be used if the electrolyte is heated to the optimal 50 °C temperature but not if room temperature is used to reduce degradation of the electrode. Management of a 1 M solution for commercial use would require the reactor to be stocked with fresh electrolyte before every use or for the electrolyte to be continuously heated overnight when the reactor is not in use to prevent crystallisation which would be a waste of energy.

The use of a saturated solution of electrolyte with a concentration higher than 1 M would cause precipitation onto the electrode's surface which evolved hydrogen bubbles can stick to and cause over saturation of hydrogen at the electrode surface, reducing the reaction rate and therefore reactor performance.

Dissolution

Current literature underlines BiVO₄ as one of the most promising materials as photoanode. Research highlights the bismuth vanadate's high theoretical conversion efficiency of 9.1 % and its small bad gap. While challenges related to long-term stability are acknowledged, they are often downplayed. Nevertheless, it appears that BiVO₄ retain his superior position in the literature.

This project presents the high degradation of BiVO₄. Regardless the optimised parameter the long-term stability issues are clearly noticeable. Firstly, the key performance indicator: photocurrent density decreases over time. Secondly, the photoelectrode's surface is visible thinner after each experiment. The ICP results would probably confirm the visible instability due to a high concentration of dissolved bismuth and vanadate. Intriguingly, the dissolution is already identifiable closely after the experiment's start even with a new produced photoelectrode. The target for PEC devices is a lifetime of ten years; however, the material is damaged after a couple of hours (KPI2). Accordingly, the selection of robust material for photoanode with a high reactivity is one of the key demands for PEC devices.

Furthermore, studies show that illumination has a promoting effect on the dissolution of BiVO₄. The dissolved material is significantly higher under the influence of illumination than in dark conditions. In a nutshell, a process-material that needs sunlight, is partly destroyed under illumination.

In conclusion, BiVO₄ has garnered acclaim for its theoretical efficiency. However, the observed challenges in long-term stability necessitate a revaluation of the material's suitability for sustained and reliable use in PEC devices.

Conclusion

The project examined the influence of flow rate, temperature, and concentration of borate buffer electrolyte on the PEC reactor performance. The short-term performance is optimal at a temperature of 50 °C and a concentration of 1 M with a flow rate of 2.38 ml s⁻¹. Tremendous mechanical degradation caused by the flow cycle prevents in-depth statements on long-term performance regarding these parameters. Flow rate solves the bubble challenge but introduces new mechanical stresses. The negative impact of mechanical stresses negates the bubble solution completely. For further research, the ultimate temperature effect on long-term runs is still vague. It would also be interesting to investigate how the optimised parameters affect other commonly used electrolytes, such as citrate and phosphate buffers.

In addition, the project highlights the stability issues of BiVO₄ as a photoelectrode. Despite the improved performance due to the optimised parameters, the degraded photoelectrode remains a limiting factor. The results underline critical long-term stability concerns. The observed challenges highlight the necessity of reevaluating BiVO₄'s suitability to reach the efficiency and durability goals for PEC devices. Current research is perhaps misguided in its focus on strategies to extend BiVO₄'s lifetime.

References

1. **International Energy Agency.** *World Energy Outlook.* 2023.
2. **Younas, Muhammad.** An Overview of Hydrogen Production: Current Status, Potential, and Challenges. *Fuel.* 2022.
3. **Xu, Xianxian, Zhou, Quan and Yu, Dehai.** The future of hydrogen energy: Bio-hydrogen production technology. *International Journal of Hydrogen Energy.* 2022.
4. **Chen, Zhebo, Dinh, Huyen N. and Miller, Eric.** *Photoelectrochemical Water Splitting.* s.l. : Springer, 2013.
5. **Melideo, Daniele et. ai.** *Life cycle assessment of Hydrogen and Fuel Cell Technologies.* s.l. : Publications Office of the European Union, 2020.
6. **Government, HM.** *Powering Up Britain.* s.l. : Department of Energy Security and Net Zero, 2023.
7. **Calise, Francesco et. al.** *Solar Hydrogen Production: Process, Systems and Technologies.* s.l. : Academic Press, 2019.
8. **Crowl, Daniel et ai.** The hazards and risks of hydrogen. *Journal of Loss Prevention in the Process Industries.* 2007.

9. **Zeng, Kai and Zhang, Dongke.** Recent progress in alkaline water electrolysis for hydrogen production and applications. *Progress in Energy and Combustion Science*. 2019.
10. **Moss, Benjamin, et al.** A Review of Inorganic Photoelectrode Developments and Reactor Scale-Up Challenges for Solar Hydrogen Production. *Advanced Energy Materials*. 2021.
11. **Wang, Mingyong.** The intensification technologies to water electrolysis for hydrogen production - A review. *Renewable and Sustainable Energy Reviews*. 2013.
12. **Dias, Paula and Mendes, Adélio.** Hydrogen Production from Photoelectrochemical Water Splitting. *Fuel Cells and Hydrogen Production*. s.l. : Springer, 2019, pp. 1003-1053.
13. **Jiang, Chaoran et. al.** Photoelectrochemical devices for solar water splitting - materials and challenges. *Chemical Society Reviews*. 2017.
14. **Zhao, Yibo, et al.** Recent Advancements in Photoelectrochemical Water Splitting for Hydrogen Production. *Electrochemical Energy Reviews*. 2023.
15. **Xiang, Chengxiang and Weber, Adam Z.** Modeling, Simulation and Implementation of Solar-Driven Water-Splitting Devices. *Angewandte Chemie*. 2016.
16. **Jacobsson, T. Jesper.** Sustainable solar hydrogen production. *Royal Society of Chemistry*. 2014.
17. **Kumar, Pushpendra and Kumar, Ashish.** Hydrogen Generation via Photoelectrochemical Splitting of Water. *Handbook of Ecomaterials*. s.l. : Springer, 2019, pp. 1807-1844.
18. **Hisatomi, Takashi, Kubota, Jun and Domen, Kazunari.** Recent advances in semiconductors for photocatalytic and photoelectrochemical water splitting. *Royal Society of Chemistry*. 2014.
19. **Sato, Norio.** *Electrochemistry at Metal and Semiconductor Electrodes*. s.l. : Elsevier, 2003.
20. **Gerischer, Heinz.** The impact of semiconductors on the concepts of electrochemistry. *Electrochimica Acta*. 1990.
21. **Creasey, George.** *Development and Scale-up of Photoelectrochemical Reactors for Hydrogen Production*. 2023.
22. **Niu, Fujun et. al.** Hybrid Photoelectrochemical Water Splitting Systems: From Interface Design to System Assembly. *Advanced Energy Materials*. 2019.
23. **Minggu, Lorna Jeffery et. al.** An overview of photocells and photoreactors for photoelectrochemical water splitting. *International Journal of Hydrogen Energy*. 2010.
24. **Tolod, Kristine Rodulfo, Hernández, Simelys and Russo, Nunzio.** Recent Advances in the BiVO₄ Photocatalyst for Sun-Driven Water Oxidation: Top-Performing Photoanodes and Scale-Up Challenges. *catalysts*. 2017.
25. **Zhang, Siyuan, et al.** Different Photostability of BiVO₄ in Near-pH-Neutral Electrolytes. *Applied Energy Materials*. 2020.
26. **Zhang, Siyuan, et al.** Dissolution of BiVO₄ Photoanodes Revealed by Time-Resolved Measurements under Photoelectrochemical Conditions. *The Journal of Physical Chemistry*. 2019.
27. **Wang, Songcan et. al.** An Electrochemically Treated BiVO₄ Photoanode for Efficient Photoelectrochemical Water Splitting. *Angewandte Chemie*. 2017.
28. **AAT Bioquest, Inc.** Sodium Borate Buffer (1 M, pH 8.5) Preparation and Recipe. *Quest Calculate*. [Online] October 9th, 2023.
<https://www.aatbio.com/resources/buffer-preparations-and-recipes/sodium-borate-buffer-ph-8-5>.
29. **Zhou, Chenyu et. al.** Temperature Effect on photoelectrochemical Water Splitting: A Model Study Based on BiVO₄ Photoanodes. *Applied Materials & Interfaces*. 2021.
30. **Xiao, Mu et. al.** Addressing the stability challenge of photo(electro) catalysts towards solar water splitting. *Royal Society of Chemistry*. 2023.
31. **Laidler, Keith J.** *The Development of the Arrhenius Equation*. Ottawa : University of Ottawa, 1984.
32. **Scheepers, Fabian, et al.** Temperature optimization for improving polymer electrolyte. *Applied Energy*. 2021.
33. **Lamers, Marlene et. al.** Formation and suppression of defects during heat treatment of BiVO₄ photoanodes for solar water splitting. *Royal Society of Chemistry*. 2018.
34. **Sun, Cheng Wei and Hsiao, Shu San.** Effect of Electrolyte Concentration Difference on Hydrogen Production during PEM Electrolysis. *Journal of Electrochemical Science and Technology*. 2018.
35. **Lin, Roger, et al.** Electrochemical Reactors for CO₂ Conversion. *catalysts*. 2020.
36. **Ahmet, Ibbi, et al.** Demonstration of a 50 cm² BiVO₄ tandem photoelectrochemical-photovoltaic water splitting device. *Sustainable Energy & Fuels*. 2019.
37. **Lu, Yuan et al.** Boosting Charge Transport in BiVO₄ Photoanode for Solar Water Oxidation. *Advanced Materials*. 2023.
38. **Yao, Xin et. al.** The Self Passivation Mechanism in Degradation of BiVO₄ Photoanode. *iScience*. 2019.
39. **Luo, Wenjun et. al.** Solar hydrogen generation from seawater with modified BiVO₄ photoanode. *Energy & Environmental Science*. 2011.
40. **Wen, Hongbiao et. al.** Dissolution behaviors of a visible-light-responsive photocatalyst BiVO₄: Measurements and chemical equilibrium modeling. *Journal of Hazardous Materials*. 2023.
41. **Toma, Francesca M. et al.** Mechanistic insights into chemical and photochemical transformations of bismuth vanadate photoanodes. *nature communications*. 2016.

Scaling up *in vitro* glycosylation reactions in cell-free systems using filtration: a preliminary assessment

Gabriela Trisno & Radhika Nyayadhis

Department of Chemical Engineering, Imperial College London, U.K.

Abstract

The growing popularity of recombinant protein therapeutics to treat clinical illnesses has led to the need to scale-up cell-free protein synthesis (CFPS) for *Pichia pastoris* (syn. *Komagataella phaffii*) which can produce glycoproteins. The scale-up of the *P. pastoris* lysate is limited by the centrifugation step that separates cell debris from its protein synthesis machinery. This paper intends to address this challenge through the identification and development of an alternative scalable method: Filtration, in particular depth filtration. The experimentation undertaken using syringe filters demonstrated filtration as a viable and scalable method for the primary clarification of the *P. pastoris* lysate. Depth filtration was found to produce lysates of high clarification but lower activity than centrifugation. The activity and protein concentration of filtered lysates may be influenced by several factors such as the filter properties, filter operation, and filter cake resistances. The depth filter method for this process is not as heavily developed and optimised as centrifugation. Depth filters were found to have low solid blocking hence good scalability. Therefore, depth filters demonstrate great potential in the primary clarification of the *P. pastoris* lysate with the correct choice of filter properties and mode of operation. Further improvements in the activities of clarified lysates could be achieved with further investigation and optimisation into the depth filtration method and the overall CFPS process.

1. Introduction

Through the years, recombinant protein therapeutics have grown in popularity to treat a variety of clinical illnesses. Proteins have a high molecular weight, as well as a complex composition and structure. Due to these features, proteins have limited thermal and proteolytic stability, as well as low solubility, which causes reduced efficacy and greater immunogenetic side effects^[1]. To ensure therapeutic efficacy, the majority of recombinant protein therapeutics are glycosylated^[2]. Glycosylation is the post-translational modification of amino acid to oligosaccharides. N- or O-linked glycosylated proteins are known to improve therapeutically relevant protein properties, such as pharmacokinetics, immunogenicity, and biological activity^[3]. Therefore, the ability to efficiently mass produce glycosylated proteins is a gap in biopharmaceutical processing.

The mass-production of glycosylated proteins *in vivo* poses multiple challenges, which can be addressed with the development of Cell-Free Protein Synthesis (CFPS). CFPS systems contain cell extracts with active translational and transcriptional machineries to produce proteins without intact cells, providing an opportunity for on-demand production. This shortens CFPS procedures compared to *in vivo*, as it alleviates lengthy cell handling protocols such as cloning and transformation. Due to its open nature, CFPS also provides the ability to control the compositions and conditions of the reaction directly and precisely^[4]. Developments have been made in CFPS to produce proteins that would otherwise exceed cellular toxicity tolerance^[5]. Furthermore,

CFPS facilitates the integration of high-throughput screening tools^[6], real-time monitoring, and automation^[7]. Collectively, these features make CFPS an attractive platform for the mass manufacturing of glycosylated proteins.

Recent developments have widened the library of CFPS systems to include prokaryotes and eukaryotes^[8]. However, prokaryotes lack the endogenous machinery to perform post-translational modifications. Meanwhile, mammalian cell lines, such as Chinese Hamster Ovary cells, also have several drawbacks: high cost, potential for propagating infectious agents, and long development time^[8]. Meanwhile, yeast-based expression systems present multiple benefits, such as robust expression, scalable fermentation, and the ability to perform post-translational modifications^[8]. The yeast *Pichia pastoris* has been reported to produce proteins in cell-free systems with the highest yield due to its ability to grow to high-cell densities^[9].

Active Endoplasmic Reticulum microsomes have protein synthesis machinery attached and contain components to perform glycosylation. Centrifugation enriches the microsomal content in the lysate. Centrifugation clarified lysate has been found to have improved glycosylation efficiency as compared to crude lysate^[10]. Despite the establishment of CFPS as a promising platform, the scale-up of the primary clarification of crude lysate presents a challenge. The aim of this study is to identify alternative scalable methods of separation and further develop a method at lab-scale.

2. Background

Although CFPS lysate production has been scaled up using various downstream techniques, there is very little literature available with respect to the scale-up of CFPS lysate using *P. pastoris*. Currently, centrifugation is the primary technique used at the lab scale to separate cell debris post-lysis. The centrifugation process is currently able to separate microsomes from the rest of the cell debris [10, 11]. Centrifugation due to its many challenges during limits the overall scale-up of CFPS using *P. pastoris*. Not only is centrifugation expensive but also difficult to scale-up, due to a lack of reliable scale down models [11]. Disc-stack centrifuges have been used in the pharmaceutical industry for the primary recovery of proteins at a pilot and industrial scale [12]. Disc-stack centrifuges have recently been implemented in a process which has linear scalability, for the cell free production of cytokine using *Escherichia coli* [13]. However, the scale-up of centrifugation from lab/pilot scale to industrial scale requires rigorous testing and modifications before the industrial scale operation can be finalised. The *P. pastoris* crude lysate contains several components of the sub-micron size. Disc-stack centrifuges cannot efficiently remove particles of submicron size, thus increasing the burden on secondary clarification operations and being unsuitable for this particular application [12]. Figure 1 shows a typical CFPS production process with alternatives to centrifugation for the primary clarification of the lysate.

Other secondary clarification methods such as chromatography are deemed unsuitable for the primary separation of the *P. pastoris* crude lysate. Chromatography requires the presence of different interactions between different components of the lysate for the separation to be feasible [12]. However, no differences in chemical interactions, protein affinities or hydrophobicity have been identified between the microsomes and the cell debris. The same applies to the use of flocculants to separate the crude lysate.

It has been identified with Dynamic Light Scattering and Nano Flow Cytometry that the *P. pastoris* cell size in the crude lysate is 2-3 μm whereas the typical microsome size is less than 0.4 μm [14]. This size difference is indicative that filtration or the use of size exclusion membranes might be possible. However, this has not yet been investigated for clarification of *P. pastoris*.

Depth filters have been commonly used within the pharmaceutical industry for the clarification of cell broths. They consist of multiple layers of media and filter-aids with different pore sizes within. It is to be noted that these pore size ratings are usually nominal and can vary in practice [12]. Depth filters are able to overcome the key problem of filter blocking

in the separation of biological components due to the multiple layers forming a pore gradient [12, 15]. This is advantageous compared to other filtration methods such as tangential flow filtration where submicron debris often causes high filter blockage and low filtration efficiencies [16]. Depth filters work by size exclusion and the adsorption of harder to remove fine colloidal particles [12]. Depth filters are easy to scale-up, inexpensive and can handle high capacities. However, they also have issues with the adsorption of some feed components, as well as leaching of impurities from the filter media [11]. Depth filtration followed by ultrafiltration has been successfully used in the post-lysis clarification of rotavirus like particles. This method was shown to have a 37% higher yield than the widely used CsCl density gradient ultracentrifugation method [17]. Depth filters have been used for recombinant adeno-associated virus (AAV) production [18]. It was found that the primary clarification of AAV lysate using Millistak® depth filters with diatomaceous earth (DE) have high virus recoveries of 84-97% despite being prone to adsorption [11, 19]. Recently, another study has demonstrated that depth filtration with DE as a filter aid is efficient for the primary clarification of the AAV lysate. It has been found that depth filtration with DE can have AAV loss limited to <2%, have higher turbidity removal and faster processing time than centrifugation followed by filtration [20]. Therefore, depth filtration may have re-applicability for the clarification of *P. pastoris* lysate to recover active microsomes, which leads to the need for a preliminary assessment of depth filters shown in this study.



Figure 1: CFPS production process with primary clarification alternatives

3. Methods

3.1 Preparation and Collection of Samples

3.1.1 Cell Culture Preparation and Lysis

An agar plate containing FHL1 *P. pastoris* culture which is a modified ribosome-overexpressing strain [21], was provided by the Polizzi group. The cell cultures used for the following experiments were prepared based on the optimised protocol for the preparation of *P. pastoris* lysate [9]. However, the cells were lysed via sonication instead of the high-pressure homogenisation described in the protocol. Sonication was performed with Fisherbrand™

Model 120 Sonic Dismembrator. The sonication probe was initially cleaned with 70% ethanol to remove any contaminants. To lyse the cells, the sonicator was utilised at a 50% amplitude setting followed by a 59-second interval, repeated 5 times.

3.1.2 Separation Methods and Operation

Lysate prepared as mentioned in Section 3.1.1, was then separately clarified with the use of a centrifuge, syringe filters and depth filters. The clarified lysates from these methods were then collected for further testing.

Lysate clarification was conducted via centrifugation with the Eppendorf™ Centrifuge 5810R. The centrifugation procedure followed the described protocol for the preparation of *P. pastoris* cell lysate^[9]. However, the sample obtained did not undergo further dialysis as described in the protocol.

Sartorius Minisart® 0.8 µm and 0.2 µm Syringe Filters (SF) were used to conduct a preliminary analysis of filtration. 2 ml of lysate was passed through the syringe filter by applying pressure by hand for 1 minute.

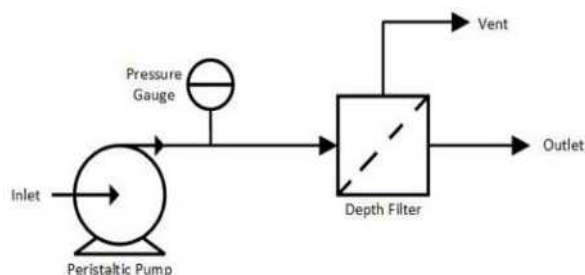


Figure 2: Schematic of the depth filter experimental set up

The depth filters used were Millistak+® HC Pro C0SP, Millistak+® HC Pro D0SP, and Clarisolve® 20MS. Note that depth filters with diatomaceous earth as filter aid were not chosen due to high operating fluxes which would not have been met due to the limitations of the project. Figure 2 shows the experimental set-up to collect samples. To operate the depth filters, a peristaltic pump was used to pump Lysis Buffer A and lysate through the filter. Details regarding Lysis Buffer A can be found in the described protocol for the preparation of *P. pastoris* cell lysate^[9]. Initially, the filters were flushed with 200 ml of Lysis Buffer A to wet the filter surface and prevent excess lysate adsorption. Afterwards, 80 ml of lysate was passed through at a pressure difference of 1 bar, in accordance with the recommended procedure in the user guide^[22]. The filtrate was collected in two fractions, one produced in the first 5 minutes to recover held-up buffer. The second fraction was collected after 10 minutes of blowdown to obtain the desired sample. The blowdown procedure can be found in the Millipore Sigma Pod Depth Filters User Guide^[22].

The samples produced by all clarification methods were immediately collected in 1.5 ml microcentrifuge tubes to be flash frozen in liquid nitrogen and stored in -80°C to maintain the activity of cell extracts.

Table 1: Syringe and Depth Filters Properties^[22,23]

Filter Name	Filter Medium	Pore Size
Millistak+® HC Pro C0SP	Polyacrylic fibre pulp, Silica filter aid	0.2 – 1.2 µm
Millistak+® HC Pro D0SP	Nonwoven, Silica filter aid, Polyacrylic fibre pulp	0.5 – 8 µm
Clarisolve® 20MS	Polypropylene and cellulose fibres, inorganic filter aid	0.5 - 20 µm
Ministart® 0.8 µm	Cellulose acetate	0.8 µm
Ministart® 0.2 µm	Cellulose acetate	0.2 µm

3.2 Analytical Methods

3.2.1 Clarification Measurement

To determine the clarification of cell lysates for each separation method, optical density measurements at 600 nm (OD600) were taken using the Eppendorf™ BioPhotometer. The OD600 of lysates was measured prior to clarification and post-clarification. The lysates were serially diluted by a factor of 100X with Lysis Buffer A to obtain an accurate measurement. Lysis Buffer A was used as a blank. The clarification percentage was calculated with the following formula:

$$\text{Clarification (\%)} = \frac{OD600_{\text{pre-clarification}} - OD600_{\text{post-clarification}}}{OD600_{\text{pre-clarification}}} \times 100\%$$

3.2.2 Protein Concentration

Bicinchoninic Acid (BCA) assay was conducted to obtain the concentration of proteins in the clarified lysate. The BCA assay works on the principle of the reduction of the Cu²⁺ ion to Cu⁺ ions by peptide bonds. The Cu⁺ ion formed reacts with bicinchoninic acid to form a purple-coloured complex which strongly absorbs light at a wavelength of 562 nm. The extent of colour change is proportional to the amount of protein in the sample, which can be measured using colorimetric methods. The Pierce™ BCA Protein Assay kit was used, the diluted albumin (BSA) standards prepared, and the BCA reaction procedure followed the protocol described for the microplate method^[24]. The CLARIOstar® Plus Microplate Reader (BMG Labtech) was used to measure the intensity of purple-coloured complex exhibited by each sample. The absorbance values

collected were subtracted by the blank absorbance value to eliminate any background absorbance. Two repeats of each sample were tested to calculate the average absorbance value. The BSA standards protein concentration were plotted against absorbance, which was linearly correlated using Microsoft Excel. By determining the equation of the linear correlation, the protein concentrations of the samples were calculated.

3.2.3 Coupled *in vitro* translation and transcription

To test the biological activity of the samples, a cell-free reaction was performed to synthesise Luciferase, an enzyme that produces bioluminescence when reacted with the reporter protein D-luciferin. An agar plate with *E. coli* culture containing recombinant Luciferase plasmid was provided by the Polizzi group. To grow the *E. coli* primary culture, a single colony was added to 5 ml Lysogeny Broth (LB) miller media and 500 μ l kanamycin. The primary culture was incubated for 8 hours at 37°C, shaking at 270 rpm. Afterwards, 500 μ l of primary culture was inoculated into a 2L baffled flask containing 500 ml LB miller media and 5 ml kanamycin. The bulk culture was incubated for 12-16 hours at the same operating conditions as previously described for the primary culture. The bacterial cells were harvested by centrifugation at 4°C and 6000g for 15 minutes and collected as 250 ml pellets. Vectors were then extracted from the *E. coli* using the HiSpeed® Plasmid Maxi Kit according to the procedure described for low-copy plasmids^[25].

To conduct the coupled *in vitro* transcription and translation protein synthesis, the reaction mix followed the optimised compositions and conditions developed for *P. pastoris*^[9]. 24.5 μ l of clarified lysate was added to an equal volume of reaction mix. To determine luciferase production, the procedure followed was in accordance with the optimised protocol developed for *P. pastoris*^[9]. Afterwards, the microplate was incubated at 21°C and the luminescence produced from the reaction was read with CLARIOstar® Plus Microplate Reader (BMG Labtech), in 1 hour 30 minutes intervals for 5 hours. A negative control of each sample was also prepared and simultaneously run with the same conditions to eliminate background luminescence.

4. Results

4.1 Pore Size Validation

Higher emission of luminescence indicates that more luciferin was produced, which correlates to higher amounts of active protein synthesis machinery present in the clarified lysate. Figure 3 shows the luminescence produced by the reaction over time, as mentioned in Section 3.2.3. “Activity”

discussed in the following sections pertains to the capacity of each clarified lysate to perform protein synthesis. The samples clarified using the 0.8 μ m and 0.2 μ m syringe filters were observed to have comparable activity to that clarified by centrifugation. This is apparent after 3 hours of reaction, in which the activity of the cell-free reaction reached peak values for all three samples. As the samples obtained from syringe filtration shows activity, this is an indication that filtration is a viable option for the primary clarification of *P. pastoris* lysate.

In Figure 3, it is observed that the activity of 0.8 μ m and 0.2 μ m syringe filter samples are comparable at the peak. However, 0.8 μ m syringe filter samples showed a slower rise in activity than 0.2 μ m syringe filter samples.

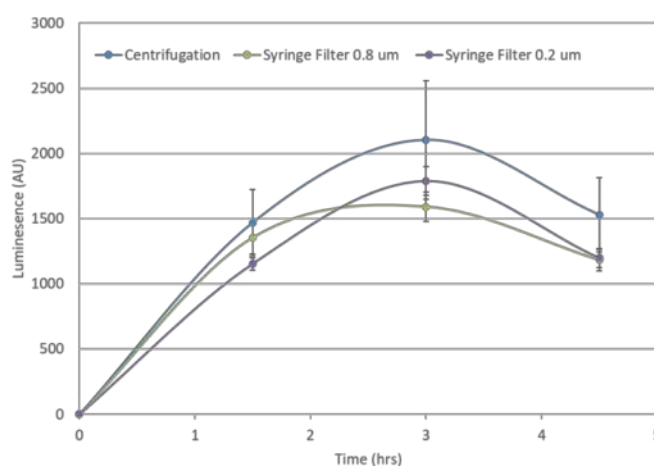


Figure 3: Luminescence produced by cell-free reaction over time for samples obtained from centrifugation and syringe filters

4.2 Depth Filter and Syringe Filter Performance

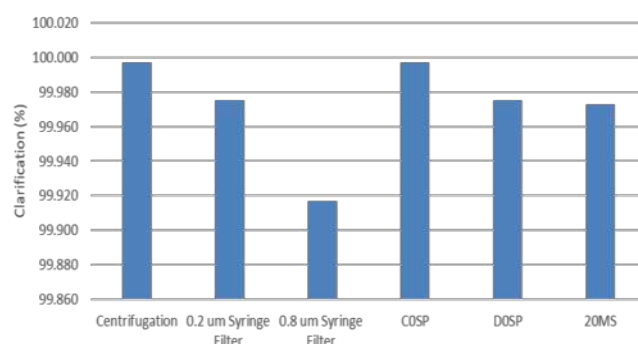


Figure 4: Percentage extent of lysate clarification for samples obtained by centrifugation, Minisart® 0.2 μ m and 0.8 μ m, Millistak+® HC Pro COSP, Millistak+® HC Pro DOSP, Clarisolve® 20MS

Figure 4 shows the clarification percentage of each separation method tested. The lysate obtained from depth filters showed high clarification comparable to that of centrifugation. However, the samples obtained from syringe filters showed lower clarification in comparison.

The protein concentration of the clarified lysate obtained from each separation method is shown in Figure 5. The protein concentration in the 0.8 μm syringe filter sample was found to be 34.7% lower than the protein concentration in the 0.2 μm syringe filter sample. As observed, clarified lysates obtained from depth filters showed lower protein concentrations compared to those obtained by centrifugation and syringe filters. Furthermore, the 20MS clarified lysate showed 56.9% higher protein concentration compared to the D0SP clarified lysate. This may be indicative of the absorption action of the filters and other effects discussed in Section 5. Based on the figures mentioned, it is apparent that no concrete conclusions can be made regarding the filter performance.

It is vital to note that the turbidity measurement was obtained through optical density, therefore, the percentage difference between the methods may not appear to be significant. However, a clear difference in colour and turbidity was observed in person. Clarified lysate obtained by centrifugation and syringe filters were slightly yellow coloured, whereas depth filters produced clear clarified lysates.

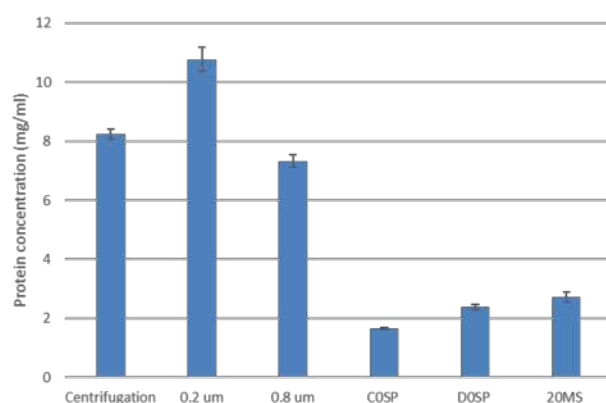


Figure 5: Protein concentration of the clarified lysate obtained from each separation method

4.3 Protein Synthesis Activity Comparison

The lysate activity for each method determined through the luciferase assay is shown in Figure 6. As observed, depth filters produce clarified lysates of lower activity than centrifugation. 20MS produced clarified lysates of 57.2% higher activity at the peak as compared to D0SP. This is interesting to observe as both filters have the same lower nominal pore size of 0.5 μm but 20MS has layers with larger pore

sizes. Meanwhile, clarified lysates produced by C0SP displayed no activity, which is interesting to observe as C0SP is constructed from the same media as D0SP but has lower nominal pore sizes. These variations in activity may be attributed to the differences in pore sizes, filter media and filter aids across all the filters, which is further discussed in Section 5.

5. Discussion

The lower activities of the depth filter clarified lysates as compared to the centrifugation lysate may at first indicate that the depth filters are not the right choice for the separation. However, it has to be noted that the protocol for the separation by centrifugation has been extensively developed and optimised for the *P. pastoris* lysate [9]. On the other hand, this study was the first time that depth filters were used in this process. The discussion section of this paper highlights potential reasons for lower activities and protein concentrations in depth filter clarified lysates and other general nuances.

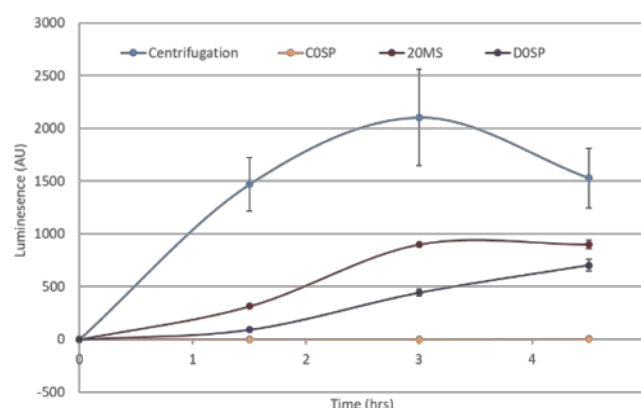


Figure 6: Luminescence produced over time by cell-free reaction for samples obtained by centrifugation and depth filtration

5.1. Filter Properties

5.1.1 Pore Size

The differences in the protein concentrations of the syringe filter samples as mentioned in Section 4.2. may be explained by the different pore sizes of the two filters. The different pore sizes may also explain the 0.8 μm syringe filter samples exhibiting a slower rise in activity than 0.2 μm syringe filter samples as mentioned in Section 4.1. It is postulated that the 0.8 μm syringe filter may allow more cell debris to pass through than the 0.2 μm syringe filter due to its higher pore size. The higher amount of cell debris present in the 0.8 μm syringe filter sample would therefore 'dilute out' the proteins. Additionally, the higher amount of cell debris present in 0.8 μm syringe filter sample may interfere with the cell-free reaction due to steric hindrance or other contributing interactions, which may lead to a slower rise in

activity. Notably, it appears in Figure 3 that the 0.8 μm syringe filter sample has a higher activity than 0.2 μm syringe filter sample. However, no solid conclusions can be made as the error bars overlap. This hypothesis should be further confirmed by testing the samples using dynamic light scattering.

It can also be observed in Figures 5 and 6 that the depth filters with layers of higher pore sizes (such as D0SP and 20MS) produce clarified lysates of higher activities and protein concentrations than those with layers of lower pore sizes (such as C0SP). This trend may appear to completely contradict the syringe filter hypothesis, however, other factors listed in Sections 5.1 and 5.2 should be considered. In addition, it is important to note that the C0SP filter was observed to have a much higher solid blockage than D0SP and 20MS during experimentation. The high solid blockage on the filter media may be attributed to the absence of layers with a pore size $>1.2 \mu\text{m}$ which was present in the D0SP and 20MS depth filters^[22]. Therefore, the high blockage levels are consistent with the lack of activity and low protein concentrations in the C0SP clarified lysate.

5.1.2 Filter Media

In Sections 4.2 and 4.3, depth filter lysates were found to be lower in activity and protein concentration than syringe filter lysates. A potential reason for these differences may be attributed to the material of the filter media. The syringe filters consist of cellulose acetate which has been found to have low protein binding properties as compared to other media. Cellulose acetate is also widely used for protein recovery applications^[23]. On the other hand, the 20MS filter is composed of polypropylene, while C0SP and D0SP filters were constructed with polyacrylic^[22]. The three filter materials have differing physico-chemical properties, for example, cellulose acetate is more polar than polypropylene^[26]. The variations in physico-chemical properties lead to materials having different component retention properties^[26, 27]. The different retention properties may explain the differences between protein concentrations in the samples.

It is postulated that the same principle may also apply for the microsomes present in the lysate, which may explain the differences in activities of the samples. However, there were no available literature for the retention of microsomes across different filter media, which should be further investigated.

5.1.3 Filter Aids

Filter aids are added to filter media to prevent pore blockage by solids and increase the porosity of the filter cake, thus improving filtration efficiency. Filter aids were present in the depth filters but not in the syringe filters. The presence of silica filter aids

in the C0SP and D0SP filters may explain the lower protein concentration in the clarified lysate described in Sections 4.2 and 4.3. Silica filter aids were found to possess high binding capacities to positively charged proteins^[28]. Therefore, it may be likely that proteins are being sorbed to the silica filter aids, which causes D0SP and C0SP samples to exhibit lower protein concentrations. Furthermore, silica filter aids may potentially retain microsomes. However, there is no literature available regarding this and should be further investigated.

5.2 Depth Filter Operation

5.2.1 Dilution

As described in Section 4.1 and 4.2, clarified lysates obtained from depth filters showed lower activity and protein concentrations as compared to those produced by centrifugation and syringe filters. This could be attributed to the operation of the depth filters at blowdown as mentioned in Section 3.1.2, which may retain the initially added Lysate Buffer A. It is known through literature that an increase in lysate dilution with buffer lowers lysate activity^[29]. During operation, it was observed that 50 ml of Lysis Buffer A was still retained despite attempts to recover it with blowdown. The lysate may have been diluted by the retained Lysis Buffer A thus lowering the concentration of active transcriptional and translational machineries present in the sample. Meanwhile, the centrifuge and syringe filters were not operated in the same manner, hence, the samples obtained with these methods were not further diluted.

5.2.2. Environment

As mentioned in Section 3.1.2, the lysate was obtained 10 minutes after operation. The temperature of the depth filters was not maintained cold, as recommended for the enrichment of microsomes to prevent protein denaturation^[9]. Thus, the lower protein concentrations and activities found in the depth filter samples in Sections 4.2. and 4.3. may be explained by protein denaturation. Meanwhile, clarified lysate from the syringe filters was obtained after applying pressure for 1 minute, which is not sufficient time to cause a large increase in the lysate temperature to cause inactivation due to thermal protein denaturation.

5.2.3 Operation Mode

As mentioned in the user guide, the depth filters are recommended to be operated continuously^[22]. However, due to limitations presented in the preparation of cell lysates, the filters were operated at blowdown. Operating the depth filters continuously may potentially improve the activity and protein concentration of the depth filter lysates. The operation of the depth filters at a pressure higher

than 1 bar may also lead to improvements in the protein concentration and activity of the depth filter samples.

5.3 Filter Cake Resistance

As the syringe filters and depth filters used for experimentation have different dimensions and configurations, it is vital to note that results obtained from both methods cannot be directly compared. During experimentation, a high flow resistance due to filter cake formation was developed in the syringe filters after 2 ml of lysate was passed through. However, this was not observed for the D0SP and 20MS filters after 80 ml of lysate was filtrated through. The ability of the two depth filters to prevent complete pore blockage may be attributed to the presence of filter aids, a pore size gradient throughout the multiple filtration layers, and a lower lysate throughput to filtration area ratio.

Although the high activity of the syringe filter samples may seem attractive, the ability of the depth filters to prevent blockage must also be considered when scaling-up. This is because the filter cake resistance not only determines the viability of filter scale-up but also affects the filtrate compositions and concentrations. From this, it can be determined that the configuration of the filter is a vital aspect which can affect the properties of the filter cake, which affects the efficiency of filtration performance for scale-up.

5.4 Loss of Activity

In this study, *P. pastoris* was prepared at a larger scale, 140g of cells were prepared as opposed to 2-5g prepared for biological experiments. The scaling-up of the CFPS process from cell preparation to primary clarification may have an effect on lysate properties. Therefore, it is postulated that the large-scale handling of cells could have led to a loss of activity of the sample.

6. Conclusion and Outlook

This study intended to develop filtration as a scalable method for the primary clarification of *P. pastoris* lysate. Experimentation undertaken with syringe filters demonstrated filtration as a feasible method to produce clarified lysates of comparable protein synthesis capabilities as those obtained via centrifugation. Depth filters produced clarified lysates with high clarifications but exhibited low activities and protein concentrations. Despite this, depth filters demonstrated the potential for scale-up with further optimisation due to low filter blockage.

In order to gain a deeper understanding into the clarification process, the clarified lysates must be tested for the presence of microsomes and their ability to undergo glycosylation.

To implement depth filters in a large-scale CFPS process, the continuous operation of the depth filters has to further developed and optimised. The depth filter operation must be conducted to reduce dilution and filtration time, which might improve clarified lysate activity. Further studies must also be conducted to explore the interactions between lysate and filter components, in order to minimise the retentions of proteins and microsomes.

Furthermore, filter blockage and flow resistance due to filter cake developed is a challenge in the use of filters in the pharmaceutical industry^[15]. Although this paper was able to qualitatively note the reduced blockage in depth filters, the filter cake properties must be determined through V_{\max} or P_{\max} testing^[20].

The V_{\max} and P_{\max} testing will give insight into changes in filtrate properties with time and will allow for further clarity on the scale-up of the depth filters to pilot/manufacturing scales. This testing would require 6 times more lysate than the amount of lysate prepared in this study.

The optimisation of the production stages prior to primary clarification of lysate is vital to producing clarified lysates of high glycosylation activities. *P. pastoris* has been used in the biopharmaceutical industry to produce therapeutics, hence, the conditions for its growth in a bioreactor have been optimised and shows potential re-applicability^[30]. The cell lysis method affects the effectiveness of cell membrane breakage, as well as the concentration of microsomes. Therefore, the procedure has to be optimised to improve the activity of lysate to produce glycoproteins. It has been shown that the use of glass beads instead of sonication for cell lysis has improved clarified lysate activity^[30]. Hence, the optimisation of *P. pastoris* lysate production process may alleviate the burden posed on depth filters. Overall, the optimisation of the entire process must be carried out to allow for further scale-up.

Acknowledgements

We would like to thank Ms. Farzana Alam for her continuous support in the completion of this project. We would also like to express our sincere appreciation to Rui Wu, Dr. Andre Ohara, Dr. Chileab Redwood-Sawyer, Clodagh Towns and Dr. Umang Shah for their help in the laboratory.

References

- [1] Ma, B., Guan, X., Li, Y., Shang, S., Li, J., & Tan, Z. (2020). Protein Glycoengineering: An Approach for Improving Protein Properties. *Frontiers in chemistry*, 8, 622. <https://doi.org/10.3389/fchem.2020.00622>
- [2] Weston Kightlinger, Katherine F. Warfel, Matthew P. DeLisa, and Michael C. Jewett (2020). *Synthetic Glycobiology: Parts, Systems, and Applications*, *ACS Synthetic Biology* **2020** 9 (7), 1534-1562. DOI: 10.1021/acssynbio.0c00210
- [3] Solá, R. J., & Griebenow, K. (2009). Effects of glycosylation on the stability of protein pharmaceuticals. *Journal of pharmaceutical sciences*, 98(4), 1223–1245. <https://doi.org/10.1002/jps.21504>
- [4] Swartz, J.R. (2012). Transforming biochemical engineering with cell-free biology. *AIChE J.*, 58: 5-13. <https://doi.org/10.1002/aic.13701>
- [5] Katzen, F., Chang, G., & Kudlicki, W. (2005). The past, present and future of cell-free protein synthesis. *Trends in biotechnology*, 23(3), 150–156. <https://doi.org/10.1016/j.tibtech.2005.01.003>
- [6] Zhang, Y. et al. (2019). Accurate high-throughput screening based on digital protein synthesis in a massively parallel femtoliter droplet array. *Sci. Adv.* 5, eaav8185(2019). DOI:10.1126/sciadv.aav8185
- [7] Georgi, V. et al. (2015). On-chip automation of cell-free protein synthesis: new opportunities due to a novel reaction mode. *Lab Chip*, 2016,16, 269-28. DOI: <https://doi.org/10.1039/C5LC00700C>
- [8] Sethuraman, N., & Stadheim, T. A. (2006). Challenges in therapeutic glycoprotein production. *Current opinion in biotechnology*, 17(4), 341–346. <https://doi.org/10.1016/j.copbio.2006.06.010>
- [9] Aw, R., Spice, A. J., & Polizzi, K. M. (2020). Methods for Expression of Recombinant Proteins Using a *Pichia pastoris* Cell-Free System. *Current protocols in protein science*, 102(1), e115. <https://doi.org/10.1002/cpps.115>
- [10] Brödel, A. K., Sonnabend, A., & Kubick, S. (2014). Cell-free protein expression based on extracts from CHO cells. *Biotechnology and bioengineering*, 111(1), 25–36. <https://doi.org/10.1002/bit.25013>
- [11] Besnard, L., Fabre, V., Fettig, M., Gousseinov, E., Kawakami, Y., Laroudie, N., Scanlan, C., & Pattnaik, P. (2016). Clarification of vaccines: An overview of filter based technology trends and best practices. *Biotechnology advances*, 34(1), 1–13. <https://doi.org/10.1016/j.biotechadv.2015.11.005>
- [12] Liu, H. F., Ma, J., Winter, C., & Bayer, R. (2010). Recovery and purification process development for monoclonal antibody production. *mAbs*, 2(5), 480–499. <https://doi.org/10.4161/mabs.2.5.12645>
- [13] Zawada, J. F., Yin, G., Steiner, A. R., Yang, J., Naresh, A., Roy, S. M., Gold, D. S., Heinsohn, H. G., & Murray, C. J. (2011). Microscale to manufacturing scale-up of cell-free cytokine production--a new approach for shortening protein production development timelines. *Biotechnology and bioengineering*, 108(7), 1570–1578. <https://doi.org/10.1002/bit.23103>
- [14] Alam, F. (2023). Communication of Ongoing Scientific Research & Results, Polizzi Group, Imperial College London.
- [15] Singh, N., Pizzelli, K., Romero, J. K., Chrostowski, J., Evangelist, G., Hamzik, J., Soice, N., & Cheng, K. S. (2013). Clarification of recombinant proteins from high cell density mammalian cell culture systems using new improved depth filters. *Biotechnology and bioengineering*, 110(7), 1964–1972. <https://doi.org/10.1002/bit.24848>
- [16] Schmidt, S.R., Wieschalka, S., Wagner, R. (2017). Single-Use Depth Filters: Application in Clarifying Industrial Cell Cultures. *BioProcess International* 15. <https://bioprocessintl.com/2017/single-use-depth-filters-application-clarifying-industrial-cell-cultures/>
- [17] Peixoto, C., Sousa, M. F., Silva, A. C., Carrondo, M. J., & Alves, P. M. (2007). Downstream processing of triple layered rotavirus like particles. *Journal of biotechnology*, 127(3), 452–461. <https://doi.org/10.1016/j.jbiotec.2006.08.002>
- [18] Cecchini, S., Virag, T., & Kotin, R. M. (2011). Reproducible high yields of recombinant adeno-associated virus produced using invertebrate cells in 0.02- to 200-liter cultures. *Human gene therapy*, 22(8), 1021–1030. <https://doi.org/10.1089/hum.2010.250>
- [19] Namatovu, H., Hsu, W., Waghmare, W., Wastler, S., McDowell, C., Butman, B.T., 2006. Evaluation of filtration products in the production of adenovirus candidates used in vaccine production: overview and case study. *Bioprocessing* 5, 67–74. <https://doi.org/10.12665/J53.Namatovu>
- [20] Meierrieks, F., Pickl, A., & Wolff, M. W. (2023). A robust and efficient alluvial filtration method for the clarification of adeno-associated viruses from crude cell lysates. *Journal of biotechnology*, 367, 31–41. <https://doi.org/10.1016/j.jbiotec.2023.03.010>

- [21] Aw, R., & Polizzi, K. M. (2019). Biosensor-assisted engineering of a high-yield *Pichia pastoris* cell-free protein synthesis platform. *Biotechnology and bioengineering*, 116(3), 656–666. <https://doi.org/10.1002/bit.26901>
- [22] Millipore Sigma. (2021). Pod Depth Filters User Guide. Millipore Sigma Aldrich. <chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.sigmaaldrich.com/deepweb/assets/sigmaaldrich/product/documents/424/763/pod-depth-filters-ug4697en-mk.pdf>
- [23] Sartorius. (2023). Sartorius Minisart® Selection Guide How to Choose the Optimal Housing and Membrane Material for Your Application. Sartorius. <https://www.sartorius.com/en/pr/lab-filtration/minisart-selection-guide>
- [24] ThermoFisher. (2020). Pierce™ BCA Protein Assay Kit. Thermo Scientific. https://assets.thermofisher.com/TFSAssets/LSG/manuals/MAN0011430_Pierce_BCA_Protein_Asy_U G.pdf
- [25] Qiagen®. (2012) HiSpeed® Plasmid Purification Handbook. QIAGEN. <https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ah UKewje1Yy-8oyDAxWSTkEAHUbFA1EQFnoECBwQAQ&url=https%3A%2F%2Fwww.qiagen.com%2Fus%2Fresources%2Fdownload.aspx%3Fid%3D5a63b6b2-7fd7-4fbf-b3cd-e50d8270fea1%26lang%3Den&usg=AOvVaw262tRm3wAVlb9426VYiV11&opi=89978449>
- [26] Klus H., Pachinger, A., Nowak, A. (2001). Comparison of the selective retention capacity of cigarette filters made from cellulose acetate and polypropylene. Cooperation Centre for Scientific Research Relative to Tobacco. <https://www.coresta.org/abstracts/comparison-selective-retention-capacity-cigarette-filters-made-cellulose-acetate-and>
- [27] Velu, S., Rmababu, K., Muruganandam, L. (2014). Preparation, Characterization and application studies of cellulose acetate – poly acrylic acid blend ultra filtration membranes. *International Journal of ChemTech Research*, 6(3), 1855-1857. [chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://sphinxnsai.com/2014/vol6pt3/10/\(1855-1857\)ICMCT14.pdf](chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://sphinxnsai.com/2014/vol6pt3/10/(1855-1857)ICMCT14.pdf)
- [28] Chu, L. K., Borujeni, E. E., Xu, X., Ghose, S., & Zydney, A. L. (2023). Comparison of host cell protein removal by depth filters with diatomaceous earth and synthetic silica filter aids using model proteins. *Biotechnology and bioengineering*, 120(7), 1882–1890. <https://doi.org/10.1002/bit.28386>
- [29] Hershewe, J. M., Warfel, K. F., Iyer, S. M., Peruzzi, J. A., Sullivan, C. J., Roth, E. W., DeLisa, M. P., Kamat, N. P., & Jewett, M. C. (2021). Improving cell-free glycoprotein synthesis by characterizing and enriching native membrane vesicles. *Nature communications*, 12(1), 2363. <https://doi.org/10.1038/s41467-021-22329-3>
- [30] Liu, WC., Gong, T., Wang, QH. *et al.* (2016). Scaling-up Fermentation of *Pichia pastoris* to demonstration-scale using new methanol-feeding strategy and increased air pressure instead of pure oxygen supplement. *Sci Rep* 6, 18439 (2016). <https://doi.org/10.1038/srep18439>

A Comparative Study of the Pure Gas Permeability of PIM-1 Membranes and Other Polymers for CO₂ Separation

Amanda Wong and Joella Diong

Department of Chemical Engineering, Imperial College London, U.K.

Abstract

Membrane-based separation technologies provide a less energy-intensive alternative to amine scrubbing for carbon capture applications. Polymers of Intrinsic Microporosity (PIMs), PIM-1 in particular, are a promising class of glassy polymeric membrane material due to its thermal stability and high intrinsic free volume, redefining the Robeson upper bound in 2008. However, PIM-1 membranes suffer from physical aging, and exhibit high permeabilities but moderate selectivities. The separation performance of PIM-1 membranes can be enhanced through alcohol soaking treatment and the functionalisation of the nitrile group into amidoxime. Here we show that methanol-soaked PIM-1 membranes exhibit higher permeabilities as compared to as-cast PIM-1 membranes. AO-PIM-1 membranes display improved CO₂/N₂ and CO₂/CH₄ selectivities but at the expense of lower permeabilities. A comparison between the rubbery PolyActive™ membrane and the glassy PIM-1 membrane confirms that PIM-1 membranes typically show higher permeabilities than their rubbery counterparts due to higher solubility coefficients arising from the non-equilibrium excess free volume as reported in literature. Methanol soaking increases the permeability of PIM-1 membranes due to higher free volume caused by the swelling of polymer chains and the removal of trapped residual solvents. Amidoxime-functionalisation increases the selectivity but decreases the permeability of PIM-1 membranes due to tightening of the microstructure. Our results demonstrate that PIM-1 membranes are well-suited for carbon capture applications. However, several challenges, such as physical aging and the lack of reproducibility and processability in scale-up, have yet to be addressed for the commercialisation of PIM-1 membranes.

Keywords: polymer membranes, CO₂ separation, PIM-1, PolyActive™, methanol soaking, amidoxime-functionalisation

Introduction

To meet the 1.5°C threshold for global warming set out by the Intergovernmental Panel for Climate Change, increasing emphasis is placed on the development of post-combustion carbon capture technologies. Amine scrubbing is currently the most robust technology for carbon capture from flue gas, though it is quite energy intensive. Membrane-based separation poses a promising alternative to chemical absorption for CO₂ sequestration and can reduce heating requirements by operating at ambient temperatures [1].

This paper focuses on the fabrication of dense film polymer membranes and compares the separation performances of rubbery and glassy polymer membranes, as represented by PolyActive™ and PIM-1 membranes respectively. Optimisation of the performance of PIM-1 membranes was further explored by investigating the effect of methanol soaking on the pre-aging separation performance of PIM-1 membranes, and the effect of the functionalisation of PIMs through a comparison between PIM-1 and AO-PIM-1 membranes.

Background

Upper Bound Relationship

The key mass transport parameters of gas separation membranes include diffusivity, solubility, permeability, and permselectivity. The relationship between diffusivity and solubility can be expressed by

$$P_i = D_i * S_i \quad \text{E.1}$$

Where P_i , D_i and S_i are the permeability, diffusion coefficient, and solubility coefficient of diffusing species i respectively [2].

The permselectivity of a membrane is defined as the ratio of the permeabilities of two diffusing species.

$$\alpha_{i/j} = \frac{P_i}{P_j} = \frac{D_i}{D_j} * \frac{S_i}{S_j} \quad \text{E.2}$$

Where $\alpha_{i/j}$ is the permselectivity of the membrane, P_x , D_x and S_x are the respective permeability, diffusion coefficient and solubility coefficient of diffusing species $x = \{i, j\}$, with species i being the more permeable component out of the gas pair i and j [2]. Using equation 1, permselectivity can be split into contributions from diffusion selectivity (D_i/D_j) and solubility selectivity (S_i/S_j).

There exists a trade-off between the permeability and permselectivity of a membrane. The limit to the highest combinations of permeability and permselectivity of known polymers is referred to as the “upper bound” and is shown by a linear log-log plot of pure gas permselectivity ($\alpha_{i/j}$) versus permeability (P_i) in figure 1. The upper bound relationship is primarily determined by the diffusion selectivity, with a modest contribution from solubility selectivity. As permeability increases, free volume increases, hence decreasing the diffusion selectivity. At the same time, when the diameter of gas j is larger than that of i , more sorption sites become available for the larger gas as free volume increases, hence decreasing the solubility selectivity [3]. The contributions of diffusion selectivity and solubility selectivity to the permselectivity-permeability trade-off are shown in figure 1.

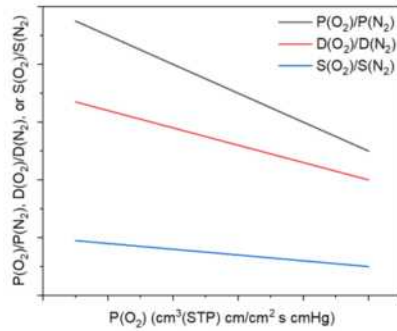


Fig. 1: Graph showing the solubility selectivity, $S(O_2)/S(N_2)$, and diffusion selectivity, $D(O_2)/D(N_2)$, contributions to the upper bound permselectivity trade-off, $P(O_2)/P(N_2)$. The trend of permselectivity is shown in black, diffusion selectivity is shown in red and solubility selectivity is shown in blue.

Rubbery and Glassy Polymers

Polymer membranes can be classified into two main categories, rubbery polymers and glassy polymers. Rubbery polymers are amorphous polymeric materials that are above the glass transition temperature (T_g). They are characterised by weaker intermolecular forces and more flexible molecular chains, and generally exhibit higher fluxes but lower permselectivities. The mass transport mechanism in rubbery polymers can be described by the solution-diffusion model and modelled using Fick's law of diffusion. The concentration of soluble gases in the rubbery polymer membrane follows Henry's law, exhibiting a linear dependence on penetrant pressure.

$$C = S * p \quad \text{E.3}$$

Where C is the concentration, S is the solubility and p is the pressure [4].

An example of a rubbery polymeric membrane material is the commercialised PolyActive™, which is a polyether oxide (PEO) based copolymer with a chemical structure as shown in figure 2 [1].

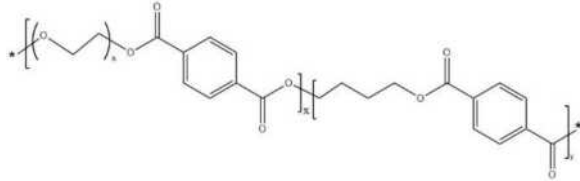


Fig. 2: Schematic showing the molecular structure of PolyActive™.

On the other hand, glassy polymers are amorphous polymeric materials that are below the glass transition temperature (T_g). Due to the non-equilibrium nature of glassy polymers, they are characterised by rigid structures with restricted molecular chain motion and excess free volume, hence typically exhibit lower fluxes but higher permselectivities. The difference in specific volume between rubbery and glassy polymers as a function of temperature is summarised in figure 3.

The mass transport mechanism in glassy polymers can be described by the nonlinear dual-sorption model [2].

$$C = C_H + C_D = \frac{C_H' b p}{1 + b p} + k_D p \quad \text{E.4}$$

Where C is the concentration, C_H and C_D are the contributions from Langmuir sorption and Henry's law sorption respectively, C_H' is the Langmuir saturation

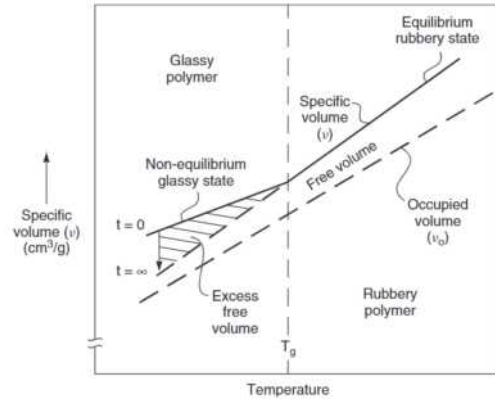


Fig. 3: Schematic of specific volume of rubbery and glassy polymers as a function of temperature. The shaded region shows the non-equilibrium excess free volume in glassy polymers. Taken from [5].

constant, b is the Langmuir affinity constant, k_D is the Henry's law solubility constant, and p is the pressure.

The Henry's law sorption contribution is identical to that of rubbery polymers, whereas the Langmuir sorption contribution can be described as a "hole-filling" process due to the excess free volume in glassy polymers. It is assumed that the gas adsorbed on the Langmuir sorption sites is immobilized, with gas diffusing through the Henry's sorption sites only. Hence, the Langmuir sorption mode dominates at low pressure and the Henry's law sorption mode dominates at higher pressures as the Langmuir sorption sites become occupied as shown in figure 4.

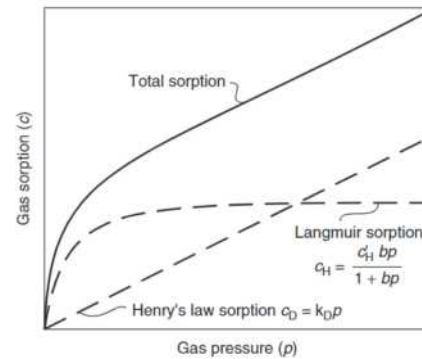


Fig. 4: Graph showing the Langmuir and Henry's Law (dashed lines) contributions to total gas sorption (solid line) representing the dual sorption model. Taken from [5].

The upper bound is dominated by glassy polymers primarily due to higher solubility coefficients than their rubbery counterparts, with a modest contribution from diffusion selectivity [6]. The excess free volume in glassy polymers leads to higher solubility coefficients, shifting the upper bound curve to the right for glassy polymers. It is also observed that at equal permeability, glassy polymers possess higher solubility coefficients and lower diffusion coefficients than rubbery polymers.

Out of all glassy polymers, Polymers of Intrinsic Microporosity (PIMs) are regarded as a promising class of polymeric material for high-performance gas separation membranes, owing to its higher intrinsic free volume as compared to other glassy polymers [7].

Polymers of Intrinsic Microporosity (PIMs)

Polymers of Intrinsic Microporosity (PIMs) were first discovered in 2003 as an attempt to design organic materials with an activated carbon-like structure but with higher surface. The microporosity is due to the rigid and contorted structure of the polymer, which prevents an efficient packing structure upon solidifying and creates free volume. Hence, the microporosity is due to the intrinsic molecular structure rather than the processing technique [8].

Generally, the macromolecular structure of PIMs consists of a structural unit which is responsible for the contorted shape of the polymer chain, and a linking group which joins together structural units and prevents their rotation about each other [9]. The structural unit of PIM-1 is spirobisindane (SBI). This first group of SBI-based PIMs showed high gas permeability and moderate permselectivity to CO₂ and in 2008 redefined the upper bound [10]. Other structural units such as triptycene (Trip), benzotriptycene (BTrip) and Tröger's base (TB) have also been used and have demonstrated improved permeabilities and selectivities, further establishing new upper bounds in 2015 [11] and 2019 [12] shown in figure 5.

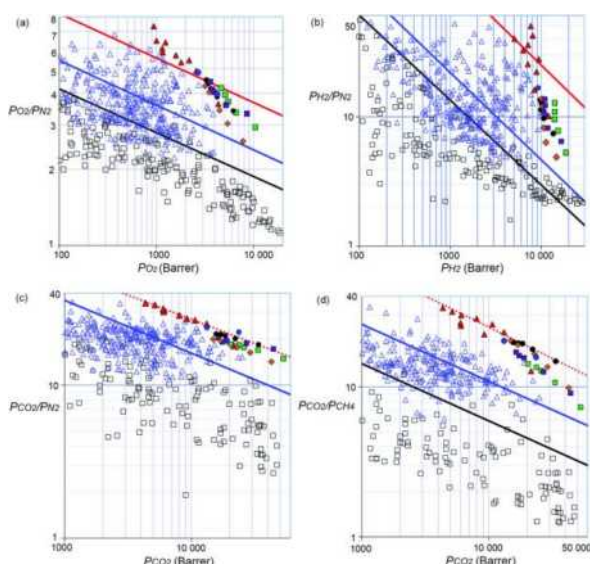


Fig. 5 Robeson graphs showing gas permeability and selectivity data of various membranes which established the 2019 upper bounds (red line) taken from [12].

PIM-1 is the most common PIM used as the selective layer in Thin Film Composite membranes [1]. Despite the moderate selectivity of PIM-1, it remains of high industrial interest for CO₂ separation due to its high gas permeability, thermal stability, and solubility in common solvents [9].

Research and development to improve the selectivity of PIM-1 membranes focuses on the modification of the intrinsic structure, for example, by incorporating porous nanofillers into the PIM-1 matrix [13], functionalisation of nitrile (-CN) groups in the polymer backbone [14], self-crosslinking [15], or UV treatment [16].

Functionalisation of PIM-1, often via the conversion of the nitrile (-CN) group, generally improves selectivity but decreases permeability.

Carboxylate-functionalised PIMs have shown significant improvements in CO₂/N₂ and CO₂/CH₄ selectivities by 267% and 129% respectively [17]. The enhanced selectivity may be attributed to both the CO₂-philic nature of the polar functional groups and their increased intermolecular attraction which tightens the microporous structure. However, the increased attraction between neighbouring chains has shown to decrease total surface area and permeability [14].

AO-PIM-1 is an amidoxime-functionalised PIM-1 with a chemical structure as shown in figure 6. Pure gas testing of AO-PIM-1 has shown an increase in CO₂/N₂ selectivity by 45% and in CO₂/CH₄ selectivity by 113%, but a decrease in CO₂ permeability to 1153 Barrer [18]. Both Langmuir and BET models show a lower apparent surface area for AO-PIM-1 in comparison to PIM-1 but a higher micropore surface area, which may be credited to the conversion of mesopores to micropores due to intermolecular hydrogen bonding [19].

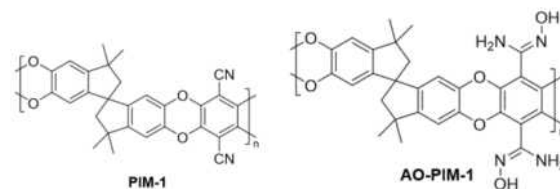


Fig. 6: Schematic showing the molecular structures of PIM-1 (left) and AO-PIM-1 (right).

Challenges

Several challenges hinder the wide commercialisation of PIM membranes for post-combustion carbon capture, including the lack of reproducibility and processability in scale-up, membrane plasticization and physical aging.

PIM-1 is insoluble in various polar solvents including Dimethylacetamide (DMAc), making it more challenging to process into hollow fiber membranes, which is the industrially preferred module configuration due to the higher packing density and surface-to-volume ratio [20]. Efforts have been made to increase the solubility of PIM-1 in polar solvents through the functionalisation of PIMs, but have resulted in a loss in permeability. Many novel membrane materials lack reproducibility due to the presence of unavoidable defects, the irregular packing of polymer chains, and the processing history dependence of glassy polymers [1]. Techniques used for casting membranes at lab or pilot scale may also not be viable at industrial scale. Moreover, limited studies have been done to assess the performance of membrane materials at flue gas compositions instead of single gas or simple mixed gas conditions, resulting in uncertainties in the commercialisation of such membrane materials.

Plasticisation refers to the irreversible swelling of polymer membranes due to the presence of highly-sorbing penetrants (e.g. CO₂), leading to a loss in selectivity, and poses an issue in post-combustion carbon capture applications and high-pressure operating conditions [20]. Common approaches used to minimise the effects of plasticisation include chemical cross-linking and sub-T_g annealing [21].

Physical aging is defined as the loss in non-equilibrium free volume in glassy polymers over time, resulting in a decrease in permeability. PIM-1 membranes exhibit a dual-aging mechanism, with the effects of lattice contraction dominating over the diffusion of pores to the surface of the membrane [22]. The rate of physical aging is both thickness and time dependent. Aging rates are higher in thinner films, and the rate of physical aging declines with time as the driving force of physical aging (i.e., the difference between the actual and equilibrium free volume) decreases [2] [23]. Physical aging can be mitigated through methanol soaking, thermal cross-linking, ultraviolet treatment, or the addition of fillers to construct mixed-matrix membranes [23].

Alcohol Soaking Treatment

Immersing PIM-1 membranes in alcohols, such as methanol, can increase the permeability of the membrane pre-aging, as well as reverse the effects of physical aging, as shown in figure 7. This can be attributed to the swelling of polymer chains via plasticisation leading to an increase in free volume [22], and the removal of past processing history and trapped residual solvents which may obstruct gas transport through the membrane. However, such alcohol-soaking method requires the dismantling of gas separation membrane modules and is not practical at an industrial scale. Almansour et al [23] have developed an alternative rejuvenation approach that involves exposing aged PIM-1 films to alcohol vapor, hence eliminating the need for membrane dismantling. Traces of alcohol could also be easily removed by passing inert gas or air through the membrane module following the vapor treatment, minimizing the downtime of the equipment.

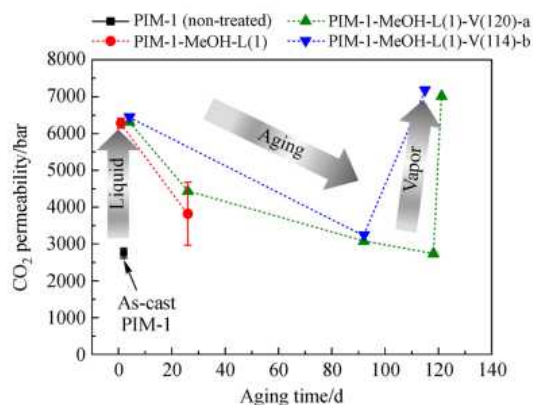


Fig. 7: Graph of CO₂ permeability as a function of aging time in methanol-soaked and unsoaked PIM-1 membranes taken from [23]. The black square represents the as-cast PIM-1 membrane, and the red circles represent the methanol-soaked PIM-1 membrane. The green triangles and the blue inverted triangles represent PIM-1 membranes that were soaked in methanol post-casting, then treated with methanol vapor post-aging. The numbers in the brackets following “L” and “V” denote the day of liquid alcohol soaking and the day of vapor alcohol treatment respectively, with the day following the collection of the cast membrane from the petri dish as day 1.

Thin Film Composite Membranes

Thin film composite (TFC) membranes are composed of four layers: a support layer, a gutter layer to smoothen the surface of the support layer and prevent the selective

layer from penetrating into the support layer, a thin selective layer to separate feed components, and a protective layer. TFC membranes are more suited for industrial applications, as the material used in each layer can be different and optimised independently to provide the desired functionality, driving the performance of TFC membranes towards the upper bound [1]. State-of-the-art TFC membranes made up of PIM-1 show a CO₂ permeance of 3200 GPU, with CO₂/N₂ and CO₂/CH₄ selectivities of 64 and 45 respectively [24]. However, for ease of comparison of the separation performances of various selective membrane materials, dense film membranes were used in this study.

Methods

Membrane Fabrication

To prepare PIM-1 and PolyActive™ membranes, 2 wt% solutions were prepared by dissolving the polymer in Chloroform. The composition of the PolyActive™ pellets used was 1500PEOT77PBT23 (1500 g/mol, 77 wt% PEO, 23 wt% PBT). Solutions were mixed overnight using a magnetic stirrer and centrifuged at 14000 rpm for 10 minutes to remove remaining impurities. Solution was then cast onto a levelled 7 cm glass petri dish and left overnight for the solvent to evaporate in a chloroform atmosphere at room temperature. A dense membrane was then obtained.

To prepare the AO-PIM-1 membrane, 3 wt% AO-PIM-1 in N,N-dimethylformamide (DMF) solution was filtered through a 0.45 µm PTFE filter, to remove possible dust or polymer gel particles, onto a levelled glass petri dish. After slow evaporation of the solvent at 60 °C, a dense membrane was formed.

The thicknesses of the dense films were measured using an electric micrometre.

Methanol Soaking

Post-casting treatment of PIM-1 and AO-PIM-1 membranes were carried out by soaking the dry membranes in methanol for 24h at ambient conditions, then leaving to air dry for 3 days before masking.

Membrane Masking

The membranes were cut using the MTI Precision Disc Cutter and masked using aluminium tape and epoxy glue. The active surface area of the membrane was measured using ImageJ software.

Membrane Testing

The membranes were tested using a single-gas constant-volume variable-pressure method as shown in figure 8. Tests were performed at 35°C, and the feed pressure for testing PolyActive™, PIM-1, and AO-PIM-1 membranes were 1.5 bar, 2 bar and 2.3 bar respectively. Masked membranes were vacuumed in the testing cell for 1h, or longer if necessary, before each run. The gases were tested in the following order: H₂, O₂, CO₂, N₂, CH₄.

The rate of increase in pressure on the permeate side was analysed based on the time-lag method to assess the membrane performance for each gas using the following equations:

$$P = \frac{V * l}{A} \frac{T_0}{p_f p_0 T} \left(\frac{dp}{dt} \right) \quad \text{E. 5}$$

Where P is the permeability of the testing gas in Barrer (1 Barrer = $10^{-10} \text{ cm}^3(\text{STP}) \text{ cm cm}^{-2} \text{ s}^{-1} \text{ cmHg}^{-1}$), V is the volume of the permeate in cm^3 , l is the thickness of the membrane in cm, A is the effective area of the membrane in cm^2 , p_f is the feed pressure in cmHg, p_0 is the pressure at standard state in cmHg, T is the operating temperature in K, T_0 is the temperature at standard state in K, dp/dt is the slope of the increase in pressure on the permeate side at pseudo-steady state in cmHg s^{-1} , as calculated by computational linear fitting of the results. The standard state pressure and temperature were taken to be 76 cmHg and 273.15K respectively.

$$D = \frac{l^2}{6\theta} \quad \text{E. 6}$$

Where D is the diffusion coefficient of the testing gas in $\text{cm}^2 \text{ s}^{-1}$, l is the thickness of the membrane in cm, θ is the time lag in s. The time lag θ can be determined from the intercept of the steady-state linear region of the pressure-time plot with the time axis.

Using the relationship between solubility and diffusivity in equation 1, solubility (in $\text{cm}^3 \text{ cm}^{-3} \text{ cmHg}$) can be back-calculated from the experimentally determined values of permeability and diffusivity. The selectivity of the membrane towards different gas pairs can also be found using equation 2.

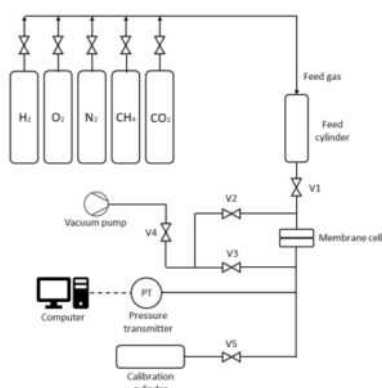


Fig. 8: Schematic of the testing rig for constant-volume variable-pressure method.

Results and Discussion

Gas Flux Measurements

Figure 11 shows the permeability of pure gases through the respective membranes as the rate of pressure change (dp/dt) is proportional to the permeability. The order of permeability of the gases for all the membranes is $\text{CO}_2 > \text{H}_2 > \text{O}_2 > \text{CH}_4 > \text{N}_2$.

Solubility Analysis

Solubility coefficient is a thermodynamic parameter that is dependent on the free volume distribution and interactive forces between the penetrant and polymer, which leads to the sorption of penetrants onto the

membrane [22]. From figure 9, the solubility of a penetrant generally shows a positive correlation with its critical temperature. The higher the critical temperature, the more condensable the gas is, and hence the higher the solubility.

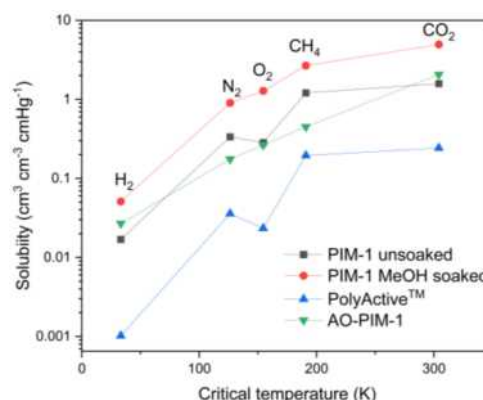


Fig. 9: Graph showing the solubility of gases (left to right) H_2 , N_2 , O_2 , CH_4 , CO_2 against critical temperature. Unsoaked PIM-1 membrane is represented by black squares, methanol-soaked PIM-1 membrane is represented by red circles, PolyActiveTM membrane is represented by blue triangles and AO-PIM-1 membrane is represented by inverted green triangles.

Diffusivity Analysis

Diffusion coefficient is a kinetic parameter that reflects the restrictions of the surrounding medium on the diffusing species [5]. It is related to the size of the penetrant, chain mobility, packing density and fractional free volume of the polymer [22]. From figure 10, the diffusivity of a penetrant generally shows a negative correlation with its effective diameter. The higher the effective diameter, the lower the diffusivity, indicating a size sieving effect of the pores in the polymer. This is because larger gases with higher effective diameters exhibit more interactions with the polymer chains than smaller gases, leading to lower mobility [25]. The same trend is observed for all polymer membranes tested.

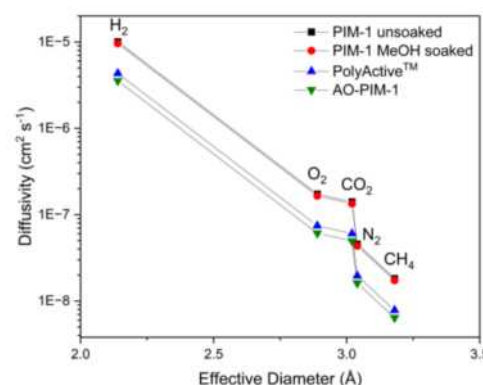


Fig. 10: Graph showing the diffusivity of gases (left to right) H_2 , O_2 , CO_2 , N_2 , CH_4 against effective diameter. Unsoaked PIM-1 membrane is represented by black squares, methanol-soaked PIM-1 membrane is represented by red circles, PolyActiveTM membrane is represented by blue triangles and AO-PIM-1 membrane is represented by inverted green triangles.

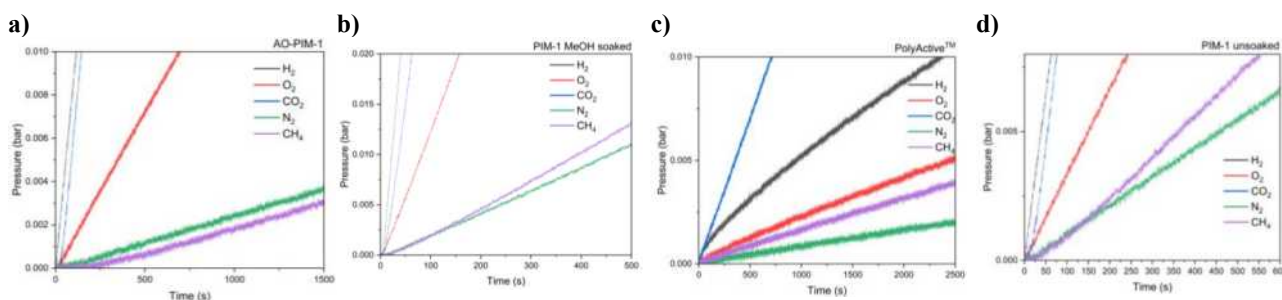


Fig. 11: Graphs showing the permeate-side pressure against time for the different feed gases H₂ (black), O₂ (red), CO₂ (blue), N₂ (green), CH₄ (purple) for membranes made from a) AO-PIM-1, b) methanol-soaked PIM-1, c) PolyActive™ and d) unsoaked PIM-1.

Permeability Analysis

According to the solution-diffusion model, permeability is a function of the solubility coefficient and the diffusion coefficient. The contributions of solubility and diffusivity to permeability are shown in figure 12 respectively. Kinetic diameters are related to the mean free paths of molecules, taking into account both molecular size and structure [25]. It can be observed that the relatively high permeability for H₂ is not because of solubility, but is instead attributed to a higher diffusion coefficient due to H₂ having the smallest kinetic diameter out of all the gases tested.

Permeability generally decreases with the kinetic diameter of the penetrant gas due to decreasing diffusivities, since molecular sieving effects are greater for larger penetrants. On the contrary, solubility increases with kinetic diameter except for the case of CO₂ which is highly condensable. This matches the positive correlation observed between the critical temperature of the gases and the solubility in figure 9, as molecules with higher kinetic diameters are typically more condensable [25]. The opposing effects of kinetic diameter on diffusivity and solubility suggest that diffusion is the dominating mechanism in gas permeability through polymer membranes, with CO₂ being the exception where solubility effect dominates.

The permeability of CO₂ is the highest among all gases for all membranes tested, though the distinction is less pronounced for PolyActive™ membranes. The selectivity for CO₂ is dominated by solubility selectivity and less so by diffusion selectivity, as various functional groups in the polymeric materials display an affinity to CO₂, resulting in higher solubilities. The polyethylene oxide group in PolyActive™ and the aryl ether linkage in PIM-1 exhibit favourable interactions and a strong affinity to CO₂ as compared to other gases [1] [10]. In AO-PIM-1, the nitrile group in PIM-1 is converted to a highly basic and polar amidoxime (AO) group, introducing CO₂-philic characteristics to the functionalised polymer [18].

Comparison of PolyActive™ and Unsoaked PIM-1 Membranes

From figure 12, PolyActive™ membranes show a lower permeability than PIM-1 membranes due to both lower solubility coefficients and lower diffusion coefficients. PIM-1 is a type of glassy polymer with intrinsic microporosities and a rigid, contorted structure due to restricted rotational motion along the polymer backbone and inefficient packing of aromatic rings in the polymer chain. This leads to high porosity and fractional free volume in PIM-1 molecules, contributing to high solubilities [20].

PolyActive™ is a rubbery polyether oxide (PEO) based block copolymer. The amorphous phase is made up of PEO blocks, whereas the crystalline phase is made up of poly(butylene terephthalate) (PBT) blocks [26]. Gas is transported through the amorphous phase, whereas the crystalline phase provides mechanical stability and structural rigidity but does not contribute to permeability [27]. The semi-crystalline structure of PolyActive™, combined with the strong crystallisation tendency and inherently low CO₂ permeability of PEO itself [1], results in a lower overall permeability of PolyActive™ as compared to the highly porous PIM-1, and is particularly evident for CO₂.

It can be concluded that glassy PIM-1 membranes typically show higher permeabilities than rubbery polymer membranes due to higher solubility coefficients arising from the non-equilibrium excess free volume.

Effects of Methanol Soaking on PIM-1 Membranes

As-cast PIM-1 membranes were soaked in methanol and the effects of the alcohol treatment on the separation performance of PIM-1 membranes were evaluated. From figure 12, soaking the membrane in methanol has shown to increase solubility and hence permeability, but has no effect on diffusivity.

Such phenomenon could be explained by two possible hypotheses: Methanol swells up the rigid polymer chains in PIM-1, resulting in an increase in free

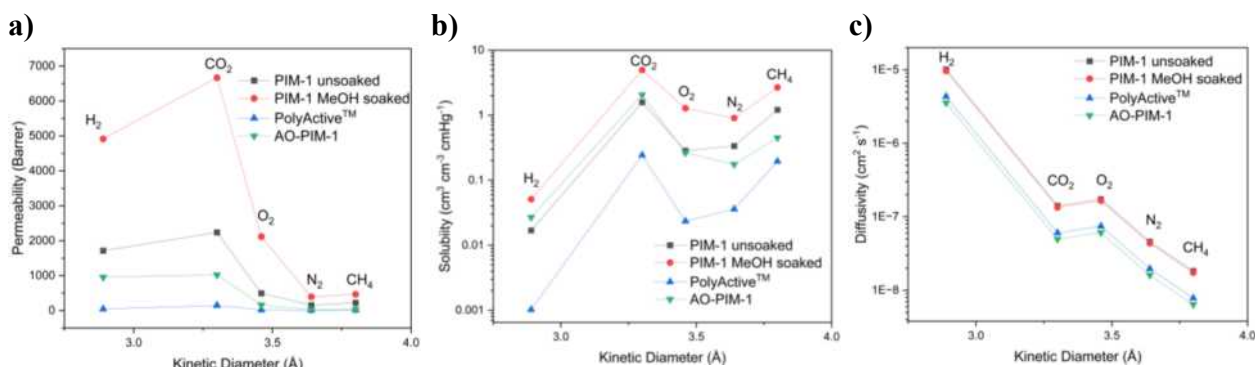


Fig. 12: Graphs showing the a) permeability, b) solubility, c) diffusivity of gases (left to right) H_2 , CO_2 , O_2 , N_2 , CH_4 against kinetic diameter. Unsoaked PIM-1 membrane is represented by black squares, methanol-soaked PIM-1 membrane is represented by red circles, PolyActive™ membrane is represented by blue triangles, and AO-PIM-1 membrane is represented by inverted green triangles.

volume and solubility coefficients [23]; Past processing history and traces of trapped residual casting solvents, which may hinder gas transport in as-cast PIM-1 membranes, are also removed through the methanol-soaking treatment [22].

Thus, due to the high intrinsic free volume of PIM-1 and the positive effects of methanol soaking, the soaked PIM-1 membrane exhibited the best overall performance in terms of permeability among all membranes tested.

Evaluation of Amidoxime-Functionalised PIM-1

AO-PIM-1 is an amidoxime-functionalised PIM-1 synthesized through the post-polymerisation modification of the nitrile group in PIM-1 [18]. Comparing the separation performance of AO-PIM-1 membrane against the methanol-soaked PIM-1 membrane in figure 12, AO-PIM-1 displayed a lower permeability than PIM-1 due to both lower solubility coefficients and lower diffusion coefficients.

By converting the nitrile group in PIM-1 into a polar and highly basic amidoxime functional group, the microstructure of the polymer is tightened due to the extensive intermolecular hydrogen bonding between amidoxime moieties [18]. This decreases the surface

area of the polymer, leading to a lower sorption capacity and hence lower solubility of AO-PIM-1. At the same time, the reduced free volume within the polymer leads to higher tortuosity and diffusional restrictions, thus lowering the diffusivity of AO-PIM-1.

Selectivity Analysis and Upper Bound Plot

PIMs tested in this work generally have a high CO_2 permeability which are above current industrially used rubbery polymers like Pebax [28], PolyActive™ and polyvinyl acetate (PVAc) [29] as shown in figure 13. This is due to glassy polymers like PIMs having higher excess free volume than rubbery polymers, which increases their solubility coefficient and therefore their permeability [2].

However, the unsoaked PIM-1 displays a moderate selectivity which is very similar to PolyActive™ for CO_2/CH_4 and much lower than PolyActive™ for CO_2/N_2 selectivity. This suggests that PolyActive™ is more size selective than PIM-1, which may be explained by the molecular sieving effect of the pores.

Methanol soaking of PIM-1 membranes increases CO_2 permeability, as well as the CO_2/N_2 and CO_2/CH_4 selectivities.

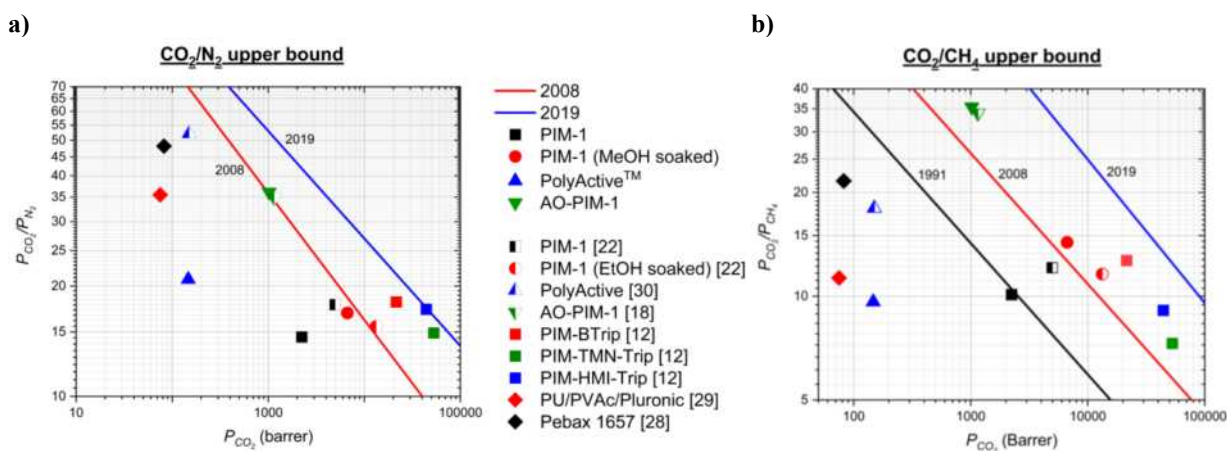


Fig. 13: Upper bound plots of CO_2 selectivity with a) nitrogen and b) methane against CO_2 permeability for tested membranes of unsoaked PIM-1 (black square), methanol soaked PIM-1 (red circle), PolyActive™ (blue triangle) and AO-PIM-1 (inverted green triangle) compared to literature values (equivalent half-filled symbols), high performance BTrip PIMs (red, green and blue squares), and commercial rubbery polymers PVAc (red diamond) and Pebax (black diamond).

Table 1: H₂, O₂, CO₂, N₂, CH₄ pure gas permeabilities, and CO₂/N₂ and CO₂/CH₄ selectivities for the tested polymer membranes of PIM-1 (unsoaked), methanol-soaked PIM-1, PolyActive™ and AO-PIM-1, and various other polymers found in literature.

Polymer	Permeability				Selectivity		
	H ₂	O ₂	CO ₂	N ₂	CH ₄	CO ₂ /N ₂	CO ₂ /CH ₄
PIM-1	1714	494	2237	154	221	14.5	10.1
PIM-1 (MeOH soaked)	4915	2118	6667	395	465	16.9	14.3
PolyActive™	44	17	147	7	15	20.9	9.6
AO-PIM-1	955	161	1021	28	29	36.1	35.4
Unsoaked PIM-1 [22]	1740	742	4970	280	410	17.8	12.1
PIM-1 (EtOH soaked) [22]	4500	2200	13300	857	1150	15.5	11.6
PolyActive™ [30]	14	-	150	3	8	52.1	17.9
AO-PIM-1 [18]	912	417	1153	33	34	34.9	33.9
Pebax 1657 [28]	-	-	82.15	1.70	3.80	48.2	21.6
PIM-Btrip [12]	12100	4330	21500	1190	1690	18.1	12.7
PIM-TMN-Trip [12]	18800	10400	52800	3540	7250	14.9	7.3
PIM-HMI-Trip [12]	16600	8540	44200	2560	4870	17.3	9.1
PU/PVAc/Pluronic [29]	-	6	75	2	7	35.5	11.3

A comparison of PIM-1 and AO-PIM-1 shows the expected permeability/selectivity trade-off as AO-PIM-1 displays a much higher selectivity but lower permeability. AO-PIM-1 is more size selective, which may be due to AO-PIM-1 having more micropores compared to PIM-1. However, this also decreases the overall surface area and free volume of the polymer, which decreases the permeability. The increased presence of micropores is due to intermolecular hydrogen bonding which tightens the microstructure, hindering the transport of larger molecules, CH₄ and N₂, hence increasing selectivity [18].

From figure 13, while PIM-1 and AO-PIM-1 are closer to the 2019 upper bound than commercially used membranes like PVAc and PolyActive™, they are farther away than the BTrip based PIMs which established the 2019 upper bound itself [12]. The upper bound is dominated by high free volume glassy polymers due to their high solubility coefficients and permeability. Among all the polymers presented, it is observed that the range of permeabilities is much greater than the range of selectivities. This is because the permeability of a material can be changed by altering its chemical structure; however, changes in chemical structure which change the transport properties of one gas will also typically change the transport properties of other gases in the same way [5], resulting in lesser changes in selectivity.

Conclusions

The separation performance of PIM-1 membranes for carbon capture applications have been evaluated and compared against commercialised rubbery polymers (PolyActive™). The effects of methanol-soaking and amidoxime-functionalisation of PIM-1 membranes have also been assessed.

PolyActive™ membranes exhibited a much lower permeability of over 15 times as compared to unsoaked PIM-1 membranes. Permeability, solubility, and diffusion data obtained for PolyActive™ membranes demonstrated that lower solubility coefficients of rubbery polymer membranes contribute to their lower permeabilities. Methanol-soaking of PIM-1 membranes has been shown to increase CO₂

permeabilities by almost two-fold. AO-PIM-1 membranes have also shown improvements in CO₂/N₂ and CO₂/CH₄ selectivities by over 100% but with decreased permeabilities due to tightening of the microstructure.

Despite the potential of PIM-1 as a high-performance gas separation polymeric material, issues such as physical aging and the lack of reproducibility and processability in scale-up present challenges for the commercialisation of PIM-1 membranes. Future research could be done on the rejuvenation of free volume in PIM-1 membranes post-aging through methanol soaking treatment. Tests could also be conducted at simple mixed gas conditions and with gas mixtures of compositions comparable to that in flue gases, in order to assess the separation performance of PIM-1 membranes more accurately for industrial post-combustion carbon capture applications. Thin Film Composite (TFC) membranes present a viable solution to enhance permeability without compromising selectivity. Gas separation tests of TFC membranes with PIM-1 as the selective layer could be conducted and their performance assessed.

Acknowledgement

The authors would like to express their sincere gratitude to Sunshine Iguodala for her patience and guidance throughout the project, Naiqi Meng for his help in the lab, and everyone in the Song group for their support and encouragement.

References

- [1] M. Liu, M. D. Nothling, S. Zhang, Q. Fu and G. G. Qiao, "Thin film composite membranes for postcombustion carbon capture: polymers and beyond," vol. 126, 2022.
- [2] M. M. Merrick, R. Sujunani and B. D. Freeman, "Glassy polymers: historical findings, membrane applications, and unresolved questions regarding physical aging," vol. 211, no. 123176, 2020.
- [3] L. M. Robeson, Z. P. Smith, B. D. Freeman and D. R. Paul, "Contributions of diffusion and

- solubility selectivity to upper bound analysis for glassy gas separation membranes,” vol. 453, 2014.
- [4] S. Wiederhorn, R. Fields, S. Low, G.-W. Bahng, A. Wehrstedt, J. Hahn, Y. Tomota, T. Miyata, H. Lin, B. Freeman, S. Aihara, Y. Hagihara and T. Tagawa, “Mechanical Properties,” in *Springer Handbook of Materials Measurement Methods*, Berlin, Springer, 2006.
- [5] R. W. Baker, *Membrane Technology and Applications*, 3rd ed., J. Chichester, Ed., Wiley & Sons, 2012.
- [6] L. M. Robeson, Q. Liu, B. D. Freeman and D. R. Paul, “Comparison of transport properties of rubbery and glassy polymers and the relevance to the upper bound relationship,” vol. 476, pp. 421-431, 2015.
- [7] Y. Wang, B. S. Ghanem, Y. Han and I. Pinnau, “State-of-the-art polymers of intrinsic microporosity for high-performance gas separation membranes,” vol. 35, no. 100755, 2022.
- [8] P. M. Budd, B. S. Ghanem, S. Makhseed, N. B. McKeown, K. J. Msayib and C. E. Tattershall, “Polymers of intrinsic microporosity (PIMs): robust, solution processable, organic nanoporous materials,” *Chemical Communications*, no. 2, pp. 230-231, 2004.
- [9] N. B. McKeown, “Polymers of Intrinsic Microporosity (PIMs),” vol. 202, no. 122736, 2020.
- [10] L. M. Robeson, “The upper bound revisited,” vol. 320, no. 1-2, 2008.
- [11] R. Swaidan, B. Ghanem and I. Pinnau, “Fine-Tuned Intrinsically Ultramicroporous Polymers Redefine the Permeability/Selectivity Upper Bounds of Membrane-Based Air and Hydrogen Separations,” vol. 4, no. 9, 2015.
- [12] B. Comesaña-Gándara, J. Chen, C. G. Bezzu, M. Carta, I. Rose, M.-C. Ferrarari, E. Esposito, A. Fuoco, J. C. Jansen and N. B. McKeown, “Redefining the Robeson upper bounds for CO₂/CH₄ and CO₂/N₂ separations using a series of ultrapermeable benzotriptycene-based polymers of intrinsic microporosity,” vol. 12, 2019.
- [13] Wang, Chenhui; Guo, Fangyuan; Li, He; Xu, Jian; Hu, Jun; Liu, Honglai; Wang, Meihong, “A porous ionic polymer bionic carrier in a mixed matrix membrane for facilitating selective CO₂ permeability,” vol. 598, no. 117677, 2020.
- [14] S. He, B. Zhu, S. Li, Y. Zhang, X. Jiang, C. H. Lau and L. Shao, “Recent progress in PIM-1 based membranes for sustainable CO₂ separations: Polymer structure manipulation and mixed matrix membrane design,” vol. 284, no. 120277, 2022.
- [15] X. Chen, Z. Zhang, L. Wu, X. Liu, S. Xu, J. E. Efome, X. Zhang and N. Li, “Polymers of Intrinsic Microporosity Having Bulky Substitutes and Cross-Linking for Gas Separation Membranes,” vol. 2, no. 2, 2020.
- [16] F. Y. Li, Y. Xiao, Y. K. Ong and T.-S. Chung, “UV-Rearranged PIM-1 Polymeric Membranes for Advanced Hydrogen Purification and Production,” vol. 2, no. 12, pp. 1456-1466, 2012.
- [17] J. W. Jeon, D.-G. Kim, E.-h. Sohn, Y. Yoo, Y. S. Kim, B. G. Kim and J.-C. Lee, “Highly Carboxylate-Functionalized Polymers of Intrinsic Microporosity for CO₂-Selective Polymer Membranes,” vol. 50, no. 20, pp. 8019-8027, 2017.
- [18] R. Swaidan, B. S. Ghanem, E. Litwiller and I. Pinnau, “Pure- and mixed-gas CO₂/CH₄ separation properties of PIM-1 and an amidoxime-functionalized PIM-1,” vol. 457, pp. 95-102, 2014.
- [19] H. A. Patela and C. T. Yavuz, “Noninvasive functionalization of polymers of intrinsic microporosity for enhanced CO₂ capture,” vol. 48, no. 80, pp. 9989-9991, 2012.
- [20] C. Ma and J. J. Urban, “Polymers of Intrinsic Microporosity (PIMs) Gas Separation Membranes: A mini Review,” vol. 2, no. 02002, 2018.
- [21] R. Swaidan, B. Ghanem, E. Litwiller and I. Pinnau, “Physical Aging, Plasticization and Their Effects on Gas Permeation in “Rigid” Polymers of Intrinsic Microporosity,” vol. 48, no. 181, pp. 6553-6561, 2015.
- [22] P. Bernardo, F. Bazzarelli, F. Tasselli, G. Clarizia, C. Mason, L. Maynard-Atem, P. Budd, M. Lanč, K. Pilnáček, O. Vopička, K. Friess, D. Fritsch, Y. Yampolskii, V. Shantarovich and J. Jansen, “Effect of physical aging on the gas transport and sorption in PIM-1 membranes,” vol. 113, no. ISSN 0032-3861, pp. 283-294, 2017.
- [23] F. Almansour, M. Alberto, R. Bhavsar, X. Fan, P. Budd and P. Gorgojo, “Recovery of free volume in PIM-1 membranes through alcohol vapor treatment,” vol. 15.

- [24] M. Yu, A. B. Foster, M. Alshurafa, J. M. Luque-Alled, P. Gorgojo, S. E. Kentish, C. A. Scholes and P. M. Budd, "CO₂ separation using thin film composite membranes of acid-hydrolyzed PIM-1," vol. 679, no. 121697, 2023.
- [25] X. Chen, S. Kaliaguine and D. Rodrigue, "Correlation between Performances of Hollow Fibers and Flat Membranes for Gas Separation," vol. 47, 2017.
- [26] A. Car, C. Stropnik, W. Yave and K.-V. Peinemann, "Tailor-made Polymeric Membranes based on Segmented Block Copolymers for CO₂ Separation," vol. 18, no. 18, pp. 2815-2823, 2008.
- [27] G. Bengtson, S. Neumann and V. Filiz, "Membranes of Polymers of Intrinsic Microporosity (PIM-1) Modified by Poly(ethylene glycol)," vol. 7, pp. 1-21, 2017.
- [28] N. Habib, Z. Shamair, N. Tara, A.-S. Nizami, F. H. Akhtar, N. M. Ahmad, M. A. Gilani, M. R. Bilad and A. L. Khan, "Development of highly permeable and selective mixed matrix membranes based on Pebax®1657 and NOTT-300 for CO₂ capture," vol. 234, 2020.
- [29] M. A. Semsarzadeh and B. Ghalei, "Characterization and gas permeability of polyurethane and polyvinyl acetate blend membranes with polyethylene oxide–polypropylene oxide block copolymer," Vols. 401-402, 2012.
- [30] M. Karunakaran, R. Shevate, M. Kumar and K.-V. Peinemann, "CO₂-selective PEO–PBT (PolyActive™)/graphene oxide composite membranes," vol. 51, no. 14187-14190, 2015.
- [31] K. M. Rodriguez, S. Lin, A. X. Wu, G. Han, J. J. Teesdale, C. M. Doherty and Z. P. Smith, "Leveraging Free Volume Manipulation to Improve the Membrane Separation Performance of Amine-Functionalized PIM-1," vol. 60, no. 12, pp. 6593-6599, 2021.

Model-based Design Space and Flexibility Analysis for Carbon Capture Adsorbent Screening

Khushali Gosain and Emma Smith

Department of Chemical Engineering, Imperial College London, U.K.

Abstract

Carbon capture via pressure-vacuum swing adsorption (PVSA) shows great promise for industrialisation due to the low energy required for solvent regeneration compared to absorption. Zeolite 13X, UTSA-16 and CALF-20 are all adsorbents that show promise for the PVSA process. These three adsorbents are screened based on capture cost, energy usage and productivity, and importantly their process flexibility, using a model-based framework. Process flexibility is a key and novel metric to consider for industrialising this process and is defined as the amount of deviation from the nominal operating point that is allowed such that the purity and recovery constraints are met. A relaxation of the recovery constraint was also analysed to investigate how this affects process flexibility and the other performance indicators. UTSA-16 outperformed the other adsorbents with respect to its energy usage, productivity and most importantly, its process flexibility owing to its high CO₂ selectivity and linear CO₂ adsorption isotherm. Regarding Zeolite 13X and CALF-20, it is desirable to decrease the recovery constraint to increase flexibility and productivity while there were only marginal improvements for UTSA-16. The results from this study further emphasise the importance of selecting an adsorbent material based on process flexibility to commercialise the PVSA process.

Keywords: Pressure- vacuum swing adsorption, process flexibility, Zeolite 13X, UTSA-16, CALF-20

1 Introduction

The drive to decarbonise the energy sector and manufacturing industries have led to increasing research interest in post-combustion carbon capture technologies with an aim to bring into action over the coming decades (IEA, 2020). As renewable energy sources cannot match the electricity demand due to their intermittent nature and lack of installations, electricity production will need to rely on fossil fuel-powered plants for the foreseeable future. Post-combustion carbon capture allows for these plants to continue operating with a significant reduction in carbon emissions.

Whilst the traditional carbon capture process via amine absorption works effectively, pressure-vacuum swing adsorption (PVSA) shows great promise within the industry as compared to absorption, it requires less energy to regenerate the solvent (Lin, et al., 2021). In the process, flue gas is exposed to a fixed bed of solid adsorbent that has a higher adsorption affinity to CO₂ than N₂, allowing for CO₂ separation from the flue gas. Carbon sequestration can thus be carried out to store the captured CO₂.

This report investigates the process flexibility and operability of different adsorbent materials within the PVSA process by assessing key performance indicators (KPIs) (as seen in Table 1) based on different parameters specified within the adsorption column. Process flexibility is defined as the amount of deviation allowed from the nominal point such that process constraints are still met. For the PVSA process, these constraints are CO₂ purity and recovery since they determine if the CO₂ stream is suitable for geological storage. Process flexibility, which has largely been absent from adsorption and PVSA literature, is a key factor to consider within post-combustion carbon capture systems as potential upstream disturbances could result in the CO₂ purity or recovery violating environmental regulations. Once the PVSA process is deployed industrially, it will need to operate with a large

flexibility. It is hence important to consider this metric from the conceptual design stage, since varying input factors (such as operating pressures, column specifications, adsorbent material and cycle times) can affect flexibility (Ward & Pini, 2022).

2 Background

2.1 Project Background

The main model used within this project is derived from Ward and Pini's study on PVSA optimisation (Ward & Pini, 2022). The study develops a dynamic detailed model for a 1D adsorption column that identifies the operating point with the lowest capture cost and calculates other KPIs, as shown in Table 1 for a specified PVSA process.

Table 1: KPI Equations where n_i^j is the moles and m_i^j is the mass of species i in step j (ads, evac, pres). \dot{m}_{CO_2} is the mass flowrate of CO₂ captured [tonne/year]

Key Performance Indicator	Equation
Capture cost [\$/tonne]	$\frac{\text{Total annual cost}}{CO_{2, recovery} \times \dot{m}_{CO_2}}$
Energy usage [kWh/tonne]	$\frac{E_T}{m_{CO_2, out}^{evac}}$
CO ₂ productivity [mol/m ³ s]	$\frac{n_{CO_2, out}^{evac}}{V_{bed} t_{cycle}}$
CO ₂ purity [%]	$\frac{n_{CO_2, out}^{evac}}{n_{CO_2, out}^{evac} + n_{N_2, out}^{evac}}$
CO ₂ recovery [%]	$\frac{n_{CO_2, out}^{evac}}{n_{CO_2, in}^{pres} + n_{N_2, in}^{ads}}$

Energy usage can be seen as the total energy (E_T) used to capture one tonne of CO₂. It is important to keep this low as this energy would be taken from the power plant that the carbon capture process is attached to and would result in decreased efficiency and profitability of the electricity generated. Productivity is the amount of CO₂ captured per unit bed volume (V_{bed}) and per total cycle time (t_{cycle}). Maximising productivity increases CO₂

captured whilst also minimising the required volume of adsorbent and cycle time. A desirable PVSA process thus would have a high productivity, low capture cost and low energy usage.

Following their work, Sachio et. al developed a design space framework that quantifies the PVSA's process flexibility whilst achieving the process constraints. The operating point with the largest flexibility can therefore be identified and the trade-off between the KPIs and process flexibility can be explored (Sachio, et al., 2023a).

Utilising both the mathematical model as well as the design space framework, Purwanto's thesis investigated the separation performance and process flexibility of the adsorbents, Zeolite 13X and ZIF-36-FRL (Purwanto, 2023). Through his global sensitivity analysis on input parameters, it was concluded that the adsorbent material has a much more significant contribution to the adsorption's performance than other operational process parameters. Based on these results, this report will incorporate a process flexibility metric to explore two adsorbent materials that have not been previously studied within this workflow.

Within this workflow, CO₂ purity must exceed 95%, to meet environmental regulations set by the US Department of Energy (Alhajaj & Vega, 2021) and its recovery target is set at 89%, to ensure this process is economically worthwhile. It should be noted that the classical recovery constraint is 90% however in this report, the maximum constraint studied is 89% to validate the results with Purwanto's thesis (Purwanto, 2023).

To this end, this report provided design spaces for a range of adsorbents and compared their suitability for upscaling in industry based on their process flexibility and KPIs. The KPI values and flexibility of the cost optimal point and the most flexible point was also quantified to better understand how these two points differ. Secondly, a KPI analysis was performed for each adsorbent comparing across the design space and then for all adsorbents to explore the trade-offs between capture cost, productivity and energy usage. Lastly, analysis on the design space based on the recovery constraint was performed to evaluate its effect on the material performance and flexibility. This was conducted because recovery is a target value that is not rigorously set based on environmental regulations like the purity constraint. It is rather to ensure sufficient CO₂ is being captured from the flue gas and therefore will be varied from 89% to 85% to explore how it affects the adsorbent's process flexibility and KPIs.

2.2 PVSA Cycle

The PVSA process can be described in four steps.

1. **Adsorption (ads)** – The feed gas flows at a steady velocity v_F over the adsorbent bed at a high-pressure P_H . The adsorbent material adsorbs CO₂ on the surface and expels the N₂ rich product to the product end.
1. **Forward Blowdown (bd)** – The feed end of the column is closed, and the pressure is reduced to an intermediate pressure P_L . The N₂ rich effluent is then collected from the product

end of the column.

2. **Reverse Evacuation (evac)** – The product end is closed, and pressure is further reduced to a low pressure P_L via a vacuum pump. The CO₂ rich stream is then collected at the feed end.
3. **Feed pressurization (pres)** – Pressure is raised again to P_H by adding flue gas to the column from the feed end.

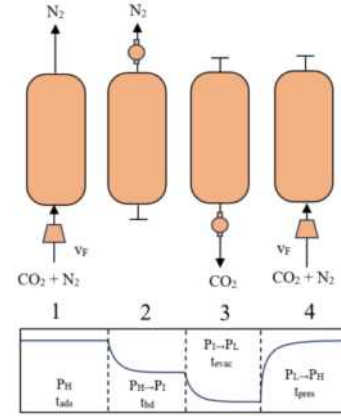


Figure 1: PVSA Schematic adapted from (Ward & Pini, 2022). The figure describes the 4 PVSA process steps – 1. adsorption, 2. forward blowdown, 3. reverse evacuation, 4. feed pressurization

Figure 1 shows the four steps as well as the pressure differences across the different stages of the process. These four steps are repeated until cyclic steady state (CSS) is achieved to obtain results. CSS is achieved when all key performance indicators have a relative error of less than 0.5% for 10 sequential adsorption cycle simulations.

3 Methodology

3.1 Proposed Workflow

Figure 2 summarises the workflow defined within the study to explore different adsorbent material choices from a feasibility perspective. Each step in the workflow is explained in this section.

3.2 Choice of Adsorbents

As discussed previously, adsorbents have a great impact on carbon capture performance. High porosity, large surface area to weight and sorption selectivity are a few key properties that affect their adsorption capacity. The most researched adsorbent materials within the carbon capture process include zeolites, activated carbon and metal-organic frameworks (MOFs) (Ma, et al., 2023). Zeolite 13X was chosen as a standard to compare other adsorbents against as it is widely used and an extensively researched adsorbent within the PVSA process. Additionally, it will be used as validation with Purwanto's thesis (Purwanto, 2023). The two remaining adsorbents this study investigated were both MOFs; CALF-20 and UTSA-16. Despite having a higher adsorbent cost, MOFs' unique properties of uniform pore distribution and high specific surface area results in a high CO₂ selectivity (Gaikwad, et al., 2020). Furthermore, MOFs are highly tailorable which enables physicochemical properties such as porosity and

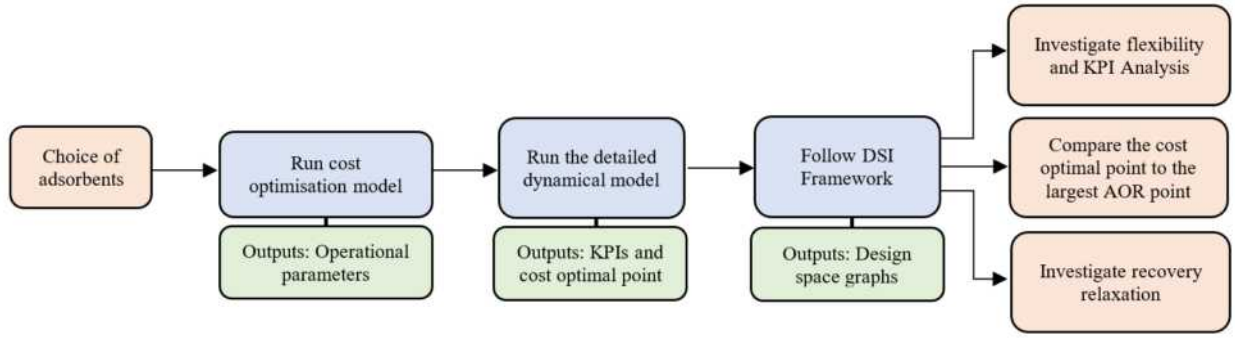


Figure 2: Methodology workflow defined along with the outputs at each stage in this study

crystallinity to be finely tuned for different adsorption parameters. (Ma, et al., 2023). In a previous study comparing UTSA-16's techno-economic performance to Zeolite 13X and other MOFs within the PVSA process, it was shown that UTSA-16 achieved exceptional KPI values and hence was chosen to be studied in this report (Alhajaj & Vega, 2021). It should be noted that Alhajaj & Vega's report did not analyse the metric of process flexibility between these adsorbents which was carried out in this paper. The main drawback for using UTSA-16 is that as it is specifically tuned, the adsorbent cost is expensive and cannot easily be upscaled for industrial purposes. Therefore, CALF-20 was also investigated as it is currently the most commercially viable MOF and has been used in Svante's industrial pilot plant since 2021 (Ozin, 2022). It is a zinc-based MOF that not only has a higher selectivity towards CO₂ than N₂, but also than water, making it applicable to processes where water is present within the flue gas stream. The synthesis of CALF-20 is a single-step process using commercially available components and therefore, shows the most promise for industry use (Lin, et al., 2021).

Table 2 shows the adsorbent's approximate cost and density. From this, the column's bed density (ρ_b) can be calculated (Equation (1)) and computed within the mathematical model. It should be noted that the cost of MOFs are an estimate based on their synthesis process.

$$\rho_b = \rho_{abs}(1 - \epsilon) \quad (1)$$

Table 2: Absolute density [kg/m³] and cost [\$ /tonne] for all studied adsorbents: Zeolite 13X, CALF-20 and UTSA-16

Adsorbent	ρ_{abs} [kg/m ³]	Cost [\$ /tonne]	Cost Reference
Zeolite 13X	1180	1,500	(Ward & Pini, 2022)
CALF-20	570	25,000	(Siegelman, et al., 2021)
UTSA-16	1180	10,000	(Peng, et al., 2022)

3.3 Adsorption Equilibrium

The dual-site Langmuir (DSL) model was used to model the CO₂ and N₂ adsorption equilibrium of the selected adsorbents to match previous studies following the same workflow. Equations (2) to (5) show how to calculate q_i^* , the moles of species i captured per kg of adsorbent at equilibrium, based on q_i , the characteristic amount adsorbed on site b and d, and c_i , the molar concentration of the gas phase fraction (y_i). $b_{i,0}$ and $d_{i,0}$, the reference

adsorption equilibrium constant for each site, are used alongside ΔU_i , the change of internal energy, R, the gas constant, and the temperature to calculate b_i and d_i , the adsorption equilibrium constant of species i for each site (Ward & Pini, 2022). The parameters for the different adsorbents' DSL model can be found in Table 3 (Khurana & Farooq, 2016).

$$q_i^* = \frac{q_{b,i} b_i c_i}{1 + \sum_{j=1}^{n_c} b_j c_j} + \frac{q_{d,i} d_i c_i}{1 + \sum_{j=1}^{n_c} d_j c_j} \quad (2)$$

$$b_i = b_{i,0} \exp\left(\frac{-\Delta U_{b,i}}{RT}\right) \quad (3)$$

$$d_i = d_{i,0} \exp\left(\frac{-\Delta U_{d,i}}{RT}\right) \quad (4)$$

$$c_i = \frac{y_i P}{RT} \quad (5)$$

From Equations (2) to (5), the adsorbents' isotherms for pure components of CO₂ and N₂ can be modelled, as shown in Figure 3a and b. With these two isotherms, a selectivity graph as shown in Figure 3c can be computed using Equation (6) for selectivity, S_{CO_2} , and flue gas compositions of 85% N₂ and 15% CO₂.

Table 3: DSL isotherm parameters for all studied adsorbents. These parameters are used in the dynamical PVSA model.

	CO ₂	N ₂	Units
Zeolite 13X			
$q_{b,i}$	3.09	5.84	mol/kg
$q_{d,i}$	2.54	0.00	mol/kg
$b_{i,0}$	8.65×10^{-7}	2.50×10^{-6}	m ³ /mol
$d_{i,0}$	2.63×10^{-8}	0.00	m ³ /mol
$\Delta U_{b,i}$	-36.64	-15.80	kJ/mol
$\Delta U_{d,i}$	-35.69	0.00	kJ/mol
UTSA-16			
$q_{b,i}$	4.40	9.32	mol/kg
$q_{d,i}$	7.34	9.32	mol/kg
$b_{i,0}$	4.11×10^{-12}	2.87×10^{-6}	m ³ /mol
$d_{i,0}$	6.33×10^{-7}	2.87×10^{-6}	m ³ /mol
$\Delta U_{b,i}$	-45.54	-9.84	kJ/mol
$\Delta U_{d,i}$	-30.54	-9.84	kJ/mol
CALF-20			
$q_{b,i}$	2.39	5.66	mol/kg
$q_{d,i}$	3.27	0.00	mol/kg
$b_{i,0}$	5.52×10^{-7}	8.14×10^{-7}	m ³ /mol
$d_{i,0}$	5.19×10^{-8}	0.00	m ³ /mol
$\Delta U_{b,i}$	-35.10	-18.00	kJ/mol
$\Delta U_{d,i}$	-29.00	0.00	kJ/mol

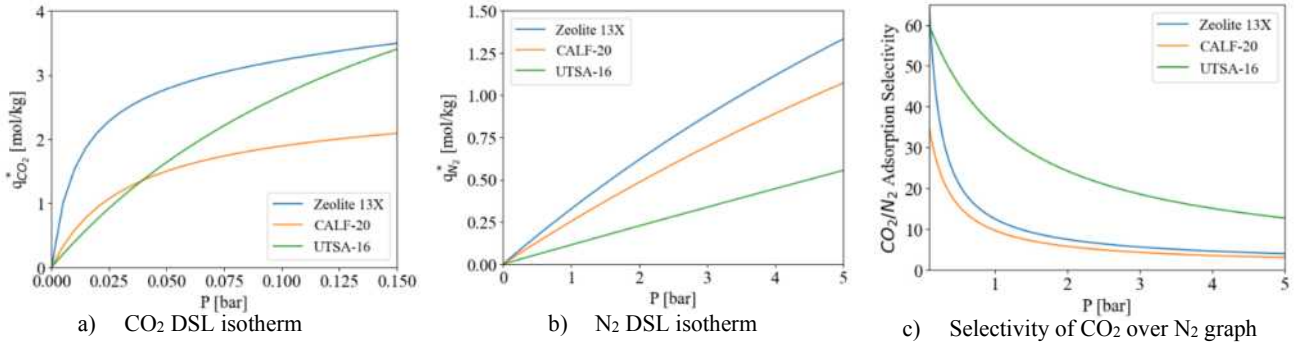


Figure 3: DSL isotherms for a) CO₂ adsorption b) N₂ adsorption. c) Selectivity graph of CO₂ over N₂ for all studied adsorbents

$$S_{CO_2} = \frac{q_{CO_2}^*}{q_{N_2}^*} \quad (6)$$

It is clear to see from Figure 3c, that UTSA-16 has the highest CO₂ selectivity followed by Zeolite 13X and then CALF-20.

3.4 Cost Optimisation Model

A cost optimisation algorithm as developed by Ward and Pini (Ward & Pini, 2022) was applied in this paper to obtain the optimal parameters for cycle times, pressures, flowrate and length of column (L) as seen in Table 4. The pressures involved in the cyclic PVSA cycle, P_H , P_I and P_L , were varied as inputs to the mathematical model as after the adsorbent material, these operational parameters, were found to have the largest effect on performance in Purwanto's global sensitivity analysis (Purwanto, 2023). The pressure ranges were calculated to be $\pm 40\%$ of the cost optimal point to stay consistent with Purwanto's work. A L/r_{in} aspect ratio of 6 was maintained along with a difference of 0.0175m between r_{in} and r_{out} which represent the column inner radius and outer radius, respectively.

It is worth noting that running the mathematical model for different adsorbents using Zeolite 13X's parameters listed in Table 4 would result in no feasible points being found. Results to this can be found in Table S1 in Supporting Information (SI). This highlights the importance of obtaining operational parameters for each adsorbent individually due to the highly synergistic nature between the operational parameters and material choice.

Table 4: Column and operational parameters obtained for all studied adsorbents from the cost optimisation model.

	Zeolite 13X	UTSA-16	CALF-20
L [m]	3.56	1.42	1.72
t_{pres} [s]	20.00	20.00	20.00
t_{ads} [s]	67.50	64.00	21.30
t_{bd} [s]	40.50	30.00	63.30
t_{evac} [s]	174.00	58.50	70.00
v_F [m/s]	1.70	0.80	1.48
$P_{H,min}$ [bar]	2.83	2.65	2.77
$P_{H,max}$ [bar]	6.61	6.17	6.47
$P_{I,min}$ [bar]	0.47	0.27	0.46
$P_{I,max}$ [bar]	1.09	0.64	1.08
$P_{L,min}$ [bar]	0.02	0.03	0.01
$P_{L,max}$ [bar]	0.04	0.07	0.03

3.5 Dynamic Model

The mathematical model used to simulate the adsorption column consists of a coupled system of partial differential and algebraic equations (PDAEs) that characterise the governing material, momentum and energy balances in an 1D adsorption column, as shown in Table 5. This adsorption model, as denoted by Haghpanah et al., formulates the process performance using extensive simulations varying in space and time and solved using *odes15s* on MATLAB (Haghpanah, et al., 2013). A techno-economic assessment is also utilised to calculate the cost per tonne of CO₂ captured.

Table 5: Partial differential and algebraic equations that represent material and energy balances where \bar{P} , \bar{T} , \bar{T}_a , \bar{T}_w , \bar{v} , τ , Z represent the non-dimensional pressure, temperature, ambient temperature, wall temperature, velocity, time and longitudinal coordinate. α_i , Π_i , σ_i , Ω_i , ψ represent dimensional groups. x_i , x_i^* signifies the nondimensional adsorbed amount of species i and the equilibrium adsorbed amount of species i , respectively. P_0 and v_0 represent the characteristic pressure and velocity respectively.

Name:	Partial differential equation:
Overall mass balance:	$\frac{\partial \bar{P}}{\partial \tau} - \frac{\bar{P}}{\bar{T}} \frac{\partial \bar{T}}{\partial \tau} = -\bar{T} \frac{\partial}{\partial Z} \left(\frac{\bar{P}\bar{v}}{\bar{T}} \right) - \psi \bar{T} \sum_{i=1}^{n_c} \frac{\partial x_i}{\partial \tau}$
Component material balance:	$\frac{\partial y_i}{\partial \tau} + \frac{y_i}{\bar{p}} \frac{\partial \bar{P}}{\partial \tau} - \frac{y_i}{\bar{T}} \frac{\partial \bar{T}}{\partial \tau} = \frac{1}{Pe} \frac{\bar{T}}{\bar{P}} \frac{\partial}{\partial Z} \left(\frac{\bar{P}}{\bar{T}} \frac{\partial y_i}{\partial Z} \right) - \frac{\bar{T}}{\bar{P}} \frac{\partial}{\partial Z} \left(\frac{\bar{P}\bar{v}y_i}{\bar{T}} \right) - \frac{\bar{T}}{\bar{P}} \psi \frac{\partial x_i}{\partial \tau}$
Solid-phase material balance:	$\frac{\partial x_i}{\partial \tau} = \alpha_i (x_i^* - x_i)$
Pressure drop:	$-\frac{\partial \bar{P}}{\partial Z} = \frac{150}{4r_p^2} \left(\frac{1-\epsilon}{\epsilon} \right)^2 \left(\frac{v_0 L}{P_0} \right) \mu \bar{v}$
Column energy balance:	$\frac{\partial \bar{T}}{\partial \tau} + \Omega_2 \frac{\partial \bar{P}}{\partial \tau} = \Omega_1 \frac{\partial^2 \bar{T}}{\partial Z^2} - \Omega_2 \frac{\partial}{\partial Z} (\bar{P}\bar{v}) + \sum_{i=1}^{n_c} \left[(\sigma_i - \Omega_3 \bar{T}) \frac{\partial x_i}{\partial \tau} \right]$
Wall energy balance:	$\frac{\partial \bar{T}_w}{\partial \tau} = \Pi_1 \frac{\partial^2 \bar{T}_w}{\partial Z^2} + \Pi_2 (\bar{T} - \bar{T}_w) - \Pi_2 (\bar{T}_w - \bar{T}_a)$

To stay consistent with Purwanto's thesis, the PVSA system is designed to capture CO₂ from a 1000MW coal power plant, assuming a molar feed gas mixture of 85% N₂ and 15% CO₂ (Purwanto, 2023). The gas mixture is also assumed to be available at a temperature of 298.15K and a pressure of 1 bar. Other parameters utilised within the mathematical model can be seen in [Table 6](#). These parameters include column specifications and assumptions made for the techno-economic assessment of the carbon capture plant. The constants highlighted here have been taken from previous papers on this study and remain consistent within this analysis. Further parameters, including the adsorbent's DSL values and operating conditions highlighted in [Section 3.3](#) and [Section 3.4](#) are also specified within the model. For this investigation, 4096 Sobol sample points were simulated on MATLAB within the P_H , P_I and P_L ranges obtained from the cost optimisation model. After running the solver, the model calculates the different KPIs for the 4096 sample points.

Table 6: Column specifications and techno-economic parameters

Parameter	Value	Units
Bed voidage (ϵ)	0.37	-
Particle voidage (ϵ_p)	0.35	-
Particle radius (r_p)	0.001	m
Particle tortuosity (τ_p)	3	-
Molecular diffusivity (D_m)	1.5	10 ⁻⁵ m ² /s
Thermal conductivity of gas (K_z)	0.09	J/m/K/s
Thermal conductivity of wall (K_w)	16	J/m/K/s
Heat capacity of gas phase ($C_{p,g}$)	30.7	J/mol/K
Heat capacity of adsorbed phase ($C_{p,a}$)	30.7	J/mol/K
Heat capacity of adsorbent ($C_{p,s}$)	1070	J/mol/K
Heat capacity of column wall ($C_{p,w}$)	502	J/mol/K
Density of wall (ρ_w)	7800	kg/m ³
Dynamic viscosity of gas (μ)	1.72	10 ⁻⁵ kg/m/s
Overall inside heat transfer coefficient (h_{in})	8.6	J/m ² /K/s
Overall outside heat transfer coefficient (h_{out})	2.5	J/m ² /K/s
Ratio of ideal gas heat capacities (γ)	1.4	-
Adiabatic efficiency (η)	0.72	-
Pressure profile time constant (λ)	0.5	s ⁻¹
Coal power plant capacity	1000	MW
Thermal efficiency	0.4	-
Lower heating value (LHV) of coal	28000	10 ³ J/kg
Carbon content of coal, DAF basis	0.7	-
Thermal power	2500	MW
Rate of coal consumption	89.3	kg/s
Molar flow of carbon in flue gas	5208	mol/s
Discount rate	0.08	-
Economic lifetime	25	Year
Electricity cost	0.06	\$/kWh

3.6 Design Space Identification (DSI)

Framework

A DSI framework is used to assess the design and flexibility of the carbon capture PVSA process simultaneously. The 4096 Sobol sampling points run through the rigorous process model and establish the knowledge space in this framework. This is the whole region of considered parameters.

From this knowledge space, a design space boundary can then be developed using the direct sampling method whereby linear polyhedral approximations are constructed to encompass all feasible points. This determines an alpha shape which represents the piece-wise geometric edges. From this, a design space can be identified and is defined as the operational parameter space where all points satisfy the process constraints, namely CO₂ purity and recovery. Within the formulation of the design space, a variable known as 'maxvp' is specified. This is the maximum percentage of violated points allowed within the design space. After analysing maxvp's effect on the different KPIs, a maxvp value of 0.01 was chosen across all adsorbents ([Table S2](#) in SI).

Within the design space, a nominal operating point (NOP) is then chosen. The choice of NOP is critical when considering the range of flexibility. A cost optimal NOP will always lie on the boundary of feasible points and therefore provides only a limited flexibility range. However, choosing the NOP with the largest range of flexibility can prove to be more costly. Therefore, both will be investigated and the cost difference between these two NOPs will be compared across the variety of adsorbent materials.

The process flexibility is determined by the size of a point's acceptable operating range (AOR). Its calculation can be conceptualised as a cuboid that expands around the NOP at its centre until one of the vertices intersect with the boundary of the design space. The AOR thus provides the operating range of a parameter such that it remains within the feasible region.

3.7 Recovery Relaxation Methodology

As mentioned above, the cost optimal NOP shows very little flexibility because it lies on an active constraint. Analysis of recovery constraint relaxation should thus be conducted to see how the KPIs are affected as the flexibility increases. The recovery constraint was chosen as it can be easily controlled by the process operators whereas the purity constraint cannot be relaxed due to strict environmental regulations. Two methods of recovery relaxation were explored within the study:

1. Obtain and compare the largest AOR NOP's KPIs and flexibility after each relaxation.
2. Obtain the cost optimal NOP from 89% and compare results from that fixed NOP to see how its flexibility and KPI changes (Sachio, et al., 2023b).

As a result, the effect of decreasing the recovery constraint on the flexibility metric and other KPIs could be studied. 85% was the lowest recovery value examined within this study as it is the lowest a process could operate at meeting government regulations (Department of Business, Energy and Industrial Strategy, 2021).

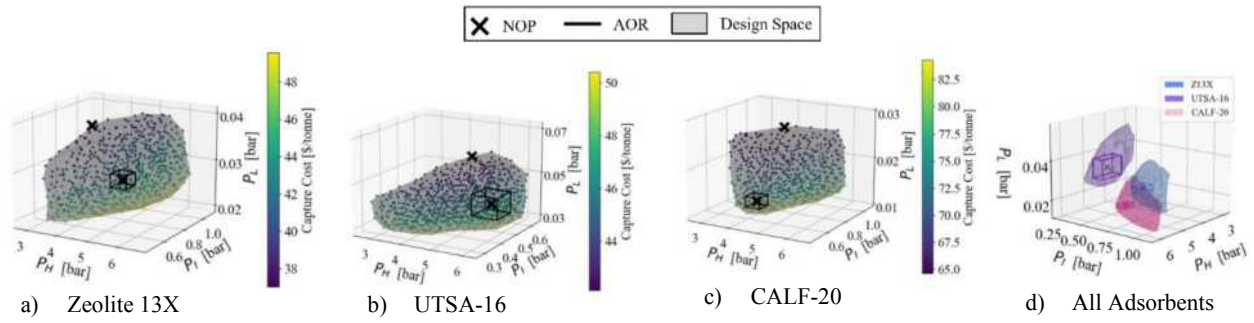


Figure 4: Design space graphs showing the largest AOR and cost optimal NOP for a) Zeolite 13X b) UTSA-16 c) CALF-20. d) Design space graphs showing the largest AOR NOP for all adsorbents

4 Results

4.1 Adsorbent Design Space

Following the methodology outlined in [Section 3.5](#), design space analysis was conducted for Zeolite 13X, CALF-20 and UTSA-16, with the Zeolite 13X design space and NOPs being validated against Purwanto's work. [Figures 4a,b and c](#) show the design space along with the largest AOR and cost optimal NOP. Upon visually examining these figures, it is clear to see that UTSA-16's design space is more densely populated than Zeolite 13X and CALF-20 (also seen in [Table S3](#)). The cost optimal NOPs of all three materials is shown to be on the boundary of the design space, and do not show an AOR space. This is concurrent with the findings of Sachio et al. that show that the cost optimal NOP shows little to no flexibility (Sachio, et al., 2023b). Lastly, the heat maps on the graphs show how capture cost varies across the design space, whereby CALF-20's capture cost is significantly higher due to the higher estimated adsorbent cost (see [Table 2](#)).

[Figure 4d](#) shows all three adsorbent design spaces and largest AOR point plotted on the same graph. It is evident that UTSA-16 has the largest design space and AOR point amongst the studied materials. Additionally, it operates at a higher P_L range and lower P_I . The CALF-20 and Zeolite 13X operating pressures are similar with design spaces that overlap slightly however, the UTSA-16 design space does not overlap at all with the other two materials.

[Table 7](#) shows the percentage difference in capture cost between the cost optimal point and the largest AOR point. This difference gives some indication of how flexibility and capture cost vary across the design space. A lower percentage difference is desirable since it indicates that a good balance can be struck between flexibility and capture cost. It is evident to see that UTSA-16 has the lowest change in cost from the two NOPs whereas CALF-20 has the largest. This can be seen visually on the graph where CALF-20's NOPs are further away from each other based on the capture cost heat map.

Table 7: Percentage difference in capture cost

Adsorbent	% difference in capture cost from the largest AOR and cost optimal NOP
Zeolite 13X	17%
CALF-20	20%
UTSA-16	9%

4.1.1 Adsorbent Flexibility Range

Adsorbent flexibility range refers to the percent deviation allowed from the NOP that still satisfies the process constraints. This is found by looking at the minimum and maximum values of the projected AOR of an operating point. These values are pictured in [Figure 5](#) for all three adsorbents.

As can be seen from the figure, UTSA-16 presents the largest allowed deviations in P_H , P_I and P_L , demonstrating that it allows for the greatest process flexibility. UTSA-16 is then followed by CALF-20 in terms of flexibility, with Zeolite 13X presenting the lowest allowed flexibility.

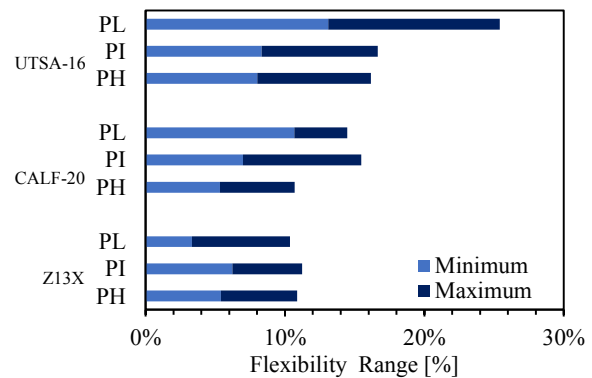


Figure 5: Pressure Flexibility Range across the AOR

4.2 KPI Analysis

KPI analysis was conducted to analyse trends within each adsorbent as well as to conduct comparisons across the studied materials. [Table 8](#) shows the operational parameters and range within the AOR space as well as the KPI values at the cost optimal NOP. These values within the AOR, are provided to indicate both the best- and worst-case scenarios with regards to KPI values and can also be visually seen in [Figures 6a,b and c](#). From [Figure 6a](#), it can be seen that CALF-20 has the highest cost whereas UTSA-16 has comparable capture cost to Zeolite 13X. From [Figure 6b](#), UTSA-16 was shown to have almost double the productivity of Zeolite 13X and CALF-20. This is highly desirable as it shows that UTSA-16's bed volume and duration of cycle can be halved to achieve the same productivity as the other adsorbents. Additionally, it was found that the cost optimal point for each adsorbent had a lower productivity than the largest AOR NOP. In [Figure 6c](#), CALF-20 has the highest energy consumption whereas

Table 8: Operational parameters range within the AOR and operational parameters point for the cost optimal NOP for all adsorbents (to 3 significant figures)

	Zeolite 13X		CALF-20		UTSA-16	
	Largest AOR NOP					
	min	max	min	max	min	max
P _H [bar]	4.67	5.21	4.31	4.79	4.96	5.83
P _I [bar]	0.731	0.819	0.493	0.575	0.495	0.585
P _L [bar]	0.026	0.029	0.013	0.016	0.033	0.042
Capture cost [\$/tonne]	42.2	44.9	75.8	79.7	44.5	47.6
Productivity [mol/m³s]	1.12	1.23	1.10	1.17	2.54	2.94
Energy Usage [kWh/tonne]	613	660	881	947	501	573
	Cost Optimal NOP					
(P _H , P _I , P _L) [bar]	(4.19, 0.700, 0.039)		(4.74, 0.750, 0.028)		(4.59, 0.567, 0.057)	
Capture cost [\$/tonne]	37.0		64.6		42.1	
Productivity [mol/m³s]	1.00		1.16		2.28	
Energy Usage [kWh/tonne]	516		708		444	

UTSA-16 has the lowest. Overall, the cost optimal points provide a lower energy usage.

Figures 7 a, b and c below show the KPI heat maps presented for UTSA-16. The heat maps show how capture cost, productivity and energy usage vary across the design space as P_H , P_I and P_L change for UTSA-16 however, these trends were observed to be the same throughout all adsorbents that were studied aside from Zeolite 13X's capture cost. This could be attributed to its low adsorbent material cost. Heat maps for the other materials can be found in Figure S1 and S2.

From Figure 7a, the cost increases as P_L decreases. This is expected as the process operates closer to vacuum, the operating cost of the column increases. Additionally, the capture cost increases as P_H increases, which is due to the higher energy required to compress the flue gas to higher pressures. Both Figures 7b and 7c, show the same trends, indicating that an

increase in productivity results in an increase in the energy usage. This reveals a trade-off between the two KPIs as a high productivity and low energy usage is preferred.

Based on the DSI and KPI analysis, UTSA-16 appears to be a very promising adsorbent for the PVSA process, with a high allowed process flexibility along with low energy usage, moderate capture cost and high productivity.

4.3 Recovery Relaxation

From the two recovery relaxation methods, it was found that the first method, whereby the results from the largest AOR NOP from each relaxation is compared, achieved different KPIs and AOR sizes upon relaxation. Results from this method are presented below. The second method where the cost optimal NOP is fixed upon relaxation showed no change to its flexibility. Results from this can be seen in Table S4.

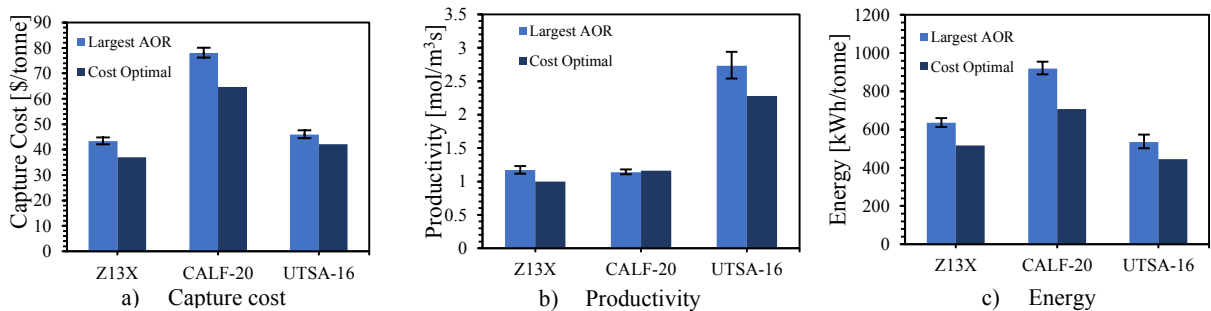


Figure 6: Bar charts showing the average KPI from the largest AOR NOP and cost optimal NOP for a) Capture cost b) Productivity c) Energy Usage. The error bars indicate the minimum and maximum KPI value found within the AOR.

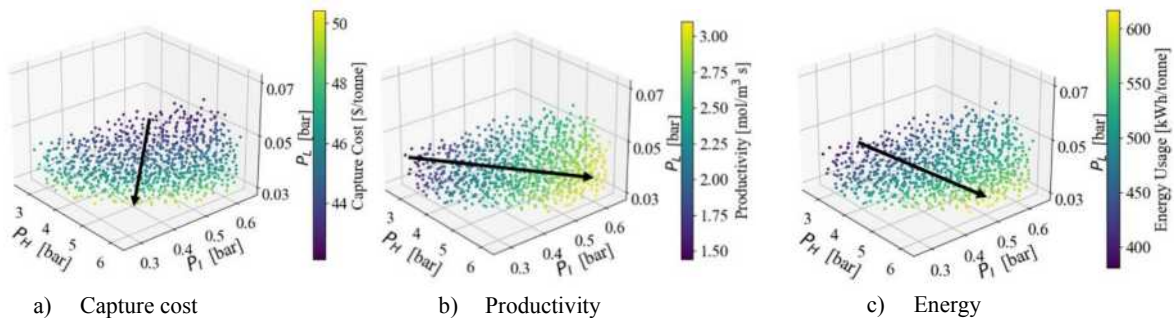


Figure 7: UTSA 16's design space graphs highlighting the heat map for a) Capture cost b) Productivity c) Energy Usage

The following violin plots pictured in [Figure 8](#) were produced to visually depict the effect of relaxing the recovery constraint on the productivity. The shaded area indicates all the points within the AOR, and the line indicates the mean of the capture cost found within the AOR.

Further violin plots for the other KPIs can be seen in [Figure S3](#). The shape of the violin indicates the distribution of points within the largest AOR. The shorter, more squat, violin indicates a smaller distribution within that AOR's productivity.

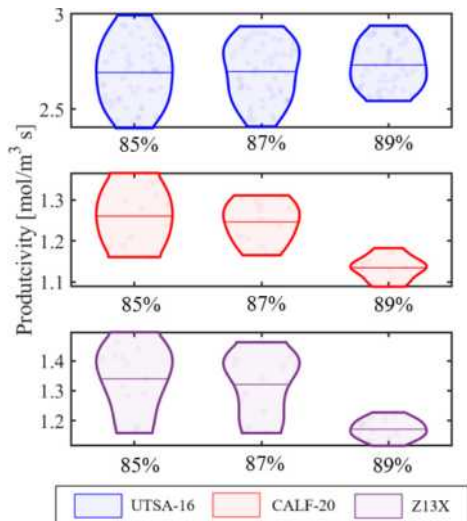


Figure 8: Violin plots highlighting the productivity range and average productivity within the AOR upon recovery relaxation for all adsorbents

As seen from [Figure 8](#), relaxing the recovery constraint has limited effect on UTSA-16's productivity as the overall shape and mean value does not significantly change. However, for both CALF-20 and Zeolite 13X, relaxing the recovery constraint improves the material's productivity range as the violin's shape has been lengthened. This trend was seen for all the KPI violin plots whereby UTSA-16 showed little difference in KPI range and Zeolite 13X and CALF-20 showed an increase. This indicates that it is favourable to lower the recovery constraint for these two materials. It is also important to note that the distribution and size of the violin plots do not change to a significant degree when lowering the recovery constraint to 85% from 87%.

Flexibility range graphs were also produced to better understand how the recovery relaxation would

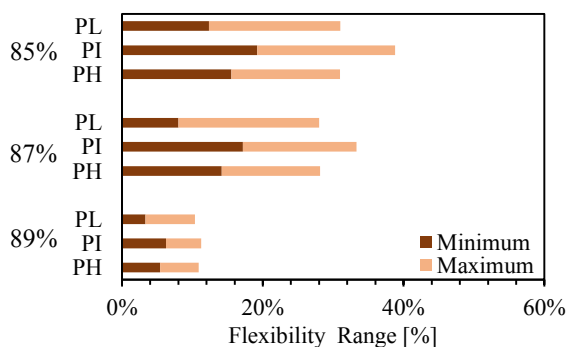


Figure 9: Zeolite 13X pressure flexibility range for recovery constraints from 89% to 85%

affect pressure flexibility. [Figure 9](#) pictures the range for Zeolite 13X. Graphs for the other materials can be found in [Figure S4](#). It appears across all adsorbents that decreasing the recovery constraint is beneficial from 89% to 87% however there are no significant benefits to reducing the recovery to 85%. It is therefore beneficial for both Zeolite 13X and CALF-20 to relax their recovery constraint to 87% to increase process flexibility as well as productivity. Since there is minimal change observed for UTSA-16, it is advised to keep the constraint at 89% to keep the process as economically and environmentally worthwhile as possible.

5 Discussion

5.1 Adsorbent Comparison

To evaluate the results from the DSI and KPI analysis, the CO₂ DSL isotherm and selectivity graph presented earlier, [Figures 3a and c](#), was consulted to better understand how the material parameters affect the results obtained.

Zeolite 13X and CALF-20 show a similar steep gradient in their CO₂ isotherm whereas UTSA-16 shows a more linear decline (seen in [Figure 3a](#)). This indicates that even with slight pressure changes, the CO₂ selectivity for both Zeolite 13X and CALF-20 would change significantly at lower pressures. As UTSA-16 has a more linear trend, there are more variations of operating pressures that can perform well within the constraints of the system, leading to a higher flexibility.

From the KPI analysis, we found that UTSA-16 showed the highest productivity and lowest energy usage. The material's high productivity could be attributed to the fact that UTSA-16 has the highest CO₂ selectivity (as seen in [Figure 3c](#)) compared to the other adsorbents. Its low energy usage could be explained via its linear selectivity, meaning the adsorbent does not require as low of a vacuum pressure to achieve sufficient working capacity and meet the CO₂ recovery target. This would also explain why UTSA-16's P_L is not as low as the other adsorbents. Due to this, UTSA-16's capture cost is similar to Zeolite 13X despite having an adsorbent cost of 10,000 \$/tonne, as less energy is required to lower the pressure.

Overall, UTSA-16 was found to be the best adsorbent due to its high CO₂ selectivity and more linear isotherm. Xiang et al., obtained similar adsorption properties comparing UTSA-16 to other MOFs and proposed that UTSA-16's abnormally high selectivity towards CO₂ could be explained by its optimal diamondoid pore cages and its overall structure (Xiang, et al., 2012). In their paper, neutron diffraction revealed that the terminal water molecules in UTSA-16's structure strongly interact and bind with the CO₂ molecules, enabling for high selectivity and hence explaining UTSA-16's excellent performance. Another paper that compares the key performance parameters of Zeolite 13X and UTSA-16 within the PVSA process, found that Zeolite 13X had the lowest capture cost and UTSA-16 had the lowest energy requirement (Alhajaj & Vega, 2021). Although their values differ greatly as different input parameters were used, this trend is concurrent with results obtained with this study.

A factor to note regarding this analysis is that a surrogate model such as a Gaussian Process (GP) (such as in (Purwanto, 2023)) or Artificial Neural Network (ANN) was not employed to interpolate between the generated sobol points. This would generate artificial points that would improve the resolution of the design space. This was not implemented, however, its use could provide a more consistent comparison between adsorbent materials by evening out the number of points found in the design space and AOR, allowing for “better predictions” based on KPI. These predictions however would be based on artificial points and introduce an uncertainty into the analysis conducted on the DSI methodology.

5.2 Adsorbent Feasibility

From the KPI and DSI results, it is evident that UTSA-16 is the best adsorbent within this model workflow. However other factors such as its synthesis and stability, should be considered before industrialising UTSA-16. Moreover, as CALF-20 is the first MOF to be used in industry, it is important to compare its properties with UTSA-16.

CALF-20's raw materials are available commercially at a large scale. Its synthesis does not require extreme conditions and is therefore favourable from a safety and environmental perspective (Lin, et al., 2021). UTSA-16 exhibits comparable characteristics as its raw materials are low cost and common. (Xiang, et al., 2012). Both adsorbents exhibit remarkable stability and through experimentation, maintain adsorption capacity over time after continuous cycling.

One of the main reasons why CALF-20 is the only MOF to be industrialised is because of its high CO₂ selectivity over water. Zeolite 13X and most MOFs have a significantly reduced adsorption performance with even just 1% relative humidity (Nguyen, et al., 2023). This could be a reason why CALF-20 does not perform well based on its poor KPI values and flexibility, as the flue gas composition studied here consists solely of CO₂ and N₂. A study that synthesised 3D printed UTSA-16 monoliths, investigated the performance of UTSA-16 with water (Grande, et al., 2020). They concluded that UTSA-16 showed a higher water adsorption capacity than it does for CO₂, meaning water can easily displace CO₂ on the adsorbent. This shows that in the presence of water, UTSA-16 is unable to adsorb CO₂. This is a key factor and should be addressed before industrialising UTSA-16. The presence of moisture in flue gas is common and to requires drying to remove which is an additional cost to consider. As this study's techno-economic assessment does not include drying costs, UTSA-16's and Zeolite 13X's capture cost could be significantly higher than what is estimated here.

The last property to consider for a feasible adsorbent concerns its economics. As highlighted in [Table 2](#), the cost of MOFs are significantly greater than zeolites. However, as UTSA-16's material costs is only an estimate, it is therefore recommended to conduct a sensitivity analysis based on adsorbent material cost to better understand its effect on the capital costs of the PVSA process.

5.3 Recovery Relaxation

The trend observed with both the KPI violin plots, and the flexibility range graphs indicate a shift in active constraints from 87% to 85% whereby there is no significant change in KPIs or flexibility range upon relaxing the recovery constraint further. This indicates that upon lowering the recovery constraint to 85%, the AOR cannot expand to a significant degree without violating the purity constraint i.e., there is a change in active constraint to the purity constraint. A similar trend was found in (Sachio, et al., 2023b) where they observed the change in AOR for the cost optimal point at 89% with relaxing recovery for Zeolite 13X. In this paper they found an increase in flexibility as the recovery constraint was relaxed.

Upon conducting a similar analysis for the method where the cost optimal points were fixed, an increase in flexibility was not observed (more details can be found in [Table S4](#)). This discrepancy could be accounted for by the difference in input parameters being used, with the paper using (P_H , P_L , v_F) This indicates that the boundary of the design space is bound by the recovery constraint with the input parameters (P_H , P_L , v_F) but not so for (P_H , P_L , P_L).

It should however be noted that lowering the recovery constraint to increase flexibility can cause a decrease in the amount of CO₂ captured from the inlet flue gas (as recovery targets are lowered). This could be unfavourable from an environmental perspective, and this trade-off should be considered while optimising for flexibility and environmental considerations.

5.4 Key Findings

To summarise the results and discussion section, UTSA-16 showed great promise as a potential adsorbent as it exhibits good process flexibility and KPI values. This could be explained directly from the adsorbent's CO₂ selectivity graph from [Figure 3c](#). Therefore, it was concluded that adsorbents presenting a high CO₂ selectivity with a linear isotherm trend perform well in the PVSA process. This means that just from obtaining the CO₂ selectivity graph, an adsorbent's performance within this PVSA system can be estimated without the mathematical model and design space framework.

To quantify the difference between the cost optimal and most flexible point, the percentage difference in cost was worked out. It was found that UTSA-16 has the least difference in cost. Additionally, it was found that the cost optimal point overall had a lower productivity and lower energy usage to the largest AOR NOP for all the adsorbents.

From our analysis of KPI trends, similar trends were found across all adsorbents with variation in our input parameters (P_H , P_L , P_L). For energy usage and productivity, a trade-off between the two KPIs has been highlighted, indicating the complexity of the optimisation problem.

It was found from the recovery relaxation analysis, that it would be beneficial for CALF-20 and Zeolite 13X to reduce the constraint to 87% in order to improve their process flexibility and KPI range. However, it is important to note that reducing recovery would result in less CO₂ being captured and from an

environmental perspective, would not be favourable. As UTSA-16's performance and flexibility were largely unaffected by the relaxation, it would be better to keep the recovery constraint at 89%.

Additionally, despite UTSA-16's performance and process flexibility shown within the workflow, the adsorbent performs very poorly when moisture is present in the inlet flue gas. Therefore, to obtain an adsorbent that is suitable for industry, other key factors such as its stability, synthesis and performance in water should be considered.

6 Conclusion

To conclude, adsorbent material choices were successfully screened based on process flexibility, a metric that had not been previously explored for the PVSA process. From this, UTSA-16 was found to have the best process flexibility, lowest energy usage and highest productivity.

Through our research, several potential avenues have been identified that could be investigated to build on the conclusions found here. There is potential to explore the inlet flue gas compositions. Rather than using the constant 15%/85% CO₂/N₂ composition used within this methodology, water and methane could be considered in the flue gas inlet to see how this would affect the adsorbents' performance.

Another factor to consider is the use of a surrogate model such as a Gaussian Process or Artificial Neural Network to increase the resolution of our design space. Use of such a method could improve the DSI analysis by increasing the design space resolution.

Lastly, since UTSA-16's cost was just an estimate, conducting a cost-based sensitivity analysis might allow better understanding of the extent that the MOF's adsorbent cost impacts the process' capture cost.

Process flexibility is an important factor to consider when industrialising the PVSA process to achieve a suitably operable and controllable process. The findings and outlook from this research present a new perspective on screening adsorbents for the PVSA process based on techno-economic factors and process flexibility and bring us one step closer to realising an industrial scale PVSA carbon capture process.

7 Acknowledgements

The authors would like to thank Adam Ward, Steven Sacchio, and Haditya Kuku Purwanto for all their help throughout this project.

8 References

Alhajaj, A. & Vega, L., 2021. Are we missing something when evaluating adsorbents for CO₂ capture at the system level?. *Energy & Environmental Science*, 14(6360).
Department of Business, Energy and Industrial Strategy, 2021. *Carbon Capture, Usage and Storage*, London: UK GOV.
Gaikwad, S., Kim, S.-J. & Han, S., 2020. Novel metal-organic framework of UTSA-16 (Zn) synthesized by a microwave method: Outstanding performance for CO₂ capture with improved stability to acid gases. *Journal*

of Industrial and Engineering Chemistry, Volume 87, pp. 250-263.

Grande, C., Middelkoop, V., Blom, R. & Matras, D., 2020. Multiscale investigation of adsorption properties of novel 3D printed UTSA-16 structures. *Chemical Engineering Journal*, Volume 402.

Haghpahan, R. et al., 2013. Multiobjective Optimization of a Four-Step Adsorption Process for Postcombustion CO₂ Capture Via Finite Volume Simulation. *Journal of Industrial Engineering Chemistry Research*, 52(11), pp. 4249-4265.

IEA, 2020. *Energy Technology Perspectives*, Paris: IEA.

Khurana, M. & Farooq, S., 2016. Adsorbent Screening for Postcombustion CO₂ Capture: A Method Relating Equilibrium Isotherm Characteristics to an Optimum Vacuum Swing Adsorption Process Performance. *Journal of Industrial Engineering Chemistry Research*, 55(8), pp. 2447-2460.

Lin, J.-B. et al., 2021. A scalable metal-organic framework as a durable physisorbent for carbon dioxide capture. 374(6574), pp. 1464-1469.

Ma, M. et al., 2023. Tailored porous structure and CO₂ adsorption capacity of Mg-MOF-74 via solvent polarity regulation. *Chemical Engineering Journal*, Volume 476.

Nguyen, T., Shimizu, G. & Rajendran, A., 2023. CO₂/N₂ separation by vacuum swing adsorption using a metal-organic framework and experimental validation. *Chemical Engineering Journal*, Volume 452.

Ozin, G., 2022. *CALF-20: A carbon capture success story*. [Online]

Available at:

<https://www.advancedsciencenews.com/calf-20-a-carbon-capture-success-story/> [Accessed 28 11 2023].

Peng, P. et al., 2022. Cost and potential of metal-organic frameworks for hydrogen back-up power supply. *Nature Energy*, Volume 7, pp. 448-458.

Purwanto, H. K., 2023. *Flexible Operation Assessment of Adsorption-based Carbon Capture Systems via Design Space Identification*, s.l.: Msc Thesis, Imperial College London.

Sachio, S., Ward, A., Pini, R. & Papathanasiou, M., 2023a. Embedding Flexibility to the Design of Pressure-Vacuum Swing Adsorption Processes for CO₂ Capture. *Proceedings of the 33rd European Symposium on Computer-aided Process Engineering*.

Sachio, S., Ward, A. P. R. & Maria, M. P., 2023b. Operability-economics trade-offs in adsorption-based CO₂ capture processes. *Preprint available at ArXiv*.

Siegelman, R., Kim, E. & Long, J., 2021. Porous materials for carbon dioxide separations. *Nature Materials*, Volume 20, pp. 1060-1072.

Ward, A. & Pini, R., 2022. Efficient Bayesian Optimization of Industrial-Scale Pressure-Vacuum. *Journal of Industrial and Engineering Chemistry*, 61(36), pp. 13650-13668.

Xiang, S. et al., 2012. Microporous metal-organic framework with potential for carbon dioxide capture at ambient conditions. *Nat Communications*, Volume 954.

Support Vector Machines Practice on Design Space Identification

Brooklyn Robinson and Defne Demirdesen

Department of Chemical Engineering, Imperial College London, U.K.

Abstract

Design space characterisation has been developed to understand the feasible region of operation and product quality attributes. This paper introduces support vector machines as an approach for classifying the design space for an industrial case study on Protein A chromatography. The presented approach is investigated through four different types of kernels: (a) linear kernel, (b) polynomial kernel, (c) Gaussian kernel, and (d) Sigmoid kernel. The two main model metrics, Matthew Correlation Coefficient and accuracy are used to evaluate imbalanced and balanced datasets. The Gaussian kernel produces the highest performance with these model evaluation metrics. It is demonstrated how much tuning hyperparameters can optimise the support vector machines model within the dataset with the Grid Search approach. The correlation between dataset size and the performance of the model is successfully explored through data reduction. While analysing the model performance, fairness, computational time, and accuracy become the fundamental categories.

Keywords: design space, support vector machines (SVM), kernel, chromatography

1. Introduction

1.1 Biopharmaceutical Industry and Challenges

The rising demand in the biopharmaceutical industry necessitates the requirement for leaner manufacturing with lower costs (Gerogiorgis, et al., 2020). The present challenges within the biopharmaceutical industry are high experimentation costs and waiting times. In addition, biopharmaceutical processes are highly complex and infeasible for disturbances, creating significant difficulties (Hong, et al., 2017).

New approaches for enhancing manufacturing processes have been investigated to overcome these challenges. Specifically, computational modelling applications have been carried out to provide a better understanding of the process and predict outputs in the whole domain (Manzon, et al., 2020).

1.2 Design Space Development

Design space investigation was the approach of determining a design solution or solutions that best match the given design requirements from a space of unsettled points. These solutions were made up of unique combinations of independent input variables (Yamada, et al., 2022). This approach was practical for explaining the critical process parameters and understanding the feasible region of product quality attributes and techno-economic performances (Gerogiorgis, et al., 2020).

The classical approach to characterise design space consisted of four steps: (a) Perform a comprehensive experimental analysis of the correlation between process parameters and critical quality attributes (CQAs). CQAs are classified as the physical and/or chemical properties that should be within suitable limits for achieving desired product quality. (b) Conduct sensitivity analysis on process parameters that affect CQAs and select the parameters that indicate the highest sensitivity.

(c) Create a graphical or mathematical representation of the design space using computational modelling. (d) Evaluate the final design space by operating additional validation experiments (Kusumo, et al., 2020). Characterising design space with this approach gave a functioning outcome. However, the obstacles within high experimentation costs and long waiting times made this approach undesirable.

The Bayesian approach was “an adaptation of nested sampling for the characterisation of a probabilistic [design space]. (Kusumo, et al., 2020)” This method tended to create greater flexibility in detecting the probability threshold. It offered a higher availability of effective strategies for producing replacement plans. The major impediments of this method were that it was expensive and mainly tractable with low-dimensional design space (Kusumo, et al., 2020). This prevented the further analysis of the process operations.

The surrogate model approach mainly focused on the presence of disjoint feasible regions. These were categorised as nonconvex problems with high computational expenses. This methodology showed effective identification of complex feasible regions. It was primarily based on constructing a surrogate to describe the feasibility function, called “surrogate-based feasibility analysis (Geremia, et al., 2023).” The main restriction of this approach was that it required a large number of sampling points and a high computational burden.

The alpha-radius approach defined an alpha shape that outlined the largest geometrical shape where the points inside of it meet all the product specifications. It was an approach to tackle nonlinear problems. Three different methodologies, tolerance-based, resolution support, and combinatorial, were used to identify a design space with an alpha-shape representation. The combinatorial method gave the best outcome out of the three. However, the alpha-radius method still required a large

amount of data points and high computational time. As a methodology to describe the design space, it presented significant challenges (Sachio, et al., 2023).

1.3 Support Vector Machines

Support Vector Machines (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems (Stecanella, 2017). It performs exceptionally well with a limited number of data points. This makes SVM suitable to resolve the challenges in previous design space investigation approaches. The supervised learning model uses labelled datasets to train algorithms that accurately classify data or predict outcomes (IBM, n.d.).

SVM is particularly applicable to the case study due to its effectiveness in classifying the dataset into two groups and outlier detection. SVM has the capability to transfer the dataset to higher dimensions, which improves accuracy in the classification (IBM, n.d.). Moreover, as SVM applies to limited data, it can apply data reductions.

1.4 Objectives

This study aimed to create an SVM model for data classification from the case study. Improved upon the selected model by tuning the hyperparameters. Tested the model performance at varying dataset sizes to evaluate and compare their impacts.

2. Methodology

2.1 Dataset Formulation

The case study introduced the design decisions and KPI of interest to select a feasible bound for the design problem. Equation 2.1 to 2.3 describes the mathematical formulation of the design parameters.

$$\mathbf{y} = f(\boldsymbol{\theta}) \quad (2.1)$$

$$\theta_L \leq \theta \leq \theta_U \quad (2.2)$$

$$\mathbf{g}(\mathbf{y}) \leq 0 \quad (2.3)$$

where $\boldsymbol{\theta}$ is the vector of design decisions with the lower and upper bounds shown as θ_L and θ_U , respectively, \mathbf{y} illustrates the vector of monitored KPIs, f is the process model, and \mathbf{g} is the vector of performance constraints concerning the selected KPIs (Sachio, et al., 2023).

Determined design decisions would be used to create the dataset by the Sobol sequence, a low discrepancy quasi-random sequence (Academy, 2020). This meant that the data was evenly distributed within the input bound constraints, reducing the risk of overfitting (Rusch & TK., 2020). Sobol sequence thoroughly explores the input design space, which makes it more suited to real-world data. Thus, it could handle variations better. Equation 2.4 formulated the generation of this dataset

within the stated lower and upper bounds of the design decisions.

$$\theta_{in} = \text{Sobol}(\text{dim}, \theta_L, \theta_U, \text{sp}) \quad (2.4)$$

Based on the desired process specifications, the constraints on the desired product outputs were specified. As SVM solved binary classification problems, the output data needed to be classified into two groups. Data points satisfying all desired product constraints would be classified as 1, while data points failing one or more would be classified as 0.

2.2 Normalisation

Normalising the input data transformed them onto the same scale between zero and one. This improved the performance and training stability of the model. The linear scaling normalisation technique was used because the inputs were uniformly distributed across a fixed range (Anon., n.d.).

$$X_{normalised} = \frac{(X - X_{min})}{(X_{max} - X_{min})} \quad (2.5)$$

Equation 2.5 took the maximum and minimum of a chosen input. Then, it substituted that input data into X to normalise each point of that set.

2.3 Split, Test and Train

The split, test, and train technique was implemented in the model to estimate the performance of the machine learning algorithm that was used to make predictions (Online, 2023). This technique involved dividing the data into two parts: training and testing sets. The training set was employed to train the model, while the testing set was utilized to evaluate the model's performance (Online, 2023). This approach enabled training the models on one set and assessing their accuracy on an independent, unseen testing set. Following best practice, 80% of the data was used to train the model, while 20% was used to test the model (Gholamy, et al., 2018). It was observed that the data generated from the Sobol sequence was ordered, which would cause a biased split. Therefore, the data was shuffled before it was split for fairness.

SVM with the specified kernel was trained on the training dataset to produce a model. Subsequently, the test dataset features were fed into this trained model, predicting outputs as 1 or 0. These predicted outputs were then compared to the actual target values from the test dataset.

If the model predicted 1 and its actual target value was 1, this was a true positive (TP). On the other hand, if the model predicted a 0 and its real target value was 0, then this was a true negative (TN). When the prediction and the target value did not match up, this was a false positive (FP) or a false negative (FN). This prediction

summary was presented in a confusion matrix form, as seen in Figure 2.1 (Kulkarni, et al., 2020). It helped to identify the classes that were confused by the produced model.

		ACTUAL VALUES	
		Positive (1)	Negative (0)
PREDICTED VALUES	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 1.1 Confusion Matrix Diagram (adapted by [Mohajon, 2020])

Two core model evaluation metrics were calculated from the confusion matrix to analyse the results. The first metric was accuracy, as formulated in Equation 2.6.

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \quad (2.6)$$

Accuracy gauged the frequency of correct predictions from the classifier model (Agrawal, 2023). It divided the number of correct predictions by the total test dataset. This metric was suitable for analysing models trained off balanced datasets (Olugbenga, 2023).

The second metric was the Matthew Correlation Coefficient (MCC), formulated in Equation 2.7.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (2.7)$$

MCC measured the difference between the actual and predicted values. It ranged between -1 to 1, where 1 represents the best agreement (Voxco, 2021). It considered all values of the confusion matrix. This metric was suitable for analysing models trained in imbalanced datasets (Encyclopedia & Community, 2022).

2.4 Support Vector Machines Kernels

The kernel function in SVM was a similarity function that operated on input vectors on the original feature space and computed a modified inner product in a higher-dimensional space employing the kernel trick allowed for a non-linear SVM without explicitly mapping data into higher dimensions. The kernel trick became crucial, especially when linear regression was not feasible in the given input space. The kernel trick enabled linear separation by transforming the data into a higher-dimensional space. This constructed a linear model in the transformed space, corresponding to a

nonlinear model, by elevating the scalar product into higher-dimensional space without explicitly computing the mapping. The scalar product was substituted by various kernel functions in Table 2.1 with their mathematical formulation. In this study, the provided kernel function was used and compared to detect the most appropriate one (Bajorath, et al., 2022).

Table 2.1 Mathematical Representation of Different SVM Kernels (Matsuzaki, 2020)

Type	Mathematical Representation
Linear	$K(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \cdot \mathbf{x}$
2 nd Degree Polynomial	$K(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^T \cdot \mathbf{x}' + c)^2$
3 rd Degree Polynomial	$K(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^T \cdot \mathbf{x}' + c)^3$
4 th Degree Polynomial	$K(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^T \cdot \mathbf{x}' + c)^4$
Gaussian (RBF)	$K(\mathbf{x}, \mathbf{x}') = e^{\left(-\frac{\ \mathbf{x} - \mathbf{x}'\ ^2}{2\sigma^2}\right)}$
Sigmoid	$K(\mathbf{x}, \mathbf{x}') = \tanh(\alpha \mathbf{x}^T \cdot \mathbf{x}' + c)$

2.5 Optimisation of SVM Kernel

After selecting the most suitable kernel, the SVM model accuracy could be optimised with the hyperparameters. Two main hyperparameters were regularisation (C) and gamma (γ). The penalty parameter (C) was used to describe the misclassification. This showed the SVM optimisation about the bearable limit of the error. It provided a monitoring of the trade-off between the decision boundary and the misclassification term. Gamma explained the influence of individual training samples on the decision boundary. In other words, the high gamma value illustrated a more localised influence, while low gamma implied a broader influence (Liu, 2020).

Tuning the hyperparameters provided the finding of the values of C and gamma that resulted in the best model performance. Grid search was a common approach for hyperparameter tuning. It methodically generated and evaluated a model for each combination of parameters mentioned in a grid (Learn, n.d.). During the tuning procedure, it was essential to address overfitting. Overly tuning hyperparameters on the training data had the potential to make a poor generalisation to new, unseen data. In order to minimise this, cross-validation (cv) was used to assess the performance of different hyperparameter combinations (Tour, 2017).

2.6 Reducing Data

The SVM model worked better with a balanced dataset (Palade, et al., 2012). The dataset generated via the Sobol sequence can be imbalanced. If the dataset was imbalanced, the first data reduction step should only reduce the classification that is dominating the dataset. Once balanced, the number of data points classified as 0s was equal to the amount classified as 1s. This prevented misguided performance evaluation and created fairness within the dataset. The balanced dataset was subsequently reduced in size to analyse its effect on the model accuracy.

3. Application of Protein A Chromatography

3.1 Protein A Chromatography

Protein A chromatography was the initial stage in monoclonal antibodies (mAbs) downstream purification. This procedure aimed to capture the product, eliminate impurities, and reduce sample volume. The resulting process yield showed the performance of this procedure. However, variability in input mixture composition from cell-based upstream production systems posed a challenge. Therefore, the design space identification should provide for the manipulation of feed variability (Sachio, et al., 2023).

A process by Steinebach et al., seen in Figure 3.1, was used to understand the case study. It was a cyclic multicolumn approach that was employed in three steps (A, B, C) to process the feed continuously. Initially (Step A), column 1 received fresh feed while column 2 got the output from column 1. In step B, column 1 was washed while column 2 received the output of column 1 and fresh feed. Finally, column 1 was regenerated in Step C, while column 2 received fresh feed (Steinebach, et al., 2016).

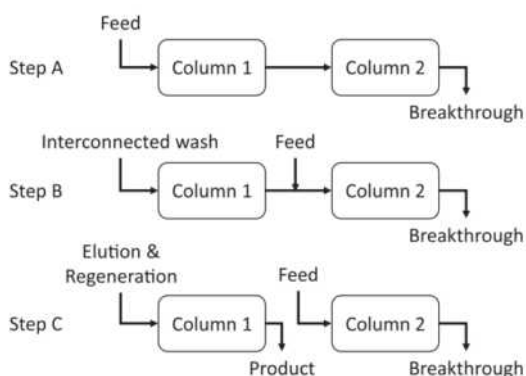


Figure 3.1 Multicolumn Protein A Chromatography Diagram (adapted by [Steinebach, et al., 2016])

The flexibility and performance of the process were investigated under three different variables: (i) mAb concentration in the feed stream (c feed), (ii) feed volumetric flow rate (Q feed), and (iii) the column

switching time (T switch). The corresponding key performance indicators (KPIs) were yield and productivity (Steinebach, et al., 2016).

3.2 Data Classification

In total, 4096 data were generated by the Sobol sequence. Each data point had three inputs and two outputs. The inputs and their respective bounds were in Table 3.1 below.

Table 3.1 Lower and Upper Bounds of Design Decisions (Inputs)

Inputs	Lower Bound	Upper Bound
Feed concentration (c feed) [mg ml^{-1}]	0.21	0.63
Feed volumetric flowrate (Q feed) [mg min^{-1}]	0.5	1.5
Switch time (T switch) [min]	40	120

The resulting outputs from these inputs were yield and productivity. The product specification constraints were that the yield and productivity must be greater than or equal to 99% and $4 \text{ mg ml}^{-1} \text{ h}^{-1}$, respectively. As SVM solved binary classification problems, the current data needed to be classified into two groups. Data points satisfying both desired product constraints would be labelled 1, while data points failing one or both would be marked as 0 via sorting code.

The 4096 dataset was imbalanced, consisting of 1269 data points classified as 1 while 2827 as 0. The MCC score was used to compare the performance of training the model with different kernels.

4. Results and Discussion

4.1 Comparison of Different SVM Kernels

Table 4.1 The MCC Results with Their Corresponding SVM Kernels

Type	MCC (%)
Linear	40.43
2 nd Degree Polynomial	62.05
3 rd Degree Polynomial	69.25
4 th Degree Polynomial	70.22
Gaussian (RBF)	94.12
Sigmoid	-41.01

The linear kernel did not perform very well, having an MCC score of 40.43%. Changing to a 2nd order polynomial, the MCC score increased by 21.62% to 62.05%. Increasing the order of the polynomial to 3rd degree increased the MCC score by 7.2% to 69.25. Increasing the order polynomial further showed a slight improvement of 0.97%.

The optimum kernel was Gaussian, with an MCC score of 94.12%. This outperformed the 4th degree polynomial by 23.9%, making it a suitable kernel to tune.

The sigmoid kernel performed poorly as it failed to classify any true values correctly. The mathematical representation of the sigmoid function shown in Table 2.1 was thus unsuitable for the dataset.

4.2 Comparison of Tuned and Untuned Kernel

Table 4.2 Optimum Hyperparameter Values at Different Dataset Sizes

Dataset Size	C	Gamma
4096	100000	1.25
2538	100000	1
1500	100000	0.5
1000	10000	0.5
200	1000	1

The investigated C parameters were 1, 10, 100, 1000, 10000, and 100,000. Optimum C for datasets 4096, 2548, and 1500 was the max value from the Grid Search approach at 100,000. It was found that for the larger dataset sizes, the Grid Search approach always favoured a higher C value. A repetition was done with the Grid search, including C equalled to 1,000,000. The larger datasets found this parameter to be the best value; however, the runtime of the model doubled as it was much more computationally expensive. Despite the doubled computational cost, the accuracy showed a negligible change. Thus, the Grid Search max C value was designed to be 100,000. The lowest optimum C chosen was 1000 for the size 200 dataset. This meant that a sufficient range of C was provided to find the optimum parameter.

The investigated gammas in the Grid Search were 0.25, 0.5, 1, 1.25, 1.5 and 2. The optimum gamma value for varying datasets was found to vary between 0.5 and 1.25. None of the optimum gamma values were at the minimum or maximum gamma provided. This meant a sufficient range of gamma values was provided to explore.

As dataset size decreased, optimum hyperparameter C decreased. This was because a lower C results in a smoother decision boundary, thus reducing overfitting. In smaller datasets, overfitting was a greater risk.

The gamma hyperparameter decreases as the dataset size decreases. Then, at the low dataset size of 200, it increased. The gamma might have increased because, at dataset size 200, the C hyperparameter was two orders of magnitude smaller than the 1500, 2538, and 4096 datasets. The higher gamma value increased the complexity, picking up on non-linearity that may have been lost with both a low C and low gamma value.

4.3 Mean Accuracy Improvement with the Tuned Gaussian Kernel

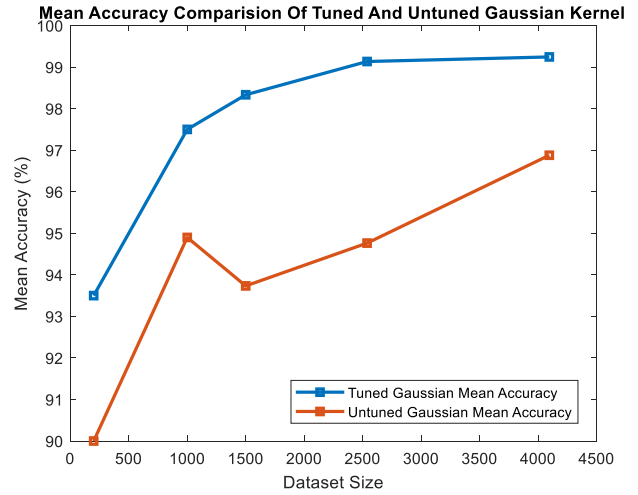


Figure 4.1 Mean Accuracy at Varying Dataset Sizes for the Tuned and Untuned Gaussian Kernel

In Figure 4.1, the tuned Gaussian kernel outperformed the untuned Gaussian kernel on every dataset. The tuned mean Gaussian accuracy followed a logarithmic profile in relation to the dataset size. The untuned Gaussian mean accuracy decreased as the dataset size decreased. Interestingly, there was a spike in performance at dataset 1000 for the untuned Gaussian kernel. This could be because the default untuned hyperparameter C equalled to 1, which was slightly suited to this dataset size. As previously discussed, C decreased as the dataset size decreased.

Table 4.3 Accuracy Results Across Five Repetitions for Varying Dataset Sizes

Dataset Size	Mean Accuracy	Variance	Maximum	Minimum
4096	99.244	0.068	99.634	98.902
2538	99.134	0.114	99.606	98.622
1500	98.333	1.851	99.667	97.000
1000	97.500	5.641	99.500	95.000
200	93.500	46.992	97.500	87.500

Across five repetitions, the dataset of 4096 data points produced a mean accuracy of 99.24% with a variance of 0.068. The dataset size was then reduced by 38% to 2538 data points. The size 2538 dataset was balanced, with an equal amount of data classified as 1 and 0. Despite the 38% reduction in data points, the mean accuracy only decreased by 0.11%. This minimum reduction was because SVM performs optimally with balanced datasets.

The 4096 dataset was reduced in size by 75.6% to the 1000 dataset. This resulted in the mean accuracy decreasing by 1.74% to 97.5%.

The 4096 dataset was reduced in size by 95.1% to the size 200 dataset. The resultant mean accuracy was

93.5%, which was 5.74% lower than the mean accuracy for the dataset size of 4096.

As shown in Figure 4.1, the accuracy declined faster at lower dataset sizes. Going from dataset size 4096 to 200 was an extra reduction of 19.5% than going from 4096 to 1000. However, the mean accuracy drop increased from 1.74% to 5.74%, which was 3.3 times as large.

4.4 Variance of Accuracy Across Repetitions for Different Dataset Sizes

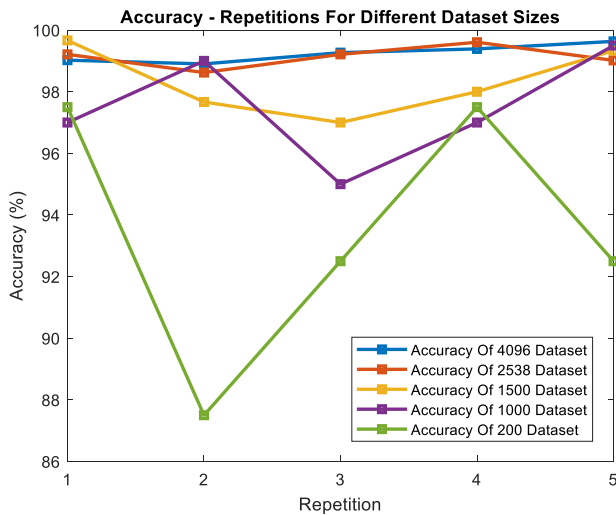


Figure 4.2 Accuracy against Repetition for Varying Dataset Sizes

Reducing the dataset size caused the variance of accuracy for that dataset to increase. As shown in Figure 4.2, the size 200 dataset varied across a larger range than all the other datasets. Figure 4.2 also indicated the independence of results between repetitions having no trend.

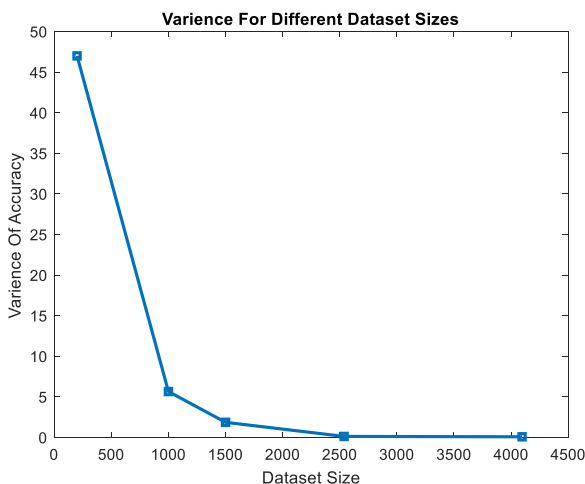


Figure 4.3 Variance of Accuracy against Dataset Sizes

Figure 4.3 showed the exponential decay of the variance of the accuracy as the dataset size increases. In dataset 1500, the variance was 1.85, and then at 2500, it was 0.11, suggesting swift decay. The variance quickly increased below the size 1000 dataset, meaning the model was unreliable and contained more randomness.

4.5 Removing Bias from Accuracy Metric

There was unfairness present in the 80/20 split accuracy score when varying dataset sizes. Splitting different size datasets 80/20 naturally resulted in them being tested against different amounts of data points. This means the tests performed across different dataset sizes were not consistent between them. The resulting accuracy score from these tests was not suitable to compare the performance between different dataset sizes. Due to the smaller test size for the smaller datasets, the randomness in the outputted accuracy was higher between repetitions. This resulted in a higher variance.

To account for this, an overall accuracy score metric was calculated. This was achieved by using all 4096 data points as test data. It was tested against the same models produced at varying dataset sizes that the original accuracy score was tested against. It was calculated using Equation 2.6, which was used to get the original accuracy score. The difference was that the test dataset consisted of the whole 4096 dataset. Thus, each model built from varying dataset sizes, was tested against all 4096 data points. The resulting accuracy score was called the overall accuracy metric. Using the same test dataset to evaluate each model improved the consistency of the outputted overall accuracy metric. This allowed better comparisons of performance between models. However, by reducing the unfairness in the calculation of accuracy from the test data dataset, a new unfairness was introduced in how the model was trained.

Testing the model against the whole dataset meant the model had been trained on some of this data. The model had a high probability of correctly predicting data it had been trained on, making it unfair. This unfairness was increased for the larger dataset size as they trained on a larger portion of the dataset than the smaller datasets. For example, the size 4096 dataset was trained on 3277 data points, while the size 200 dataset was trained on 160 data points.

To resolve this issue, the test accuracy score metric was calculated. A new test dataset consisting of 3927 data points was generated via the Sobol sequence. This test dataset was within the same bounds as the original 4096 dataset. Similar to the overall accuracy metric, it was tested against the same models produced at varying dataset sizes and used Equation 2.6 to get its accuracy score. The test accuracy metric was fairer across datasets as each model was tested against the same amount of data and had not been trained on it. This minimised variance of the test accuracy metric between repetitions and removed bias.

4.6 Overall Accuracy Metric

Table 4.4 Overall Accuracy Results Across Five Repetitions for Varying Dataset Sizes

Dataset Size	Mean Overall Accuracy	Variance	Maximum	Minimum
4096	99.707	0.022	99.854	99.463
2538	99.331	0.032	99.585	99.072
1500	98.667	0.005	98.731	98.584
1000	98.189	0.165	98.657	97.559
200	94.170	1.268	95.459	92.505

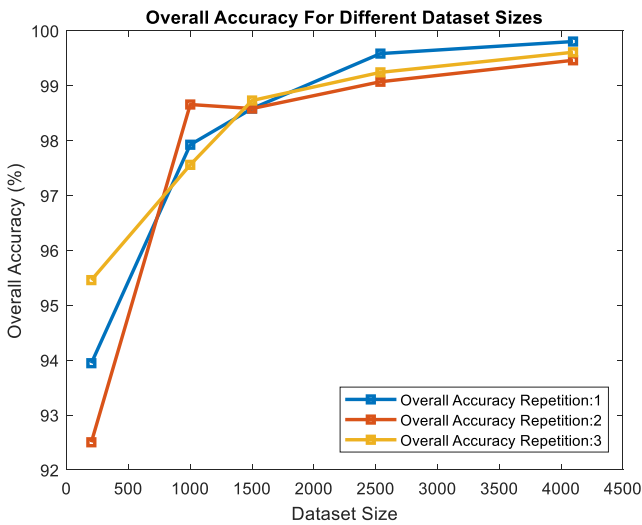


Figure 4.4 Overall Accuracy Against Dataset Size for Three Repetitions

The natural profile without taking a mean across repetitions was logarithmic as opposed to the more sporadic accuracy outputs shown in Figure 4.2.

The overall accuracy outputs from 3 repetitions for the size 4096 dataset were very close together, illustrating slight variance between repetitions. By contrast, the overall accuracy outputs for the size 200 dataset were more spread apart, evidencing a higher variance between repetitions. For the overall accuracy, as the dataset size increased, the variance decreased. This trend was also present for the accuracy. However, the variance in each dataset was smaller for the overall accuracy metric than for the accuracy metric. At dataset 200, the overall accuracy and its variance were 1.268 and 46.992, respectively. At a higher dataset of 1000, the overall accuracy and its variance were 0.165 and 5.641, respectively. For both datasets, the variance in accuracy has been reduced by over 30 times. This evidences that the overall accuracy metric has improved the consistency between repetitions by decreasing the variance at all dataset sizes.

4.7 Test Accuracy Metric

Table 4.5 Test Accuracy Results Across Five Repetitions for Varying Dataset Sizes

Dataset Size	Mean Test Accuracy	Variance	Maximum	Minimum
4096	99.598	0.002	99.669	99.516
2538	99.302	0.012	99.440	99.109
1500	98.737	0.019	98.981	99.559
1000	98.146	0.506	98.905	97.123
200	94.006	1.496	95.544	92.284

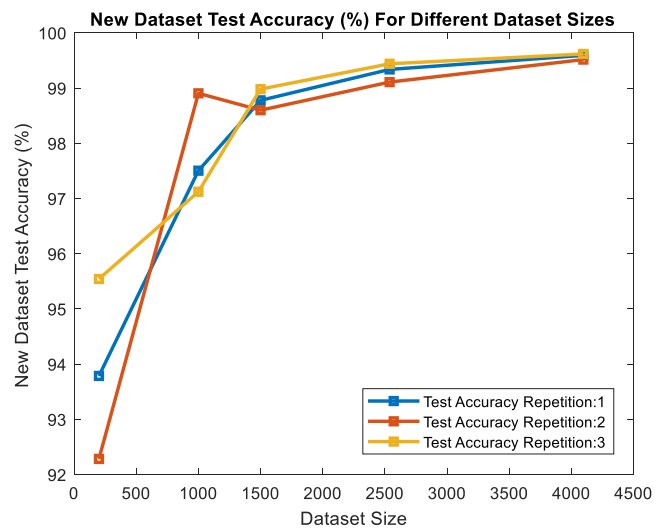


Figure 4.5 Test Accuracy against Dataset Size for Three Repetitions

The results in Figure 4.5 of the test accuracy were fair. This was because the test accuracy dataset was unique. Despite this, the test accuracy exhibited similar trends to the overall accuracy. It showed logarithmic behaviour with the dataset size, and the variance decreased as the dataset size increased.

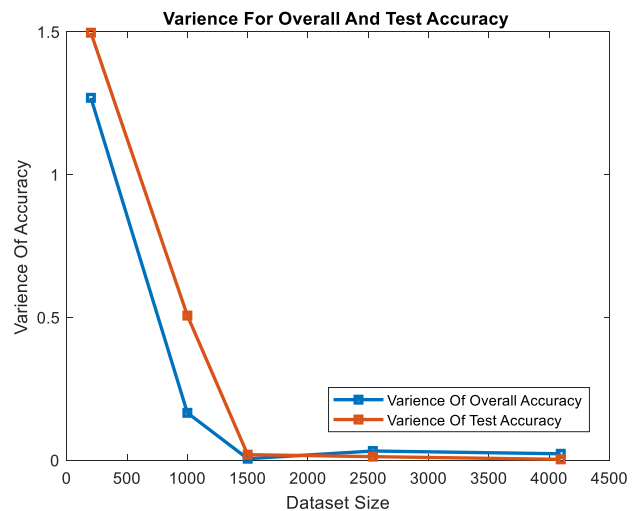


Figure 4.6 Comparison of Variance Between Overall Accuracy and Test Accuracy

Interestingly, the test accuracy, when compared to overall accuracy had a smaller variance at the higher dataset sizes but a larger variance at the smaller dataset sizes. As shown in Figure 4.6, this switch occurred at the size 1500 dataset. The overall accuracy below that point had lower variance due to its inherent bias in how it was trained. Above this point, the variance was essential the same since the higher datasets produced more consistent results.

4.8 Accuracy Metrics Discussion

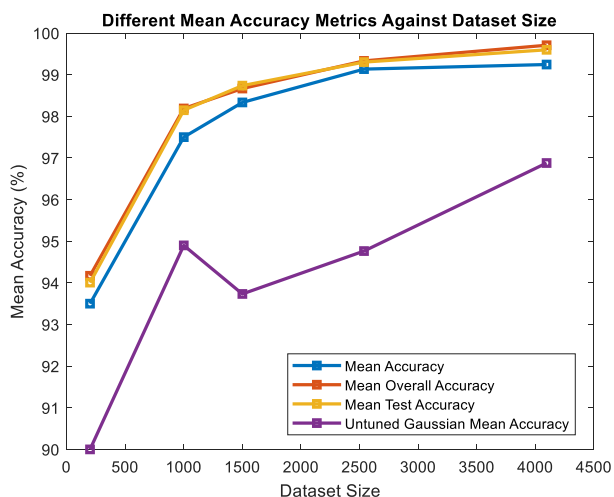


Figure 4.7 Different Mean Accuracy Metrics against Dataset Size

In Figure 4.7, the mean overall accuracy was consistently larger than the mean test accuracy except at dataset size 1500. This trend evidenced the predicted training bias in the overall accuracy score. From Table 4.4 and Table 4.5, the largest increase in mean accuracy from the test accuracy to the overall accuracy was 0.164% at dataset 200. The subsequent largest increase was 0.104% for dataset 4096. Thus, the found bias was minimal. This was shown in Figure 4.7 by the overall accuracy values only being slightly above the test accuracy values.

In Figure 4.7, the test accuracy and overall accuracy were consistently higher than the mean accuracy score. This was because the variance was higher for the original accuracy outputs, causing the resultant mean to be lower. This variance was a result of being tested on less data. The maximum test data for the original accuracy calculation used 820 data points from the size 4096 dataset. This test data set size then decreased as the actual dataset the split was performed on decreased. In contrast, the test data set size for the overall and test accuracy metrics was consistently 4096 and 3927, respectively.

The mean test accuracy was 99.598% at dataset 4096. Reducing the dataset by 75.6% to a dataset of 1000 caused the mean test accuracy to reduce by 1.452% to 98.146%. Reducing the size 4096 dataset by 95.1% to a

dataset of 200 data points caused the mean test accuracy to reduce by 5.592%. In Section 3.3, this same reduction in dataset size from a 4096 dataset to a size 200 dataset resulted in the mean accuracy reducing by 5.74%. For the same decrease in dataset size, the mean test accuracy declined by 0.148% less than the mean accuracy.

5. Conclusion

This study focused on the primary challenges of the design space approach for the biopharmaceutical industry. It successfully implemented a support vector machines approach for characterising the design space. This approach was demonstrated through an industrial case study of Protein A chromatography used in biopharmaceutical manufacturing. The design decisions – feed composition, flow rate, and switching time – were considered simultaneously with their impact on product quality. The design space was generated by the inputs that follow the constraints on productivity and yield.

As limited data availability was a primary setback within the industry, a data reduction analysis was conducted. The Gaussian kernel was found to perform the best with the design space dataset, producing a high Matthew Correlation Coefficient. A suitable exploratory range for the hyperparameters C and gamma was provided, allowing the optimum hyperparameters to be found specific to each dataset size. The tuned Gaussian kernel significantly improved the model's performance, producing high accuracy values for the varying dataset sizes.

However, for the accuracy metric, as the dataset size decreased, the variance increased, making the model less reliable at lower dataset sizes. To account for this the performance of the model was later assessed by feeding in the overall dataset and a new dataset created from a Sobol sequence. The variance in the accuracy was minimised by using the test accuracy metric. This metric also allowed for a fair performance comparison between the models trained at their respective dataset sizes. The overall accuracy delivered a similar high accuracy and low variance but was slightly bias as it was tested with data it had trained on.

From the mean test accuracy, it can be established that support vector machines is a good model approach for design space characterisation in the biopharmaceutical industry. With minimal randomness, its high accuracy results, suitability with smaller datasets, and short computational time make it a great technique. This high performance, however, quickly begins to decrease at the much lower dataset sizes.

6. Outlook

Expanding this study further, one could investigate to eliminate randomness within the model entirely. Additional analysis on the optimisation hyperparameters could be done to minimise its randomness. During this study, the data was reduced to 200 points. This could be decreased further while keeping the accuracy within the desired range. Therefore, an advanced model could be developed to make more robust predictions, resulting in high accuracy with smaller datasets. Furthermore, the support vector machines model approach could be assessed to understand its contribution to risk analysis.

7. Acknowledgements

We would like to express our genuine appreciation to Steven Sachio for his assistance, time, and wisdom throughout his research project.

8. References

- Academy, G. E., 2020. *SELECTION OF QUASI-RANDOM POINTS BASED ON SOBOLE*. [Online] Available at: <https://academy.gannetsolutions.com/wp-content/uploads/2020/04/SOBOLE-SEQUENCES.pdf> [Accessed 5 12 2023].
- Agrawal, S. K., 2023. *Metrics to Evaluate your Classification Model to take the right decisions*. [Online] Available at: <https://www.analyticsvidhya.com/blog/2021/07/metrics-to-evaluate-your-classification-model-to-take-the-right-decisions/#h-accuracy> [Accessed 8 12 2023].
- Anon., n.d. *Machine Learning Normalisation*. [Online] Available at: <https://developers.google.com/machine-learning/data-prep/transform/normalization#:~:text=The%20goal%20of%20normalization%20is,training%20stability%20of%20the%20model> [Accessed 7 12 2023].
- Bajorath, Rodríguez-Pérez, R. & Jürgen, 2022. Evolution of Support Vector Machine and Regression Modeling in Chemoinformatics and Drug Discovery. *Journal of Computer-Aided Molecular Design*.
- Encyclopedia & Community, S., 2022. *Matthews Correlation Coefficient*. [Online] Available at: <https://encyclopedia.pub/entry/35211> [Accessed 9 12 2023].
- Geremia, M., Bezzo, F. & Ierapetritou, M. G., 2023. A novel framework for the identification of complex feasible space. *Elsevier*.
- Gerogiorgis, Diab, S. & I., D., 2020. Design Space Identification and Visualization for Continuous Pharmaceutical Manufacturing. *PubMed Central (PMC)*.
- Gholamy, A., Kreinovich, V. & Kosheleva, O., 2018. Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation. *UTEP*.
- Hong, M. S. et al., 2017. Challenges and opportunities in biopharmaceutical manufacturing control. *Elsevier*, pp. 106-114.
- IBM, n.d. *What is supervised learning?*. [Online] Available at: <https://www.ibm.com/topics/supervised-learning#:~:text=the%20next%20step-What%20is%20supervised%20learning%3F,data%20or%20predict%20outcomes%20accurately> [Accessed 2 12 2023].
- Kulkarni, A., Chong, D. & Batarseh, F. A., 2020. Foundations of data imbalance and solutions for a data democracy. *Academic Press*, pp. 83-106.
- Kusumo, K. P., Gomoescu, L., Paulen, R. & Muñoz, S. G., 2020. Bayesian Approach to Probabilistic Design Space Characterization. *I&EC Research*.
- Learn, S., n.d. *Tuning the hyper-parameters of an estimator*. [Online] Available at: https://scikit-learn.org/stable/modules/grid_search.html [Accessed 10 12 2023].
- Liu, C., 2020. *SVM Hyperparameter Tuning using GridSearchCV*. [Online] Available at: <https://www.vebuso.com/2020/03/svm-hyperparameter-tuning-using-gridsearchcv/> [Accessed 9 12 2023].
- Manzon, D. et al., 2020. Quality by Design: Comparison of Design Space construction methods in the case of Design of Experiments. *Elsevier*, Volume 200.
- Matsuzaki, T., 2020. *Mathematical Introduction to SVM and Kernel Functions*. [Online] Available at: <https://tsmatz.wordpress.com/2020/06/01/svm-and-kernel-functions-mathematics/> [Accessed 9 12 2023].
- Matsuzaki, T., 2020. *Mathematical Introduction to SVM and Kernel Functions*. [Online] Available at: <https://tsmatz.wordpress.com/2020/06/01/svm-and-kernel-functions-mathematics/> [Accessed 7 12 2023].
- Mohajon, J., 2020. *Confusion Matrix for Your Multi-Class Machine Learning Model*. [Online] Available at: <https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model->

ff9aa3bf7826

[Accessed 4 12 2023].

Olugbenga, M., 2023. *Balanced Accuracy: When Should You Use It?*. [Online] Available at: <https://neptune.ai/blog/balanced-accuracy#:~:text=Accuracy%20can%20be%20a%20useful,related%20to%20the%20accuracy%20paradox>.

[Accessed 9 12 2023].

Online, S., 2023. *All About Train Test Split*. [Online] Available at: <https://www.shiksha.com/online-courses/articles/train-test-split/#:~:text=A%20train%20test%20split%20is,on%20the%20unseen%20testing%20set>.

[Accessed 7 12 2023].

Palade, Batuwita, R. & Vasile, 2012. Class Imbalance Learning Methods for Support Vector Machines.

Ralf Otto, A. S. a. U. S., 2014. *Rapid growth in biopharma: Challenges and opportunities*. [Online] Available at: <https://www.mckinsey.com/industries/life-sciences/our-insights/rapid-growth-in-biopharma>

[Accessed 22 November 2023].

Rusch, S. M. & TK., 2020. Enhancing accuracy of deep learning algorithms by training with low-discrepancy sequences. *Seminar für Angewandte Mathematik*.

Sachio, S., Kontoravdi, C. & Papathanasiou, M. M., 2023. A model-based approach towards accelerated. *Elsevier*.

Stecanella, B., 2017. *Support Vector Machines (SVM) Algorithm Explained*. [Online] Available at: <https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/>

[Accessed 4 12 2023].

Steinebach, F. et al., 2016. Model based adaptive control of a continuous capture process for monoclonal antibodies production. *Elsevier*, Volume 1444, pp. 50-56.

Tour, T. D. I., 2017. *Grid-search and cross-validation*. [Online] Available at: https://pactools.github.io/auto_examples/plot_grid_search.html

[Accessed 10 12 2023].

Voxco, 2021. *Matthews's correlation coefficient: Definition, Formula and advantages*. [Online] Available at: <https://www.voxco.com/blog/matthewss-correlation-coefficient-definition-formula-and-advantages/>

[Accessed 9 12 2023].

Yamada, Y. et al., 2022. Reduce Environmental Impact and Carbon Footprint for Cost Competitive Process Plant Design: Integrating AVEVATM Process Simulation with modeFRONTIER®. *Elsevier*, Volume 49, pp. 211-216.

Impact of Predicted Data on ML-QSPR Predictions of Lower Flammability Limits for Pure Compounds

Bing Zheng Cheong and Rouxuan Chow

Department of Chemical Engineering, Imperial College London, U.K.

Abstract

This study explores the impact of incorporating predicted data into the training sets of Machine Learning-based Quantitative Structure-Property Relationships (ML-QSPR) models, focusing on the prediction of Lower Flammability Limits (LFLs) in pure compounds which is crucial for safety in the oil, gas, and petrochemical industries. Utilizing data from the DIPPR 801 Project database, two distinct subsets were created: a mixed subset containing both experimental and predicted LFL values, encompassing 1,506 compounds, and an experimental subset comprising exclusively experimental data, with 299 compounds. Various regression models, including Ridge Regression, Lasso, Partial Least Squares (PLS), Support Vector Machines (SVM), and Kernel-based Partial Least Squares (KPLS), were employed. The robustness, internal predictivity, and stability of these models were rigorously assessed through Q_{LMO}^2 validation and statistical tests such as y-scrambling and pseudo-descriptor tests. External predictivity was evaluated using distinct test sets and quantified via the Q_{Ext}^2 metric. The performance of the developed models was further analyzed using a range of metrics including R^2 , RMSE, MAE, and MAPE. Additionally, the applicability domain for the LASSO and PLS models was explored, providing insights into the models' reliability within their respective chemical space. Findings indicate that whilst the inclusion of predicted data in training sets improved feature selection and robustness against overfitting, it reduces predictivity on unseen data. This research provides insights into the benefits and trade-offs of using predicted data in ML-QSPR models, suggesting its potential utility in preliminary screening processes where maximum predictivity is not paramount.

1. Introduction

The lower flammability limit (LFL), usually expressed in percentage volume (vol%), is defined as the minimum concentration of fuel in air required to sustain flame propagation [1]. It is an important physicochemical property that characterises the flammability and combustibility of substances, knowledge of which is crucial to maintaining process safety within the oil, gas, and petrochemical industry [2]. Common uses of LFL values include input parameters for risk assessments or informing the start-up of a reactor outside of flammable ranges [1]. As such, reliable and accurate LFL values are required to maximize the process safety in design and operational procedures.

Traditionally, LFLs are determined experimentally through various standard testing methods such as: (i) US Bureau of Mines Tube Method, (ii) ASTM E-681/E-918, (iii) DIN 51648/EN 1839, amongst many others [1]. However, experimental determination of LFLs can be expensive, time consuming and pose safety risks for especially reactive substances. Additionally, the experimental LFL is not absolute, but depend on several factors, such as the geometry of the apparatus, the type and strength of the ignition source, the test pressure and temperature and the degree of mixing [3]. Coupled with the rapid introduction of new chemicals in the process industry, the development of alternative methods to predict LFLs with reasonable accuracy is required.

In response to these challenges, Machine learning based quantitative structure property relationships (ML-QSPR) have emerged as a popular alternative for predicting LFLs. This approach leverages ML algorithms to establish a mathematical relationship between molecular descriptors and the physicochemical properties of a molecule where molecular descriptors are used to numerically encode various structural aspects of

a molecule. This method, when compared to approaches such as calculated adiabatic flame temperature (CAFT) estimation of LFLs, demonstrates a distinct advantage: a reliable QSPR model only requires the molecular structure of a chemical to predict LFLs, which enables evaluation of even unprepared chemicals.

To highlight studies in this field, in a seminal study presented in 2008, Gharagheizi [4] utilized data from the DIPPR 801 database to develop a predictive model for LFLs of organic compounds. Employing a dataset comprising 1056 LFL data points, he crafted a multiple linear regression (MLR) model consisting of four descriptors. The dataset was partitioned into a training set, consisting of 845 compounds (80%), and a test set with 211 compounds (20%). The model demonstrated robust predictive power, evidenced by a high goodness-of-fit (R^2) of 0.9698, a mean absolute percentage error (MAPE) of 7.68% and a root mean square error (RMSE) of 0.1561 across the entire set of compounds.

Following this, Pan et al. [5] in their 2009 study developed a MLR model with four descriptors to predict the LFL of pure hydrocarbons. This model was based on a dataset of 354 pure hydrocarbons sourced from the DIPPR 801 project, with an 80/20 split into 284 compounds for training and 70 for testing. The model's performance was quantified using several metrics: it achieved a R^2 of 0.9967, RMSE of 0.07, MAPE of 6.07 %, and a mean absolute error (MAE) of 0.043 across the entire dataset.

It is important to note that all the studies above developed QSPR models using LFL data sourced from the DIPPR 801 Project database. However, due to the lack of reliable and consistent experimental values [1], a large majority of LFLs reported are predicted values rather than being experimentally derived. This goes

against the common notion that prediction methods should be exclusively built off experimental data [6].

In a paper authored by Chen et.al [7], a MLR model of four descriptors was built on the same DIPPR 801 database. However, only experimental data was used, where an 80/20 train/test split was used leading to 366 molecules in the training set and 92 molecules in the test set. This model achieved an R^2 of 0.9302 and a MAE of 0.3036 on the test set.

The current lack of research into the effects of incorporating predicted data in the training set for QSAR model development necessitates a thorough investigation. This gap is significant, considering the potential benefits of using predicted data, such as enlarging the dataset to combat issues such as imbalanced data and model overfitting [8]. Overfitting is particularly problematic given the vast array of molecular descriptors available for calculation, which can number in the hundreds or thousands [9]. Moreover, the prevalence of predicted data in many esteemed "gold standard" databases underscores the need to explore this approach. Accordingly, this study aims to determine the impact of using a mixed dataset, comprising of both experimental and predicted data within the training set, on the reliability and predictive accuracy of a QSPR model, specifically in relation to its ability to predict Lower Flammability Limits (LFL).

2. Materials and Methodology

The following section describes the methodology used for model development and validation in this paper, which has been summarized in Figure 1.

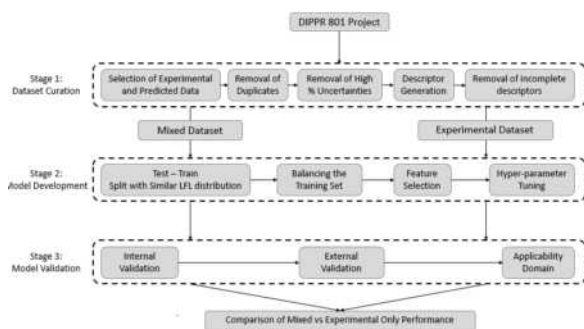


Figure 1: QSPR modelling workflow for comparison of models developed on a mixed versus experimental dataset.

2.1. Dataset Curation

The dataset used in this study was retrieved from the 2021 edition DIPPR 801 Project database which is referred to as the "Gold Standard" for pure component property values. DIPPR 801 is the largest collection of critically evaluated, pure component thermophysical properties and is maintained by the Design Institute for Physical Properties (DIPPR), which was established by the American Institute of Chemical Engineers (AIChE) in 1980. To ensure the "completeness" of recorded pure component properties, prediction methods have been used to obtain recommended values where no reliable experimental data is available. Additionally, the "accuracy" or quality of the values are dependent on a

well-developed and stringent evaluation process, which aims to publish results that are complete, consistent with literature, as well as self-consistent [6]. This ensures that predicted values still achieve the same standards of consistency and reliability as experimental values, which is vital when utilizing predicted data for model development.

The DIPPR 801 database, which contains Hazard and Safety Properties for 1,965 compounds, has reported LFLs for 1,896 of these compounds. Among these, only 474 are based on experimental measurements, accounting for just 25% of the LFL data in 2021 Edition DIPPR 801. The remaining 1,422 compounds were classified under predicted, smoothed, not specified or unknown. For this study, only predicted and experimental LFL values were considered, resulting in a total of 1,831 compounds. To ensure data integrity, duplicated molecules were identified and discarded based on their canonical SMILES generated using the RDKit software.

Additionally, it's essential to address experimental errors when working with data, as they can negatively impact the quality of QSAR models and subsequently affect the prediction of new compounds [10]. The DIPPR 801 database categorizes its property values with specific levels of quantized uncertainty. A study done by Wenlock and Carlsson [11] found that QSPR models built on data with low experimental uncertainties gave rise to prediction improvements ranging from 3.3% to 27.5%. As such, this concept was applied to experimental and predicted values by including compounds from the first 6 quantized uncertainty levels, which are <0.2%, <1%, <3%, <5%, <10% and <25%. This led to a total of 1,764 compounds in the dataset.

2.2. Generation and Pre-processing of Descriptors

Molecular descriptors were computed using Mordred, an open-source descriptor calculation software. Mordred can generate up to 1,825 descriptors, a figure comparable to that of the PaDEL-Descriptor software, which calculates up to 1875 descriptors and fingerprints [12]. The choice of Mordred was influenced by its ability to calculate descriptors twice as fast, handle descriptor generation for large molecules, and its compatibility with both Python 2 and 3 [12].

Calculations of descriptors were based on the minimum energy molecular geometries of the compounds as optimized by DataWarrior based on the MMFF94 force field. Resulting structures were used as inputs to Mordred for computation of 2D and 3D descriptors.

Descriptors that failed to generate for more than 95% of compounds were removed from our dataset. Furthermore, to enhance data quality, compounds with over 200 missing descriptors were also excluded. Following these exclusions, any remaining descriptors that failed to generate were removed. This was done to retain the maximum the number of compounds and calculated descriptors available for model development.

Two distinct subsets were then created with one comprising a mix of both experimental and predicted LFL values in the training set while the other solely contains experimental LFL values in the training set. Both test sets for each subset consists of only experimental values of different sizes. For both subsets, a 90/10 train/test split was used while ensuring the distribution of LFLs between the training and testing set are similar.

Finally, the training set was balanced according to the following best practices [13]:

- i) The uncertainty of LFLs should be five times smaller than the total range of LFLs in the dataset.
- ii) No large gaps that exceed 15% of the entire range of LFLs are allowed between two consecutive LFLs ordered by magnitude.

Figure 2 and 3 presents the histograms for LFLs categorized by use in model training or testing for the final mixed and experimental subset. From the figures, it can be observed that the distributions of the train and test sets are similar, hence the test set is suitable for validation of models developed using the training set.

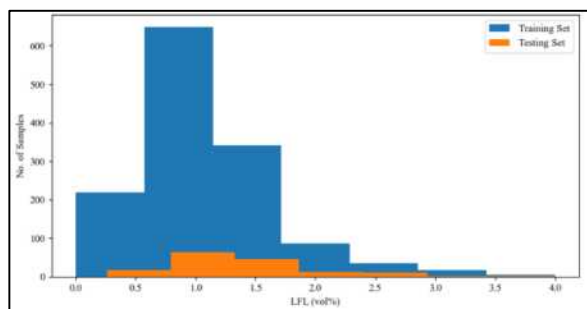


Figure 2: Histogram of LFLs for the mixed subset categorized as the train set (1354 observations) and test set (152 observations).

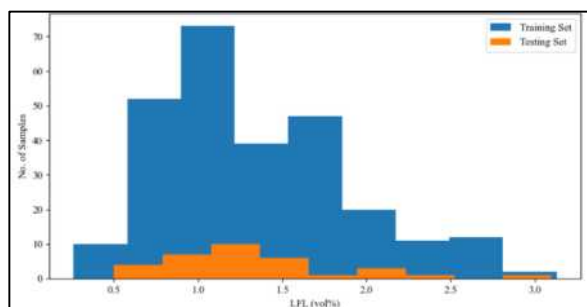


Figure 3: Histogram of LFLs for the experimental subset categorized as the train set (266 observations) and the test set (33 observations).

Table 1: Range of LFLs in the train and test set for the mixed and experimental subset.

Subset	Training Set Range (vol%)	Testing Set Range (vol%)
Mixed	0.00001 – 4	0.26 – 4
Experimental	0.26 – 3.13	0.5 – 3.1

Table 1 presents the LFL ranges for both train and test sets of each subset. As the range of LFL values in the test set for both the mixed and experimental subset fall within their respective training set ranges, provided the

correct descriptors and model parameters were selected, no significant extrapolation for test set predictions should occur.

2.3. Feature Selection

Feature selection is also critical for developing interpretable and accurate predictive models as it can reduce the risk of overfitting and identify important features with meaningful property relationships in the data. For this study, feature selection was carried out via the filter method using Pearson's correlation coefficients due to it being commonly used and ease of adaptability [14]. Descriptor reduction for this study followed a three-step process:

- i) Removal of descriptors that were constant across all compounds, as the descriptors did not encode the structural differences between compounds that contributed to their differing LFLs.

- ii) Elimination of descriptors with low correlation to the target variable (defined as having an absolute correlation coefficient less than 0.1). These features are unlikely to be significant predictors for the target variable.

- iii) Removal of one descriptor from each pair of highly correlated descriptors (defined as absolute pairwise correlation above 0.9). This step was to minimize multicollinearity, a condition where two or more descriptors are highly correlated, leading to redundancy in information and potential instability in model predictions.

2.4. Machine Learning Algorithms Used

Regression methods are essential in QSAR (Quantitative Structure-Activity Relationship) models. At the heart of these regression techniques is the least squares criterion, which aims to minimize the sum of the squared differences between the observed LFL values and those predicted by the model. This can be mathematically represented as seen in Equation 1.

$$\min_{\beta} \sum_{i=1}^n \left(y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2 \quad (1)$$

In this equation, n represents the number of samples, p is the number of molecular descriptors, y represents the response vector (LFL values in this context), β are the regression coefficients, and x denotes the molecular descriptors.

In this study, three linear regression methods, consisting of Ridge, LASSO and PLS, alongside two non-linear regression methods, KPLS and SVR, were investigated.

Ridge: Ridge regression reduces the coefficients of regression by applying a penalty on their squared sum (L2 norm). The formula combines the standard least squares criterion with an additional term where λ acts as

a shrinkage factor on the squared regression coefficients as seen in Equation 2.

$$\min_{\beta} \sum_{i=1}^n \left(y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p \beta_j^2 \quad (2)$$

The value of λ determines the severity of the penalty: higher values result in greater shrinkage of the coefficients towards zero. Unlike some other methods, ridge regression retains all predictor variables in the model because it does not reduce the coefficients to exactly zero, meaning no variables are excluded from the regression model.

LASSO: LASSO (Least Absolute Shrinkage and Selection Operator) is a regularization technique that applies a penalty to the sum of the absolute values of the regression coefficients (L1 norm).

$$\min_{\beta} \sum_{i=1}^n \left(y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p |\beta_j| \quad (3)$$

Mathematically, this involves adding a term penalizing the absolute values of the coefficients to the usual least squares criterion as seen in Equation 3. The shrinkage parameter (λ) influences the degree of penalty. A key feature of LASSO is that it can force some of the regression coefficients to become exactly zero, effectively excluding those variables from the model (feature selection). The maximum number of variables selected is limited by the number of compounds (n). In cases of highly correlated variables, LASSO tends to select one variable from the group and ignores the others.

PLS: Partial Least Squares (PLS) is utilized as an effective method for predicting the LFL of compounds using molecular descriptors. Unlike Principal Component Analysis (PCA), PLS is not just a dimension reduction technique; it specifically aims to maximize the covariance between latent variables and the LFL response variable. This makes PLS particularly suitable for our purpose, as it often requires fewer components for accurate prediction compared to methods like Principal Component Regression (PCR), which we did not examine in this context for this reason. Like PCR, PLS ensures that the latent variables are uncorrelated, which is crucial for the reliable prediction of LFL using molecular descriptors.

RBF-KPLS: Radial Basis Function – Kernel-based Partial Least Squares method combines the concepts of Kernel methods and PLS regression. It employs the Radial Basis Function (RBF) kernel to map input data into a high-dimensional feature space. Similar to traditional PLS, RBF-KPLS seeks to find the multidimensional direction in the feature space that explains the maximum multidimensional variance of the input data while also having the highest covariance with the target variable. The use of RBF kernel allows the

capture of complex, non-linear relationships between the predictors and the response variable.

RBF-SVR: Like RBF-KPLS, RBF-SVR is also adept at managing non-linear data relationships. Central to RBF-SVR are support vectors, the critical data points that define the hyperplane boundaries in the model. RBF-SVR's optimization objective balances margin maximization and error penalty. This balance is governed by three key parameters: the regularization parameter, C , which dictates the margin-error trade-off; the kernel coefficient, γ , controlling the influence of individual training examples; and the epsilon parameter, ϵ , which sets the width of the epsilon insensitive tube, thus determining the model's tolerance to prediction errors.

2.5. Model Development

As part of model development, model hyperparameters were determined using a 9-fold cross validation. The training set was randomly divided into 9 approximately equal folds, 8 of which were used to train the model (training set) and the remaining fold used to validate the trained model (validation set). This process was repeated 9 times until each fold was used as a validation set exactly once. A grid search method was used to identify the optimal hyperparameter(s) that yielded the lowest root mean squared error (RMSE) averaged across 9 folds.

Descriptors were also standardized and mean-centered prior to model training to ensure each descriptor had a mean of 0 and standard deviation of 1 within the training set. This is to ensure that all features contribute equally to the model, preventing descriptors with larger scales from dominating the model's behaviour.

2.6. Performance Metrics of ML Algorithms

For a QSPR model to be accepted for use in regulatory purposes (such as risk assessments), it must be validated to ensure it produces accurate and reliable estimates [15]. Following model development, the performance of the QSPR models were evaluated using the following metrics: the coefficient of determination (R^2), root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE), which are calculated using Equation 4-7:

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y}_i)^2} \quad (4)$$

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (5)$$

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (6)$$

$$\text{MAPE} = \frac{1}{N} \sum_{i=1}^N \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (7)$$

Where y_i , \hat{y}_i , \bar{y}_i represents the observed, predicted and mean observed LFLs respectively, whilst N represents

the number of compounds in the train, validation, or test set.

2.7. Internal Validation

The internal performance of the model can be characterized by its ability to accurately reproduce data in the training set, a factor often referred to as “goodness-of-fit”. This is quantitatively represented by the coefficient of determination on the training set (henceforth referred to as r^2). However, r^2 alone is not a sufficient measure of robustness nor predictivity, as it tends to increase as the number of descriptors increase. Hence, its leave-many-out cross-validated counterpart (Q_{LMO}^2) was also calculated using Equation 8 [16]:

$$Q_{LMO}^2 = 1 - \frac{\sum_{j=1}^M \sum_{i=1}^N (y_i - \hat{y}_{i/j})^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (8)$$

Where $\hat{y}_{i/j}$ is the predicted value when j^{th} part of the dataset is left out with the training set split into M folds. In contrast to r^2 , Q_{LMO}^2 improves only when useful descriptors are added to predict compounds that are left out. Thus, a high Q_{LMO}^2 is required for a robust model. For this study, a 9-fold cross validation procedure was used to determine Q_{LMO}^2 .

Additionally, the model’s stability (statistical significance) was tested using a y-scrambling and pseudo-descriptor test [9]. This involves comparing the performances of developed models against those generated by random chance correlations. In conducting these tests, the original feature selection and model building processes were strictly adhered to, thus preventing selection bias. Further details regarding these tests are available in [9]. Both the internal performance and stability of the model were considered in parallel to ensure that developed models had sufficient internal predictivity (defined as having $Q_{LMO}^2 > 0.7$) and were statistically significant [16].

2.8. External Validation

The external predictivity and generalizability of developed QSPR models were estimated by evaluating their performance on the test set, which contained compounds not involved in model development. For this study, external predictivity (Q_{Ext}^2) is calculated using Equation 9 [17]:

$$Q_{Ext}^2 = 1 - \frac{[\sum_{i=1}^{N_{test}} (\hat{y}_i - y_i)^2] / N_{test}}{[\sum_{i=1}^{N_{train}} (y_i - \bar{y}_{train})^2] / N_{train}} \quad (9)$$

Q_{Ext}^2 was derived as a better alternative for determining external predictivity as compared to R^2 and Q^2 as it is evaluated independently of test set composition, hence reducing the effect of a different test set size [17]. Additional performance metrics introduced in Section 2.6 were also evaluated on the test set, which were denoted as R^2 , $RMSE_{test}$, MAE_{test} , and $MAPE_{test}$ respectively.

2.9. Applicability Domain

Defining the Applicability Domain (AD) of a QSPR model is essential, as it delineates the chemical structure and response space where reliable predictions can be made. The AD is intrinsically connected to the composition of the model’s training set, necessitating that a new chemical be structurally similar to those in the training set for its prediction to be considered as an interpolated value with reduced uncertainty [18].

A practical method to visualize the AD is through a Williams Plot [15], which maps standardized residuals (y-axis) against leverage values (x-axis) of the dataset’s chemicals. This plot helps to identify chemicals that are significantly different from those in the training set, indicating potential extrapolation in predictions.

In this study, the model space is represented by the descriptor matrix (\mathbf{X}), comprising of N samples and k variables (descriptor values). Leverage values (h) were calculated for each compound within the dataset using the following equation:

$$h_i = \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i \quad (10)$$

Where h_i is the leverage value of chemical i within the descriptor space and \mathbf{x}_i is the descriptor row-vector of the chemical. Leverage value gauge the extent to which a compound’s descriptor values deviate from the other compounds within the dataset.

A key parameter in assessing leverage is the warning leverage (h^*), which is typically defined as:

$$h^* = \frac{3p}{n} \quad (11)$$

Where p is the number of model variables plus one and n is the number of training chemicals. A chemical’s leverage exceeding h^* suggests that its prediction stems from significant extrapolation, potentially leading to less reliable results. Therefore, the Williams Plot and leverage calculations provide a framework to evaluate the AD, ensuring that predictions made are within a reliable and relevant chemical space.

3. Results & Discussion

3.1. Results of Feature Selection

Table 2 summarizes the results of feature selection on the mixed and experimental subsets.

Table 2: Summary of number of training point (N_{train}), testing points (N_{test}), and retained molecular descriptors after feature selection for mixed and experimental subsets.

Subset	N_{train}	N_{test}	No. of Molecular Descriptors
Mixed	1354	152	153
Experimental	266	33	178

Notably, within the mixed subset, only 152 out of the 1,354 samples in the train set are experimental values. Furthermore, despite undergoing the same feature selection procedure, more features were retained for the experimental subset when compared to the mixed subset. The discrepancy could be attributed to the

experimental training set being smaller than the mixed training set, resulting in less information available per descriptor. Thus, the feature selection method used was unable to fully capture patterns within the subset.

3.2. Results of Internal Validation of Models

The internal validation of the developed models was meticulously conducted using their respective training sets, either from the mixed or experimental subset. The comparative results of r^2 and Q_{LMO}^2 for these models are illustrated in Figure 4 and detailed in Table S1.

Evaluation of the internal validation results of the model performances revealed that all models were deemed to be internally predictive ($Q_{LMO}^2 \geq 0.7$), regardless of the subset used to train them.

An important observation from this validation process was that higher Q_{LMO}^2 values alongside smaller gaps between r^2 and Q_{LMO}^2 were observed for models trained on the mixed subset. These results suggests that models developed using the mixed subset were more robust than their counterparts trained on the experimental subset as variations in the train set did not significantly impact goodness of fit.

A notable exception to this trend was the PLS algorithm, for which the model developed on the experimental subset exhibited a higher Q_{LMO}^2 compared to its mixed subset counterpart (0.9220 vs 0.8950). Furthermore, it also had the smallest gaps between r^2 and Q_{LMO}^2 across both subsets. These results suggest that the PLS algorithm produced more robust models compared to other algorithms.

To evaluate the statistical significance of developed models, the r^2 was compared against the average

r_{random}^2 derived by models developed from 50 iterations of y-scrambling and pseudo-descriptors tests. The results are tabulated in Table 3.

Results of y-scrambling suggest that all models, regardless of the subset used to train them, performed significantly better than chance, having r^2 values that exceeded r_{random}^2 by more than 2.3 standard deviations (SD), a threshold indicative of 1% level significance. Application of the pseudo-descriptor tests, which tend to yield higher r_{random}^2 values due to intercorrelation among real descriptors, still showed most models remained significant [9]. However, Table 3 shows that models trained on the mixed subset achieved r^2 values significantly higher than $r_{random}^2 \pm 2.3SD$ when compared to their counterparts trained on the experimental subset.

An exception was observed in the RBF-SVR trained on the experimental subset. The developed model presented an $r^2 \leq r_{random}^2 \pm 2.3SD$ for the more demanding pseudo-descriptor test, suggesting that the significance of the model was not beyond doubt, as pure chance alone created equivalent or better models to describe the given data despite being based on a new set of random number pseudo-descriptors. This could be due to the inherent flexibility of SVR, which forms a regression model defined by several support vectors. As the number of support vectors approaches the number of samples, this method can replicate the training set almost perfectly despite randomized data being used.

Overall, the internal validation results indicate an enhanced resilience against random chance correlations when models were trained using the mixed subset, as demonstrated by the comparatively lower r_{random}^2 values across y-scrambling and pseudo-descriptor test.

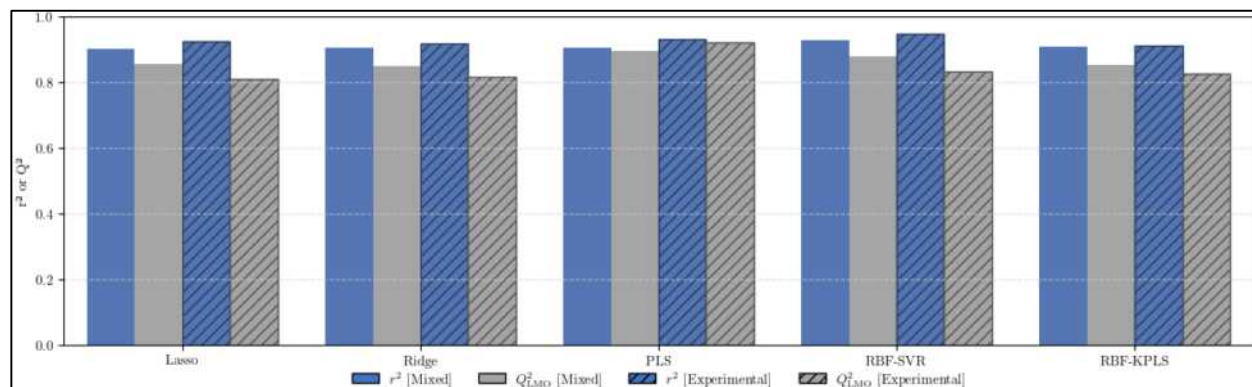


Figure 4: Bar chart of r^2 and Q_{LMO}^2 results across different machine learning algorithms for the mixed and experimental subset.

Table 3: Results of y-scrambling and pseudo-descriptors tests across different machine learning algorithms for the mixed and experimental subset.

Model	It.	Test	Mixed Subset			Experimental Subset		
			r_{random}^2	+2.3 SD	r^2	r_{random}^2	+2.3 SD	r^2
LASSO	50	Pseudo-descriptors	0.004 ± 0.006	0.0181	0.904	0.832 ± 0.024	0.8871	0.926
		Y-Scrambling	0.001 ± 0.002	0.0048		0.007 ± 0.024	0.0616	
RIDGE	50	Pseudo-descriptors	0.006 ± 0.009	0.0268	0.9060	0.775 ± 0.029	0.8423	0.919
		Y-Scrambling	0.0005 ± 0.001	0.0028		0.006 ± 0.018	0.0469	
PLS	50	Pseudo-descriptors	0.003 ± 0.005	0.0144	0.906	0.729 ± 0.027	0.7910	0.932
		Y-Scrambling	0.001 ± 0.001	0.0027		0.004 ± 0.009	0.0232	
RBF-SVR	50	Pseudo-descriptors	-0.001 ± 0.004	0.0075	0.9282	0.935 ± 0.086	1.1336	0.948
		Y-Scrambling	-0.004 ± 0.001	-0.0009		0.005 ± 0.023	0.0584	
RBF-KPLS	50	Pseudo-descriptors	0.006 ± 0.008	0.0230	0.896	0.697 ± 0.058	0.8315	0.917
		Y-Scrambling	0.002 ± 0.004	0.0109		0.011 ± 0.02	0.0573	

3.3. Results from External Validation

External validation of the developed models was carried out on their respective test sets (mixed or experimental subset). Figure 5 presents the performance metrics described in Section 2.6 for each model, and the results are also tabulated in Table S2.

From Figure 5, models trained on the experimental subset generally exhibited higher Q_{Ext}^2 values than those trained on the mixed subset, implying superior external predictivity which resulted in lower prediction error metrics (RMSE, MAE, and MAPE). In contrast, models trained on the mixed subset, displayed a reduction in Q_{Ext}^2 values, which ranges from 0.4 – 4.3% decrease when compared to their counterparts trained on the experimental subset. These findings suggest that the inclusion of predicted data adversely affected the external predictivity of the model.

A notable exception to this trend was the PLS algorithm, which demonstrated a lower Q_{Ext}^2 when the model was trained on an experimental subset as compared to its counterpart trained on the mixed subset. Additionally, models trained on the experimental subset exhibited more variation in Q_{Ext}^2 across different ML algorithms, unlike models trained on the mixed subset which exhibited relatively stable Q_{Ext}^2 .

Figure 6 presents the r^2 and R^2 values obtained for the developed models, which represent the goodness-of-fit metric for the train and test set respectively, with results tabulated in Table S3. When r^2 is significantly higher than R^2 , there is good evidence to suggest that the model is overfitted. Overfitted models tend to capture the noise

within the dataset rather than learning the underlying patterns of the dataset, leading to worse predictivity on unseen data.

It can be observed that models trained on a mixed subset exhibited more consistent gaps between r^2 and R^2 , which were smaller in comparison to models trained on the experimental subset. This suggest that models trained on the experimental subset experienced a higher degree of overfitting, which is unsurprising considering the smaller training sample size combined with the high number of features present in the experimental subset. Unexpectedly, despite experiencing higher degrees of overfitting, models trained on the experimental subset still exhibited higher external predictivity (higher Q_{Ext}^2).

A plausible explanation could be that predicted values introduced noise into the mixed subset, despite the same upper limit of threshold uncertainty (<25%) being used to filter out errors in LFL values. This led to lower Q_{Ext}^2 values being obtained for models trained on a mixed subset. Additionally, the consistent yet lower Q_{Ext}^2 values across mixed subset models could imply that the QSPR model is unable to produce predictions of higher quality than their training set.

Additionally, an important observation is that the PLS model trained on the experimental subset exhibited the largest gap between the two values, thus exhibiting significant overfitting. This likely accounts for the steep decrease in Q_{Ext}^2 observed. This implies that despite the apparent quality of the experimental subset (less noise), it was outweighed by the effect of overfitting.

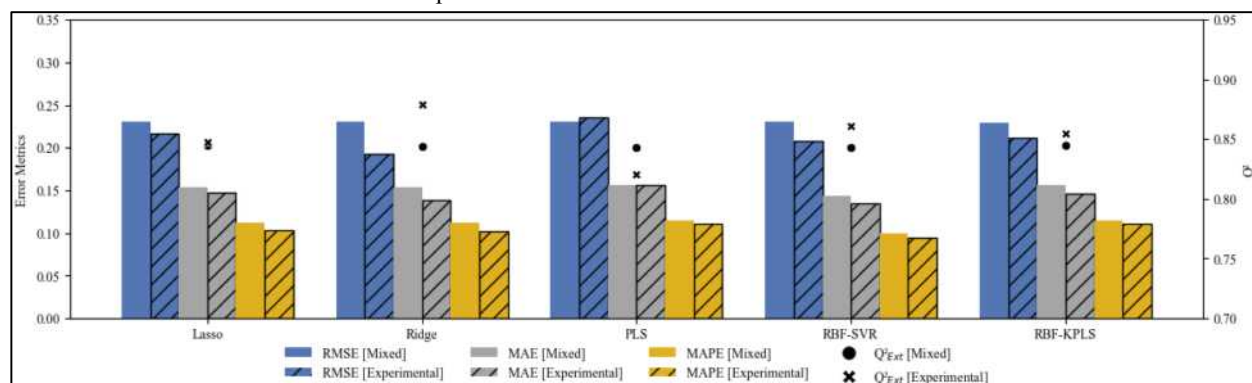


Figure 5: Bar chart of performance metrics across different machine learning algorithms for the mixed and experimental subset.

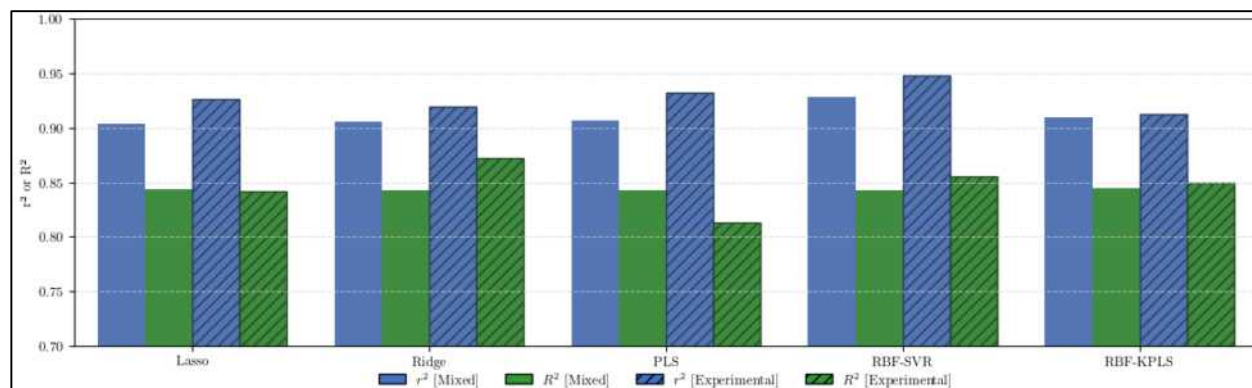


Figure 6: Bar chart of r^2 and R^2 results across different machine learning algorithms for the mixed and experimental subset.

3.4. Williams Plot for Applicable Models

Lastly, the applicability domain (AD) of the developed models is defined as according to the methodology outlined in Section 2.9. As the LASSO algorithm can perform its own feature selection, the number of model variables, p , was taken to be equal to the number of non-zero coefficients in the resulting regression equation. Conversely, PLS models utilize latent variables for regression, hence p was defined to be equal to the number of latent variables and the descriptor matrix was similarly transformed prior to calculation of leverages using the hat matrix. Due to the differing number of samples within the mixed and experimental subsets, analysis of the Williams Plot was done by expressing the number of points as a percentage of the total respective subset. These findings are detailed in Table 4. Additionally, Figure 7 showcases the Williams Plots for LASSO and PLS models using both experimental and mixed subsets respectively.

Table 4 indicates that models trained on the experimental subset consistently showed higher percentages of training points that were predicted correctly (standardized residual $< 3SD$) but lower percentages of testing points that were predicted correctly, regardless of which ML algorithm was used. This suggests a more pronounced overfitting in models trained on the experimental subset, which corroborates with our previous findings from Section 3.3.

Further analysis also reveals that models trained on the experimental subset exhibited higher percentages of influential training points ($h \geq h^*$) within the dataset (73.68% for LASSO, 5.26% for PLS) as compared to models trained on the mixed subset (22.67% for LASSO, 4.36% for PLS). These results suggest that the experimental subset was impacted by a higher degree of influential points as compared to the mixed subset. A possible explanation could be due to smaller training sample sizes, influential points were more prominent in the experimental subset as compared to the mixed

subset, where the effect of individual influential points were averaged out by the remaining samples.

Similarly, a higher percentage of high leverage training points predicted correctly was observed for the models trained on the experimental subset. This could suggest that these models were prone to overfitting to influential data points which compromised their robustness. This is supported by our previous findings in Section 3.2, where models trained using the experimental subset demonstrated lower Q_{LMO}^2 overall.

Interestingly, the LASSO model trained on an experimental subset resulted in a higher percentage of high leverage test points predicted correctly, likely contributing to the higher Q_{Ext} observed in Section 3.3, despite having lower percentages of testing points that were predicted correctly (which are inclusive of high leverage test points). A possible explanation could be due to the model overfitting to influential training points, consequently improving its predictivity for testing points far from the structural domain of the training set used. This finding cannot be extrapolated to the PLS models, as no compounds within the test set were identified as high leverage points for the model trained on the experimental subset.

Due to limitations of the methodology used to define the AD, ADs for kernel-based ML algorithms (RBF-SVR and KPLS) were unable to be defined [19]. Additionally, the methodology applied resulted in a warning leverage greater than 1 (h^* of 2.01) for the Ridge model trained on an experimental subset. As leverage values are bounded by an upper limit of 1, this would suggest that the model be applicable to all known and unknown compounds, which is in direct conflict for the purposes of identifying an AD to begin with. Hence, it is recommended that future studies implement an alternative method to define their ADs to gain better understanding of the impact of predicted data on a model's applicability.

Table 4: Analysis of Williams Plots for LASSO and PLS models trained on a mixed and experimental subset.

Model	LASSO		PLS	
Dataset	Mixed	Experimental	Mixed	Experimental
Correctly Predicted Train Points (%)	98.01	98.87	98.23	99.25
Correctly Predicted Test Points (%)	96.71	93.94	96.05	90.91
High Leverage Train Points (%)	22.67	73.68	4.36	5.26
High Leverage Test Points (%)	11.18	69.70	1.32	0.00
High Leverage Train Predicted Correctly (%)	19.32	64.55	3.85	4.68
High Leverage Test Predicted Correctly (%)	0.93	7.02	0.13	0.00
Incorrectly Predicted Train Points (%)	0.81	0.00	1.70	0.75
Incorrectly Predicted Test Points (%)	1.32	0.00	3.95	9.09

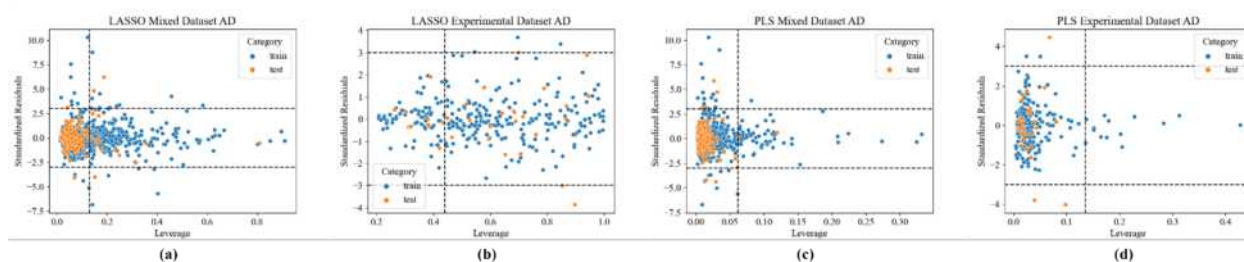


Figure 7: Williams Plots for: (a) LASSO trained on mixed subset, (b) LASSO trained on experimental subset, (c) PLS trained on mixed subset and (d) PLS trained on experimental subset.

4. Conclusion

This study investigated the impact of incorporating predicted values in the training set for developing ML-QSPR models, specifically for predicting LFLs of pure compounds. Two LFL subsets were curated from the DIPPR 801 Project: one consisting of both predicted and experimental values (mixed subset) and the other consisting of only experimental values (experimental subset). An external test set was split off for each subset, which comprised off experimental values that did not participate in model development. This was done to assess whether models trained on the mixed subset went beyond replicating the effectiveness of past prediction methods used to obtain the predicted values to begin with.

4.1. Summary of findings

Our findings suggest that the inclusion of predicted data within the training set demonstrated several advantages, including improved feature selection performance, enhanced robustness, better protection against chance correlation and increased robustness against overfitting. However, these benefits were accompanied by a noticeable trade-off in the form of reduced predictivity on unseen data.

In light of these observations, the implications of this study could be significant for the field. Despite a large majority (88.77%) of the training set being composed of predicted values, the impact on the predictivity of the models was relatively modest, showing a decrease of only 0.4 to 4.3%. This suggest that the inclusion of a well-curated predicted data, which is critically evaluated at the source to ensure consistency and reliability, can be beneficial for models by expanding the training set size, albeit at the cost of slightly lower predictivity. While this approach may not be advisable for models intended for regulatory or safety purposes, where high predictivity is essential, it could prove useful in applications such as preliminary screening process where maximum predictivity is not the primary objective.

4.2. Limitations

The main weakness of our study was due to the feature selection procedure proposed. Our study's feature selection procedure, while comprehensive, was not entirely effective in identifying the optimal set of features for building the most efficient models. This limitation led to a degree of overfitting in all developed models. Overfitting is a critical concern as it can severely impact the model's predictivity, potentially being a major contributor to the observed decrease in predictivity when predicted values were included in the training set. The influence of overfitting, therefore, cannot be overlooked as it likely played a significant role in shaping the study's findings, particularly in terms of the modest reduction in predictivity.

Furthermore, the proposed methodology exhibited a notable over-reliance on the published uncertainty levels for predicted Lower Flammability Limits (LFLs) in the

DIPPR database. This over-reliance stems from the fact that uncertainty levels are assigned based on the data informing the regression models rather than on the inherent uncertainties associated with the regression methods themselves. This distinction is crucial as it directly impacts the reliability of the predicted data.

Additionally, the prediction methods employed to obtain these LFL values vary significantly, ranging from group/atomic contribution models to empirical correlations. A major limitation arises from the lack of clarity regarding the proportion of predicted data derived from each of these methods, introducing a potential bias in the training set composition and, by extension, the model's performance.

Overall, this study sheds light on the nuanced role of predicted data in the development of ML-QSPR models, especially in sectors like the petrochemical industry where obtaining experimental data can be challenging. The findings advocate for a balanced approach in training set composition, emphasizing the need for future research to refine methods for quantifying and accommodating the uncertainties inherent in predicted data. Such advancements could lead to more effective and efficient predictive models across various chemical and process engineering applications.

4.3. Future work

In light of the findings and limitations of this study, a key area for future research would be the exploration of using identical test sets or introducing an additional validation set consisting of identical components. This approach would provide a more rigorous assessment of the model's performance and its generalizability. By employing identical test sets across different training subsets, the direct comparability of model performances can be enhanced, allowing for a clearer understanding of the impact of training set composition on model predictivity. This method would serve as a robust check against overfitting, ensuring that the model's predictive capabilities are not solely tailored to the specific characteristics of the training set. Such a strategy would not only bolster the reliability of ML-QSPR models but also provide a more comprehensive framework for evaluating the effectiveness of including predicted data in training sets. The outcomes of this future work could significantly contribute to the optimization of predictive modelling in fields where data availability and quality are critical considerations.

References

- [1] R. Rowley J, E. Bruce-Black J. PROPER APPLICATION OF FLAMMABILITY LIMIT DATA IN CONSEQUENCE STUDIES. *SYMPOSIUM SERIES*. 2012; (NO. 158): <https://www.icheme.org/media/9189/paper58-hazards-23.pdf>.
- [2] R. Rowley J. *Flammability Limits, Flash Points, and Their Consanguinity: Critical Analysis, Experimental Exploration, and Prediction*. Doctor of Philosophy. Brigham Young University; 2010.
- [3] Vidal M, Rogers WJ, Holste JC, Mannan MS. A review of estimation methods for flash points and flammability limits. *Process safety progress*. 2004; 23 (1): 47-55. 10.1002/prs.10004.
- [4] Gharagheizi F. Quantitative Structure–Property Relationship for Prediction of the Lower Flammability Limit of Pure Compounds. *Energy & fuels*. 2008; 22 (5): 3037-3039. 10.1021/ef800375b.
- [5] Pan Y, Jiang J, Ding X, Wang R, Jiang J. Prediction of flammability characteristics of pure hydrocarbons from molecular structures. *AIChE journal*. 2010; 56 (3): 690-701. 10.1002/aic.12007.
- [6] Bloxham JC, Redd ME, Giles NF, Knotts TA, Wilding WV. Proper Use of the DIPPR 801 Database for Creation of Models, Methods, and Processes. *Journal of chemical and engineering data*. 2021; 66 (1): 3-10. 10.1021/acs.jced.0c00641.
- [7] Chen C, Lai C, Guo Y. A novel model for predicting lower flammability limits using Quantitative Structure Activity Relationship approach. *Journal of loss prevention in the process industries*. 2017; 49 240-247. 10.1016/j.jlp.2017.07.007.
- [8] Xu P, Ji X, Li M, Lu W. Small data machine learning in materials science. *npj computational materials*. 2023; 9 (1): 42-15. 10.1038/s41524-023-01000-z.
- [9] Rücker C, Rücker G, Meringer M. y-Randomization and Its Variants in QSPR/QSAR. *Journal of chemical information and modeling*. 2007; 47 (6): 2345-2357. 10.1021/ci700157b.
- [10] Zhao L, Wang W, Sedykh A, Zhu H. Experimental Errors in QSAR Modeling Sets: What We Can Do and What We Cannot Do. *ACS Omega*. 2017; 2 (6): 2805-2812. 10.1021/acsomega.7b00274.
- [11] Wenlock MC, Carlsson LA, Ognichenko L, Hromov A, Kosinskaya A, Stelmakh S, et al. How Experimental Errors Influence Drug Metabolism and Pharmacokinetic QSAR/QSPR Models. *Journal of chemical information and modeling*. 2021; 55 (1): 125-134. 10.1021/ci500535s.
- [12] Moriwaki H, Tian Y, Kawashita N, Takagi T. Mordred: a molecular descriptor calculator. *Journal of Cheminformatics*. 2018; 10 (1): 4. 10.1186/s13321-018-0258-y.
- [13] Tropsha A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular informatics*. 2010; 29 (6-7): 476-488. 10.1002/minf.201000061.
- [14] Roubehie Fissa M, Lahiouel Y, Khaouane L, Hanini S. QSPR estimation models of normal boiling point and relative liquid density of pure hydrocarbons using MLR and MLP-ANN methods. *Journal of molecular graphics & modelling*. 2019; 87 109-120. 10.1016/j.jmgm.2018.11.013.
- [15] OECD. *GUIDANCE DOCUMENT ON THE VALIDATION OF (QUANTITATIVE)STRUCTURE-ACTIVITY RELATIONSHIPS [(Q)SAR] MODELS*. Series on Testing and Assessment Organisation for Economic Co-operation and Development; 2007.
- [16] Gramatica P. Principles of QSAR models validation: internal and external. *QSAR & combinatorial science*. 2007; 26 (5): 694-701. 10.1002/qsar.200610151.
- [17] Consonni V, Ballabio D, Todeschini R. Comments on the Definition of the Q 2 Parameter for QSAR Validation. *Journal of Chemical Information and Modeling*. 2009; 49 (7): 1669-1678. 10.1021/ci900115y.
- [18] NETZEVA TI, WORTH AP, MYATT G, NIKOLOVA-JELIAZKOVA N, PATLEWICZ GY, PERKINS R, et al. Current status of methods for defining the applicability domain of (Quantitative) structure-activity relationships : The report and recommendations of ECVAM workshop 52. *Alternatives to laboratory animals*. 2005; 33 (2): 155-173. 10.1177/026119290503300209.
- [19] Fechner N, Jahn A, Hinselmann G, Zell A. Estimation of the applicability domain of kernel-based machine learning models for virtual screening. *Journal of Cheminformatics*. 2010; 2 (1): 2. 10.1186/1758-2946-2-2.

Techno-economic Assessment of a Novel Hybrid PV-T and Heat Pump System for Household Heating

Adnan Hakim and Kabishan Sivarasan

Department of Chemical Engineering, Imperial College London, U.K.

Abstract

A techno-economic analysis is undertaken to assess a novel hybrid system that integrates photovoltaic-thermal (PV-T) collectors and a vapour compression heat pump for household heating applications. The proposed system uses a configuration such that the outlet PV-T water is the heat sink of an air source heat pump. This aims to replace the use of natural gas boilers for residential heating, which is a significant contributor to global greenhouse gas emissions. The paper investigates the payback time of this system across four locations: London (UK), Rome (Italy), Tokyo (Japan), and Los Angeles (USA). An in-house MATLAB code was used to conduct simulation studies, incorporating hourly wind speed, irradiance and ambient temperature profiles of each location, whilst varying the number and arrangement of collectors. Results indicate a payback time of 8.6 years in London, 3.4 years in Italy, 6.4 years in Tokyo and 7 years in Los Angeles, which is achieved through the annual cost savings of using this system as opposed to a gas boiler to fulfil the heating demands of an average household. Shorter payback times compared to existing literature is attributed to lower capital costs of the proposed system as well as higher annual cost savings as result of the recent rise in natural gas prices.

Introduction

The current global energy supply is marked by an increasing demand for power, dwindling fossil fuel resources, and increasing environmental issues associated with its use. Reliance on fossil fuels, which accounted for 82% of the global primary energy use in 2022 [1], poses a threat to long-term energy security and contributed to 75% of total greenhouse gas emissions in the same year [2].

Projections show that failure to curb greenhouse gas emissions will result in a global temperature increase of 4.3°C by 2100 [3]. Disruption to ecosystems, increased frequency and intensity of extreme weather events, and food shortages are a few of the numerous consequences as a result of this increase. Growing concerns regarding the impact of fossil fuels has prompted an international imperative to transition towards renewable energy technologies, which offer a sustainable means to meet energy requirements.

In this context, solar energy is inherently the most abundant and inexhaustible source of renewable energy to date. The Earth intercepts $1.8 \times 10^{11} \text{ MW}$ from the sun, which is many orders of magnitude larger than the present rate of global energy consumption [4]. Historically, the primary obstacle to widespread adoption of solar power has been the large initial capital cost. However, the rise of gas prices accelerated by recent crises, such as the pandemic and Ukraine war, has provided developed nations with incentives to adopt alternatives [5].

Photovoltaic-Thermal (PV-T) collectors present a promising avenue for solar energy utilisation, by integrating solar Photovoltaic (PV) technology for electricity production with solar thermal technology for heat production. This can be combined with a vapour compression heat pump to fulfil household heating demand, whilst offsetting the required electricity consumption.

This report aims to provide a holistic techno-economic assessment of the proposed system across four locations: London (UK), Rome (Italy), Tokyo (Japan), and Los Angeles (USA). An in-house MATLAB code will be used to model its operation, using relevant

weather data from each location. A payback time (PBT) analysis will be conducted based on the annual cost savings of using this system, as opposed to a natural gas boiler, to fulfil the heating demand of an average household.

Background

In recent years, PV collectors have emerged as a sustainable and promising technology for harnessing solar energy. It has witnessed widespread adoption in various sectors, including building integrated systems, desalination plants and solar home systems [6]. The global cumulative installed capacity of PV systems has increased from 100.9 GW at the end of 2012 to 400 GW by the end of 2017 [7]. This figure has since increased to over 1000 GW [8], indicating rapid growth in the use of PV systems. This growth can be attributed to the increasing efficiency of solar cells as a result of extensive research, government support in the form of financial incentives, and the increasing greenhouse gas emissions associated with the use of fossil fuels [9]. Literature suggest that this growth is greater in more economically developed countries, with the rationale that these governments are able to invest in more financial incentives. A study published in 2022 supports this, concluding that there is a statistically significant correlation between the production of PV energy per person in EU countries and the GDP per capita [10].

Furthermore, developments in Spain's regulatory framework have been analysed to result in PV facilities becoming a profitable investment for domestic consumers, even without the use of any government subsidies [11]. This is largely the result of consumers being given incentives to sell their excess produced electricity back to the electrical grid. This income stream allows for a greatly reduced payback time for the PV panels, making the investment much more attractive to private households.

Although PV collectors have been widely used for solar energy utilisation, there are some challenges that limit its potential. Under higher ambient temperatures and stronger solar radiation intensity, PV collectors experience a loss in electrical efficiency due to a rise in

temperature of the PV module, causing a high electrical resistance [12]. PV-T collectors aim to address some of these issues by removing heat from the PV module, which can subsequently be utilised, depending on the type of system it is coupled up with [13].

PV modules and PV-T collectors have been compared in various locations, some even outside of the developed world. A study in Ghana conducted a comparative performance valuation of water-based PV-T against conventional PV modules, both of which were made up of mono-crystalline Silicon PV technology [14]. The annual combined electrical and thermal energy output of the PV-T system was 1237.71 kWh/m², which was over 6 times that of the PV system at 194.79 kWh/m². However, when considering the electrical output alone, the PV system had an electrical energy yield which was 30% higher than the PV-T. Furthermore, the monthly average electrical efficiency values were 1.2-1.6% higher for the PV system. This is a result of Ghana's wet season between April and October, characterised by lower solar irradiance levels and lower ambient temperatures, which both favour PV electrical performance. The study, however, did not carry out an economic analysis and emphasised the importance of investigation under different climatic conditions.

According to the European Environment Agency, 35% of energy-related EU emissions in 2021 was from the buildings sector [15]. Residential buildings make up a large percentage of the buildings sector, with UK households accounting for 26% of total building emissions in 2020 [16]. As global efforts intensify to combat increasing greenhouse gas emissions, the exploration of sustainable alternatives to the traditional natural gas boiler system for space heating is imperative.

The use of PV panels in combination with an electric heating system has been explored as an alternative to traditional space and water heating. This hybrid system reduces the heat pump's reliance on grid supplied power, subsequently lowering the carbon footprint. A study investigated the cost-effectiveness of four different low carbon technologies: air source heat pumps, ground source heat pumps, PV panels combined with an electric heating system and biomass boilers. It found that whilst a gas-based heating system remained cost-effective in all scenarios, the most viable mitigation technology was PV – resulting in a reduction of 0.015 Mt CO₂ at a cost of abatement of £160/tonne CO₂ [17]. The success of combining PV panels with the heating system is what set the general direction for this paper, albeit using PV-T collectors instead.

The distinguishing characteristic of PV-T collectors compared to PV, is their ability to extract thermal energy generated on the panel, as well as electrical energy. In water-type PV-T systems, this is done using water to flow over the PV module, extracting heat from it. It is reported that PV-T systems also have consistently higher electrical efficiency compared to PV systems in equivalent weather conditions. [18]. The thermal energy extracted is of particular interest, as systems have been designed to utilise this energy for domestic heating. Using the case study of a university building, Herrando *et al* found that using a cooling system combined with

PV-T collectors would be able to cover 16.3% more of the electricity demand than when using conventional PV panels [19]. The system used an AbCH unit to provide heating and cooling from the thermal energy, which brought with it a considerably high capital cost.

An alternative to this system would be to couple the PV-T collectors with a heat pump. These solar assisted heat pumps (SAHPs) have been found to offer higher performance values for heat and power provision compared to stand-alone heat pump systems [20], with a coefficient of performance (COP) of up to 5.5 for heating. Such systems have had techno-economic analyses conducted on them before. Obalanlege *et al.* found that, for a household in Belfast, the most economically viable system configuration was 12 PV-T modules with a total area of 16.3 m². The system was calculated to produce 2.4 MWh of electricity and 2.0 MWh of hot water per year. It was found to be able to cover half of the electrical demand of the household and a third of the heating demand, at a cost of £11,550 with a payback time of 14 years [21]. A similar SAHP was installed in a test facility for a different study and field-tested for a period of eight months [22]. This seven-panel system was found to have an average coefficient of performance close to 4, depending on the mode the system was running in. The methodology involved the installation of a test facility consisted of a solar-assisted heat pump system, which was monitored using a dedicated data acquisition and control system.

The significance of this paper lies within its objective to fill a gap that is present within the existing literature surrounding solar assisted heat pumps. Typically, the outlet of a PV-T collector is connected to the evaporator of the heat pump system as a cold source [22]; it has almost exclusively been researched in European locations, as shown through this literature review. A novel system is proposed in which the outlet PV-T water is used as a heat sink at the condenser, absorbing the heat rejected by the working fluid. This paper aims to conduct a payback time analysis of this novel configuration, opening the door to further research in this area if proven viable. Locations outside of Europe in the developed world will also be explored, particularly in nations where natural gas is the main source of residential heating.

Methods

A hybrid system was proposed, with the aim of replacing the conventional natural gas boiler system used by the majority of households for space and water heating [23]. The proposed model connects a given number of PV-T collectors with an air source heat pump in one continuous system, as shown in Figure 1. The outlet water from the PV-T collectors, is fed directly into the condenser, to be heated further to a desired temperature (T_{desired}) of 55°C [24] using a vapour compression cycle.

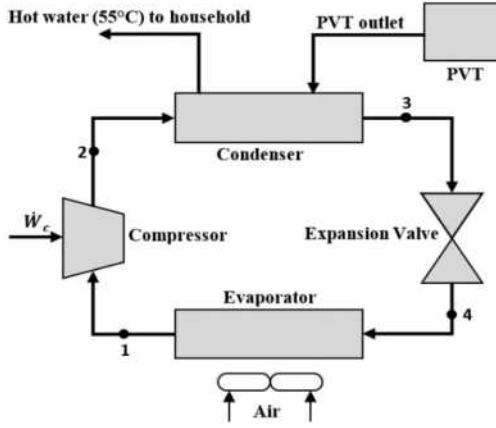


Figure 1. Schematic of the system combining the PV-T collectors with a vapour compression heat pump.

The mass flowrate and outlet temperature of the PV-T collectors is taken as the heat sink (water) inlet conditions at the condenser. The components in the cycle include a single-stage rotary type compressor, condenser, expansion valve and an evaporator. The working fluid is R32, and the cold source is air at 7°C.

The overall size of the heating system, Q_{sys} , was selected to be 10kW, based on the recommended heat pump size for a three-bedroom household [25]. Based on this, the mass flowrate of the water entering the system, $\dot{m}_{fl,sys}$, is calculated:

$$\dot{m}_{fl,sys} = \frac{Q_{sys}}{(T_{desired} - T_{tap})} \quad (1)$$

where T_{tap} is the tap water temperature, set at 7°C [26]. This flowrate would be used to calculate the flowrate of water entering a singular PV-T collector, $\dot{m}_{fl,in}$, based on the configuration of collectors used:

$$\dot{m}_{fl,in} = \frac{\dot{m}_{fl,sys}}{N_{rows}} \quad (2)$$

where N_{rows} is the number of rows in the collector configuration. Parallel configurations would have a value of N_{rows} equal to the number of collectors, N_c .

Verification

An in-house MATLAB code was utilised to model the operation of PV-T collectors. The code was verified against the work of Han et al. [27], which studied the effect of 13 different spectral splitting fluids on the electrical efficiency of a collector. Figure 2 shows a close match between the electrical efficiencies determined experimentally and the those obtained via simulation of the present model, demonstrated by a root mean squared deviation of 0.79%.

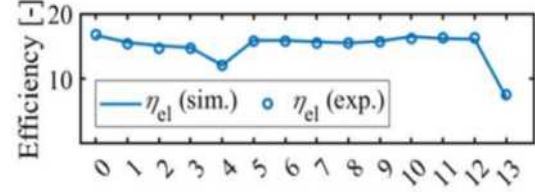


Figure 2. Electrical efficiency of the PV-T collector using 13 different spectral splitting fluids, determined experimentally by Han et al. plotted against the values obtained using the present model.

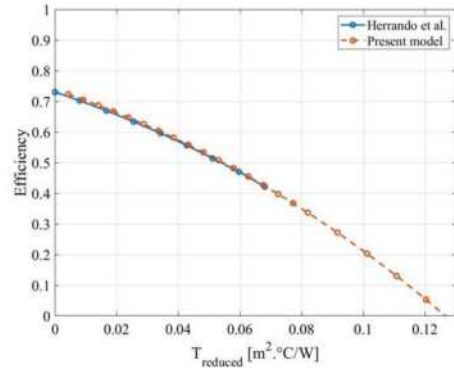


Figure 3. Comparison between the characteristic thermal efficiency curve of a conventional PV-T collector with a box-type heat exchanger, as proposed by Herrando et al., and the curve predicted by the present model.

In order to verify that the developed code accurately predicts performance, it was used to make a comparison to the performance of a PV-T collector proposed by Herrando et al. [28], for which results are available in the literature. This was done by modifying the code to use the same geometric configuration as the collector in literature. A characteristic thermal efficiency curve was plotted using the present model and compared to literature as presented by Figure 3. A root mean squared error value of 0.52% indicated minimal deviation, thus verifying the present model for use.

PV-T Collectors

The structure of the PV-T collector used, and the dimensions are shown in Figure 4 and Table 1 respectively.

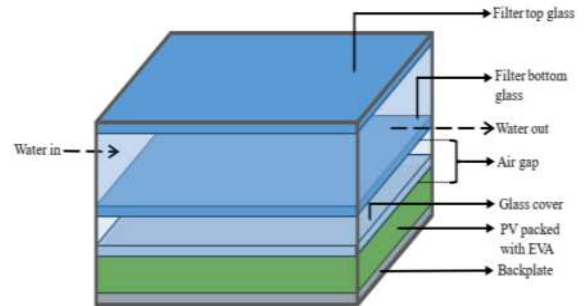


Figure 4. Schematic of PV-T collector; a cover glass above the filter is not used. The filter selective liquid is water. The top glass of the filter is anti-reflective thin glass, and bottom glass of the filter is short glass.

Table 1. Dimensions of the PV-T collector used.

Property	Value
Aperture Length (m)	1.2
Aperture Width (m)	0.6
Filter channel size (m)	0.01
Thickness of air gap between filter and PV-T (m)	0.01
Thickness of glass on PV-T (m)	0.004
PV Absorptivity	0.93
$c_{p,fl}$ of fluid in filter (J/kg/K)	4180

The energy balance of each layer was coded into the MATLAB script, such that they could be simultaneously solved to display each intermediary temperature. The balance for the top glass of the filter is given by:

$$A_c \cdot h_{c,filter} \cdot (0.5 \cdot T_{fl,in} + 0.5 \cdot T_{fl,out} - T_{flg1}) + A_c \cdot G_{2,abs} + A_c \cdot h_{wind} \cdot (T_{flg1} - T_a) + A_c \cdot h_{r,flg2sky} \cdot (T_{sky} - T_{flg1}) = 0 \quad (3)$$

where $T_{fl,in}$ and $T_{fl,out}$ are the filter coolant inlet outlet temperature respectively; $h_{c,filter}$ and h_{wind} are the convective heat transfer coefficient in the filter channel and between the cover glass and ambient air respectively; T_{flg1} corresponds to the temperature of the top glass of the filter. $G_{2,abs}$ is the solar radiation absorbed by the top glass of the filter, $h_{r,flg2sky}$ is radiation from filter top glass to sky, and A_c is the total area covered by the collector.

The sky temperature is given as:

$$T_{sky} = 0.0552 \cdot T_a^{1.5} \quad (4)$$

where T_{sky} is the sky temperature and T_a is the ambient temperature [27].

The energy balance for the selective liquid layer is given as:

$$A_c \cdot h_{c,filter} \cdot (T_{flg1} - 0.5 \cdot T_{fl,in} - 0.5 \cdot T_{fl,out}) + A_c \cdot h_{c,filter} \cdot (T_{flg2} - 0.5 \cdot T_{fl,in} - 0.5 \cdot T_{fl,out}) + A_c \cdot G_{3,abs} - \dot{m}_{fl} \cdot c_{p,fl} \cdot (T_{fl,out} - T_{fl,in}) = 0 \quad (5)$$

where \dot{m}_{fl} is the mass flowrate of the selective liquid and $c_{p,fl}$ is the specific heat capacity of it; $G_{3,abs}$ is the solar radiation absorbed by the liquid.

Extending to the bottom glass of the filter, the energy balance introduces convective and radiative heat transfer coefficients, $h_{c,fl2pv}$ and $h_{r,fl2pv}$, between the filter and the PV glass:

$$A_c \cdot h_{c,filter} \cdot (0.5 \cdot T_{fl,in} + 0.5 \cdot T_{fl,out} - T_{flg2}) + A_c \cdot (h_{r,fl2pv} + h_{c,fl2pv}) \cdot (T_{gpv} - T_{flg2}) + A_c \cdot G_{4abs} = 0 \quad (6)$$

where T_{flg2} is the filter bottom glass temperature, T_{gpv} is PV glass temperature, G_{4abs} is the solar radiation absorbed by the filter bottom glass.

The energy balance for the PV glass cover is given by:

$$A_c \cdot (h_{r,fl2pv} + h_{c,fl2pv}) \cdot (T_{flg2} - T_{gpv}) + A_c \cdot h_{c,pvg2pv} \cdot (T_{pv} - T_{gpv}) + A_c \cdot G_{5abs} = 0 \quad (7)$$

where T_{pv} is the PV temperature, G_{5abs} is the solar radiation absorbed by the PV glass, and $h_{c,pvg2pv}$ is the radiation from the PV glass to PV. Radiation from the PV glass to the sky is considered negligible due to the filter being closely covered on the PV.

The energy balance of the PV layer introduces convective heat transfer coefficients between the PV glass and PV layer ($h_{c,pvg2pv}$) and between the PV layer and ambient air ($h_{c,pv2amb}$), as well as the solar radiation absorbed by the PV layer, G_{pvabs} :

$$A_c \cdot h_{c,pvg2pv} \cdot (T_{gpv} - T_{pv}) + A_c \cdot h_{c,pv2amb} \cdot (T_a - T_{pv}) + A_c \cdot G_{pvabs} = 0 \quad (8)$$

The electrical efficiency of the PV-T collector, $\eta_{e,pvt}$, is dependent on cell temperature, T_{pv} , and is given by:

$$\eta_{e,pvt} = \eta_{e,25} \cdot [1 - T_{coeff} \cdot (T_{pv} - 298)] \quad (9)$$

where $\eta_{e,25}$ is the module efficiency at 298K and T_{coeff} is the temperature coefficient, both of which can be found in the manufacturer specification of the PV-T collector.

In order to generate values of $\eta_{e,pvt}$ and $T_{fl,out}$, which could then be used to calculate energy output of the collectors, three parameters dependent on location were required: solar irradiance (G), wind speed (V_w) and T_a . For each location, these values were sourced from the European Commission's Photovoltaic Geographical Information System (PVGIS) [29], which contained the hourly logged data for these variables, extracted over the entire year of 2020. 2020 was chosen as it was the most recent year that complete data was accessible. Despite the end goal being a single value for annual energy output of the array of collectors, the variation of sunlight throughout the day meant that the code would initially have to be run at each hour to calculate the different thermal and electrical efficiencies through the day. Thus, an hourly profile of an average day's conditions would be necessary.

The profile consisted of the 12 hours where the sun is out, disregarding night hours. The conditions for each hour was calculated by averaging the value of every required variable at that given hour, in the year 2020. Due to variations in weather conditions between months, it was considered that carrying out this averaging calculation at the scale of each month at first would be more accurate, such that a final average could be obtained for the year by averaging the monthly profiles.

To determine to what extent these two approaches of averaging would yield different values, they were both carried out for the location of London. In order to ensure accuracy and efficiency, both methods of data handling was done using a Python script that was developed over the course of the project. The script was fed with large dataset from PVGIS and calculated average ambient conditions either on a monthly basis, which were then

combined to get an annual profile, or directly over the entire year, as shown in Table 2.

Table 2. Comparison between two methods of calculating daily irradiance profile in London

Hour	Direct Yearly Average G (W/m ²)	Monthly Annual Average G (W/m ²)
6	66.3	66.1
7	134.8	134.4
8	219.6	219.2
9	302.2	301.8
10	354.8	354.7
11	392.3	392.1
12	385.6	385.4
13	370.6	370.3
14	321.2	320.9
15	257.6	257.4
16	178.4	178.0
17	107.3	107.0

The root mean square deviation (RMSD) value calculated for solar irradiance, ambient temperature and wind speed were 0.19%, 0.08% and 0.18% respectively. Due to insignificant difference between the two methods, it was concluded that a direct yearly average profile is an accurate representation of the variations in weather conditions across the year.

Simulations

The average hourly irradiance, wind speed, and ambient temperature profiles were fed as input into the code and ran to obtain a PV-T outlet water temperature profile and an electrical efficiency profile across the day. This was repeated for the same conditions at different number of collectors based on a roof area of 75 m² [30]. Parallel and series combinations were also investigated for each number of collectors. This simulation was carried out for each location, with the respective daily profiles.

The electrical power output of the PV-T collectors at each hour, $\dot{W}_{pvt,hr}$, is calculated using:

$$\dot{W}_{pvt,hr} = \eta_{e,pvt} \cdot G \cdot A_c \quad (10)$$

The hourly thermal power output of the PV-T collectors, $\dot{Q}_{pvt,hr}$, is calculated using:

$$\dot{Q}_{pvt,hr} = \dot{m}_f \cdot c_{p,f} \cdot (T_{f,out} - T_{f,in}) \quad (11)$$

Heat Pump

The model assumes no heat losses in the condenser, thus a perfect temperature exchange between the working fluid and PV-T water. In addition, steady state operation of the heat pump components and isenthalpic expansion through the valve is assumed. Figure 4 shows a temperature-specific entropy obtained from running the code.

The condenser is comprised of the desuperheating zone (Process 2-2a), condensing zone (Process 2a-2b), and subcooling zone (Process 2b-3). The evaporator is comprised of the evaporating zone (Process 4-4a) and superheating zone (Process 4a-1). Splitting the heat exchangers into these zones was necessary because the

heat transfer coefficient are significantly affected by the type of flow.

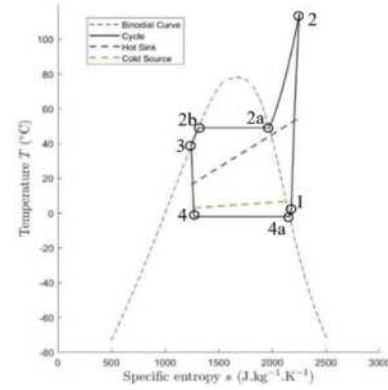


Figure 4. Temperature-specific entropy diagram for the heat pump cycle.

The working fluid enters the compressor at temperature, T_1 , and specific enthalpy, h_1 . It then undergoes compression, represented by Process 1-2, resulting in a temperature increase to T_2 and a compressor outlet specific enthalpy of h_2 , which is given by equation 9:

$$h_2 = \frac{h_{2,ideal} - h_1}{\eta_{is,eff}} \quad (12)$$

where $h_{2,ideal}$ is the specific enthalpy at the outlet for a perfectly isentropic compression and $\eta_{is,eff}$ is the isentropic efficiency.

After undergoing compression, the working fluid enters the condenser as a vapour at a higher temperature than the heat sink. As a result of the temperature difference between the working fluid and heat sink in the condenser, heat is transferred to the heat sink, via a heat exchanger in the condenser (process 2-3). The working fluid is condensed and subcooled (process 2b-3) to temperature, T_3 , by a degree of subcooling, T_{sc} :

$$T_{sc} = T_{2b} - T_3 \quad (13)$$

Subcooling is necessary to ensure only a single-phase liquid enters the expansion valve. This prevents the unwanted phenomena, such as flash gas, in the valve due to the presence of vapour [31].

The heat transfer rate to the heat sink, \dot{Q}_{out} , is given by:

$$\dot{Q}_{out} = \dot{m}_{hs} \cdot c_{p,hs} \cdot (T_{hs,out} - T_{hs,in}) \quad (14)$$

where \dot{m}_{hs} is the mass flowrate of the heat sink, $c_{p,hs}$ is the specific heat capacity of the heat sink at the outlet conditions and $T_{hs,out} - T_{hs,in}$ is the temperature increase of the heat sink. This is equal to the enthalpy reduction of the working fluid, thus the mass flowrate of the working fluid, \dot{m}_{wf} , can be calculated using:

$$\dot{m}_{wf} = \frac{\dot{Q}_{out}}{h_2 - h_3} \quad (15)$$

where $h_2 - h_3$ is the enthalpy reduction as a result of the process 2-3.

The condensed working fluid is subsequently passed through an expansion valve, which is assumed to be isenthalpic (process 3-4):

$$h_3 = h_4 \quad (16)$$

As a result of the expansion, the temperature of working fluid entering the evaporator reduces to T_4 . T_4 is lower than the temperature of the cold source, hence the working fluid absorbs heat from the cold source (process 4-1). Consequently, the working fluid evaporates and is superheated (process 4a-1) by a degree of superheating, T_{sh} :

$$T_{sh} = T_1 - T_4 \quad (17)$$

The working fluid is superheated to ensure it is a single-phase vapour when it enters the compressor, which would be subject to mechanical damage if liquid were to enter it [32].

The rate at which heat is added to the working fluid, \dot{Q}_{add} , can be calculated using:

$$\dot{Q}_{add} = \dot{m}_{wf} \cdot (h_1 - h_4) \quad (18)$$

where $h_1 - h_4$ is the enthalpy increase of the working fluid. This is equivalent to the rate at which heat is rejected from the cold source, given by:

$$\dot{Q}_{add} = \dot{m}_{cs} \cdot c_{p,cs} \cdot (T_{cs,in} - T_{cs,out}) \quad (19)$$

where \dot{m}_{cs} is the mass flowrate of the cold source, $c_{p,cs}$ is the specific heat capacity at the cold source outlet conditions and $T_{cs,in} - T_{cs,out}$ is the temperature decrease of the cold source stream.

For the purposes of this investigation, there were three main parameters that were calculated and collected using the code. The first variable is the required electricity input for the compressor, \dot{W}_c , which given by:

$$\dot{W}_c = \dot{m}_{wf} \cdot (h_2 - h_1) \quad (20)$$

The second variable was \dot{Q}_{out} , which satisfied the following energy balance:

$$\dot{Q}_{out} = \dot{Q}_{add} + \dot{W}_c \quad (21)$$

Using the first two variables, the coefficient of performance of the heat pump, COP , was able to be calculated. This is a ratio of the heat output at the condenser to the required electricity input in order to run the compressor, given by:

$$COP = \frac{\dot{Q}_{out}}{\dot{W}_c} \quad (22)$$

Payback Time Analysis

The proposed system is based on the operation of PV-T collectors for 12 hours per day, from 6:00 a.m to 6:00 p.m in the local time zone of a given location. This duration was selected since solar irradiance levels and ambient temperatures were found to be the greatest in this period. The annual electricity generated by the PV-T system, on this basis, is given by:

$$W_{pvt,yr} = 365 \sum_{6am}^{6pm} \dot{W}_{pvt,hr} \quad (23)$$

The annual thermal energy generated by the PV-T collectors can be calculated using:

$$Q_{pvt,yr} = 365 \sum_{6am}^{6pm} \dot{Q}_{pvt,hr} \quad (24)$$

The PV-T collectors partially cover the total annual household heating demand (Q_{cov}). The difference between these two values is the annual heating required by the pump, $Q_{pump,yr}$:

$$Q_{pump,yr} = Q_{cov} - Q_{pvt,yr} \quad (25)$$

The average number of hours the heat pump must run per day to fulfil the heating required, N_{hr} , is given by:

$$N_{hr} = \frac{Q_{pump,yr}}{\dot{Q}_{out,avg}} \cdot 365 \quad (26)$$

where $\dot{Q}_{out,avg}$ is the heat transfer rate to the heat sink, averaged across the day. Therefore, the annual heat pump electricity requirement, $W_{c,yr}$, can be calculated:

$$W_{c,yr} = \dot{W}_c \cdot N_{hr} \cdot 365 \quad (27)$$

Depending on the number and configuration of PV-T collectors in the system, the heat pump electricity requirement is either partially or totally accounted for by the electricity generated by the collectors. If it is partially covered, the remaining electricity is bought from the grid at a price (c_e); if it is totally covered, the extra electricity, W_{grid} , is sold back to the grid at a feed-in tariff (FIT) price.

The total capital cost of the system, C_0 , sums the cost of PV-T collectors and the heat pump:

$$C_0 = c_{pvt} \cdot N_c + c_{pump,avg} \quad (28)$$

where c_{pvt} is the cost per PV-T collector, valued at £330 [33] and N_c is the number of collectors. The heat pump model is used to calculate the cost of the heat pump based on hourly conditions and averaged across the day to get $c_{pump,avg}$.

The annual cost savings, C_s , is calculated by considering the natural gas and electricity savings as a result of using this system as opposed to a natural gas

boiler to cover the average annual heating demand, running costs and any electricity sold back to the grid:

$$C_s = W_{cov} \cdot c_e + \frac{Q_{cov}}{\eta_{boil}} c_{ng} + E_{grid} \cdot FIT + W_{c,yr} \cdot c_e \quad (29)$$

where W_{cov} is the annual electricity covered by the PV--T collectors capped at the annual heat pump electricity requirement, $W_{c,yr}$. The cost of natural gas corresponds to c_{ng} and η_{boil} is the boiler efficiency set at 0.9 [16]. Table 3 shows the relevant cost of fuels and heating demand at each location investigated:

Table 3. Average annual household heating demand, cost of natural gas, cost of electricity, feed-in tariff price, fuel inflation rate and discount rates across London, Rome, Tokyo and Los Angeles. [34 - 48]

Location	Q_{cov} (kWh/yr)	C_{ng} (£/kWh)	C_e (£/kWh)	FIT (£/kWh)	I_F (%)	d (%)
London	11,500	0.0742	0.2862	0.064	6.9	3.5
Rome	16,000	0.18	0.33	0.087	15.5	4.9
Tokyo	10,800	0.1	0.16	0.043	3.21	0.9
LA	20,000	0.048	0.228	0.15	3.2	5.5

The payback time (PBT) of the whole system is defined as the period of time required to recover the total capital cost of the system. It is calculated using [15]:

$$PBT = \frac{\ln \left[\frac{C_0 \cdot (i_F - d)}{C_s} + 1 \right]}{\ln \left[\frac{1 + i_F}{1 + d} \right]} \quad (30)$$

where i_F is the fuel inflation rate and d is the discount rate; both values are shown in Table 3 at each location.

Results and Discussion

Simulations were run with configurations of PV-T collectors in series arrangements, and in parallel. Figure 5 shows that a parallel arrangement of 10 collectors yields an electrical efficiency of approximately 0.173 during midday. 10 collectors compounded into two rows of five collectors in series each is in the realm of being 5% more efficient. Plots at different locations exhibited the same pattern, thus rows of collectors connected in series was the configuration selected at all locations.

This is the consequence of the flowrate of water going through the collector in a parallel arrangement being significantly lower since the overall mass flowrate is split between each parallel collector. When arranged in series, the PV-T collectors share the same flow of water within that row, resulting in a larger flowrate cooling each collector. Electrical efficiency increases

with a larger cooling effect and therefore a series arrangement is selected.

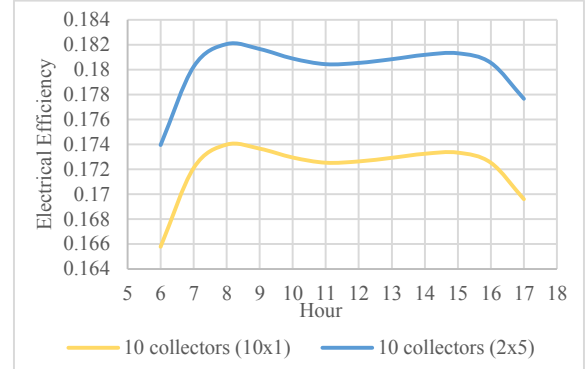


Figure 5. Electrical efficiency plot of different configurations of PV-T collectors across the day in Rome. Configurations are given in the form: number of rows x number of collectors in series per row.

Figure 5 also shows a dip in electrical efficiency during the midday. Higher levels of solar irradiance during the day increases the temperature of the PV cell, consequently decreasing its electrical efficiency. However, a drop a significant drop in efficiency is experienced towards the beginning and end of this 12-hour period. This can be attributed to the optical losses during these times being proportionately higher than during the middle of the day.

To generate results, the code was run many times for each location, with multiple arrangements of PV-T collectors connected to the heat pump. For each configuration, the key performance indicators were noted to be the PBT and C_s . Table 4 shows the results of the PBT analysis for two arrangements of PV-T collectors connected to the heat pump.

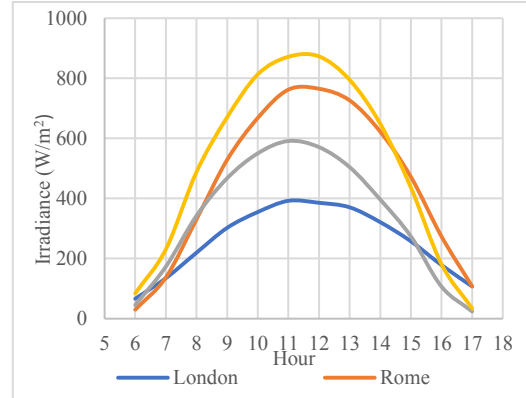


Figure 6. Average daily irradiance profile across all four locations.

Table 4. Comparison of two different arrangements of PV-T collectors connected to heat pump for each location.

Location	Collector Configuration (number of rows x number of collectors in series per row)	CoP	C_o (£)	C_s (£)	PBT (years)
London	3 × 4	3.9	8,500	890	8.6
	2 × 40	3.8	31,000	1,700	15
Rome	3 × 10	3.6	14,000	3,900	3.4
	2 × 5	3.8	7,700	3,200	2.3
Tokyo	3 × 4	3.8	8,400	1,300	6.4
	2 × 10	3.8	11,000	1,400	7.6
Los Angeles	2 × 10	3.7	11,000	1,800	7.0
	2 × 3	3.8	6,400	640	12

In London, the number of collectors tested in the model ranged from 6 to 80. Relative to their initial investment, every arrangement brings relatively small annual cost savings, hence resulting in payback times of up to 15 years. The arrangement of 3 rows of 4 collectors has the lowest payback time of 8.6 years, being the most viable option, but still not preferable when compared to the results from other locations. This can be attributed to the relatively lower solar irradiance levels compared to other locations, as shown in Figure 6, as well as lower ambient temperatures.

In Rome, the number of collectors ranged from 6 to 30. Every arrangement had a shorter *PBT* and better returns than their equivalents in other locations, with considerably higher annual cost savings. The arrangement of 3 rows of 10 has the highest returns over their lifetime, with a short payback time of only 3.4 years. The reason all options seem viable in Rome is not solely because of the hotter weather conditions in this Mediterranean city. Italy also has much higher natural gas prices [23] than similar western nations, which makes the running costs of gas boilers uniquely high. Thus, the PV-T and heat pump system that substitutes the gas boiler becomes considerably more economically viable, as running costs are much reduced.

In Tokyo, a smaller number of collectors were tested, from 6 to 20 collectors. Each arrangement had a relatively short payback time, ranging from 5.6 to 7.6 years. The arrangements of 3 rows of 4 and 2 rows of 10 panels had almost an equal annual cost savings, despite the latter having a longer payback time.

Similar to Tokyo, the number of collectors were varied between 6 and 20 for Los Angeles. However, in contrast, increasing the number of collectors yielded a lower payback time, despite higher capital costs. In contrast, Tokyo had lower payback times for smaller number of collectors. This can be attributed to the fact that the cost of electricity is 30% higher in LA than Tokyo. Therefore, arrangements whereby the electricity requirement of the heat pump was covered in full by the collectors was more favourable in LA.

However, the payback times of the system in LA was generally higher than that of Rome and Tokyo, despite having the most favourable weather conditions, as shown in Figure 6. This can be attributed to the fact that the natural gas prices in the USA are significantly lower than those in Europe or Japan. As such, any system aiming to replace gas boilers would be less economically viable, since the annual cost savings are not as large. It is possible that future government incentives may be put in place to induce a societal shift away from natural gas, such as placing a tax on its use or increasing subsidies on heat pump and solar technology, but at present no policies were found to have been enacted that would have a significant effect.

The novel hybrid system investigated has demonstrated potential for further research, with payback times broadly shorter than those found in existing literature. A paper reported a payback time of a PV-T/ground dual source heat pump as low as 5 years in Northern Italy [49], which agrees with the payback time calculated in Rome. Literature of similar systems

in the UK suggest a significantly higher payback times of up to 14 years [21].

The reason behind a lower payback time in this paper is twofold. Firstly, the capital costs used by existing literature is significantly higher. Lazzarin *et al.* used an investment cost of £11,550 for a 12-panel system in the UK [49], which is 35% higher than a similar size system in London proposed by this paper. This difference is mainly due to installation costs that weren't included during payback time analysis as well as a lower cost per PV-T collector used by this system. Secondly, due to the recent rise in natural gas prices, the annual cost savings calculated were higher than that in literature, further reducing *PBT*. Retrospective calculations using the same capital cost and natural gas price increased the payback time of the 12-collector system in London from 8.6 years to 12 years. Despite a 40% increase in *PBT*, the proposed system still outperforms similar systems in literature.

Conclusion

This paper has delivered on the aims of this project to carry out a techno-economic analysis on various proposed configurations on a PV-T and heat pump system, a sustainable alternative to gas boilers. For each chosen location, data on ambient conditions were extracted to create average daily profiles, which were fed into MATLAB models that were validated against existing models in literature. The outcome of this process are promising results for the system in Rome and Tokyo, delivering low payback times and high yearly benefits across their lifetime. In Rome, the system installed with 3 rows of 10 was calculated to only have a 3.4-year *PBT*, with the system in Tokyo (3 rows of 4) giving a *PBT* of 6.4 years.

In the other locations evaluated, the results were less promising. Payback times for Los Angeles and London, were longer, mainly due to lower natural gas prices in the United States and poor solar irradiance in London. Other locations may be considered in their stead, for further research.

Outlook

As part of continued investigation in future, the viable configurations should be further tested in two different models: one with the PV-T collectors connected to the evaporator and one with them connected to the condenser. This will allow for a comparison of the promising results of payback time found here, connected to the condenser, with the results of the more common variation, to observe which system would perform better when all else is held equal.

The heating system could also be reconfigured, with the heat pump cycle reversed, for cooling applications. This would allow for a use case in hotter locations with stronger sunlight, such as in the Middle East.

Acknowledgements

We are very thankful for Dr Chandan Pandey and Dr Jian Song for their continuous support and insight throughout the project.

References

- [1] Nasralla S. Renewables growth did not dent fossil fuel dominance in 2022, report says [Internet]. 2023 [cited 2023 Nov 9]. Available from: <https://www.reuters.com/business/energy/renewables-growth-did-not-dent-fossil-fuel-dominance-2022-statistical-review-2023-06-25/>
- [2] United Nations. Causes and effects of climate change [Internet]. 2022 [cited 2023 Nov 27]. Available from: <https://www.un.org/en/climatechange/science/causes-effects-climate-change>
- [3] Lindwall C. What are the effects of climate change? [Internet]. 2022 [cited 2023 Nov 27]. Available from: <https://www.nrdc.org/stories/what-are-effects-climate-change#animals>
- [4] Parida B, Iniyar S, Goic R. A review of Solar Photovoltaic Technologies. *Renewable and Sustainable Energy Reviews*. 2011;15(3):1625–36. doi:10.1016/j.rser.2010.11.032
- [5] Barrett N. Why are global gas prices so high? [Internet]. BBC; 2022 [cited 2023 Nov 27]. Available from: <https://www.bbc.co.uk/news/explainers-62644537>
- [6] Parida, B., Iniyar, S. and Goic, R. (2011) ‘A review of Solar Photovoltaic Technologies’, *Renewable and Sustainable Energy Reviews*, 15(3), pp. 1625–1636. doi:10.1016/j.rser.2010.11.032.
- [7] Chowdhury, Md.S. *et al.* (2020) ‘An overview of solar photovoltaic panels’ end-of-life material recycling’, *Energy Strategy Reviews*, 27, p. 100431. doi:10.1016/j.esr.2019.100431.
- [8] Aghaei, M. (2022) ‘Autonomous Monitoring and analysis of Photovoltaic Systems’, *Energies*, 15(14), p. 5011. doi:10.3390/en15145011.
- [9] Tyagi, V.V. *et al.* (2013) ‘Progress in solar PV technology: Research and achievement’, *Renewable and Sustainable Energy Reviews*, 20, pp. 443–461. doi:10.1016/j.rser.2012.09.028.
- [10] Wolniak, R. and Skotnicka-Zasadzień, B. (2022) ‘Development of photovoltaic energy in EU countries as an alternative to fossil fuels’, *Energies*, 15(2), p. 662. doi:10.3390/en15020662.
- [11] Codina, E. *et al.* (2023) ‘Is switching to solar energy a feasible investment? A techno-economic analysis of domestic consumers in Spain’, *Energy Policy*, 183, p. 113834. doi:10.1016/j.enpol.2023.113834.
- [12] Han, Z. *et al.* (2021) ‘Electrical and thermal performance comparison between PVT-St and PV-st systems’, *Energy*, 237, p. 121589. doi:10.1016/j.energy.2021.121589.
- [13] Sevela, P. and Olesen, B.W., 2013. Development and benefits of using PVT compared to PV. *Sustainable Building Technologies*, pp.90–97.
- [14] Abdul-Ganiyu, S. *et al.* (2020) ‘Investigation of Solar Photovoltaic-Thermal (Pvt) and Solar Photovoltaic (PV) performance: A case study in Ghana’, *Energies*, 13(11), p. 2701. doi:10.3390/en13112701.
- [15] EEA (2023) *Greenhouse gas emissions from energy use in buildings in Europe*, European Environment Agency. Available at: <https://www.eea.europa.eu/en/analysis/indicators/greenhouse-gas-emissions-from-energy> (Accessed: 03 December 2023).
- [16] Watkins A. Climate change insights, families and households, UK: August 2022 [Internet]. Office for National Statistics; 2022 [cited 2023 Dec14]. Available: <https://www.ons.gov.uk/economy/environmentalaccounts/articles/climatechangeinsightsuk>
- [17] Rafique, A. and Williams, A.P. (2021) ‘Reducing household greenhouse gas emissions from space and water heating through low-carbon technology: Identifying cost-effective approaches’, *Energy and Buildings*, 248, p. 111162. doi:10.1016/j.enbuild.2021.111162.
- [18] Joshi, S.S. and Dhoble, A.S. (2018) ‘Photovoltaic -Thermal Systems (pvt): Technology review and future trends’, *Renewable and Sustainable Energy Reviews*, 92, pp. 848–882. doi:10.1016/j.rser.2018.04.067.
- [19] Herrando, M. *et al.* (2019) ‘Solar combined cooling, heating and Power Systems based on hybrid PVT, PV or solar-thermal collectors for building applications’, *Renewable Energy*, 143, pp. 637–647. doi:10.1016/j.renene.2019.05.004.
- [20] Badiel, A. *et al.* (2020) ‘A chronological review of advances in solar assisted heat pump technology in 21st Century’, *Renewable and Sustainable Energy Reviews*, 132, p. 110132. doi:10.1016/j.rser.2020.110132.
- [21] Obalanlege MA, Xu J, Markides CN, Mahmoudi Y. Techno-economic analysis of a hybrid photovoltaic-thermal solar-assisted heat pump system for domestic hot water and power generation. *Renewable Energy*. 2022;196:720–36.
- [22] Leonforte, F. *et al.* (2022) ‘Design and performance monitoring of a novel photovoltaic-thermal solar-assisted heat pump system for residential applications’, *Applied Thermal Engineering*, 210, p. 118304. doi:10.1016/j.applthermaleng.2022.118304.
- [23] Orso L. What does the latest census data reveal about how homes are heated across England and Wales? [Internet]. [cited 2023 Nov 29]. Available from: <https://www.nesta.org.uk/blog/what-does-the-latest-census-data-reveal-about-how-homes-are-heated-across-england-and-wales/>
- [24] Lévesque B, Lavoie M, Joly J. Residential Water Heater temperature. *Canadian Journal of Infectious Diseases*. 2004;15(1):11–2. doi:10.1155/2004/109051
- [25] Creevy O. What Size Air source heat pump do I need for my home? [Internet]. 2023 [cited 2023 Nov 29]. Available from: [9](https://heat-

</div>
<div data-bbox=)

- pumps.org.uk/what-size-air-source-heat-pump-do-i-need-for-my-home/
- [26] Brita. What is the best temperature for drinking water? [Internet]. [cited 2023 Nov 26]. Available from: <https://www.brita.co.uk/news-stories/dispenser/what-is-the-best-temperature-for-drinking-water>
 - [27] Swinbank WC. Long-wave radiation from clear skies. Quarterly Journal of the Royal Meteorological Society. 1963;89(381):339–48. doi:10.1002/qj.49708938105
 - [28] Herrando M, Ramos A, Zabalza I, Markides CN. A comprehensive assessment of alternative absorber-exchanger designs for hybrid pvt-water collectors. Applied Energy. 2019;235:1583–602. doi:10.1016/j.apenergy.2018.11.024
 - [29] JRC Photovoltaic Geographical Information System (PVGIS) - european commission. [cited 2023 Dec 14] Available at: https://re.jrc.ec.europa.eu/pvg_tools/en/
 - [30] MTB. How much does a new roof cost? [Internet]. 2023 [cited 2023 Nov 29]. Available from: <https://mylocaltoolbox.co.uk/cost-guides/how-much-does-a-new-roof-cost/>
 - [31] Chapp T. Tech tip: Liquid Subcooling [Internet]. IIAR. 2016 [cited 2023 Dec 7]. Available from: https://www.iiar.org/IIAR/IIAR/IIAR_News/Tech_Tip/Tech_Tip_Liquid_Subcooling.aspx
 - [32] Selbas R, Kizilkcan O, Sencan A. Thermoeconomic optimization of subcooled and superheated vapor compression refrigeration cycle. Energy. 2006;31(12):2108–28.
 - [33] Herrando M, Ramos A, Freeman J, Zabalza I, Markides CN. Technoeconomic modelling and optimisation of solar combined heat and power systems based on flat-box pvt collectors for domestic applications. Energy Conversion and Management. 2019;175:67–85.
 - [34] Ofgem. Average gas and electricity usage. [cited 2023 Dec 1]. Available from: <https://www.ofgem.gov.uk/information-consumers/energy-advice-households/average-gas-and-electricity-use-explained>
 - [35] Ofgem. Energy price cap. [cited 2023 Nov 29]. Available from: <https://www.ofgem.gov.uk/energy-price-cap>
 - [36] CTM. Feed in tariff scheme . [cited 2023 Dec 2]. Available from: <https://www.comparethemarket.com/energy/content/feed-in-energy-tariffs/>
 - [37] Department of Health & Social Care, editor. DHSC Group Accounting Manual 2022 to 2023: Additional guidance, version 3 [Internet]. GOV.UK. 2023 [cited 2023 Dec 3]. Available: <https://www.gov.uk/government/publications/dhsc-group-accounting-manual-2022-to-2023/dhsc-group-accounting-manual-2022-to-2023-additional-guidance-version-1>
 - [38] Beckett D. Consumer price inflation, UK: October 2023 [Internet]. Consumer price inflation, UK - Office for National Statistics. Office for National Statistics; 2023 [cited 2023 Dec 2]. Available from: <https://www.ons.gov.uk/economy/inflationandpriceindices/bulletins/consumerpriceinflation/october2023>
 - [39] Italy long term interest rate (I:ILTIRNM) [Internet]. Italy Long Term Interest Rate. [cited 2023 Dec 2]. Available from: https://ycharts.com/indicators/italy_long_term_interest_rates
 - [40] Statista . Italy: Pre-tax natural gas bill by tariff area 2023. 2023 [cited 2023 Dec 2]. Available from: <https://www.statista.com/statistics/1339913/annual-natural-gas-cost-for-average-domestic-consumers-by-tariff-area-italy/>
 - [41] O'Neill A. Japan: Inflation rate 2028 [Internet]. Statista. 2023 [cited 2023 Dec 14]. Available: <https://www.statista.com/statistics/270095/inflation-rate-in-japan/>
 - [42] Klein C. Japan: Energy consumption per household by region 2022 [Internet]. Statista. 2023 [cited 2023 Dec 14]. Available from: <https://www.statista.com/statistics/1291424/japan-energy-consumption-per-household-by-region/>
 - [43] Japan Energy Prices [Internet]. GlobalPetrolPrices.com. [cited 2023 Dec 14]. Available from: <https://www.globalpetrolprices.com/Japan/>
 - [44] 1. N. Sönnichsen. Household natural gas prices by country 2022 [Internet]. Statista. 2023 [cited 2023 Dec 14]. Available from: <https://www.statista.com/statistics/702735/household-natural-gas-prices-in-selected-countries/>
 - [45] U.S. Bureau of Labor Statistics. Consumer price index, Los Angeles Area - November 2023 : Western Information Office. 2023 [cited 2023 Dec 4]. Available from: https://www.bls.gov/regions/west/news-release/consumerpriceindex_losangeles.htm
 - [46] The Federal Reserve. Current discount rates. 2023 [cited 2023 Dec 5]. Available from: <https://www.frbdiscountwindow.org/Pages/Discount-Rates/Current-Discount-Rates>
 - [47] U.S. Bureau of Labor Statistics. Average energy prices, Los Angeles: Western Information Office. 2023 [cited 2023 Dec 5]. Available from: https://www.bls.gov/regions/west/news-release/2023/averageenergyprices_losangeles_20231120.htm
 - [48] Staff W. Average natural gas usage per month - 2023 [Internet]. Shrink That Footprint. 2023 [cited 2023 Dec 14]. Available from: <https://shrinkthatfootprint.com/average-natural-gas-usage-per-month>
 - [49] Lazzarin R, Noro M, Chow TT, Sathe TM, Joshi SS, Wolf M, et al. Photovoltaic/thermal (PV/t)/ground dual source heat pump: Optimum Energy and economic sizing based on performance analysis [Internet]. Energy and Buildings. Elsevier; 2020 [cited 2023 Dec 14]. Available: <https://www.sciencedirect.com/science/article/pii/S378778819331767>

Supply Chain Optimisation for Plasmid DNA

Cheryl Lum and Yu Hang Yang

Department of Chemical Engineering, Imperial College London, U.K.

Abstract: In the evolving field of genetic engineering, plasmid DNA (pDNA) has emerged as a frontier of innovation and therapeutic potential. The surge in demand for pDNA was catalysed by the development of DNA vaccines against SARS-CoV-2 and has since exacerbated a need to design a robust supply chain network to offer pharmaceutical-grade pDNA accessibility globally. This paper first reviews the overarching challenges hindering the performance of the current supply chain. Then, it proposes a feasible solution to address the disparity between the pDNA supply and demand by modelling the supply chain using actual industry data. Two Mixed-Integer Linear Programming (MILP) optimisation models were developed on Python and deployed using Pyomo to minimise total costs and total time in the supply chain. Results from candidate supply chains propose that establishing a medium-scale manufacturing site offers the optimal supply chain network to cater for different demand scales. Moreover, demand scales ranging from 1,500 doses to 750,000 doses of pDNA were evaluated to assess the model's resilience. Finally, suggestions on future areas were proposed to relieve the supply chain bottleneck and promote the global applications of advanced therapy medicinal products (ATMPs).

Keywords: plasmid DNA, cost-benefit analysis, supply chain, MILP, optimization

1. Introduction:

Plasmid DNA (pDNA) is a small, circular, double-stranded DNA molecule. It naturally exists as an extrachromosomal DNA capable of autonomous replication of the host's chromosome via a segment of the plasmid called the replicon [1].

pDNA is an essential instrument for genetic manipulation and therapies due to its distinctive feature of acting as a vector in molecular cloning. It is a fundamental starting material in developing DNA vaccines, recombinant antibody therapies and cell therapies. [2]. Nevertheless, the success of developing the DNA vaccines against SARS-CoV-2 [3] has accelerated the ATMPs industry.

The increase in usage in novel ATMPs represents a paradigm shift in treating diseases and offering personalised and highly effective solutions to patients in modern days. Thus, it underpins the significance of biotechnological innovation and therapeutic potential in our current landscape. The increased usage of novel ATMPs represents a paradigm shift in treating diseases and offering personalised and highly effective solutions to patients in modern days. Thus, it underpins the significance of biotechnological innovation and therapeutic potential in our current landscape.

Our study focuses on a keystone development in the field of ATMPs – pDNA, the molecule that has emerged to be at the frontier of genetic engineering. Hence, the need for large-scale production of Good Manufacturing Practice (GMP) grade materials for pDNA has never been greater. Marked by significant advancements and expanding applications in biotechnology and medicine, the growth trajectory of pDNA is expected to increase exponentially as it continues to shape the future of genetic research and therapeutic development. Such growth can be further demonstrated by the pDNA manufacturing market size being valued at USD 540 billion in 2023 [4]. This reflects pDNA's pivotal position in the field of genetic engineering as research continues to unravel new applications and to improve existing methodologies with an overall industry CAGR of 21.7% from 2022 - 2030. [4]

Despite pDNA's promising growth prospects, the current supply chain is constantly battling through backlogs and bottlenecks. The pDNA supply chain is very intricate and complex, both intrinsically and extrinsically. This is where our study comes in place: to delve into the substantial challenges by analysing the supply chain structure, as well as to fill in the gap in the research landscape by developing a MILP model to assess the robustness of the pDNA supply chain.

This was done by further dissecting node by node to ensure that any of our candidate supply chains is comprise of one manufacturing site, one airport, one storage unit, and one demand zone coupled with two modes of transportation - air and truck.

Firstly, we begin by synthesising key insights from literature reviews. Secondly, two MILP formulations were developed to enable comparisons to be made between candidate pDNA supply chains. Finally, a cost-benefit analysis on the average costs per dose of pDNA and facility utilisation rates was conducted. Findings were evaluated and compared across candidate supply chains.

Our study focuses on developing an optimisation-based mathematical model to address the prevailing challenges in the pDNA industry. As such, mathematical formulations and constraints are employed by optimisation algorithms to identify the best candidate supply chain. It also aids the decision-making process to design an optimal supply chain network to minimise costs and maximise capacity to meet global demand.

2. Background:

2.1 Overview

The growing prospects of the applications of ATMPs in the field of genetic engineering has generated a surge in demand for genetic therapies. This calls for a need to develop a robust supply chain framework. Previous works by D Ibrahim on the supply chain optimization for viral vectors and RNA vaccines [5] proposes that the configuration of any given ATMPs supply chain should include manufacturing plants, fill and finish plants, quality control sites, storage warehouses and a selection of routes and transportation regimes.

Furthermore, literature suggests that supply chain designs, capacity and investment optimisation models should focus on the long-term decision by selecting strategic locations of the plants to offer global reach [6]. Given the size and intricacies of the optimisation problems, the adoption of computational tools and modelling framework have been highly effective in the supply chain optimisation field. Combining a MILP optimisation model with computational tools in the pharmaceutical supply chain optimisation space yields multiple advantages [7]. Firstly, feasible candidate supply chain structures can be offered [8]. Such structures are proposed after investigating the desired direction (either minimising or maximising) of the Key Performance Indicators (KPIs). Secondly, it allows for better maintenance across therapy quality and reduces inherent human errors.

2.2 Challenges in pDNA supply chain

The post COVID-19 landscape is currently battling through a pDNA supply chain bottleneck as the industry calls for more and more pharmaceutical-grade pDNA. Challenges remain unsolved and are derived from I. scaling-up pDNA manufacturing capacity and the pDNA supply chain to meet the global demand [9] II. developing robust supply chains that are capable of handling pDNA products under cold chain logistics [10] III. Producing clinical grade pDNA that adheres to Current Good Manufacturing Practices (cGMP) [11] and other regulatory protocols such as the U.S. Food and Drug Administration (FDA), which can be stringent and vary across regions. IV. High costs associated with expanding and operating facilities for manufacturing and storage, which limits the overall capacity of pDNA being produced. All of these present a challenging landscape that may hinder progress in the rapidly advancing field of cell and gene therapy.

A major challenge in pDNA industry revolves around an unprecedented surge in demand, largely attributable to the significant and consistent growth experienced by the cell and gene therapy sector[9]. This upswing is intricately linked to the critical role of plasmids in viral vector production. The industry currently grapples with bottlenecks, as the substantial growth in the gene therapy field was not entirely foreseeable, resulting in a backlog in plasmid production and prolonged waiting lists. The unforeseen urgency for plasmids to be readily available on demand has led to challenges such as production backlogs and extended waiting times. The foremost concern is the potential inability to meet industry demands promptly, potentially impeding research and development pipelines globally. This delay could adversely impact market expectations and, more critically, disappoint patients awaiting solutions tailored to their specific needs. Despite recent investments in new facilities and capacity by some players, the gap between the escalating need for high-quality plasmids and the current supply capacity remains pronounced. [9] Reports of delays from suppliers underscore the severity of the issue, emphasising the urgent need for an alignment between the demand for commercial-quality plasmids and the industry's

production capabilities. This discrepancy presents a critical challenge that may hinder progress in the rapidly advancing field of cell and gene therapy.

2.3 In-vivo and ex-vivo

Applications of pDNA could be categorised by in-vivo or ex-vivo therapies. In-vivo is a non-patient specific mechanism which refers to the direct introduction of pDNA into the body. pDNA will enter cells and exert its effects, commonly found in DNA vaccines. [12]

Ex-vivo is a patient specific mechanism which removes cells from the body. Cells are then treated with pDNA in a laboratory setting, and then reintroduced into the body. This method is often used in gene therapies where specific cells (like T-cells) are modified outside the body before being returned to the patient. [13]

2.4 pDNA manufacturing

pDNA manufacturing is the most time-consuming and expensive step in the pDNA supply chain. Besides that, it also involves several sophisticated biotechnological processes, which includes fermentation, cell harvesting, and purification. Each aforementioned step requires meticulous quality control and adherence to cleanroom standards. [14].

The primary upstream manufacturing starts with an E. coli culture medium. It first undergoes a single batch fermentation in a large stainless-steel bioreactor or single-use fermenters to scale-up [15] Next, the primary downstream purification begins with harvesting of the cells, lysis, and clarification. Purification can be very challenging as most of the critical impurities (RNA, genomic DNA, endotoxins) are negatively charged and are also similar in size as circular pDNA, making it difficult to separate [16].

Bulk production of pDNA is challenging, as this requires continuous monitoring of the important production parameters including agitation, pH, temperature, dissolved carbon dioxide, pressure, and foam formulation. A study by O. Ruiz, Miladys, Jorge, M. Pupo, and Eduardo, shows that, the average amount of pDNA recovered from 1 kilogram of wet biomass ranges from 0.5 to 1.0 grams final product [17]. Therefore, there is a substantial manufacturing challenge that causes production backlogs and prolonged waitlists for pDNA in the market.

2.5 Quality control (QC)

Ensuring the quality control of pDNA is essential to guarantee the identity, purity, and efficacy of the product. This necessitates rigorous compliance with cGMP regulations [18].

The quantification of yield serves as a foundational step in assessing the efficiency of the pDNA extraction and purification process. The concentration of DNA is typically determined using agarose gel electrophoresis, which detects the presence of other substances in the DNA. The process utilises a spectrophotometer, which measures absorbance at where DNA absorbs, 260 nm [19]. Primarily in labs, the A260/A280 ratio is used to assess the purity of the molecule. A ratio nearing 1.8 signifies purity; however,

a ratio of 1.6 or lower suggests the presence of protein, phenol, or other contaminants, which exhibit an absorbance at 280 nm [19].

Endotoxin detection is another crucial part of pDNA quality control. It can significantly reduce transfection efficiencies, influencing the uptake of pDNA during in-vivo therapies; but when in bloodstream, lead to hypotension and respiratory failure [20]. However, it is also non-detectable on agarose gels and by optical density [21]. An extra test of using LAL (Limulus ameocyte lysate) is done to detect the presence of endotoxin. If present, LAL will coagulate [22].

As it is an important aspect of manufacturing, it is therefore necessary to address the supply chain problem to relief the sector from the upsurge of the pDNA demand in the post COVID-19 horizon, to increase pDNA production capacity without compromising on quality and costs.

2.6 Storage

Two types of storage units are widely used in the genetic therapy industry. This includes small scale regional stores and large-scale warehouses. Small-scale regional stores are typically compact facilities strategically located in proximity to research institutions, clinics, and biotechnology companies. These storage units are characterised by their limited size, making them suitable for short-term storage and rapid access to genetic materials. Large-scale warehouses are extensive storage facilities designed to accommodate vast quantities of genetic materials for long-term preservation. These facilities are essential for biobanks, gene therapy manufacturers, and institutions with extensive genetic material collections.

Short-term storage usually refers to a maximum period of 18 days at room temperature–[23]. This time frame is strategically designated to accommodate flexibility in the transportation process, providing a buffer period during which logistical adjustments and organisational considerations can be addressed.

Long-term storage of pDNA is done in TE (10mM Tris-HCl, 1mM EDTA, pH 8.0) buffer at a controlled temperature of -20°C [24]. The utilisation of this specific buffer composition serves as a protective feature to ensure the enduring stability of the genetic material over an extended period. The pH of the TE buffer is maintained at 8.0, a critical factor in safeguarding the integrity of DNA. This is particularly significant due to the sensitivity of DNA molecules to changes in pH. The stabilising role of Tris-HCl as a buffering agent ensures that the pH remains within the optimal range, preventing any deviations that might compromise the chemical structure and functionality of the stored genetic material [25]. It can be stored for a maximum of 270 days under -20°C conditions. [23]

2.7 Fill and finish

Fill and finish stage stands as the final step in the complex manufacturing process of pDNA. Under GMP guidelines, an electroporation buffer is prepared, mixed and dispensed into individual, single-use aliquots [26].

Such meticulous procedures are instrumental in maintaining the integrity of the pDNA, ensuring the safe and effective application across therapeutic uses.

2.8 Cold chain logistics

Cold chain temperature-controlled logistics ensures that pDNA remains at the required temperature throughout the distribution process. Utilising specialised packaging with ice packs or dry ice for shipments that need to be kept at -20°C or -80°C [27]. Monitoring temperature during transit using temperature loggers is essential.

3. Methods:

3.1 Data Collection

3.1.1 Demand Scale

In order to mimic global demand scales, three key demand scales as follows were investigated and chosen for our studies: Phase I/II clinical trials – 200; Phase III clinical trials – 1000 patients and 2000 patients and commercial scale – 32,000 patients. The corresponding number of doses pDNA required globally for each scale is denoted in *table 1*. Such demand scales were modelled after VGXI's paper [28].

Table 1- the demand scales of focus of our studies and its corresponding data

Demand scale	Number of patients in each region	Number of doses required globally
Phase I/II clinical trials	200	3,000
Phase III clinical trials	1000	15,000
Phase III clinical trials	2000	30,000
Commercial	32,000	475,000

3.1.2 Manufacturing Sites

A total of six manufacturing sites locations across the US and Europe were selected, of which includes three types of manufacturing sites with varying production capacities.

Data shown in the *Supplementary Information section* outlines the parameters related to the types of manufacturing sites being investigated. [7] [29]. It showcases the production capacities and all relevant cost parameters required for our model's inputs.

Table 2- Node notations and locations of all selected facilities to be investigated upon

Manufacturing sites (j)	Location	Airport (ak)	Location
j1	USA Madison	ak1	USA/Dane County Regional Airport
j2	USA San Diego	ak2	USA/San Diego International Airport
j3	USA Texas	ak3	USA/George Bush Intercontinental Airport
j4	EU-ESP San Sebastian	ak4	EU/Bilbao Airport
j5	EU-SWE Sundsvall	ak5	EU/Sundsvall-Timrå Airport
j6	EU-DEU Marburg	ak6	EU/Frankfurt Airport
Storage sites (k)	Location	Demand points (l)	Location
k1	USA/Cleveland	l1	UK
k2	UK/Glasgow	l2	USA

To ensure that our model can provide a comprehensive representation of the real-world scenarios, all 6 of our candidate manufacturing sites displayed in table 2 were selected across 2 regions: the US and Europe. Three distinct production scales of

manufacturing sites were investigated with their corresponding data shown in the supplementary information section that are required for our model's inputs. Each region contains one of each production size as the foundational design.

All candidate manufacturing sites are assumed to be a black box; meaning operations listed as follows were grouped together to be evaluated as one: primary manufacturing; secondary manufacturing; fill and finish; quality control. Quality control (QC) and raw material costs were assumed to be 1,242 \$/dose and 1,397 \$/dose according to literature values [7].

Other assumptions are outlined as follows:

- Single production line at any manufacturing sites
- Single product- non-viral pDNA being produced throughout
- Batch process
- Continuous year-round operation without interruption at 100% efficiency

3.1.3 Airport

The logistical framework of the model is designed for full geographical reach across both the US and West Europe. Eight identified airports shown in table 2, distributed across the cardinal directions, North, South, East, and West of each region serve as nodes for the transportation network. Figure 12 shows a clear representation of our proposed flight path distribution from the selection of airports across the USA and Europe.

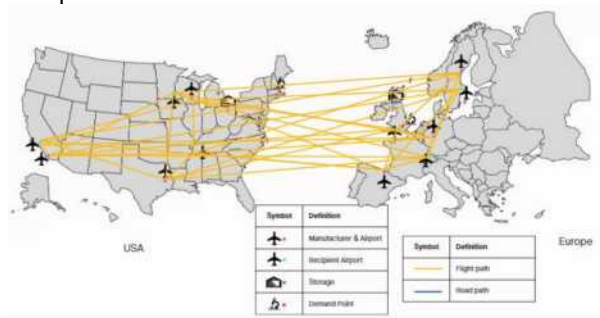


Figure 12: Flight path distribution of the selected airports across the USA and Europe

3.1.4 Storage

Both storage locations shown in table 2 were selected based on existing industry storage sites. Both were assumed to be warehouses which offers a total storage space of 150m³.

The economic considerations associated with pDNA storage involve capital and operational costs. The capital cost component encompasses the installation of the storage infrastructure, with a charge of \$28,700 per facility per week [29]. The operational cost component comprises ongoing expenses related to utilities, labour and maintenance. Operational costs were quantified at \$0.0098 per dose [29], accounting for day-to-day activities necessary for the efficient functioning of the storage facilities.

3.1.5 Transport

Plasmid DNA is transported through two distinct channels: 1) airlifted and 2) refrigerated truck. The

airfreight rate is approximately \$9 per kg [30], with a maximum weight per trip fixed at 64,480 kg [31]. Transportation by refrigerated truck incurs a cost of \$0.26 per kilometer, at a maximum weight per trip of 1,500 kg per trip, equivalent to 1.75 million doses per trip. Travel distances between entities were estimated taking the shortest driving distance available on Google Maps. travel speeds for airfreight and roads are 926

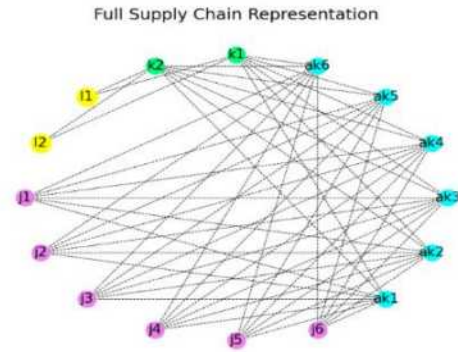


Figure 1 - Full representation of the proposed pDNA supply chain

km/h [32] and 95 km/h respectively. [33] Travel times were hence calculated by dividing the travel distance by its respective travel speeds.

3.2 Mathematical Formulations

3.2.1 Problem Statement

The overarching aim of the supply chain optimisation problem is to ensure pDNA products are adequately distributed to the demand zones to meet global demands in a timely manner. It is hence necessary to design a viable and cost-effective supply chain to address the current challenges within the pDNA industry. A full representation of our proposed pDNA supply chain can be seen in Figure 1, where all possible nodes to be established could be visualised.

It is therefore vital to address the KPIs in the pDNA supply chain to make informed decisions to better optimise the existing supply chain frameworks. The underlying models were therefore formulated as MILP optimisation problems and further constructed on Python 3.12.0 and PYOMO 6.7 to be solved using the IBM ILOG CPLEX solver. These were chosen as because of its efficiency in handling supply chain optimisation-based MILP problems and are widely utilised in industry [34].

This problem encompasses identifying existing good network structures while accounting for constraints and parameters.

3.2.2 Objective Function

Two models in aims to minimise the KPIs: 1) total costs and 2) total time of the pDNA supply chain were developed.

3.2.2.1 Minimising Costs

The overarching problem was further broken down into two independent models. Our first model addresses the KPI on total costs. One core facet of the problem stems

from the need to optimise capacity planning to meet global demands. It was necessary to then implement a robust product flow allocation system. In principle, minimising the total costs associated with the production of pDNA to maximise its returns from sales is a crucial element to be considered when evaluating the efficacy of a supply chain.

Another core facet of the problem addresses the need to identify effective network structures to be established within each candidate supply chain. Minimising costs associated with pDNA manufacturing sites and storage sites; which includes capital costs (initial installment, equipment), operational costs (labour and raw material) and transportation rates is central to solving the formulated objective as seen in *equation 45*. Trade-offs were also considered within the optimisation model whereby it is necessary to determine the lowest total cost of operations to ensure the best return on investments. *Equations 1-35* were utilised in this model.

Objective Function for Minimising Costs

Minimising costs associated with the manufacturing of pDNA. This includes manufacturing sites, storage units and transportation between nodes.	$\min f(x) = \sum_j Y_j(CI_j + OI_j) + (QC + \text{raw material})P_j + \sum_k Y_k(CI_k + OI_k) + \sum_j \sum_{ak} C_{j,ak}^{\text{transport}} X_{j,ak} + \sum_{ak} \sum_k C_{ak,k}^{\text{transport}} X_{ak,k} + \sum_k \sum_l C_{k,l}^{\text{transport}} X_{k,l}$	45
--	--	----

3.2.2.2 Minimising Time

The model addresses the second KPI - total time of the supply chain. Time is thereby the main variable to be considered in this formulation. *Equation 46* aims to minimise time associated with the whole supply chain network.

The model's parameters and constraints were based off the initial cost model, whereby *equations 1-35* were also utilised in this model. *Equations 36-44* were also formulated and adopted in this model. Cost variables were no longer included in the objective function shown in *equation 46* to be evaluated.

Objective Function for Minimising Time

Minimising time associated with total manufacturing time - which includes batch time; transportation time- which includes time for transportation between nodes	$\min f(x) = \sum_j T_j Y_j + \sum_{ak} \sum_j T_{j,ak} X_{j,ak} + \sum_k \sum_{ak} T_{ak,k} X_{ak,k} + \sum_l \sum_k T_{k,l} X_{k,l}$	46
---	--	----

3.2.3 Decision variables

Decision variables are values to be decided by the MILP optimisation problem. All decision variables encompass

an equal importance to obtain the optimal solution for the supply chain optimisation network problem. Decisions made determine I. which of the facilities $j \in j, k \in k$ were selected to be commissioned within the time horizon II. which of the transport nodes between facilities were to be established to complete the supply chain. III. The total costs associated with the supply chain which is further broken down into capital costs which includes a. rent b. initial capital costs for building the facility; operational costs which outlines the operational costs which includes a. electricity, water and bills b. raw material; transportation costs, which includes a. air transport (\$/kg) b. truck transport (\$/km).

3.2.4 Constraints

Constraints formulated could be categorised as follows I. Material balances, II. Logic constraints outlining the existence of a facility or node, III. Transport and storage capacity constraints, IV. Global demand of the demand zones for pDNA, V. Time constraints VI. Cost constraints

Description	Material Balances	
Total amount of products of pDNA produced	$P_j = \sum_{ak} Q_{j,ak}, \quad \forall j \in J$	1
Sample mass balance of the flow of products	$\sum_j Q_{j,ak} = \sum_k Q_{ak,k}, \quad \forall ak \in AK$	2
Sample mass balance of the flow of products	$\sum_{ak} Q_{ak,k} = \sum_l Q_{k,l}, \quad \forall k \in K$	3
Sample mass balance of the flow of products	$\sum_k Q_{k,l} \geq D_l, \quad \forall l \in L$	4
Number of products produced at manufacturing site j is equal to the multiple of the number of batches utilised and batch size at the corresponding site	$P_j = B_j^{\text{size}} B_j^{\text{number of}}, \quad \forall j \in J$	5
Description	Logic Constraints	
Ensures transportation links can only exist with existing facilities	$X_{j,ak} \leq Y_j, \quad \forall j \in J, ak \in AK$	6
	$X_{ak,k} \leq Y_{ak}, \quad \forall ak \in AK, k \in K$	7
	$X_{k,l} \leq Y_k, \quad \forall k \in K, l \in L$	8
Ensures transportation links arriving at the (cont.) facilities must exist for the existing facilities	$\sum_j X_{j,ak} \geq Y_{ak}, \quad \forall ak \in AK$	9
	$\sum_{ak} X_{ak,k} \geq Y_k, \quad \forall k \in K$	10
	$\sum_k X_{k,l} \geq 1, \quad \forall l$	11
Ensures that Europe facilities goes to Europe only and US facilities goes to US only	$\sum_k \sum_l X_{k,l} = 2$	12
	$X_{j_1,ak_4} + X_{j_1,ak_5} + X_{j_1,ak_6} = 0$	13
	$X_{j_2,ak_4} + X_{j_2,ak_5} + X_{j_2,ak_6} = 0$	14
	$X_{j_3,ak_4} + X_{j_3,ak_5} + X_{j_3,ak_6} = 0$	15
	$X_{j_4,ak_1} + X_{j_4,ak_2} + X_{j_4,ak_3} = 0$	16
	$X_{j_5,ak_1} + X_{j_5,ak_2} + X_{j_5,ak_3} = 0$	17
	$X_{j_6,ak_1} + X_{j_6,ak_2} + X_{j_6,ak_3} = 0$	18

$$X_{ak_1,k_2} + X_{ak_2,k_2} + X_{ak_3,k_2} + X_{ak_4,k_1} + X_{ak_5,k_1} + X_{ak_6,k_1} = 0 \quad 19$$

Description	Capacity Constraints		
Ensures the flow of products between each transportation link is within the minimum and maximum flow of products capacity for refrigerated trucks	$Q_{j,ak} \geq Q^{MIN} X_{j,ak},$	$\forall j \in J, ak \in AK$	20
	$Q_{j,ak} \leq Q^{MAX} X_{j,ak},$	$\forall j \in J, ak \in AK$	21
	$Q_{ak,k} \geq Q^{MIN} X_{ak,k},$	$\forall ak \in AK, k \in K$	22
	$Q_{ak,k} \leq Q^{MAX} X_{ak,k},$	$\forall ak \in AK, k \in K$	23
	$Q_{k,l} \geq Q^{MIN} X_{k,l},$	$\forall k \in K, l \in L$	24
	$Q_{k,l} \leq Q^{MAX} X_{k,l},$	$\forall k \in K, l \in L$	25
Ensures the total flow of products arriving at each storage site is below the maximum storage capacity	$\sum_{ak} Q_{ak,k} \leq K^{MAX} Y_k,$	$\forall k \in K$	26

Description	Cost Equations		
Transport costs from manufacturing site j to airport location ak for transportation link between j, ak , where both refrigerated truck and air freight costs are taken into account	$C_{j,ak}^{transport} = Q_{j,ak} DIS_{j,ak} UT,$ $+ Q_{j,ak} weightAUT,$	$\forall j \in J, ak \in AK$	27
Transportation costs to deliver products between nodes for each refrigerated truck as the only mode of transport	$C_{ak,k}^{transport} = Q_{ak,k} DIS_{ak,k} UT,$	$\forall ak \in AK, k \in K$	28
	$C_{k,l}^{transport} = Q_{k,l} DIS_{k,l} UT,$	$\forall k \in K, l \in L$	29
Cost equations for manufacturing sites	$C_j^{CAPEX} = (CI_j B_j^{number\ of} + R_j) \times Y_j,$	$\forall j \in J$	30
	$C_j^{OPEX} = OI_j B_j^{number\ of},$	$\forall j \in J$	31
Cost equations for storage units	$C_k^{CAPEX} = R_k,$	$\forall k \in K$	32
	$C_k^{OPEX} = OI_k \sum_{ak} Q_{ak,k},$	$\forall k \in K$	33

Description	Time equations		
Time equations for manufacturing sites	$T_j = B_j^{number\ of} B_j^{time},$	$\forall j \in J$	34
	$T_j \leq wait\ time,$	$\forall j \in J$	35
Time equation for transportation time between nodes	$T_T = \sum_{ak} \sum_j T_{j,ak} X_{j,ak} + \sum_k \sum_{ak} T_{ak,k} X_{ak,k}$ $+ \sum_l \sum_k T_{k,l} X_{k,l}$		36

Description	Cost equations		
Returns the total costs for 1) transport 2) facilities	$C^{total\ transport} = \sum_{j,ak} C_{j,ak}^{transport} + \sum_{ak,k} C_{ak,k}^{transport}$ $+ \sum_{k,l} C_{k,l}^{transport}$		37
	$C^{total\ facilities} = \sum_j (C_j^{CapEx} + C_j^{OpEx} + (QC + raw\ materials) P_j) + \sum_k (C_k^{CapEx} + C_k^{OpEx})$		38
Big M and small ε cost constraints: de/activation of constraints when needed	$\sum_j \sum_{ak} C_{j,ak}^{transport} \geq \varepsilon X_{j,ak}$		39
	$\sum_j \sum_{ak} C_{j,ak}^{transport} \leq M X_{j,ak}$		40
	$\sum_{ak} \sum_k C_{ak,k}^{transport} \geq \varepsilon X_{ak,k}$		41
	$\sum_{ak} \sum_k C_{ak,k}^{transport} \leq M X_{ak,k}$		42
	$\sum_k \sum_l C_{k,l}^{transport} \geq \varepsilon X_{k,l}$		43
	$\sum_k \sum_l C_{k,l}^{transport} \leq M X_{k,l}$		44

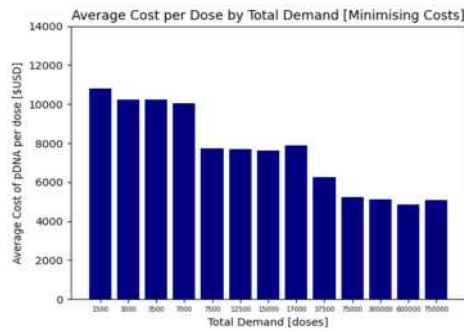


Figure 2-Results from the minimising costs model on the average cost per dose of pDNA by total global demand

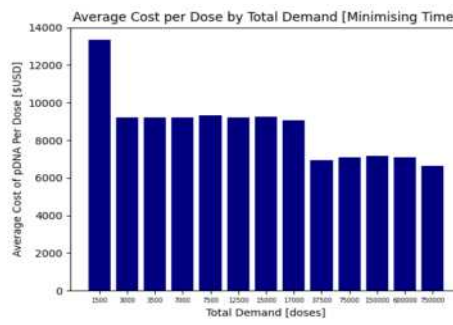


Figure 3- Results from the minimising time model on the average cost per dose of pDNA by total global demand

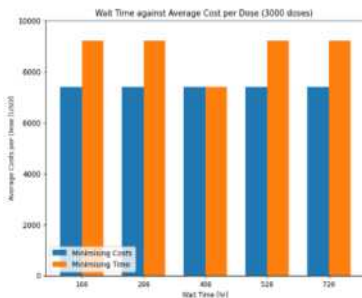


Figure 4- Wait time against cost per dose (3000 doses)
Average Manufacturing Sites Utilization Rate for 3000 Doses

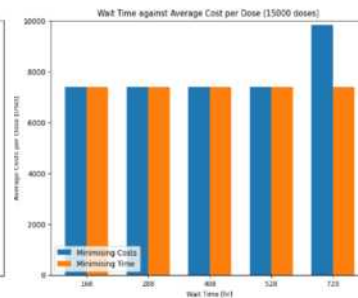


Figure 5- Wait time against cost per dose (15,000 doses)
Average Manufacturing Sites Utilization Rate for 15000 Doses

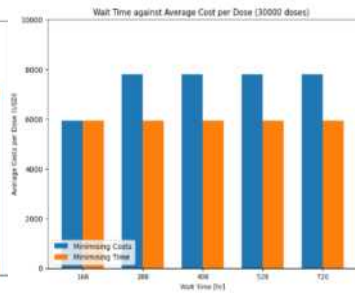


Figure 6- Wait time against cost per dose (30,000 doses)
Average Manufacturing Sites Utilization Rate for 30000 Doses

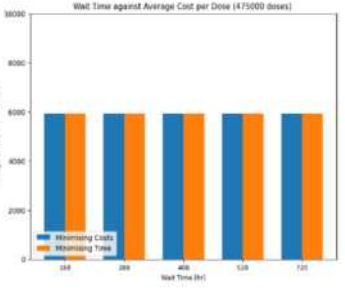


Figure 7- Wait time against cost per dose (475,000 doses)
Average Manufacturing Sites Utilization Rate for 475000 Doses

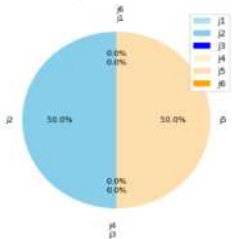


Figure 8- Facility utilisation rate (3000 doses)

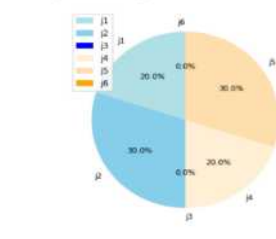


Figure 9- Facility utilisation rate (15,000 doses)

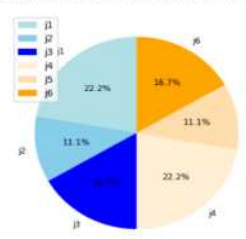


Figure 10- Facility utilisation rate (30,000 doses)

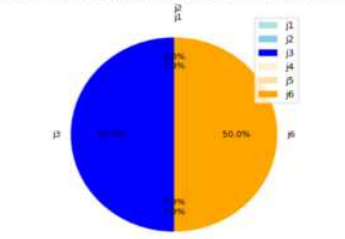


Figure 11- Facility utilisation rate (475,000 doses)

4. Results:

4.1 Average cost per dose against global demand

Figure 3 illustrates the results of the change in average cost per dose over a range of demand with the objective of minimising cost. The graph shows a decreasing trend with a large drop at 7000 doses, from \$10,797 to, and at 17,000 doses, from \$7,754 to \$6,272. Between 1500 doses to 7,000 doses and 7500 doses to 17,000 doses, a flat trend can be observed. Beyond 17,000 doses it continues to decrease till it reaches \$5082 per dose at the greatest demand scale.

Figure 2 illustrates the average cost per dose over the objective of minimising time. Similar to figure 2, a decreasing trend is noticed throughout with severe drops at 1,500 doses, reaching roughly \$9,206 per dose and 17,000 doses, reaching around \$6,946 per dose.

4.2 Wait time and facility utilisation rate

Amongst all the scalar parameters in our models, wait time was the most sensitive variable. For such, an analysis on varying wait times in increments of 5 days: 7, 12, 17 and 22 days were conducted.

4.2.1 3,000 doses

Figure 4 shows the sensitivity analysis for minimising cost and minimising time against wait time. Minimising

cost graphs show a constant trend throughout all wait times, with a constant average cost per dose of \$7,387. For minimising time, the average cost per dose is \$9,206 for all wait times except at 408 hours, plunging to \$7,500. Figure 8 also depicts that the utilisation rate is symmetrical across both USA and EU regions, both utilising small-scale manufacturers for 33% and medium-scale manufacturers for 17% and none in large-scale manufacturing.

4.2.2 15,000 doses

Figure 5 shows that for the cost minimisation model, the average cost per dose plateaus at around \$7,387 for all wait times except for 720 hr, peaking at \$9,838 per dose. On the other hand, the time minimisation maintains a constant at \$7,387 throughout all wait times. Figure 9 also displays that the utilisation rate is symmetrical in both the USA and EU regions, with both regions employing small-scale manufacturers for 30% and medium-scale manufacturers for 20%, and no utilisation in large-scale manufacturing was concluded.

4.2.3 30,000 doses

Figure 6 shows that the costs per dose whilst minimising time remains constant throughout at \$5,943. Figure 10 shows an equal split of the utilisation rate of manufacturing sites in the USA and Europe. Both regions utilizes all three types of manufacturing sites,

with small scale manufacturing sites contributing the greatest proportion of utilisation rate.

4.2.4 475,000 doses

Figure 7 illustrates that both charts display a value of \$5,943 per dose throughout all wait times for both objectives. Hence no variations in costs were observed. 100% of the facilities utilised were large-scale manufacturing sites as seen in figure 11.

5. Discussion:

5.1 Model validation

The aim of our study was to investigate three key demand scales denoted by table 1. Taking one step further, we assess the validity of our model by simulating scenarios covering a range of demands from smaller to small scale to larger than commercial scale. Results displayed in figure 2 and figure 3 provide a clear representation that both models could be used to model different demand scales; while yielding promising results that follow the descending average cost per dose trend as demand increases. This suggests that the models are able to account for demand discrepancies within the ranges of 1,500-750,000 if necessary. This directly acknowledges and offers solutions to the challenge on demand uncertainties and production backlogs.

5.2 Average costs per dose of pDNA

Further to that, the observed descending trend in figure 2 and figure 3 tallies up with the concept of the economies of scales [35], which from microeconomic perspective it means it benefits from the cost advantage when it increases its level of output. The jumps observed at 1500 against 3000 doses and 17,000 against 375,000 doses are interestingly incoherent with the rest of the demand scales. Such discrepancies could be explained by our input parameter on batch sizes whereby. The sudden decrease in average cost per dose again benefited from the economics of scale as the total cost of production was then shared out to more doses being produced. Furthermore, although it is intuitive to assume the results for the minimising cost objective would produce pDNA at a lower cost per dose, it was proven wrong in figure 2 for the demands ranging from 3,000-7,000. This suggests that the minimising time model is better suited when evaluating in this given range as it minimises time while returning a better cost.

5.3 Sensitivity analysis

Amongst all the scalar parameters in our models, wait time was the most sensitive variable. For such, an analysis on varying wait times in increments of 5 days: 7, 12, 17 and 22 days were conducted.

From figures 5,6,7, it could be deduced that varying wait time has no effects on the average cost per dose for each given demand scale under the minimising time model. This is because, time, as the KPI of this model has already been minimised under the model's evaluation on the objective function equation 46. Hence, a sensitivity analysis on increasing wait time is rather redundant, but necessary to validate the model in the case of minimising time. In general, results for both

models did not display a lot of differences under the sensitivity analysis. This further signifies that our model could be thought to be too constrained hence the investigation to establish findings on average cost per dose against varying wait times were deemed rather unsuccessful.

However, the sudden spike in average cost per dose in figure 6 for the wait times of >7 days could be explained. It was observed that the total transportation time required was two-times greater than the other demand scales. To explain this, it is important to infer findings from the established manufacturing sites. It is indicated that the model is constrained by the production capacity and batch times at a small wait time of 168hrs. This is because yielding 30,000 doses could either be done by going through 11 batches in each small scale, 2 batches in each medium scale and 1 batch in each large scale. Under the circumstances of a maximum wait time of 168hrs, the first two scenarios were instantly deemed impossible by the model. Furthermore, the distances between the large-scale manufacturing sites to the airports were greater than that of the small and medium scale to the airports. For such, the model has compromised on benefiting from the economy of scale by overproducing pDNA to keep the costs low while having to take up greater costs on transportation. This is a prime example of a trade-off between time and costs.

5.4 Facility utilisation rates

Figures 8,9,10,11 highlighted a significant difference in utilisation rate from facilities to facilities. Small and medium scale facilities were most widely installed, especially for 200-2000 patients. Looking into the absolute utilisation percentages infers that it is the most cost and time efficient for medium scale manufacturing sites to be commissioned, as this would be able to cater for the pre-clinical/phase I/II clinical trial scales without overproducing too much and could undergo multiple batches to provide for the commercial scale if wait time is not a topic of interests for the given candidate supply chain. Interestingly, this investigation also illustrates that the Phase III clinical trial scale is the only demand scale that utilises a combination of manufacturing sites.

Results showcased that under either of the two scenarios 1) by increasing wait times >528 hr or 2) evaluating using the minimising time model would suggest that it is better to install multiple manufacturing sites as opposed to having multiple batches at one site. This grants the supply chain a greater flexibility on switching facilities and can produce a better reflection of the real-life supply chains whereby multiple manufacturing sites are established simultaneously at a given time. However, one should note that the aforementioned suggestions do not apply to the large demand scale. Multiple simulations were conducted on the large demand scale whilst varying its corresponding wait time. Yet it is indicative to claim that the only type of manufacturing sites to be commissioned ought to be the commercial scale ones. This finding is backed up by figure 11 whereby commercial scale manufacturing sites took up 100% of the facility utilisation rate. One should note that both models generated the same output for all

ranges of wait time, this strongly validates the reliability of the results.

5.5 Limitations

While the developed model exhibits considerable resilience, its effectiveness is constrained by certain limitations.

One limitation in the model is the implementation of a single manufacturing line, a simplification that deviates from the complexity inherent in real-life scenarios. In practice, manufacturing processes often entail the operation of multiple parallel lines, each catering to differences in demand scales. [36]. The model's oversimplification may compromise its ability to capture the complex dynamics of diverse manufacturing lines and the simultaneous production of different products.

Furthermore, the need for a feedback mechanism is a significant limitation in the current model. In the dynamic landscape of supply chains, unforeseen events are inevitable, ranging from production delays and transportation disruptions to sudden shifts in demand [37]. The absence of a feedback loop hinders the model's capacity to respond effectively to these unpredictable factors.

6. Conclusion:

To summarise, two models aim to minimise the KPIs – total costs and total time deployed and developed. These KPIs were selected based on the level of importance between variables on a relative basis. The models developed have incorporated the complexities in real-life supply chains while being evaluated under assumptions to simplify the structure.

Manufacturing pDNA at a larger demand scale is generally cheaper, with a lower average cost per dose, as it benefits from economy of scale. However, demand scales that coincide with the batch sizes at any manufacturing site could also benefit from a lower average manufacturing cost per dose. With the assumptions in our model and the facilities in the supply chain, our models can accurately depict the supply chain landscape for demand scales that are greater or smaller than what we intended to focus on. Both could account for demand discrepancies within the demand scales of 1,500 doses to 750,000 doses if necessary.

A sensitivity conducted on varying wait times has generated results that displayed few differences. It could then be inferred that our models might be too constrained; hence, the investigation to establish findings on average cost per dose against varying wait times could have been more successful.

Finally, our analysis of the utilisation rates of the manufacturing sites showed that the best-optimised supply chain would consider commissioning the medium-scale facility as it generates the lowest costs whilst minimising time. The batch size of this facility can cater to the pre-clinical/phase I/II clinical trial scales without overproduction. It could undergo multiple batches to provide for the commercial scale if wait time is not a topic of interest for the given candidate supply chain. Furthermore, multiple simulations were conducted on the commercial demand scale. Results

claimed that the only type of manufacturing sites to be commissioned ought to be the commercial scale ones in Texas, USA and Marburg, Germany.

Although our models were able to reflect on the supply chains out of the demand scales of interest in this study, it is also important to note that the assumptions and constraints have brought some overarching limitations that should be addressed in the future. Resolving such limitations is as urgent as it is vital within ATMPs supply chain optimisation. Solutions to the limitations will effectively be able to model the pDNA supply chain while accounting for uncertainties, which could then be deployed in the industry to tackle real-life challenges to reshape the pDNA supply chain landscape.

7. Outlook:

This study focuses on minimising the total costs and time associated with pDNA production as the primary objective function. However, it is imperative to acknowledge that other facets of the supply chain are needed for a comprehensive analysis.

By exploring the utilisation of parallel production lines in further studies, efficiency and flexibility issues can be tackled. Parallel production lines allow for simultaneous manufacturing of batches, increasing overall production capacity and meeting higher demand levels [38]. Furthermore, parallel production offers enhanced flexibility in responding to fluctuations in demand or product variations [38]. This could reduce costs while serving as a strategic tool for managing uncertainties.

The increased number of nodes within the regions offers several strategic advantages. Expanded geographical coverage enhances flexibility in addressing diverse demands across different regions, heightening the adaptability and responsiveness of the overall system. Moreover, it contributes to a reduction in long-distance transportation as the distribution network becomes more localised. This reduction mitigates logistical challenges and enhances operational efficiency by minimising time and resource wastage [39]. The incorporation of more nodes also allows for the development of models that better simulate real-life scenarios. This provides a more accurate representation of the practical supply chain operations, enhancing the robustness and accuracy of the model. Potential challenges include increased computational complexities and heightened demands on data management systems.

In future work, redefining the objective function into a multi-objective function is necessary. This involves concurrently minimising both time and cost, enabling more comprehensive optimisation. MOO allows for the simultaneous consideration of conflicting objectives within the optimisation framework. Adopting the Multi-Objective Optimisation (MOO) algorithm becomes pivotal in determining the shortest path and time between two nodes while using the least cost. Trade-off analysis is an essential aspect within the MOO; this could be evaluated by plotting a Pareto front to visualise the impacts on different desired outcomes. This enables decision-makers to assess the compromises

and synergies between conflicting objectives. By explicitly recognising trade-offs, the multi-objective function provides a structured framework for decision-making, offering insights into the complex interdependencies between time, cost, and distance.

Future work utilising a mixed-integer non-programming model is a strategic choice to model and optimise for unpredictable events such as a pandemic. Striking a balance between lean and agile supply chain models becomes vital. While lean models emphasise efficiency and cost reduction, agile models focus on flexibility and responsiveness to dynamic changes. An equilibrium between these models is crucial in navigating uncertainties, ensuring efficiency and adaptability during unpredictable disruptions. Leveraging predictive analytics and machine learning algorithms enhances the capacity to anticipate future demand patterns. This aids in devising resilient supply chain strategies that can dynamically respond to fluctuations in demand, contributing to a more adaptive supply chain.

8. Acknowledgements:

We would like to express our gratitude to Dr Maria Papathanasiou and Niki Triantafyllou for offering their expertise, insights, and guidance over the course of this project. We would also like to thank Dr Maria Papathanasiou's research group for their support.

References:

- [1] h.-t.-a.-p.-s.-d. (D. 2. 2. Extrachromosomal DNA - an overview | ScienceDirect Topics.
- [2] A. R. L. a. O. T. Ramirez, "Plasmid DNA production for therapeutic applications," *Recombinant Gene Expression*, 2011. doi:10.1007/978-1-61779-433-9_35".
- [3] F. Meng, Fast construction of SARS-COV-2 associated plasmid library using parallel cloning method.
- [4] "Plasmid DNA manufacturing market size to hit US\$2,156.58 MN by 2030, Precedence Research, <https://www.precedenceresearch.com/plasmid-dna-manufacturing-market> (accessed Dec. 20, 2023)."
- [5] D. I. al., "Optimal design and planning of supply chains for viral vectors and RNA vaccines," *Computer Aided Chemical Engineering*, pp. 1633–1638, 2022. doi:10.1016/b978-0-323-95879-0.50273-3".
- [6] M. Sarkis, "Decision support tools for next-generation vaccines and advanced therapy medicinal products: Present and future," doi:10.1016".
- [7] N. Triantafyllou, "A digital platform for the design of patient-centric supply chains," *Scientific Reports*, vol. 12, no. 1, 2022. doi:10.1038/s41598-022-21290-5".
- [8] Hosseini-Motlagh, "Robust and stable flexible blood supply chain network design under motivational initiatives," *Socio-Economic Planning Sciences*, 2020. doi:10.1016/j.seps.2019.07.001
- [9] J. Ohlson, "Plasmid manufacture is the bottleneck of the Genetic Medicine Revolution," 2020. doi:10.1016/j.drudis.2020.09.040".
- [10] R. J. Juba, A. Samawova, K. Seals, P. Kinney, and E. J. Brandreth, "2140. ino-4800, a DNA plasmid vaccine encoding the ancestral SARS-COV-2 spike protein, is stable over a range of temperatures not requiring Ultra-Cold chain storage," *Open Forum Infect*".
- [11] R. H. Durland and E. M. Eastman, "Manufacturing, and quality control of plasmid-based gene expression systems," *Advanced Drug Delivery Reviews*,. doi:10.1016/s0169-409x(97)00105-1".
- [12] F. D. Ledley,b" "Non-viral gene therapy," doi:10.1016/0958-1669(94)90085-x".
- [13] N. Tiwari, J. Beilowitz, C. Sampson, and D. Peterson,, "675. high quality plasmid DNA manufacturing for ex-vivo protein synthesis and viral vector production for gene therapy," *Molecular Therapy*, vol. 23, 2015. doi:10.1016/s1525-0016(16)34284-8".
- [14] "Clean Room Classifications & ISO Standards | American Cleanrooms Systems." American Cleanroom Systems.
- [15] J. Ohlson, "Plasmid manufacture is the bottleneck of the genetic medicine revolution", 2020.".
- [16] C. Pacini, S. Cencig, and A. Giordano,, " "Gene Editing/Gene Therapies: DESIGNING AN EFFICIENT PLASMID DNA DOWNSTREAM PURIFICATION PRODUCTION PROCESS", *Cytotherapy*, May 2023."
- [17] O. Ruiz, Miladys, Jorge, M. Pupo, and Eduardo, " "Scalable Technology to Produce Pharmaceutical Grade Plasmid DNA for Gene Therapy", in *Gene Therapy - Developments and Future Perspectives*. InTech, 2011.
- [18] R. H. Durland and E. M. Eastman,, " "Manufacturing and quality control of plasmid-based gene expression systems", *Adv. Drug Del. Rev.*, vol. 30, no. 1-3, pp. 33–48, Mar. 1998.
- [19] "https://www.labxchange.org/library/items/lb:LabXchange:a03c81b4.html:1," [Online].
- [20] "https://www.bmglabtech.com/en/blog/what-are-endotoxins/#:~:text=The%20presence%20of%20endotoxins%20in,respond%20predominantly%20to%20Lipid%20A.)," [Online].
- [21] "https://www.qiagen.com/us/knowledge-and-support/knowledge-hub/bench-guide/plasmid/working-with-plasmids/endotoxins-and-what-to-consider," [Online].
- [22] " (https://www.biotage.com/blog/how-to-check-the-quality-of-plasmid-dna," [Online].
- [23] M. Murakami, " "Evaluation of DNA Plasmid Storage Conditions", 2013.Available: <https://doi.org/10.2174/1874070701307010010>".
- [24] H. N. e. al, " "Long-Term Stability and Integrity of Plasmid-Based DNA Data Storage", *Polymers*, vol. 10, no. 1, p. 28, Jan. 2018. Accessed: Dec. 21, 2023. [Online]
- [25] P. Colombi, "https://blog.addgene.org/ways-to-elute-and-store-plasmid-dna#:~:text=TE%20(10%20mM%20Tris%20DHCl,preventing%20the%20activity%20of%20DNase,"
- [26] N. A. M. Bakker et al., , " "Small-scale GMP production of plasmid DNA using a simplified and fully disposable production method", *J. Biotechnol.*: X, vol. 2, p. 100007, Jun. 2019. Accessed: Dec. 21, 2023.
- [27] D. Ibrahim et al., , " "Model-Based Planning and Delivery of Mass Vaccination Campaigns against Infectious Disease: Application to the COVID-19 Pandemic in the UK", *Vaccines*, vol. 9, no. 12, p. 1460, Dec. 2021.
- [28] VGXI, "How Much Plasmid DNA is Needed for Your Clinical Trial?"
- [29] N. Triantafyllou, M. Sarkis, A. Krassakopoulou, N. Shah, M. M. Papathanasiou, and C. Kontoravdi,, " "Uncertainty quantification for gene delivery methods: A roadmap for pDNA manufacturing from phase I clinical trials to commercialisation", *Biotechnol. J.*, Oc".
- [30] "https://www.bookairfreight.com/air-freight-calculator#:~:text=Prior%20to%20the%20pandemic%2C%20air,to%20US%24%2020%2Fkg!)," [Online].
- [31] "https://www.turkishcargo.com/en/about-us/about-turkish-cargo/fleet/cargo-aircrafts," [Online].
- [32] IPCC, "https://archive.ipcc.ch/ipccreports/sres/aviation/index.php?idp=92#:~:text=At%20the%20start%20of%20the,Transport%20aircraft%20cruise%20speed%20progress.,," [Online].
- [33] UK Department for Transport, "Road congestion and travel time statistics."
- [34] D. N. E. K. a. R. V. B. Ivanov, " "A MILP approach of optimal design of a sustainable combined dairy and biodiesel supply chain using dairy waste scum generated from dairy production", *Comput. & Chem. Eng.*, p. 107976, Sep. 2022. Accessee".
- [35] F. Badorf, S. M. Wagner, K. Hoberg, and F. Papier, "How Supplier Economies of Scale Drive Supplier Selection Decisions", *J. Supply Chain Manage.*, vol. 55, no. 3, pp. 45–67, Apr. 2019. Accessed: Dec. 21, 2023. [Online]. Available: <https://doi.org/10.1111/j>".
- [36] H. Moradlou, A. Boffelli, D. E. Mwesumio, A. Benstead, S. Roscoe, and S. Khayyam,, " "Building Parallel Supply Chains: How the Manufacturing Location Decision Influences Supply Chain Ambidexterity", *Brit. J. Manage.*, Aug. 2023.
- [37] Tan Miller and Matthew J. Liberatore, "What Are Logistics Feedback Loops, and Why Should You Care about Them?"
- [38] T. Freiheit, Y. Koren, and S. J. Hu,, " "Productivity of Parallel Production Lines With Unreliable Machines and Material Handling", *IEEE Trans. Automat. Sci. Eng.*, vol. 1, no. 1, pp. 98–103, Jul. 2004. Accessed: Dec. 21, 2023. [Online]. Available: <https://doi.org/10.1109/9.1111111>".
- [39] "TVS Supply Chain Solutions, Rising Supply Chains Complexity & How to Manage It".

Analysis of Morphology and Microstructure of the Lignin-derived Mesoporous Anode for Sodium-ion Batteries and Sodium Storage Mechanism

Yijun Yang and Qiyun Chen

Department of Chemical Engineering, Imperial College London, U.K.

Abstract

Sodium-ion batteries (NIBs) showed its capability as a cheaper substitution of Lithium-ion batteries (LIBs), as NIBs were material wise cheaper and could reach acceptable capacities. In this research, carbon materials derived from lignin and phloroglucinol polymer carbonized at different temperatures were templated using various concentrations of Pluronic F127 as soft template. Two different pore sizes of 1.68 and 3.30 nm were successfully templated. X-ray diffraction, Raman spectroscopy, nitrogen physisorption, transmission electron microscopy were used to characterize the material from microstructure to pores. The battery performances were tested for the carbon materials, and sodium cation capacity of large micropores and small mesopores within material of controlled microstructure were qualitatively studied.

Keyword: Sodium-ion Battery, Sodium-ion Pore-filling, Soft Templating, Lignin Derived Carbon Material

1. Introduction

Achieving carbon neutrality requires the development of storage technologies. Lithium-ion batteries (LIBs) are known as the most commercially successful battery. Considering the limited resource, uneven global distribution, and the energy density had not been promoted by 3% in the last 25 years^[8], sodium-ion batteries (NIBs) with very similar physicochemical properties to lithium had shown a potential alternative because of sustainability and low cost. Instead of copper being used as a current collector for LIBs, aluminum as the current collector in the NIBs offered an advantage in cost and weight for large-scale applications. Whilst lithium battery anode graphite material was not applicable for NIBs. Turbostratic carbon material (soft carbon and hard carbon) has been studied as NIB anode materials. The current shortcoming of the NIBs anode carbon materials are concentrated to: 1) lower specific capacity 2) lower cycling performance 3) lower charge and discharge rate. Thus, understanding the sodium ion storage behavior plays a significant role in optimizing the storage performance.

In this work, Galvanostatic performance of carbon anodes made of a series of lignin-derived closed porous carbon materials were studied. TEM, nitrogen physisorption and SAXS were conducted to check porosities and pore sizes, while the microstructure was characterized by XRD and Raman spectroscopy. Qualitative relation between pore sizes and sodium ion capacity were studied at the transition from micropore to mesopore, giving the idea that the small mesopores could store more sodium ions than big micropores.

2. Experimental Method

2.1 Preparation of hard carbon materials

The Pluronics F127 (Sigma Life Science) templated hard carbon samples were carbonized from precursors with different template concentrations (baseline, 50% baseline, 35% baseline, 25% baseline, and 12.5% baseline). These template concentrations within the precursor were obtained by varying the amount of lignin (Fraunhofer MODUL II, KO92) and phloroglucinol (Sigma Aldrich Phloroglucinol > 99.0%) added to the solution which had constant template concentration. For the Pluronics F127 templated baseline precursor, first, 1.125 g of Pluronics F127 was dispersed into 32.5 ml of acetone (VWR Chemicals) by overnight stirring with the beaker carefully sealed. Then, a total weight of 1.3 g of 1:1 ratioed organosolv lignin and phloroglucinol were added to the solution, one after another had been fully dissolved. The solution was then stirred overnight again before the addition of 1.25 ml glyoxal (Sigma Aldrich 40 wt% in H₂O) as a cross-linker. After the addition of the cross-linker, the solution was stirred for another 5 minutes, then the solution was placed into a fume hood to have most of the solvent evaporated, the evaporation process took 2-4 days. After evaporation induced self-assembly was done, the precursor was placed overnight into an oven set to 85 °C for cross-linking. The carbonization of the resulting polymer precursor took place in a furnace (Carbolite STF 16/450 220-240V 1PH) under nitrogen atmosphere, with the nitrogen flowrate of 500 ml/min. The furnace heated up to the template removal temperature of 350 °C at 1 °C/min heating rate, and dwelled for 1 hour for template removal, followed by the furnace heated up to carbonization temperature (1000 - 1300 °C) at 5 °C/min, and dwelled for 2 hours for carbonization. Finally, after natural cooling down, the hard carbon material was yielded.

2.2 Preparation of carbon anode

The carbon samples were ground with isopropanol by a mortar and a pestle for 30 min to reduce the particle size. The resulting powder suspension in isopropanol was collected and dried overnight in an oven set to 85 °C until isopropanol was fully evaporated. The dried carbon powder was weighed and then mixed with a binder of 5 wt.% carboxymethyl cellulose in a mass ratio of 1:2.22 to prepare a slurry in which the active material took 90% of carbon and CMC in total. The slurry was further ground using the mortar and pestle for 30 min. The slurry was coated onto an aluminum foil with a total thickness of 200 micrometers. The coated foil was dried and cut into 1 x 1 cm² square anode pieces, finally, the anodes were dried overnight in a vacuum oven before assembly into semi-coin cells.

2.3 Electrochemical measurements.

Semi-coin cells were assembled in an argon-filled glove box, with oxygen and moisture level < 0.5 ppm. Glass fiber separator between the anode and sodium metal was in 16 mm diameter, grade A. 100 microliter of 1 M NaPF₆ in EC: DMC = 1:1 electrolyte was used (KLD-NF04) per battery assembled. The battery cup, bottom, spring, and spacer were dried in an 85 °C vacuum oven overnight after ultrasonication in isopropanol for 30 min. Galvanostatic discharge/charge profile was measured on a NEWARE battery cycler tested at a constant temperature with a potential window from 0.005 - 2.5 V for all semi-coin cells. The first two cycles were run at 0.033 C, followed by 0.1 C, 0.2 C, 0.5 C, 1 C, 5 C, and finally, 0.1 C again, running five cycles for each C-rate.

2.4 Characterization of Carbon Material

Nitrogen physisorption was done on Micromeritics Tristar II Plus, with the sample ex-situ degassed at 250 °C, 16 hours, on a Micromeritics Smart VacPrep. Raman spectroscopy was carried out on a Bruker Senterra II, 20x zoom was used, with 50 x 1000

micrometer aperture and 1.5 cm⁻¹ resolution. The power, scanning time, and coaddition were varied accordingly to produce an acceptable signal-to-noise ratio. X-ray diffraction was done on a PANalytical Aeris, the scanning axis was set to 2 θ , the scanning angle was from 4.999 to 114.9674

degrees, step angle was 0.086932 degrees with a time per step of 103.53. Cu K alpha-1 and Cu K alpha-2 radiation mixed in 1:1, 40 V generator and 15A tube current were used. Electrodes weight was measured on an AND BM-20 balance.

3. Result and Discussion

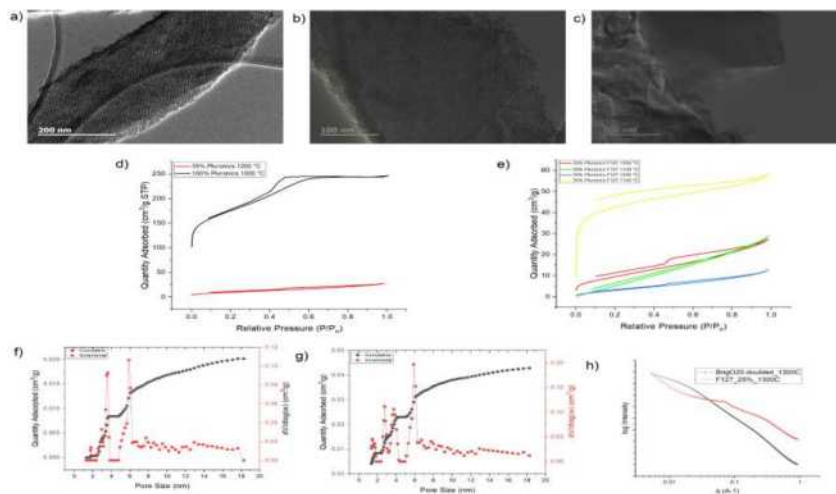


Figure 1. TEM images display Pluronic F127 templated carbon with a) 50%; b) 35%; c) 25%; nitrogen physisorption of 35% and 100% Pluronic F127 templated carbon d) and e); NLDFT analysis of Pluronic F127 templated carbon 50% - 1200 °C f) and 35% - 1200°C g); h) SAXS of 25% Pluronic F127 templated and BrijO20 double at 1300°C

3.1 Pore Size and Specific Surface Area

It was experimented by Flatthaar et al. that PEO₄₂₈-b-PHA₂₆₅ block co-polymer under 1.731:1 weight ratio with carbon precursor templated closed mesopores (37±6 nm) in lignin phloroglucinol derived carbon material, however Pluronic F127 in replacement of PEO₄₂₈-b-PHA₂₆₅ templated open mesopore channels (5±1nm) in the same material^[3]. Because of identical amount of block co-polymer was used, template removal theoretically both resulted in open channels, but in the case of PEO₄₂₈-b-PHA₂₆₅, the channels from template removal were closed during carbonization, indicating that template removal and closed pore formation were decoupled. To examine the performance of carbon anode with small closed mesopores, Pluronic F127 was used in decreased concentrations from the 1.731:1 ratioed baseline

set. 50%, 35%, 25%, and 12.5% of the baseline corresponded to block co-polymer to carbon precursor weight ratios of 1.731:1, 1.731:2, 1.731:2.86, and 1.731:4 respectively. TEM was used to identify the existence and sizes of pores, while BET analysis from nitrogen physisorption was used to identify accessibility of pores.

The samples for nitrogen physisorption were grinded, making more mesopores exposed for adsorption and desorption, thus generating hysteresis isotherms (figure 1d, e). By applying NLDFT analysis to the sample data which had hysteresis behavior, more pore size data could be obtained for cross-checking with the TEM images. SAXS was carried out on some samples for qualitative pore size estimation applying Bragg's law, which also gave ideas on pore size distribution.

35% and 50% Pluronics templated carbon had showed ordered pores from their respective TEM images [figure 1a, b], while 25% Pluronics F127 templated carbon existed non-templated pieces under TEM [figure 1c], indicating the soft templating might have failed to reach pore saturation in a low concentration 25% Pluronics F127 templates carbon.

Pore diameters were measured from the TEM images in Origin Lab, with the diameter averaged between 30 pores. It was measured that the 50% Pluronics F127 templated carbon had 3.3 ± 0.181 nm pores and 35% Pluronics F127 templated carbon had 1.68 ± 0.067 nm pores, in this research, the pore sizes were taken as measured. , though the pore size variation with template to precursor ratio would better be more carefully examined. In previous research from Libbrecht et al., the pore sizes as well as symmetry were found to varied with template to precursor ratios at low template concentrations and high amount of furfural to resorcinol ratio.

As the pores on a TEM image were organized and aligned in a row, the number of pores per nm row were estimated. There were 0.15 pore/nm row for 3.3 nm pore and 0.28 pore/ nm row for 1.68 nm pore. Considering the 3D geometry of the pores, these numbers were not representative on pore density in a unit volume of carbon material, but it could indicate that the number of 1.68 nm pores were at least not less than the number of 3.30 nm pores in the carbon material.

Following the pore sizes, closed porosity was examined with nitrogen physisorption by comparing the nitrogen volume adsorbed of one of the low concentrations (35%, 1200 °C) Pluronics F127 templated carbons with the baseline Pluronics F127 templated carbon carbonized at 1000 °C with known open porosity

[figure 1d], and comparison between 35% and 50% Pluronics F127 templated carbons at 1100 and 1200 °C in parallel [figure 1e]. The low nitrogen volume adsorbed by the low concentration Pluronics F127 templated carbons could indicate closed porosity, as the pores were visualized under TEM but did not result in the same nitrogen adsorption volume as the open porous baseline Pluronics F127 templated carbon.

Then, nitrogen physisorption surface areas of the low concentration Pluronics F127 templated carbons were given from BET analysis [table 1]. Generally, the surface area decreased with higher carbonization temperature, however, the 35% Pluronics F127 templated carbon carbonized at 1200 °C had less surface area compared to both neighboring carbonization temperatures. The measurements on 35% Pluronics F127 templated carbon were done twice, outputting precise results. It was noticed that hysteresis adsorption desorption curve could be seen from 35% and 50% Pluronics F127 templated carbons at 1200 °C, indicating that nitrogen probed the mesopores exposed at the surface. Hence, non-local density distribution functional theory was applied to the two samples for pore size distributions from nitrogen physisorption (figure 1f, 1g). The 50% Pluronics F127 templated carbon had a distribution around 3.5 nm, agreeing with the pore size averaged from the TEM image. For the 35% Pluronics F127 carbon, it was not expected to have micropores of 1.68 nm probed by nitrogen, as nitrogen was not capable accessing pores under 1.9 nm. The mesopores detected was from a distribution at 6.5 nm, which was also detected in the 50% Pluronics F127 sample, the adsorbed volume dropped from approximately 0.2 to 0.1 from 35% to 50% Pluronics F127 templated carbon at 1200 °C, indicating less mesopore of 6.5

	35%	35%	35%	35%	50%	50%
	1000°C	1100°C	1200°C	1300°C	1100°C	1200°C
Pore Size (nm)	1.68	1.68	1.68	1.68	3.30	3.30
BET Surface Area (m ² /g)	499.5	159.3	35.9	106.5	26.7	14.1
Interlayer Spacing (Å)	3.8	3.7	3.6	3.4	3.7	3.7
R ratio	1.98	2.02	2.13	5.37	2.10	2.14
C-axis Crystallite Length (Å)	9.86	9.84	9.96	58.50	10.23	10.49
A-axis Crystallite Length (Å)	14.10	15.45	15.76	>20	15.60	15.86

Table 1 Comparison of characterized data of different carbon samples

nm was detected from the surface. The non-monolithic distribution on the pore size from the NLDFT indicated SAXS or WANS would be necessary for detecting all the pores at the interior of the carbon materials.

Unfortunately, small-angle X-ray diffraction was only available at the early stage of the research, only the 25% Pluronics F127 templated carbon with unevenly distributed porosity was measured, bringing more uncertainty to the pore size distribution [figure 1h]. The peak position from SAXS of 25% Pluronics F127 templated carbon was at scattering vector of 0.08 to 0.09, by applying Bragg's law, a pore size of 6.98 to 7.85 nm could be calculated, agreeing with the NLDFT result regarding the bigger mesopore size.

In this stage, TEM image was used as a standard for pore sizes though the pore distribution was not monolithic, the number of bigger mesopores of approximately 0.65 nm decreased with increasing template concentrations could be observed from NLDFT. The uncertainty brought by the non-monolithic pore distribution needed to be minded in further analysis.

3.2 Carbon Microstructure

Performance of NIB carbon anodes is known to be sensitive to carbon microstructures^[2,7]. Hence, the sodium ion pore filling mechanism needs to be studied from comparison between carbon materials which are structurally similar. For characterizations of carbon material microstructure, XRD analysis applying Bragg's law (a_3), Scherrer equation (L_c), and R ratio^[5] defined by Ni et al. was carried out. Flatthaar mentioned that Bragg's law could not calculate data in good accuracy, it was still applied in this research as a scale for comparisons because WANS or XRD based geometry prediction algorithm was not available^[3,6]. Following the XRD, Raman spectroscopy was performed for graphene layer extent (a-axis crystallite length) calculation^[1,7] and cross checking with XRD data

for data validity.

Introduction of copolymer template into carbon precursor brought disorder to the microstructure, impacting the degree of graphitization^[3]. Flatthaar et al. experimented that only 1300 °C carbonized non-templated lignin phloroglucinol precursor showed a high degree of graphitization in the XRD data. The data showed more graphitized features such as clear separation of the (004) peak from (10) peak and narrowed and sharpened (002) peak^[3]. Similar XRD line shape could be observed in the 1300 °C carbonized 35% Pluronics F127 templated sample (Figure 2a), indicating it was easier for precursor with less template addition to reach higher structural order, thus the microstructure needed to be carefully studied by multiple characterizations methods for parameter estimations.

For XRD, the (002) peaks participated in multiple calculations (figure 2a, b). The background of the (002) peak was identified applying the method used for R ratio^[5] calculation [table 1]. The same background was subtracted in full-width half-peak maxima calculations and peak positioning^[7], which were further applied in Scherrer's equation for crystallite c-axis length calculations [table 1]. Within the XRD data of the 35% Pluronics F127 set, a red shift of (002) peak position could be identified from 1000 °C to 1300 °C carbonization temperature (figure 2a), indicating decreasing interlayer spacing a_3 , which was a sign of increasing degree of structural order. Then, R ratio was calculated for understanding the single layer fractions. The R ratio is defined as (002) peak height compared to background level, and higher R ratio was a sign of increasing structural order^[5], thus the less the R ratio, the more the single layer fraction. Within the samples, the 1300 °C carbonized 35% Pluronics F127 sample had the highest R ratio, which was more than doubled the rest, indicating a transition toward more graphitic features occurred between 1200 °C and 1300 °C for 35% Pluronics F127

samples. This transition was not favorable for NIBs anode material as it meant the interlayer spacing would likely be too small and the pathways were fewer for sodium ion diffusion^[2]. The R ratios of 1000 and 1100 °C carbonized 35% Pluronics F127 sample had slightly more single layer fraction highlighting the more disordered structure of the two samples. The 1100 °C samples of different template concentrations were having similar interlayer spacings [table 1].

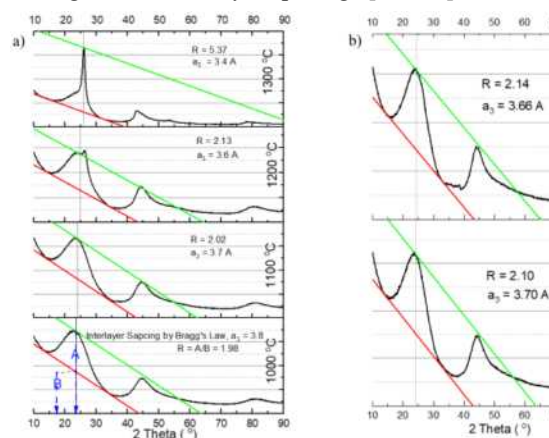


Figure 2. XRD data of 35% (a) and 50% (b) Pluronics F127 templated carbon.

Following the R ratio, c-axis crystallite length was calculated from (002) peak parameters^[7] [table 1]. L_c was used for qualitatively understanding^[3] the vertical stacking order of the graphene layers due to only (002) peak alone had participated in the calculation, and the effect of other reflections were not eliminated by band separation, but it could still show a bigger picture of the L_c . 35% Pluronics F127 sample carbonized at 1300 °C had L_c more than 5 times of the rest of the samples, highlighting its much stronger graphitic feature. As the variation of the L_c was small for the rest of the samples [table 1], the 1100 °C and 1200 °C carbonized 35% and 50% Pluronics F127 samples and the 1000 °C carbonized 35% Pluronics F127 samples could be considered qualitatively similar.

Unlike the L_c , the a-axis crystallite length L_a ^[7] could be approximated in a much higher accuracy from Raman spectroscopy spectral bands fitting. In Raman spectroscopy of the samples, the first

order band occurred within 1000-1750 cm^{-1} (figure 3 a b) and the second order band occurred within 2000-3500 cm^{-1} (figure 3 c d), the samples were analyzed with an excitation wavelength of 532 nm. The measured first order band was broken down into the well-known D1, D2, D3, D4, and G bands to carbonaceous material^[7], which were located at 1350, 1620, 1500, 1200, and 1580 cm^{-1} respectively using Lorentzian line shape fitting (figure 3 a b). The major peaks in the spectral breaking down was D1 and G band, correlating to A_{1g} and E_{2g} vibration modes respectively^[1]. According to Sadezky et. al^[1], the D1 band was assigned to defection from edges of graphene layers, and G band was assigned to graphitic lattice vibrations^[1]. It was also mentioned that the D2 band appeared at the shoulder of G band could be caused by graphene layer intercalations, D3 band was caused by amorphous carbon, and D4 band could arise from $\text{sp}^2\text{-sp}^3$ bond stretching vibration.

35% Pluronics F127 templated carbon showed increasing trend in I_D/I_G from 1000 °C to 1200 °C which was the feature of turbostratic carbon material^[3]. However, there was a rapid drop in I_D/I_G from 1200 °C to 1300 °C that the height of the two peaks lied evenly [figure3 a], this transition aligned well with the suddenly increased L_c and R ratio from XRD analysis [table 1]. Besides, correspondence between L_a and R ratio was reasonable because with increasing a-axis crystallite lengths, the graphene layers were more ordered causing lower intercalation occurrences^[5]. The estimated parameters were then confirmed from Raman second order band shape.

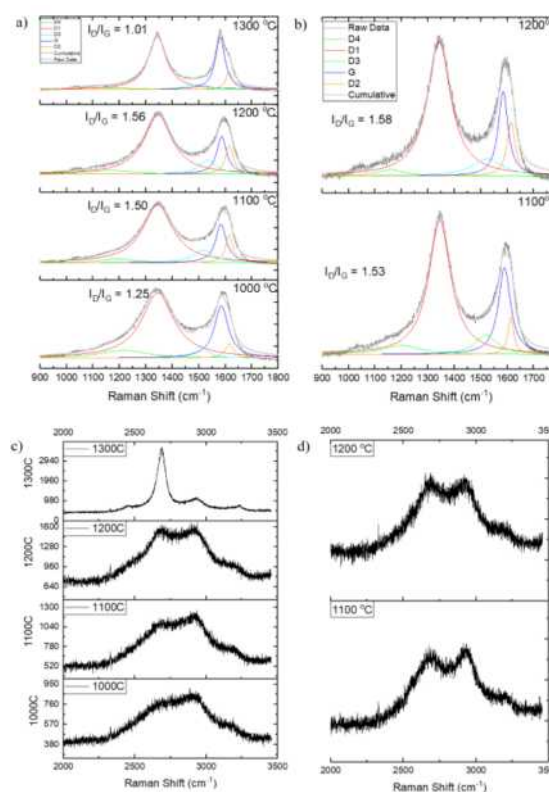


Figure 3. Raman spectra of 50% Pluronics F127 templated carbon a) first-order band; b) second-order band; 35% Pluronics F127 templated carbon c) first-order band; d) second-order band.

The second order band from 2000 to 3500 cm^{-1} had its significance in highlighting structural order of the carbon materials^[3]. The major information from second order D1 band ($2 \times \text{D1}$) was extracted from observing the 2700 cm^{-1} peak height in comparison to the 2900 cm^{-1} peak. The higher the 2700 cm^{-1} compared the 2900 cm^{-1} peak, the higher the $2 \times \text{D1}$ band, indicating higher structural order. Among the second order bands of the samples [figure 3 c d], structural order increased with increasing carbonization temperature. By comparing the two general peaks in the second order bands, the L_a calculated from the first order bands and R ratio calculated from XRD analysis were matched with the trend. The 1200 °C carbonized Pluronics F127 samples had similar parameters [Table 1] and second order band shapes, while the 1100 °C showed differences in both Raman second order bands and some of the parameters.

In conclusion of the microstructure analysis, for the microstructure to be controlled, it would be the best to compare in between 35% and 50% Pluronics F127 templated carbons at 1200 °C.

3.3 Galvanostatic Performance

The battery cycling tests started from a low C-rate of 0.033, because increasing capacities by cycles were observed in earlier 0.1 C discharge and charge cycles. Chantal et. al also had a similar issue of slight increment in capacities by cycles in the voltage vs. specific capacity plots^[3]. To tackle this problem, 0.033 C was used for the first two cycles (figure 4a). Although there were still increasement in capacities, the performance was much better than the 0.1 C cycle.

In half cells assembled from the anode coated with 35% Pluronics F127 templated carbon at 1000 to 1300 °C, and 50% Pluronic F127 templated carbon at 1100 and 1200 °C. The reversible capacity of 70-293 mAh/g was achieved from 9.9 mA Galvanostatic cycles (figure 4a). 50% Pluronics F127 templated carbon at 1200 °C had top performance of 293 mAh/g, and 35% Pluronics F127 templated carbon at 1100 °C had second best performance of 250 mAh/g (figure 4b). 35% Pluronics F127 templated carbon at 1000 °C only had 25 mAh/g plateau capacity. Initial coulombic efficiencies were the highest for 50% Pluronics F127 carbonized at 1200 °C, and then the 35% F127 carbonized at 1100 °C, specific values were recorded in table [2].

3.4 Sodium Storage Mechanism

Pore filling of sodium ion into pores was previously studied by Heather et. al^[2], claiming that pore filling occurred at low voltage region (< 0.1 V). Sodium storage in hydrothermal carbon derived from various carbonization temperatures were carefully studied, major finding was that bigger pores might result in better sodium ion capacity but optimizing the microstructure for sodium ion diffusion needed to be done

simultaneously for improving performance. Pore sizes from 1-5 nm were achieved from increasing the carbonization temperatures, however, because of increasing temperature, fewer diffusion pathways and interlayer spacing were resulted together with larger pore diameter^[2]. In this research, different pore sizes were achieved with similar microstructures, which was discussed in the microstructure section. Specific capacity and plateau capacity vs. slope capacity was concluded from the second 0.033 C cycle for all samples[Table 2], with plateau region taken from below 0.1 V.

Sodiation capacity would be compared

less 50% Pluronics F127 templated carbon with 3.30 nm pores. The plateau region capacity of 50% Pluronics F127 templated carbon was 1.4 times that of 35% Pluronics F127 templated carbon, indicating the 3.3 nm pores had better capacity than 1.68 nm pores, taking into consideration the higher number of 6.5 nm pores within the 35% Pluronics F127 templated carbon could overestimate the capacity for 1.68 nm pores. This comparison might indicate that in similar microstructure, 3.3 nm mesopores stored greater amount of sodium ions than 1.68 nm micropores, with the influence of extra 6.5 nm pores, the finding needed to be further examined.

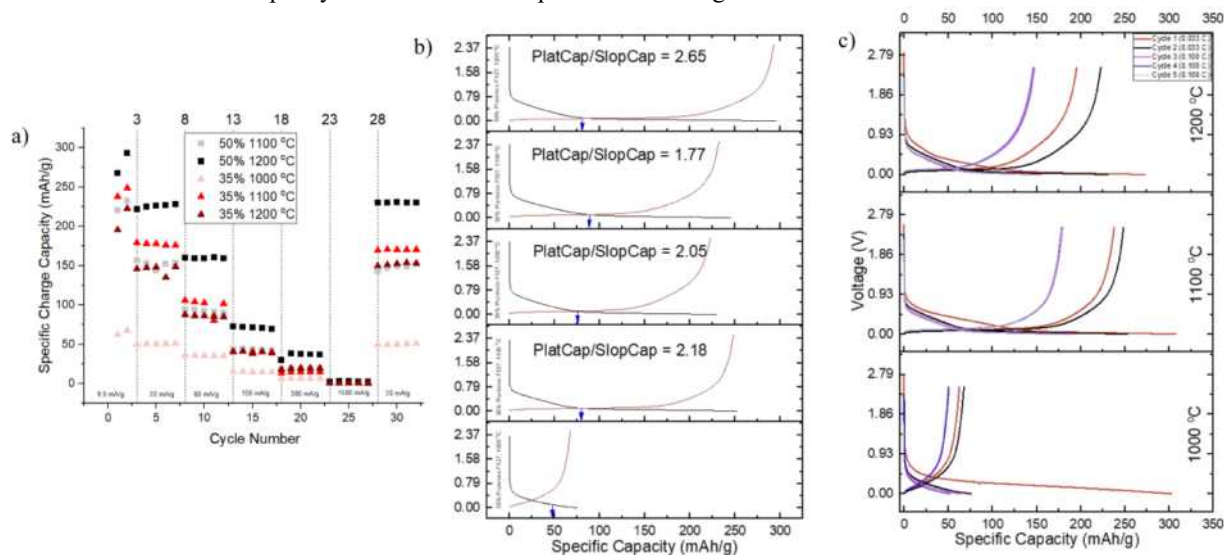


Figure 4. a) Specific Capacity of the carbon anode at different current density; b) Galvanostatic discharge/charge profile of samples from 1000 to 1200 °C in the second cycle; c) Galvanostatic discharge/charge profile of the 1st cycle to 5th cycle from 1000 to 1200 °C

between pore diameters of 1.68 ± 0.067 nm and 3.30 ± 0.181 nm, from 1200 °C carbonized 35% and 50% Pluronics F127 samples. The microstructure analysis from available characterizations indicated the two carbon materials were structurally similar. In the pore morphology section, it was known that the number of 1.68 nm pores would not be less than 3.3 nm pores, and the number of 6.5 nm pores was

The 1100 °C 35% Pluronics F127 templated carbon showed promising capacity of 250 mAh/g, but it could not be compared with Pluronics F127 templated carbon carbonized at the same temperature because they were not structurally similar. To utilize this data, more characterizations would be needed.

	35%	35%	35%	35%	50%	50%
	1000°C	1100°C	1200°C	1300°C	1100°C	1200°C
Specific Capacity (mAh/g)	67.6	248.6	222.7	N/A	232.6	293.1
Plateau Capacity / Sloping Capacity	N/A	3.52	3.16	N/A	2.77	3.95

Table 2. Specific Capacity of different carbon samples

Conclusion and Outlook

Carbon materials templated with Pluronics F127 generated different templated pore sizes. Different carbonization temperatures were used to carbonize the precursors. Carbon materials of similar microstructures were compared by galvanostatic cycling performance. Although there was the existence of an extra pore size, it could still give an idea of sodium ion pore filling preferred mesopores. More characterizations on the materials were suggested to provide better pore morphology information because the characterizations were lack of consistency in that section.

Acknowledgements

We would like to appreciate Mengnan Wang for providing trainings on different characterization methods, providing TEM, SAXS results, and supervision during the whole project. We also appreciate Zhenyu Guo on suggestions for Raman Spectroscopy data interpretation, Kaitian Zheng and Yichen Huang on training anode coating as well as battery assembly.

Reference

- 1 06/01706 Raman microspectroscopy of soot and related carbonaceous materials: Spectral Analysis and structural information. (2006). *Fuel and Energy Abstracts*, 47(4), 261. [https://doi.org/10.1016/s0140-6701\(06\)81712-1](https://doi.org/10.1016/s0140-6701(06)81712-1)
- 2 Au, H., Alptekin, H., Jensen, A. C., Olsson, E., O'Keefe, C. A., Smith, T., Crespo-Ribadeneyra, M., Headen, T. F., Grey, C. P., Cai, Q., Drew, A. J., & Titirici, M.-M. (2021). Correction: A revised mechanistic model for sodium insertion in hard carbons. *Energy & Environmental Science*, 14(5), 3216–3216. <https://doi.org/10.1039/d1ee90018h>
- 3 Flatthaar, C., & Wang, M. (n.d.). Lignin-derived mesoporous carbon for sodium-ion batteries: Block Copolymer Soft Templating and carbon microstructure analysis. *Chemistry of Materials*. <https://doi.org/10.1021/acs.chemmater.3c01520.s001>
- 4 Libbrecht, W., Verberckmoes, A., Thybaut, J. W., Van Der Voort, P., & De Clercq, J. (2017). Soft templated mesoporous carbons: Tuning the porosity for the adsorption of large organic pollutants. *Carbon*, 116, 528–546. <https://doi.org/10.1016/j.carbon.2017.02.016>
- 5 Ni, J., Huang, Y., & Gao, L. (2013). A high-performance hard carbon for Li-ion batteries and supercapacitors application. *Journal of Power Sources*, 223, 306–311. <https://doi.org/10.1016/j.jpowsour.2012.09.047>
- 6 Osswald, O., & Smarsly, B. M. (2022). OctCarb—a GNU octave script for the analysis and evaluation of wide-angle scattering data of non-graphitic carbons. *C*, 8(4), 78. <https://doi.org/10.3390/c8040078>
- 7 Xu, Z., Guo, Z., & Titirici, M.-M. (2020). Investigating the Superior Performance of Hard Carbon Anodes in Sodium-Ion Compared With. *Progress in Energy*, 2(4), 042002. <https://doi.org/10.1088/2516-1083/aba5f5>
- 8 Li, H. (2019). Practical evaluation of Li-Ion Batteries. *Joule*, 3(4), 911–914. <https://doi.org/10.1016/j.joule.2019.03.028>

Utilising Graph Neural Networks in the Glycomic Analysis of N-Glycan Biomarkers for the Diagnosis of Colorectal Cancer

Khan Kanjanabult and Serene Boonnasitha

Department of Chemical Engineering, Imperial College London, U.K.

Abstract: Colorectal cancer (CRC) is the third most common cancer worldwide. However, current screening techniques often employ invasive methods. The existing literature supports the use of glycomic biomarker analysis as a minimally invasive alternative to these methods. As alterations in the glycosylation patterns of N-glycans in immunoglobulin G (IgG) have been associated with CRC progression, they have been proposed as the biomarkers to serve this purpose. This study aims to introduce the novel methodology of employing a Graph Neural Network (GNN) model to glycomic data for binary classification to aid in CRC diagnosis, overcoming a limitation of prior studies whereby the biochemical relationships between glycans, in the form of the glycosylation reaction network, were not captured in the model. This proposed model yielded an ROC-AUC (Receiver Operating Characteristic - Area Under the Curve) of 0.604, slightly outperforming the scores of benchmark comparison algorithms: Random Forest (0.583) and Linear Regression (0.593). However, the resulting AUC is still lower than those obtained in previous studies. Overall, this study establishes a promising foundation for the integration of GNNs into glycomic analysis, although further improvement is required to solidify GNNs as the primary algorithm of choice.

Keywords: Colorectal Cancer, Biomarker, IgG N-Glycan, N-Linked Glycosylation, Machine Learning, Graph Neural Network, Graph Convolutional Network

Introduction

Colorectal Cancer and its Detection

In 2020, an estimated 1.9 million cases of CRC and over 930,000 deaths caused by CRC, occurred worldwide. This places CRC 2nd to lung cancer as the leading cause of cancer-related deaths. CRC also accounts for 10% of all cancer cases, placing it as the 3rd most common type of cancer. In comparison to these figures, by 2024, the number of annual new CRC cases is projected to increase 63%, up to 3.2 million, along with a 73% increase in the number of deaths caused by CRC, up to 1.6 million.^[1] Since visible symptoms may not be apparent in the early stages of CRC, it is often diagnosed at progressed stages when treatment options become limited and survival rates become diminished. Thus, it is crucial for the screening of CRC to be carried out in a timely manner, particularly if CRC exists in the individual's family medical history.^[2] The main drawback of currently employed screening techniques is their invasive nature, ranging from a digital rectal exam where a physical examination is carried out by a doctor to detect any abnormal rectum mass, to a biopsy where a small tissue sample is removed during a colonoscopy and studied in a lab. It is therefore of interest to investigate techniques which are less invasive and of lower costs through the utilisation of biomarker analysis, such as those obtained from blood samples.^[3] Glycobiology, a subset of molecular biology focusing on the study of glycans, has shown promise in serving this purpose in the field of diagnostic medicine.^[4]

IgG N-Glycans

Glycans are molecules composed of complex carbohydrate structures that play crucial roles in

diverse biological processes and serve as integral components of numerous biomolecules.^[5] N-glycans are a particular type of glycan which are covalently attached to specific asparagine (Asn) amino acid residues in a protein's polypeptide chain. In the last decade, with the development of high-throughput glycan analysis techniques, N-glycans have been extensively studied in IgG, a class of antibodies that aid in the immune system's response to infections and other foreign substances. This is due to IgG's status as one of the most promising potential novel biomarkers for a person's health status. IgG's higher concentration, homeostatic stability, and half-life relative to other types of immunoglobulins in the bloodstream also make it a suitable choice of protein to perform glycomic biomarker analysis on.^[6] The structure of N-glycans on IgG displays high responsiveness to environmental changes, and can provide not only insights into an individual's lifestyle, but also the progression of various health conditions.^[7]

Glycosylation reaction networks include a series of enzymatic reactions,^[8] Deviations in glycosylation patterns have significant effects on the structure and function of IgG, and consequently, on its cancer immunosurveillance capabilities. Although the exact relationship between IgG N-glycans and CRC is not yet fully understood, specific abnormalities have been linked to CRC.^[6]^[7]

By pairing this glycomic biomarker analysis with a novel implementation of a GNN model for binary classification of CRC, and analysing the model's performance, this study aims to explore its potential to aid in the diagnosis of CRC.

Background

Machine Learning and CRC Identification

The recent emergence of artificial intelligence (AI) has brought about an interest in the potentiality of machine learning (ML) being used in the field of healthcare, addressing the complexity of various disease mechanisms that have previously posed challenges in the development of diagnostic tools which are cost-effective and time-efficient. [9] Neural networks, a class of models used in the field of ML, excel at recognising complex and non-linear patterns in data. Notable examples where neural networks have been employed include using Convolutional Neural Networks (CNN) to detect abnormal cardiac sounds that would have otherwise gone unnoticed by physicians, and using Feedforward Neural Networks (FNN) to diagnose chronic kidney disease (CKD) with near-perfect performance metrics. [10] [11]

The Existing Literature

Supervised ML models learn from training data comprising of features (inputs) and their corresponding labels (outputs); this enables them to learn to make label predictions on unseen data. [12] In the field of CRC, there have been many prior implementations of supervised ML models: investigating the use of different features such as colonoscopy images, genome sequences, and blood protein concentrations. However, these approaches to CRC diagnosis have drawbacks, namely: high invasiveness, significant costs, and low specificities to CRC, respectively. [13][14][15]. These challenges can potentially be addressed by using data obtained from glycomic analysis as the features of the model instead.

While glycobiology is an emerging field of research, studies employing ML models to predict the link between glycosylation patterns and the incidence of CRC remain limited. Vučković et al [16] employed a regularised Logistic Regression model to discern between CRC patients and healthy controls using glycomic data from a subset of 760 individuals from the Study of Colorectal Cancer in Scotland (SOCCS). Davies and Nakai [17] built upon this by proposing an alternative approach, considering multiple ML algorithms including a Soft-Voting Ensemble binary classifier, an XGBoost (eXtreme Gradient Boosting) model, and a Random Forest multiclass model. Data augmentation was also explored, utilising data scaling algorithms such as MinMaxScaler, StandardScaler, and RobustScaler.

Embedding Biological Knowledge

A limitation of the aforementioned studies of coupling ML models with glycobiology is the nature of the independent tabulated data the models were trained with. This tabulated data used for the models' inputs contained only the abundances of

each glycan: failing to encapsulate the biochemical relationships within the biological systems that gave rise to that data in the first place. [18] The utilisation of GNNs can address this limitation by representing these relationships in a graph structure: each edge is mapped to an enzymatic reaction pathway within the glycosylation reaction network [19], and each node feature represents the relative abundance of one of the unique IgG N-glycans. These graphs are then passed into the supervised ML model as the input instead. This methodology embeds an implicit biological understanding of the glycosylation reaction network into the model, which was previously absent.

Dataset Overview

Data Source

The data used in this study is a subset of the SOCCS dataset that was used in the previously mentioned studies by Vučković et al. [16] and Davies and Nakai [17]. This contained the glycomic data of 1413 CRC patients and 538 matching controls of age (± 1 year), gender, and region of residence. During population sampling, the study aimed to prospectively recruit incident cases of CRC in patients ranging from the ages of 16 to 79 who presented to surgical units of Scottish hospitals. To limit survival bias, this recruitment occurred within 90 days of diagnosis. Meanwhile, controls were randomly drawn and invited from the Community Health Index, a register of all patients in the Scotland NHS (National Health Service), to participate in the study. [20]

Blood samples were then collected and dispatched to the research centre within a 72-hour window from the time of collection. Centrifugation was employed to separate plasma from whole blood, and the isolation of IgG was carried out using monolithic plates featuring immobilised G Protein, a bacterial cell wall protein with a selective affinity for IgG. The remaining unbound proteins were then washed away from the plates with phosphate-buffered saline (PBS). An intricate series of drying and incubation steps followed, releasing N-glycans from the surface of the IgG samples. These N-glycans were then fluorescently labelled with 2-aminobenzamide (2AB). Finally, using HILIC-UPLC (Hydrophilic Interaction Ultra Performance Liquid Chromatography), in conjunction with a conventional integration algorithm, the abundances of 24 distinctive glycans were determined. [16]

However, to prevent bias and improve generalisation to unseen data, not all of this data was used in this current study. [21] This is due to a significant data imbalance for individuals over the age of 60 (719 CRC patients and 4 matching controls). Hence, these samples were discarded, leaving a remainder of 1228 samples (694 CRC patients and 534 matching controls). Data regarding BMI (body mass index) was also unavailable for 238

individuals, and therefore was not used as a part of this study.

Graphical Construction

GlyCompare is a glycomic data analysis package which allows data independence in the data source to be corrected by producing intermediate glycan structures that may have been missed during the data collection phase, generating a fully-connected glycosylation reaction network.^[22] Glycwork, a multipurpose package for glycan data science, was then utilised to visualise the resulting glycosylation reaction network.^[23] These packages were used to transform the tabulated data subset from the SOCCS into a graph-structured dataset for use in this study. PyTorch Geometric, a library built upon PyTorch then provides the necessary tools to work with GNNs in Python, the programming language of choice.

Methodology

Data Splitting

The pre-processed data was divided into training, evaluation, and testing datasets, as shown in **Figure 1**. It is imperative to split the data used to build any supervised ML model so that it can be later tested with the unseen portion of the data, providing a realistic assessment of how well the model is likely perform in practice. An 80-20 split was used to produce the training and testing datasets, a commonly adopted ratio in ML.^[24] This facilitates the model's access to a substantial dataset for learning, which increases its performance, without sacrificing the availability of data for testing.^[25] A further 80-20 split was performed on the full training dataset to produce a smaller training dataset and the validation dataset to be used in hyperparameter optimisation. The validation dataset serves as an independent dataset to assess the model's performance with different hyperparameter configurations.^[22] This split was also stratified, where the class proportion is preserved in each split. This ensures exposure to a representative sample of each class during both training and testing.^[26]

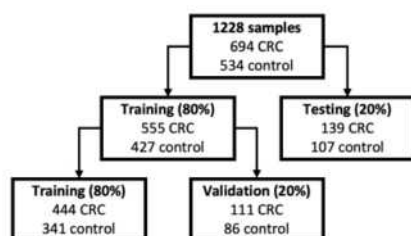


Figure 1. Overview of the proportions used during stratified dataset splitting into training, validation, and testing subsets, including the number of CRC patients and matching controls.

Model Building

To handle graph-structured data, a GNN model was a prerequisite. Various GNN models are available, including GATs (graph attention

networks) and GRNs (graph recurrent networks), among others. The GNN model of choice, however, was a GCN. Convolutional operations, a defining feature of GCNs, allows local and global relationships between the glycans of each graph structure to be captured, as information is aggregated and propagated through each region of the graph. In addition, GCNs are conceptually simpler and computationally efficient in comparison to other GNNs, making them a logical choice to base a novel exploration into leveraging graphs for CRC diagnosis on.^[27]

A GCN contains 4 main distinct types of layers: input layers, graph convolutional layers, fully connected layers, and output layers, where everything in between the input and output layers are referred to as hidden layers. The input layer captures the features of individual nodes within the provided graph and organises this information into feature vectors. Graph convolutional layers then perform the aforementioned convolutional operations on this data, capturing information from a node's neighbours and updating that node's feature representation within the model. The fully connected layer serves as a global information integrator, connecting every node's updated feature vector to all others.^[28] The output layer then transforms this high-dimensional representation into the desired output format, in this case, using a SoftMax function for soft binary classification

Hyperparameter Optimisation

To find the best model configuration, hyperparameter optimisation was performed using Optuna^[30] with the objective function to be maximised being the AUC and the decision variables to be tuned being the hyperparameters listed in **Table 1**. Categorical suggestions were employed for all hyperparameters, except for the dropout rate, for which a float suggestion was utilised due to its continuous search space.

25 trials were performed, each with 100 epochs, to find the determine the optimal hyperparameters. These numbers were selected to serve as a middle ground between allowing sufficient iterations and limiting computational demand. A learning curve was also later plotted, by confirming that the chosen epoch number sits on the plateau of the curve, the choice of 100 epochs was confirmed to be suitable with no overfitting or underfitting.^[31]

This entire process of 25 trials was repeated with 3 different samplers, each with their own algorithm for locating the optima. This included the TPESampler: a sampler well-regarded for its consistently reliable performance^[32], RandomSampler: a sampler based on independent random number generation to serve as a baseline, and NSGA-II: an attempt to investigate the use of an evolutionary genetic algorithm.^[33]

Furthermore, 3 different variants of NSGA-II were evaluated, with a population size of 50, 100, and 500, respectively. Population size in genetic algorithms refers to the number of candidate solutions present in each generation; these solutions evolve over each iteration of the optimisation

process. ^[33] Upon determining the optimal hyperparameters for each sampler, the resulting AUCs are compared. Subsequently, the hyperparameters associated with the overall best AUC are implemented into the model.

Table 1. Overview of the hyperparameters optimised, the suggested values provided to Optuna, and a description of each of their functions.

Hyperparameter	Suggestions	Function
Learning rate	10^{-4} , 10^{-3} , 10^{-2}	Controls step size optimisation step size
No. of graph convolutional layers	2, 3, 4	Balances learning potential and feature capacity with overfitting
No. of neurons in hidden layers	64, 128, 256	
Batch size	32, 64, 128	Number of graphs used in each training iteration
Activation function	ReLU, Leaky ReLU	Introduces non-linearity to capture complex relationships
Dropout rate	0 – 0.5	Proportion of neurons to drop during training to prevent overfitting
Aggregation method	Mean, Add, Max	How information is aggregated during convolutional operations

Model Training

For the training of the model, Adaptive Moment Estimation (Adam) was chosen as the optimiser. Adam brings together ideas from two popular optimisation methods: Momentum and RMSprop (Root Mean Squared Propagation). It uses a moving average of past gradients from Momentum, helping the optimiser move consistently and navigate through flat or noisy search spaces. ^[34] Additionally, it adopts the concept of RMSprop by tracking the average of squared gradients. This adaptive feature adjusts learning rates for individual parameters based on their historical gradients, making Adam effective in handling varying gradients. ^[35] By combining these approaches, Adam became the versatile and robust optimiser of choice for this study.

Evaluation Method

To quantify the model's performance during both hyperparameter optimisation and model evaluation, a performance metric must be chosen. AUC, a commonly used metric for binary classification of balanced datasets, was selected as the performance metric for this study. ^[36]

The AUC is calculated from the metrics TP (true positive), TN (true negative), FP (false positive), and FN (false negative). TP and TN represent the number of correct predictions generated by the model, while FP and FN represent the incorrect positive and negative predictions, respectively. These 4 metrics can be visualised using the confusion matrix shown in **Figure 2**.

To calculate AUC, **Equations 1, 2, and 3** are used.

$$\text{Sensitivity} = TPR = \frac{TP}{TP + FN} \quad (\text{Eq. 1})$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (\text{Eq. 2})$$

$$FPR = \frac{FP}{TP + FN} = 1 - \text{Specificity} \quad (\text{Eq. 3})$$

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 2. A confusion matrix illustrating the classification outcomes (TP , TN , FP , or FN) based on actual and predicted positive or negative values. ^[37]

Sensitivity gives the proportion of positive cases that were detected, and specificity indicates the proportion of correctly identified negative samples. Thus, a high TPR (true positive rate) and a low FPR (false positive rate) are desired. However, at different thresholds, the model yields varying values TPR and FPR . A threshold is a value which an ML model uses to determine the confidence level at which a prediction is assigned as positive. The choice of thresholds can be arbitrary and depends on

the level of strictness the model builder wants to impose on the model. For this study, a stricter threshold may lead to some CRC cases going undetected, while a more relaxed threshold will lead to a greater number of incorrect positive diagnoses. Thus, due to the intricate nature of the threshold, especially in the medial field, AUC was chosen.

The final step in calculating the AUC involves taking the area under the ROC curve obtained by plotting the *TPR* against the *FPR* values at different thresholds. Thus, negating the need for a specified threshold. Combined with the popular usage of AUC, this approach also facilitates more accurate comparisons with other pre-existing studies.

AUC has a value ranging from 0 to 1: where 0 indicates the worst performance, 1 signifies that all label predictions are correct, and 0.5 is equivalent to random classification. This is shown in **Figure 3**. Although there are no fixed criteria for AUC, generally, an AUC of 0.7 - 0.8 is generally deemed acceptable, a value of 0.8 - 0.9 is good, and a value over 0.9 suggests an outstanding model that can reliably perform binary classification. [38]

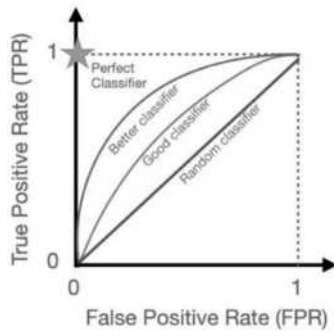


Figure 3. ROC curves corresponding to binary classifiers of varying performances. The position of the star corresponds to a classifier with perfect predictive performance. [39]

Model Assessment

With the GCN model configured with the tuned hyperparameters, the model was trained on a combined dataset comprising both training and validation data. The model was then tested on the unseen test dataset, evaluating its performance using the AUC.

Training ML models is an inherently stochastic process. Depending on the way the dataset is split into training and testing subsets, processes such as weight initialisation and data shuffling can differ. Thus, the same model with the same dataset could yield different results each time it is run. [40] While this stochasticity can enhance the model's ability to generalise its performance [41], it prevents the conditions in which the model was trained and tested from being recreated, which is important for research purposes.

To mitigate this and ensure the obtained AUCs are reliable and reproducible, seeding was introduced, fixing the initial conditions for the

random processes present in the model, including data splitting and weight initialisation. [40]

Nevertheless, even with seeding, there can still be some variations in performance from run to run. To obtain a representative AUC, for each seed, the configured model was run 5 times and the median AUC across these runs was taken as the AUC for that particular seed.

The choice of using the median as the measure of central tendency was because, unlike the mean, the median is less sensitive to outliers and extreme values. Thus, making the median the more robust measure when dealing with a limited number of runs that may not fully characterise a normal distribution. [42]

To account for the reproducibility of the results and the variability in the model's performance, the model was also run separately across 8 different seeds, creating distinct training and testing datasets for each seed. The median of the AUCs calculated for each of the 8 different seeds is then reported as the final AUC for the model.

Results & Discussion

Obtained AUC of the proposed GCN model

By running the model 5 times for each seed, and reporting the AUC of each run, the median AUC of each seed and their interquartile ranges (IQR) were calculated. These values are detailed in **Table 2**.

Table 2. Overview of the median AUC and IQR obtained from the GCN model for 8 different seeds.

Seed	Median AUC	IQR
A	0.631	0.0047
B	0.572	0.1461
C	0.606	0.0001
D	0.604	0.0011
E	0.588	0.1876
F	0.648	0.0005
G	0.581	0.1658

The final representative AUC for the model is taken as the median of these AUCs: a value of 0.604, obtained from seed D. The interquartile range (IQR) of the median AUCs across the 8 seeds was also calculated to be 0.0404.

With the final AUC of 0.604, the model has some predictive power ($AUC > 0.5$), but still warrants further improvement in order to be considered as a reliable model ($AUC > 0.7$). The large IQR of 0.404 between the seeds also implies high sensitivity to changes in the dataset used and poor generalisation (ability to perform well on different datasets).

It can also be noted that seeds with higher AUCs have lower IQRs, and vice-versa. This signifies a high sensitivity to the stochasticity of the learning process and poor robustness, being unaffected by outliers and unexpected conditions, when encountering specific data splits and model

configurations. [43] Thus, methods to improve the model's stability may need further exploration.

Comparative Analysis with Benchmark Algorithms

To gain a better understanding of the performance of the GCN model and compare its performance with models using tabulated data, 2 other ML algorithms, Logistic Regression (LR) and Random Forest (RF), were trained using the same dataset and seeding. Both of these algorithms are commonly used in ML and are simpler in their structure, providing good benchmarks for the GCN model. [44] RF has also been recognised for its effectiveness in disease identification. [45][46] The results are summarised in **Table 3**.

Table 3. Comparison of the AUC and IQR obtained from the GCN model to the benchmark algorithms (LR and RF).

Algorithm	Final AUC	IQR
GCN	0.604	0.0408
LR	0.593	0.0573
RF	0.582	0.0808

Comparable performance was observed between all 3 algorithms, with the GCN model marginally outperforming the others. The performance improvement attained by the GCN model, even if slight, shows the potential leverage of incorporating biological knowledge into ML models. A more sophisticated model would be assumed to further take advantage of the graphical nature of the dataset.

Nevertheless, it is important to acknowledge the large IQRs in all 3 algorithms. LR and RF are not trained stochastically, unlike GNNs. Therefore, this may hint at the fact that the poor generalisability originated from the dataset used in training and testing, and not the models themselves. [47][48] Inadequacy of the training dataset in representing the testing dataset's characteristics results in poor performance in supervised ML; a larger dataset with sufficient diversity can mitigate this. [49]

It is also crucial to note that both LR and RF are expected to outperform GNNs when handling smaller datasets such as this. [44] Had the dataset been larger, the GCN model would be expected to outperform LR and RF by a greater margin. In addition, a defining characteristic of LR is the underlying assumption of feature independence associated with LR [45]; an assumption that does not hold with glycomic data, which is multicollinear and interdependent. The fact that LR still performed as well as others further solidified the fact that the dataset itself had a large influence on the result.

Comparative Analysis with Previous Studies

To compare the performance of the proposed GCN model with other models in the existing literature, a comparison was conducted with the models from the studies by Vučković et al [16] and Davies and Nakai [17], as shown in **Table 4**.

The results indicate that the proposed GCN model exhibits lower performance in comparison to those in these previous studies. Since the difference was unlikely to have been caused by the algorithm choice, as established in the previous section, the difference in methodologies between the studies was explored.

Table 4. Comparison of the AUC and IQR obtained from the proposed GCN model to other algorithms utilising the same glycomic dataset from the SOCCS.

Algorithms	Final AUC
LR [16]	0.755
LR [17]	0.696
SVE (Soft-Voting Ensemble) [17]	0.727
XGB (XGBoost) [17]	0.723
RF [17]	0.722
SVM (Support Vector Machines) [17]	0.702
GCN	0.604

A difference in performance can be attributed to the differences in the features used: age, sex, and BMI were included by Davies and Nakai. [17] While investigating a model based on age and sex alone did not show significant discriminative power, BMI is known to be correlated with CRC risk [51]. Another consideration is the use of data imputation to compensate for missing BMI values in some samples, a potential source of classification bias. [52]

Instead of the 694 CRC patients and 534 matching controls used in this study, a larger and different subset of the SOCCS dataset was used by Vučković et al. [16] 760 patients and 538 matching controls were used, which included the samples neglected in this study due to an imbalance in labels for individuals over 60 years of age. The inclusion and exclusion criteria were not mentioned in the study, and therefore, the specifics of any possible class imbalances, another source of classification bias, are unknown. [21] A validation technique known as non-nested folding was also used which may introduce an optimistic bias in the final AUC calculation. [53]

Comparative Analysis with Models of Other Features

Table 5 compares the performance of the proposed glycomics-based model with alternative biomarkers for CRC identification. These alternatives include more established methods such as Zhou et al.'s [13] utilisation of CNNs in large-scale colonoscopic screening, and cutting-edge biomarkers such as multi-platform transcriptomics and whole-genome sequencing used in Long et al.'s. [54] and Wan et al.'s studies. [14] Other emerging CRC biomarkers have also been studied. The use of blood count and urinary polyamines by Hornbrook [55] and Nakajima [56] explored the use of these features in

identifying high-risk patients for early CRC detection, although only 59 samples were used by Nakajima for result validation. Li et al.'s [15] and Zhao et al.'s [57] studies of blood protein and gut bacteria have also shown promising results.

Despite the current substantial performance gap observed between the glycomics-based model and its alternatives as shown in **Table 5**, there are unique

advantages inherent to glycomics that place it above the others in one way or another, from its non-invasiveness compared to colonoscopy images, simplicity compared to multi-platform transcriptomics, to higher specificity to CRC compared to blood count - making glycomics a valuable area of research in the field of CRC. [7]

Table 5. Comparison of the AUC obtained from the proposed GCN model to other CRC classification models with non-glycomic biomarkers, adapted with modifications from Li et al. [15]

Features	Algorithms	Final AUC
Colonoscopy images [27]	CNN	0.930
Multi-platform transcriptomics [30]	RF	0.998
Whole-genome sequencing [31]	LR + SVM	0.92
Age, sex, and blood count [32]	GBM (Gradient Boosting) + RF	0.80
Urinary polyamines [29]	ADTree (Alternating Decision Tree)	0.961
Blood proteins concentration [28]	LR	0.849
Age, BMI, gut bacteria [33]	LR + SVM	0.942
Glycomics	GCN	0.604

Conclusions

Key Takeaways

This study is a first attempt at incorporating biological knowledge into a glycomics-based ML model, improving upon prior models by considering the enzymatic relationships between the glycans, and thus, sets a foundation for this novel method of CRC identification. The final obtained AUC of the GCN model was 0.604 with an IQR of 0.0404, between 8 different seeds. This AUC implies a degree of predictive power for CRC identification, although, an AUC of 0.70 would be needed to be achieved for the model to be considered somewhat reliable and its lack of robustness and generalisability still need to be addressed.

Upon comparison with the benchmark algorithms, the GCN model slightly outperforming RF and LR, despite the small dataset, showcases the potential of incorporating biological knowledge into ML models. However, in comparison to models utilising non-glycomic data, it is evident that the AUC of this GCN model falls short, signifying a need for substantial improvements in its predictive power if glycomics were to become the predominant biomarker of choice to diagnose CRC.

Overall, this study supports the continued research of using GNNs, with larger datasets and more rigorous methods, to fully explore the capabilities of this novel methodology as well as to address its limitations. The utilisation of GNNs with glycomics for CRC diagnosis is still in its early stages, and there is still ample room for exploration and refinement. The results and different comparisons outlined in this study show promising potential for further advancements in the future.

Future Considerations

Several limitations and potential improvements in the dataset have been identified in this study. An increase in dataset size and data diversity is linked to better GNN performance and generalisability. [58] [59] [60]. Comparisons with the model of other studies also suggested that usage of other features in conjunction with glycomic data may help improve the predicting power of the model as well.

The model can also be further refined. Cross-validation allows for better generalisability and soft-voting can give a more accurate measure of the model's performance. [61] [62] Further exploring other optimisers, different GNN models and hyperparameters for tuning can also be done.

Acknowledgements

We would like to sincerely thank Konstantinos Flearis for his continuous support throughout the course of this study.

Supplementary Materials

The Python code of the GCN model and the accompanying dataset for this study is available upon request.

References

1. World Health Organization. 2023. "Colorectal Cancer." Available: <https://www.who.int/news-room/fact-sheets/detail/colorectal-cancer> (Accessed: 1 December 2023)
2. Centers for Disease Control and Prevention. 2023. "Colorectal (Colon) Cancer." Available: <https://www.cdc.gov/cancer/colorectal/> (Accessed: 1 December 2023)
3. Balog, C. I.; Stavenhagen, K.; Fung, W. L.; et al. 2012. "N-Glycosylation of Colorectal Cancer Tissues: A Liquid Chromatography

- and Mass Spectrometry-Based Investigation." *Mol Cell Proteomics* 11 (9): 571–585. <https://doi.org/10.1074/mcp.M111.011601>.
4. Lauc G, Pezer M, Rudan I, Campbell H. 2016. "Mechanisms of disease: The human N-glycome." *Biochim Biophys Acta* 1860(8): 1574-1582. <https://doi.org/10.1016/j.bbagen.2015.10.016>.
5. Varki A, Kornfeld S. 2022. "Historical Background and Overview." In: Varki A, Cummings RD, Esko JD, et al., editors. *Essentials of Glycobiology* [Internet]. 4th edition. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press. Chapter 1. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK579927/> doi: 10.1101/glycobiology.4e.1.
6. Schroeder HW Jr, Cavacini L. 2010. "Structure and function of immunoglobulins." *J Allergy Clin Immunol* 125(2 Suppl 2): S41-S52. <https://doi.org/10.1016/j.jaci.2009.09.046>.
7. Sofia Shkunnikova, Anika Mijakovac, Lucija Sironic, Maja Hanic, Gordan Lauc, Marina Martinic Kavur. 2023. "IgG glycans in health and disease: Prediction, intervention, prognosis, and therapy." *Biotechnology Advances*, Volume 67, 108169. <https://doi.org/10.1016/j.biotechadv.2023.108169>.
8. Kiyoko F Aoki-Kinoshita. 2021. "Glycome informatics: using systems biology to gain mechanistic insights into glycan biosynthesis." *Current Opinion in Chemical Engineering*, Volume 32, 100683. ISSN 2211-3398.
9. Ahsan MM, Luna SA, Siddique Z. 2022. "Machine-Learning-Based Disease Diagnosis: A Comprehensive Review." *Healthcare (Basel)* 10(3): 541. <https://doi.org/10.3390/healthcare10030541>.
10. Rubin, J.; Abreu, R.; Ganguli, A.; Nelaturi, S.; Matei, I.; Sricharan, K. 2017. "Recognizing Abnormal Heart Sounds Using Deep Learning." *CoRR* abs/1707.04642. <http://arxiv.org/abs/1707.04642>.
11. A. Imran, M. N. Amin and F. T. Johora. 2018. "Classification of Chronic Kidney Disease using Logistic Regression, Feedforward Neural Network and Wide & Deep Learning." 2018 International Conference on Innovation in Engineering and Technology (ICIET), Dhaka, Bangladesh, pp. 1-6. doi: 10.1109/CIET.2018.8660844.
12. *Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in EHealth, HCI, Information Retrieval and Pervasive Technologies*. 2007. Netherlands: IOS Press.
13. Zhou, D., Tian, F., Tian, X., et al. 2020. "Diagnostic evaluation of a deep learning model for optical diagnosis of colorectal cancer." *Nat Commun* 11: 2961. DOI: 10.1038/s41467-020-16777-6.
14. Wan N, Weinberg D, Liu T-Y, et al. 2019. "Machine learning enables detection of early-stage colorectal cancer by whole-genome sequencing of plasma cell-free DNA." *BMC Cancer* 19(1): 832.
15. Li, H., Lin, J., Xiao, Y., Zheng, W., Zhao, L., Yang, X., Zhong, M., & Liu, H. 2021. "Colorectal Cancer Detected by Machine Learning Models Using Conventional Laboratory Test Data." *Technology in Cancer Research & Treatment* 20. DOI: 10.1177/15330338211058352.
16. Vučković F, Theodoratou E, Thaçi K, Timofeeva M, Vojta A, Štambuk J, Pučić-Baković M, Rudd PM, Derek L, Servis D, Wennerström A, Farrington SM, Perola M, Aulchenko Y, Dunlop MG, Campbell H, Lauc G. 2016. "IgG Glycome in Colorectal Cancer." *Clin Cancer Res* 22(12): 3078-3086. <https://doi.org/10.1158/1078-0432.CCR-15-1867>.
17. Davies, J., Nakai, S. 2022. "Extracting the Biomarker Potential of N-glycans with a Machine Learning Framework Applied to Colorectal Cancer." Department of Chemical Engineering, Imperial College London, U.K.
18. Ahmedt-Aristizabal D, Armin MA, Denman S, Fookes C, Petersson L. 2021. "Graph-Based Deep Learning for Medical Diagnosis and Analysis: Past, Present and Future." *Sensors (Basel)* 21(14): 4758. <https://doi.org/10.3390/s21144758>.
19. Puri, A.; Neelamegham, S. 2012. "Understanding Glycomechanics Using Mathematical Modeling: A Review of Current Approaches to Simulate Cellular Glycosylation Reaction Networks." *Ann. Biomed. Eng.* 40(12): 816–827. <https://doi.org/10.1080/07328303.2011.630835>.
20. Theodoratou E, Kyle J, Cetnarskyj R, Farrington SM, Tenesa A, Barnetson R, Porteous M, Dunlop M, Campbell H. 2007. "Dietary flavonoids and the risk of colorectal cancer." *Cancer Epidemiol Biomarkers Prev* 16(4): 684-693. <https://doi.org/10.1158/1055-9965.EPI-06-0785>. PMID: 17416758.
21. Kumar, P., Bhatnagar, R., Gaur, K., & Bhatnagar, A. 2020. "Classification of Imbalanced Data: Review of Methods and Applications." *IOP Conference Series: Materials Science and Engineering* 1099. DOI: 10.1088/1757-899X/1099/1/012077
22. Bao, B., Kellman, B.P., Chiang, A.W.T. et al. 2021. "Correcting for sparsity and interdependence in glycomics by accounting for glycan biosynthesis." *Nat Commun* 12: 4988. <https://doi.org/10.1038/s41467-021-25183-5>.

23. Luc Thomès, Rebekka Burkholz, Daniel Bojar. 2021. "Glycowork: A Python package for glycan data science and machine learning." *Glycobiology* 31(10): 1240–1244. <https://doi.org/10.1093/glycob/cwab067>.
24. V. R. Joseph. 2022. "Optimal ratio for data splitting." *Stat. Anal. Data Min.: ASA Data Sci.* <https://doi.org/10.1002/sam.11583>.
25. Gholamy, A., Kreinovich, V., & Kosheleva, O. 2018. "Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation."
26. Igareta, A. 2021. "Stratified sampling: You may have been splitting your dataset all wrong." Medium. Available at: <https://towardsdatascience.com/stratified-sampling-you-may-have-been-splitting-your-dataset-all-wrong-8cfd0d32502> (Accessed: 10 December 2023).
27. Zhou, J.; Cui, G.; Hu, S.; Zhang, Z.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; Sun, M. 2020. "Graph Neural Networks: A Review of Methods and Applications." *AI Open* 1: 57-81. <https://doi.org/10.1016/j.aiopen.2021.01.001>
28. Zhang, S., Tong, H., Xu, J. et al. 2019. "Graph convolutional networks: a comprehensive review." *Comput Soc Netw* 6: 11. <https://doi.org/10.1186/s40649-019-0069-y>.
29. Banerjee, K.; Prasad, V.; Gupta, R. R.; Vyas, K.; Anushree, H.; Mishra, B. 2020. "Exploring Alternatives to Softmax Function." <https://arxiv.org/abs/2011.11538>.
30. Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. "Optuna: A Next-generation Hyperparameter Optimization Framework." In *KDD*.
31. Pramoditha, R. (2022, September 29). "Plotting the Learning Curve to Analyse the Training Performance of a Neural Network." Medium. Data Science 365.
32. Sipper, M. 2021. "Neural Networks with A La Carte Selection of Activation Functions." *SN COMPUT. SCI.* 2: 470. <https://doi.org/10.1007/s42979-021-00885-1>.
33. K. Deb, A. Pratap, S. Agarwal and T. Meyarivan. 2002. "A fast and elitist multiobjective genetic algorithm: NSGA-II." *IEEE Transactions on Evolutionary Computation* 6(2): 182-197. doi: 10.1109/4235.996017.
34. Dehghani, M., Samet, H. 2020. "Momentum search algorithm: a new meta-heuristic optimization algorithm inspired by momentum conservation law." *SN Appl. Sci.* 2: 1720. <https://doi.org/10.1007/s42452-020-03511-6>.
35. Reddy, S. V. G.; Thammi Reddy, K.; ValliKumari, V. 2018. "Optimization of Deep Learning Using Various Optimizers, Loss Functions, and Dropout." *Int. J. Recent Technol. Eng.* 7: 448-455.
36. Jin Huang and C. X. Ling. 2005. "Using AUC and accuracy in evaluating learning algorithms." *IEEE Transactions on Knowledge and Data Engineering* 17(3): 299-310. doi: 10.1109/TKDE.2005.50.
37. Shrivastav, N. 2021. "Confusion matrix(TPR,FPR,FNR,TNR), Precision, recall, F1-score." Medium. Available at: <https://medium.datadriveninvestor.com/confusion-matrix-tp-r-fpr-fnr-tnr-precision-recall-f1-score-73efa162a25f> (Accessed: 3 Dec2023).
38. DW Hosmer, S Lemeshow. 2000. "Applied Logistic Regression, 2nd Ed. Chapter 5." John Wiley and Sons, New York, NY, pp. 160-164.
39. Sipio, R.D. 2021. "A quick guide to AUC-roc in machine learning models." Medium. Available at: <https://towardsdatascience.com/a-quick-guide-to-auc-roc-in-machine-learning-models-f0aedb78fbad> (Accessed: 7 December 2023).
40. Bansal, J. 2020. "How to use random seeds effectively." Medium. Available at: <https://towardsdatascience.com/how-to-use-random-seeds-effectively-54a4cd855a79> (Accessed: 12 December 2023).
41. Sabuncu, M.R. 2020. "Intelligence plays dice: Stochasticity is essential for machine learning." *ArXiv*, abs/2008.07496.
42. Forthofer, R. N.; Lee, E. S.; Hernandez, M. 2007. "Descriptive Methods." In *Biostatistics*, 2nd ed.; Forthofer, R. N.; Lee, E. S.; Hernandez, M., Eds.; Academic Press; pp 21-69. ISBN 9780123694928. DOI: 10.1016/B978-0-12-369492-8.50008-X.
43. Brownlee, J. 2020. "Why do I get different results each time in machine learning?" *MachineLearningMastery.com*. Available at: <https://machinelearningmastery.com/different-results-each-time-in-machine-learning/> (Accessed: 3 December 2023).
44. Kirasich, K.; Smith, T.; Sadler, B. 2018. "Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets." *SMU Data Science Review* 1(3): Article 9. Available at: <https://scholar-smu.edu/datasciencereview/vol1/iss3/9>
45. Morris, J.; Lieberman, M. G. 2014. "The Precise Effect of Multicollinearity on Classification Prediction." *Florida Atlantic University*.
46. Uddin, S.; Khan, A.; Hossain, M., et al. 2019. "Comparing different supervised machine learning algorithms for disease prediction." *BMC Med Inform Decis Mak* 19: 281. DOI: 10.1186/s12911-019-1004-8
47. Gustavo. 2019. "Understanding logistic regression step by step." Medium. Available at: <https://towardsdatascience.com/understanding>

- logistic-regression-step-by-step-704a78be7e0a (Accessed: 3 December 2023).
48. Biau, G.; Scornet, E. 2016. "A Random Forest Guided Tour." *TEST* 25: 197–227. DOI: 10.1007/s11749-016-0481-7.
 49. Yu Yu, Shahram Khadivi, and Jia Xu. 2022. "Can Data Diversity Enhance Learning Generalization?" In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4933–4945, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
 50. Stoltzfus, J. C. 2011. "Logistic regression: a brief primer." *Academic emergency medicine: official journal of the Society for Academic Emergency Medicine* 18(10): 1099–1104. DOI: 10.1111/j.1553-2712.2011.01185.x
 51. Mandic, M.; Li, H.; Safizadeh, F.; Niedermaier, T.; Hoffmeister, M.; Brenner, H. 2023. "Is the association of overweight and obesity with colorectal cancer underestimated? An umbrella review of systematic reviews and meta-analyses." *European Journal of Epidemiology* 38(2): 135-144. DOI: 10.1007/s10654-022-00954-6
 52. Salgado, C. M.; Azevedo, C.; Proença, H.; Vieira, S. M. 2016. "Missing Data." In: *Secondary Analysis of Electronic Health Records*. Springer, Cham. DOI: 10.1007/978-3-319-43742-2_13
 53. Tsamardinos, I., Greasidou, E., & Bouboudakis, G. 2018. "Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation." *Machine Learning* 107(12): 1895–1922. DOI: 10.1007/s10994-018-5714-4
 54. Long NP, Park S, Anh NH, et al. 2019. "High-Throughput omics and statistical learning integration for the discovery and validation of novel diagnostic signatures in colorectal cancer." *Int J Mol Sci* 20(2): 296.
 55. Hornbrook MC, Goshen R, Choman E, et al. 2017. "Early colorectal cancer detected by machine learning model using gender, Age, and complete blood count data." *Dig Dis Sci* 62(10): 2719-2727.
 56. Nakajima T, Katsumata K, Kuwabara H, et al. 2018. "Urinary polyamine biomarker panels with machine-learning differentiated colorectal cancers, benign disease, and healthy controls." *Int J Mol Sci* 19(3): 756.
 57. Zhao D, Liu H, Zheng Y, et al. 2019. "A reliable method for colorectal cancer prediction based on feature selection and support vector machine." *Med Biol Eng Comput* 57(4): 901–912.
 58. Thian, Y. L., Ng, D. W., Patrick Decourcy Hallinan, J. T., Jagmohan, P., Sia, S. Y., Adnan Mohamed, J. S., Quek, S. T., & Feng, M. 2022. "Effect of Training Data Volume on Performance of Convolutional Neural Network Pneumothorax Classifiers." *Journal of Digital Imaging* 35(4): 881-892. DOI: 10.1007/s10278-022-00594-y
 59. Good machine learning practice for medical device development: Guiding principles (no date) GOV.UK. Available at: <https://www.gov.uk/government/publications/good-machine-learning-practice-for-medical-device-development-guiding-principles>. (Accessed: 11 December 2023).
 60. Gong, Z., Zhong, P., & Hu, W. 2018. "Diversity in Machine Learning." *ArXiv*. <https://doi.org/10.1109/ACCESS.2019.2917620>
 61. Science, O.-O.D. 2019. "Properly setting the random seed in ML experiments. not as simple as you might imagine." *Medium*. Available at: <https://odsc.medium.com/properly-setting-the-random-seed-in-ml-experiments-not-as-simple-as-you-might-imagine-219969c84752> (Accessed: 11 December 2023).
 62. Ahmed, I. 2023. "What is hard and soft voting in machine learning?" *Medium*. Available at: <https://ilyasbinsalih.medium.com/what-is-hard-and-soft-voting-in-machine-learning-2652676b6a32> (Accessed: 11 December 2023)

Differentiable Equations of State for Machine Learning Thermodynamic-Property Prediction

Michael Gadaloﬀ, Luc Paoli

Department of Chemical Engineering, Imperial College London, U.K.

Abstract

Reliable equations of state (EoS) for complex fluids are necessary for fields ranging from chemical-process simulation to computer-aided molecular design, but the need for extensive experimental data limits the applicability of EoS, particularly the Statistical Associating Fluid Theory (SAFT). We aim to address the challenges of obtaining accurate SAFT EoS parameters, employing the Julia SciML ecosystem to create a machine-learning (ML) model using a differentiable SAFT EoS. Trained on saturation pressure and saturated liquid density data, this model uses a molecular fingerprint to output SAFT-VR Mie parameters, incorporating a physical foundation into its learning process. We validate the performance of our model using 80 unseen alkanes and demonstrate that the parameters generated align with expected physical trends. We also assess the predictive performance of our model for unseen properties such as isobaric heat capacity, highlighting the value of embedding physical laws in ML thermodynamics. In this work, we present the first physics-informed neural network to use SAFT-VR Mie and provide the first open-source ML model incorporating a SAFT-type EoS.

1. Introduction

Accurate thermodynamic models are a fundamental component of molecular design [1] and chemical-process simulation [2]. Obtaining reliable parameters for an Equation of State (EoS) requires a large quantity of experimental data, which may be challenging to gather. Predictive methods where thermodynamic properties are obtained from molecular structures have seen much interest over the last few decades [3–6].

In group-contribution approaches, molecules are “coarse-grained” into groups (Figure 1), where each group contributes a pre-determined amount to the overall EoS parameters [5]. Coarse-graining methods present challenges, however, notably the difficulties of subdividing a species into components. Accurately accounting for intramolecular effects such as steric hindrance requires the definition of larger second-order groups, the contribution of which must

be separately quantified.

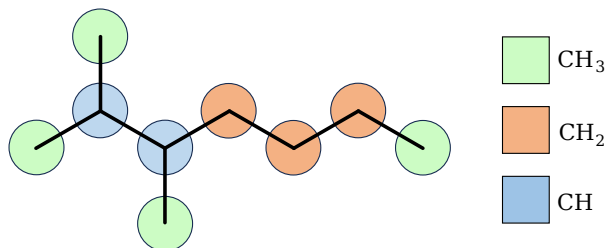


Figure 1: 2,3-Dimethylheptane, coarse-grained for a group-contribution approach.

The adaptability of machine-learning (ML) models may help overcome some of the limitations of group-contribution methods; entire molecular structures may be inputted into an ML model, eliminating the need for coarse-graining. There has been a significant rise in the popularity of ML over the last decade, but

Approach Type	Training data	Description
Cubic EoS [7]	Experimental	Critical properties and acentric factor
Helmholtz free energy [8]	Molecular dynamics	Learning A for a Mie fluid
PCP-SAFT [9]	Experimental	Parameter regression
PCP-SAFT [10]	Pseudo-experimental	Physics-informed (surrogate)
PCP-SAFT [11]	Experimental	Physics-informed (direct)

Table 1: Summary of recent approaches to machine-learning thermodynamics.

reliably extrapolating beyond the data range on which a model is trained is challenging. If used for predictive thermodynamics, the model accuracy may be enhanced if a physical basis is incorporated into the training. "Physics-informed" ML may output more-accurate estimates of thermodynamic properties.

Incorporating the Statistical Associating Fluid Theory (SAFT) into the ML model would introduce physical constraints into the training. SAFT is a powerful class of EoS in which fluids are modelled as chains of segments with associating sites in a specified potential, and thermodynamic properties may be estimated given the parameters of this representation of a species. As such, an ML model trained to output SAFT parameters corresponding to high-accuracy estimates of properties would benefit from the physical basis of a SAFT-type EoS.

In Table 1, we summarise recent work in which ML is applied to predictive thermodynamics. Biswas et al. [7] presents a graph neural network that outputs the critical point and the acentric factor, which may be used to parameterise many types of cubic EoS. Chaparro and Müller [8] have taken a different approach, developing an ML EoS by training a neural network on molecular-dynamics simulation data. Regarding SAFT, three recent works investigate the use of Perturbed-Chain-Polar SAFT (PCP-SAFT) with ML. Felton et al. [9] train on a database of PCP-SAFT parameters, while Habicht, Sadowski, and Brandenbusch [10] used an ML-model substitute for PCP-SAFT trained to accept parameters and output thermodynamic properties. Winter et al. [11] incorporated PCP-SAFT into their training, encoding

a strong physical basis into their ML model.

We elaborate on these works by integrating SAFT with a variable-range Mie potential (SAFT-VR Mie) into an ML workflow, using a molecular fingerprint as our model input.

The remainder of this study consists of a methods section, in which we present the components of our workflow. We then analyse the performance of our model, discussing the implications of our results.

2. Methods

2.1. Overview

Our workflow, illustrated in Figure 2, consists of two main components: a pre-training step involving PCP-SAFT parameters, and a main loop considering pseudo-experimental data. We detail the parts of our workflow in this section. The relative mean-square deviation (RMSD) is defined by Equation 1, where $y_{i,\text{reference}}$ is the training data and $y_{i,\text{predicted}}$ is the output from SAFT-VR Mie.

$$\text{RMSD} = \frac{1}{N} \sum_{i=1}^N \left(\frac{y_{i,\text{predicted}} - y_{i,\text{reference}}}{y_{i,\text{reference}}} \right)^2 \quad (1)$$

2.2. Molecular Fingerprints

Molecules are inputted into our model as fingerprints, vectors of ones and zeros encoding their structure. We use an atom-pair algorithm to generate these vectors [12], which enumerates the number of atoms of a given type along the shortest path between all pairs of atoms in a molecule. We remove any elements common

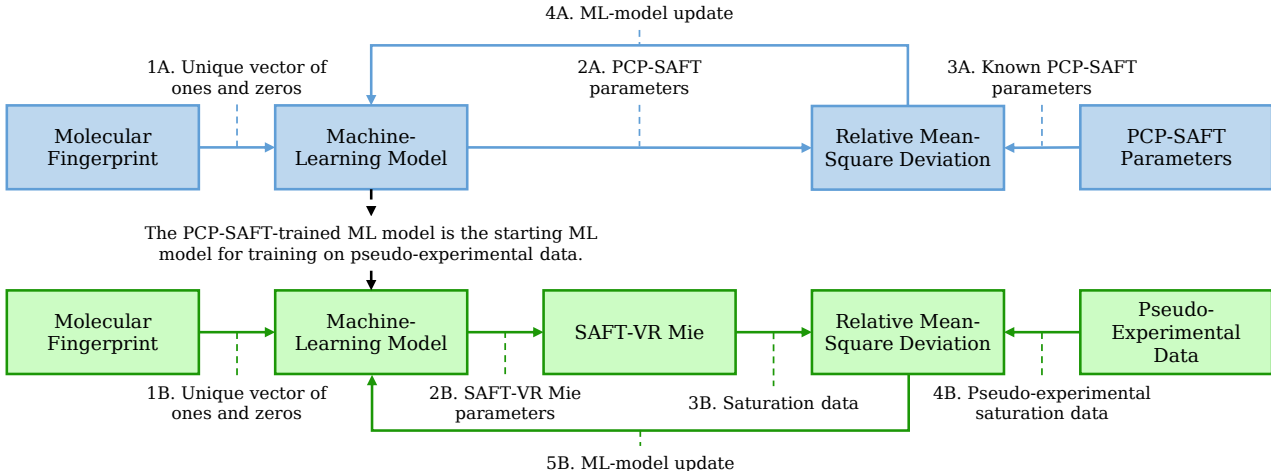


Figure 2: Workflow overview. The ML model is first pre-trained on PCP-SAFT parameters (blue), before being trained on pseudo-experimental data (green).

to all fingerprints at each position, as these do not contribute any meaningful information for distinguishing between molecules.

2.3. Neural Networks

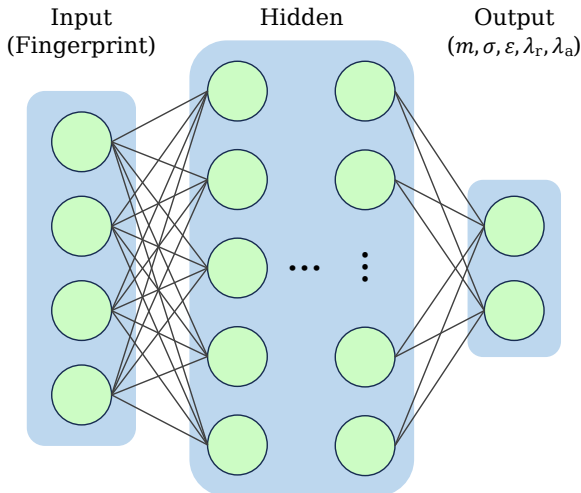


Figure 3: Illustration of a multilayer perceptron, where the green dots and lines represent neuron and connections, respectively.

We input our molecular fingerprints into a multilayer perceptron, a neural network characterized by its layered, fully connected architecture (Figure 3). Each layer consists of a set of neurons, where each neuron in a layer is connected to every neuron in the next layer. These connections are weighted, where training a model involves adjusting these weights. Within each neuron, a non-linear function is applied to the sum of the inputs, generating the output. We use the scaled exponential linear unit as our nonlinear “activation function”, which increases learning robustness and avoids vanishing gradients [13].

The ML architecture we use has a decreasing number of nodes per hidden layer (Table 2). Reducing model complexity mitigates the possibility of overfitting the network to data. Additionally, the 32-neuron layer outputs an information-rich vector that encodes the most important properties of the training molecules, increasing the likelihood that the model outputs accurate predictions of molecular properties.

We use a learning rate of 5e-6 and train the neural network using adaptive moment estimation (ADAM), an extension of gradient descent that incorporates momentum and per-parameter scaling [14]. By adjusting the ef-

fective learning rate based on the gradient history, ADAM enables faster convergence and increased stability.

Hyperparameter	Value
hidden layers	4
hidden size 1	1024
hidden size 2	512
hidden size 3	128
hidden size 4	32
activation function	SELU
optimizer	Adam
learning rate	5×10^{-6}

Table 2: Structure of the neural networks trained in this work.

2.4. SAFT-VR Mie

The neural network outputs are the parameters for SAFT-VR Mie, a flavour of SAFT developed by Lafitte et al. [15]. Its parameters include m and σ , representing the number of segments per chain and segment size respectively. SAFT-VR Mie also uses the variable-range Mie potential (Equation 3),

$$C = \frac{\lambda_r}{\lambda_r - \lambda_a} \left(\frac{\lambda_r}{\lambda_a} \right)^{\frac{\lambda_a}{\lambda_r - \lambda_a}}, \quad (2)$$

$$u^{\text{Mie}}(r) = C\epsilon \left(\left(\frac{\sigma}{r} \right)^{\lambda_r} - \left(\frac{\sigma}{r} \right)^{\lambda_a} \right), \quad (3)$$

where ϵ , λ_r , and λ_a , correspond to the potential depth, and the repulsive and attractive exponents respectively. The Helmholtz-explicit form of SAFT-VR Mie is given by,

$$\frac{A}{Nk_B T} = \frac{A_{\text{ideal}}}{Nk_B T} + \frac{A_{\text{mono.}}}{Nk_B T} + \frac{A_{\text{chain}}}{Nk_B T} + \frac{A_{\text{assoc.}}}{Nk_B T}, \quad (4)$$

where A is the Helmholtz free energy, N is the number of particles, T is temperature, and k_B is the Boltzmann constant. “mono.” refers to the contribution from monomers interacting via a Mie forcefield, “chain” the contribution from the formation of chains of length m , and “assoc.” the interactions between molecules with association sites. As we only consider alkanes, we do not consider the association term and fix λ_a to 6. We use the BasicIdeal model provided by Clapeyron.jl for the ideal contribution, which only considers translational modes.

2.5. Physics-Informed Loss Function

The optimisation objective, or “loss function”, maps from the output of the ML model to a

scalar error metric to be minimised. Typically, automatic differentiation (AD) is used to evaluate the gradient of the error metric with respect to the ML model parameters, informing a gradient-step, or model update.

2.5.1 Property Solvers

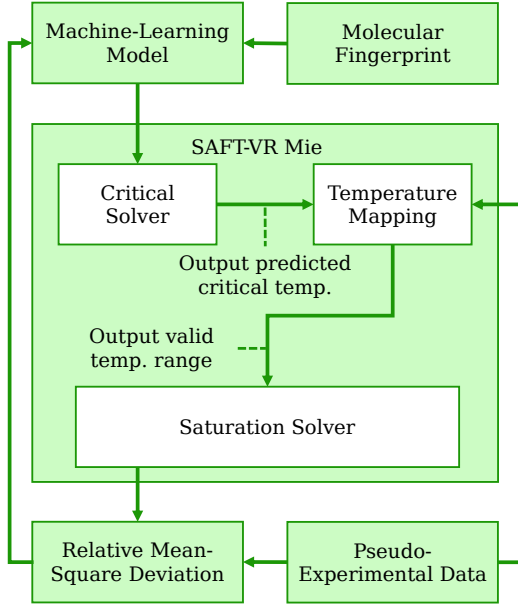


Figure 4: Internal structure of the SAFT-VR Mie block.

Much of the challenge of our approach is robustly and quickly solving for the saturation envelope in an AD-compatible manner, given an arbitrary set of SAFT-VR Mie parameters. Each epoch in our workflow involves 80 critical-point solves and 4000 saturation calculations. Should these begin to fail, the model training may become unstable, increasing the epoch times and slowing the convergence.

To solve for the critical and saturation properties, we use the functions provided by Clapeyron.jl [16], `crit_pure` and `saturation_pressure`. The critical point is defined using Equations 5 and 6, and the saturation envelope using Equations 7 and 8. μ is the chemical potential, V is the volume, V_L is the liquid volume, and V_V is the vapour volume.

$$\left(\frac{\partial^2 A}{\partial V^2}\right)_T = 0 \quad (5)$$

$$\left(\frac{\partial^3 A}{\partial V^3}\right)_T = 0 \quad (6)$$

$$p(V_L, T) = p(V_V, T) \quad (7)$$

$$\mu(V_L, T) = \mu(V_V, T) \quad (8)$$

To enhance the convergence characteristics of the property solvers, we implement a cache of the critical point and the lowest temperature on the saturation envelope. The data obtained for each temperature are then used as the initial guess for the next saturation-property calculation (Algorithm 1).

Algorithm 1 Calculate saturation envelope

Require: ‘mol_cache’ dictionary, molecule name, SAFT parameters

```

1:  $\Pi \leftarrow$  SAFT parameters
2:  $x_0 \leftarrow$  mol_cache[mol_name]
3:  $\mathbf{v} \leftarrow []$ 
4: for  $T = T_{\min}$  to  $T_{\max}$  do
5:    $\text{result} \leftarrow \text{saturation\_pressure}(\Pi, T, x_0)$ 
6:    $x_0 \leftarrow \text{result}$ 
7:    $\text{push\_back}(\mathbf{v}, \text{result})$ 
8: end for
9: return  $\mathbf{v}$ 

```

To obtain the derivatives of the saturation pressure and liquid density, we follow the method described by Winter et al. [11] where a final iteration of a property solver is defined as a “perfect Newton step” (Equation 9 and 10). Using the optimality conditions defined above, we extend this approach to the critical-property solver (Equation 11), enabling us to include the critical temperature in the loss function. V is the specific volume, sat denotes the property is evaluated at saturated conditions, p is pressure, crit. denotes the property is evaluated at the critical point and $*$ denotes a converged value.

$$V_{\text{sat,L}} = V_{\text{sat,L}}^* - \frac{p(V_{\text{sat,L}}^*, T) - p_{\text{sat}}}{\frac{\partial p(V_{\text{sat,L}}^*, T)}{\partial V}} \quad (9)$$

$$p_{\text{sat}} = \frac{A(V_{\text{sat,L}}^*, T) - A(V_{\text{sat,V}}^*, T)}{V_{\text{sat,V}}^* - V_{\text{sat,L}}^*} \quad (10)$$

$$T_{\text{crit.}} = T_{\text{crit.}}^* - \frac{\frac{\partial^2 A(V_{\text{crit.}}^*, T_{\text{crit.}}^*)}{\partial V^2}}{\frac{\partial^3 A(V_{\text{crit.}}^*, T_{\text{crit.}}^*)}{\partial^2 V \partial T}} \quad (11)$$

Equations 9, 10, and 11 enable us to solve for all required thermodynamic properties outside of the loss function, where writing AD-compatible code is not necessary. A vector of converged solutions is then passed into the loss function.

2.5.2 Temperature Mapping

A common hurdle encountered when fitting EoS parameters to saturation data is inconsistency between the experimental and predicted-property temperature ranges. If experimental saturation data exists above the critical temperature of the predicted phase envelope, it is necessary to introduce a mapping if the data are to be compared. We explored three approaches to this mapping.

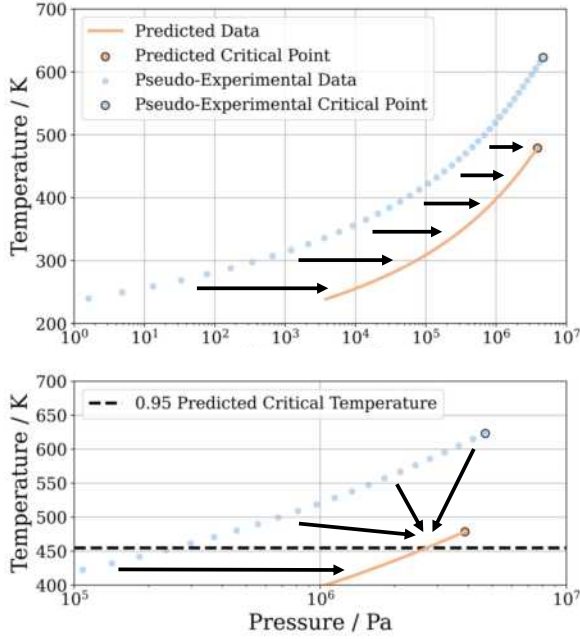


Figure 5: Illustration of temperature mapping applied during training (Equation 13)

In the first approach, all experimental data points above the critical point of the current predicted phase envelope are ignored (Equation 12). Per Equation 13, we also map all temperatures greater than $0.95T_{c,iter}$ to $0.95T_{c,iter}$ (Figure 5). Finally, we compare saturation pressures and saturated liquid densities at the same reduced temperature (Equation 14). As the mapping between real and reduced temperature is non-unique, we also include the critical point of each compound in the error calculation.

$$T_{\text{ignore}} = \begin{cases} T_{\text{exp}} & \text{if } T_{\text{exp}} < T_{c,iter} \\ \text{nothing} & \text{otherwise} \end{cases} \quad (12)$$

$$T_{\text{redirect}} = \begin{cases} T_{\text{exp}} & \text{if } T_{\text{exp}} < T_{c,iter} \\ T_{c,iter} & \text{otherwise} \end{cases} \quad (13)$$

$$T_{\text{reduced compare}} = T_{r,\text{exp}} \cdot T_{c,iter} \quad (14)$$

Here, T_{exp} is the temperature at which the experimental data is sampled, $T_{c,iter}$ is the critical temperature calculated from the parameters predicted by the current iteration of the model, and $T_{r,\text{exp}}$ is the reduced temperature at which the experimental data is sampled, defined with the experimental critical point.

2.6. Pseudo-Experimental Data

Given the predicted phase envelope from the SAFT-VR Mie block, the RMSD is computed relative to a reference. As we aim to demonstrate the utility of our workflow, we train our model on pseudo-experimental data generated from PCP-SAFT parameters [17]. We consider the saturation pressures and saturated liquid densities [18] of 80 alkanes.

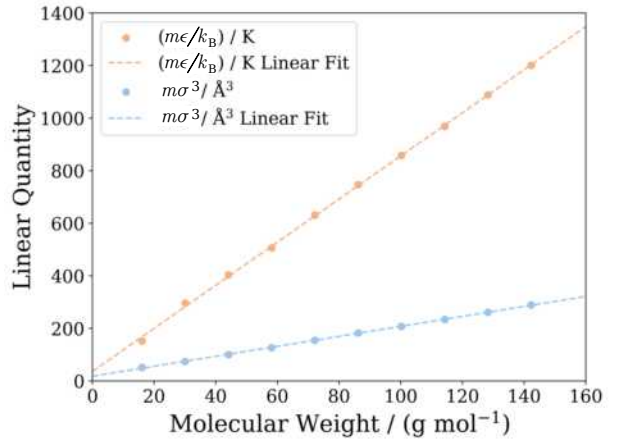


Figure 6: SAFT-VR Mie parameter trends for linear alkanes [15, 19].

We train our model on alkanes as we are interested in obtaining physically significant parameters. The SAFT parameters for alkanes form linear trends in molecular weight, so comparison with these reference curves presents a means of validating the parameters output by our ML model. Two of these trends are illustrated in Figure 6, and Equations 15, 16, and 17 are the fitted lines for each parameter combination, where M is the molecular weight. α_{Mie} is a modified van der Waals-like attractive constant [15] defined by Equation 18.

$$m\sigma^3 = 1.898M + 18.28 \quad (15)$$

$$m\frac{\epsilon}{k_B} = 8.203M + 36.25 \quad (16)$$

$$m \frac{\alpha_{\text{Mie}}}{k_{\text{B}}} = 3601M - 25020 \quad (17)$$

$$\alpha_{\text{Mie}} = -2\pi C\epsilon\sigma^3 \left(\frac{1}{\lambda_r - 3} - \frac{1}{\lambda_a - 3} \right) \quad (18)$$

2.7. Pre-Training on PCP-SAFT Parameters

As each iteration of the training algorithm may be quite slow, we accelerate convergence by “pre-training” our ML model on PCP-SAFT parameters presented by Esper et al. [17]. This significantly improved the initial guess of the neural network state and is a form of transfer learning [20]. λ_r is not a parameter of PCP-SAFT, so we set its value to be normally distributed about 25 during pre-training, with a standard deviation of 5.

2.8. Model Validation

Interested in the predictive capabilities of our ML model, we conduct a five-fold cross-validation to evaluate its performance. This means that we divided the 80 alkanes into five groups (folds), four of which are used to train the model. The fifth fold is reserved for model testing. In repeating this process five times, a “validation set” of the 80 alkanes is obtained, where all data are extrapolated from a set of four training folds.

We distribute alkanes among the five folds using three methods. Ordering by molecular weight, we select the groups randomly, choose each group of sixteen adjacent species (“stratified”) or choose species at regular intervals in this list (“interlaced”). The species at positions one, six, eleven, etc. formed a group, for example.

Unless otherwise indicated, we present results from the validation set. Additionally, we omit methane from our analysis as it does not

contain any CH_3 or CH_2 groups, predicting methane’s properties from the other alkanes is consequently very difficult. We find that the validation data for methane is always an extreme outlier.

The accuracy of the properties outputted by the model was assessed using the Average Percentage Deviation (APD) and Average Absolute Deviation (AAD), as defined by Equation 19 and 20, respectively. Explain variables.

$$\text{APD} = \frac{100}{N} \sum_{i=1}^N \left| \frac{x_{i,\text{pred.}} - x_{i,\text{ref.}}}{x_{i,\text{ref.}}} \right| \quad (19)$$

$$\text{AAD} = \frac{1}{N} \sum_{i=1}^N |x_{i,\text{pred.}} - x_{i,\text{ref.}}| \quad (20)$$

3. Results and Discussion

We present the validation and batch losses for the three temperature mappings in Table 3. The losses for the reduced approach were low, but the APDs for three of the four computed properties are the highest of the three methods. The ignore-points and redirect-points methods were very similar in their performance, possibly because we tended to overshoot the pseudo-experimental critical point throughout training. When this happens, there is no need to introduce a temperature mapping, so the progression of the two models is near-identical. Redirecting the points produced a slightly better validation batch loss. As such, all future results use this mapping.

The evolution of the model parameters as training progresses is shown in Figure 7. As shown, the parameters converge to the reference lines determined in section 2.6, highlighting their consistency with known parameters by the end of training. This suggests our ML model successfully “learnt” the mapping between a molecular fingerprint and the SAFT-VR Mie parameters.

Metric	Ignore Points	Reduced Temperature	Redirect Points
Training Batch Loss	0.0319	0.0108	0.0319
Validation Batch Loss	0.726	0.0507	0.721
Sat. Pressure APD / %	69.7	134	69.7
Sat. Liq. Density APD / %	8.82	8.21	8.82
Sat. Vap. Density APD / %	72.5	138	72.5
Isobaric Heat Capacity APD / %	9.93	10.3	9.93

Table 3: Comparison of the temperature-mapping approaches after 500 training epochs. Methane was omitted from this analysis, the data are averages across all folds, and the interlaced cross-validation was used.

Metric	Sat. Pressure / bar	Sat. Liq. Density / (mol l ⁻¹)	Sat. Vap. Density / (mol m ⁻³)	Isobaric Heat Capacity / (J mol ⁻¹ K ⁻¹)
Mean AAD	0.758	0.428	42.5	12.0
Median AAD	0.0675	0.253	2.76	9.19
Mean APD / %	58.7	6.80	62.2	14.0
Median APD / %	40.4	5.23	45.4	12.1

Table 4: AADs and APDs for properties obtained from the model SAFT-VR Mie parameters. These data are averages over the five validation folds, excluding methane. The isobaric heat capacity was computed at 1 bar.

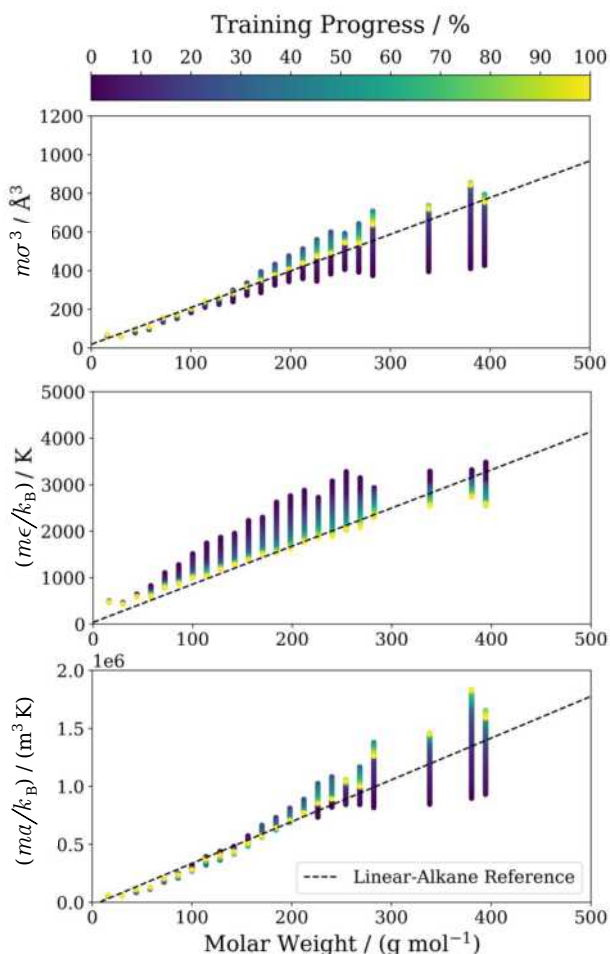


Figure 7: Linear-alkane parameter evolution during training compared with known trends. The R^2 values of the final predicted parameters relative to the reference lines were 0.943, 0.974, and 0.866 for $m\epsilon$, $m\sigma^3$, and $m\alpha$, respectively.

Given the validity of the predicted parameters, we present Clausius-Clapeyron plots for a selection of alkanes in Figure 8. The predicted curves generally align with the pseudo-experimental data, substantiating the success

of our workflow from the perspective of the thermodynamic properties.

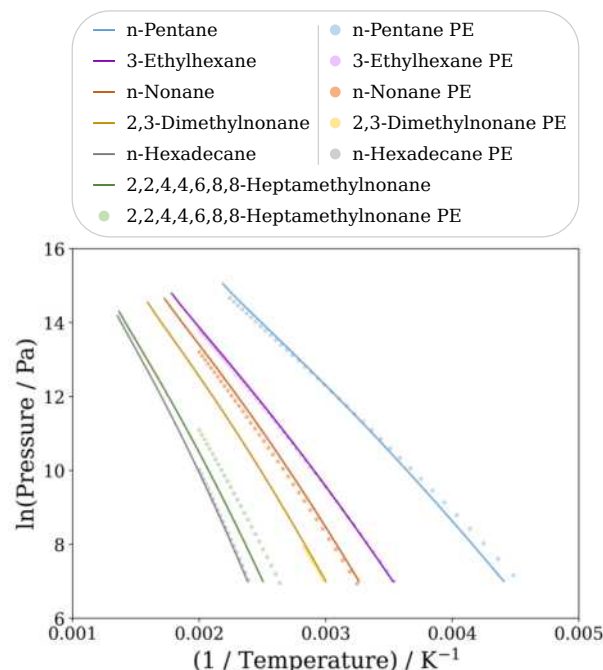


Figure 8: Clausius-Clapeyron plots for a selection of species in the validation set (PE: Pseudo-Experimental Data).

The AADs and APDs for saturation pressures, saturated vapour and liquid densities, and isobaric heat capacities are presented in Figure 9 as box plots. The mean APDs of the saturated liquid density and saturation pressure are 6.90% and 58.7%, respectively. The saturation pressure APD is high. However, looking at its AAD in Table 4, the median is just 0.0675 bar. This suggests that the APD was distorted by outliers in the low-pressure range, which is seen in Figure 9. The saturation pressure mean AAD is also about an order of magnitude larger than the median, similarly indicating distortion due to outliers.

The accuracy of the predicted saturated liq-

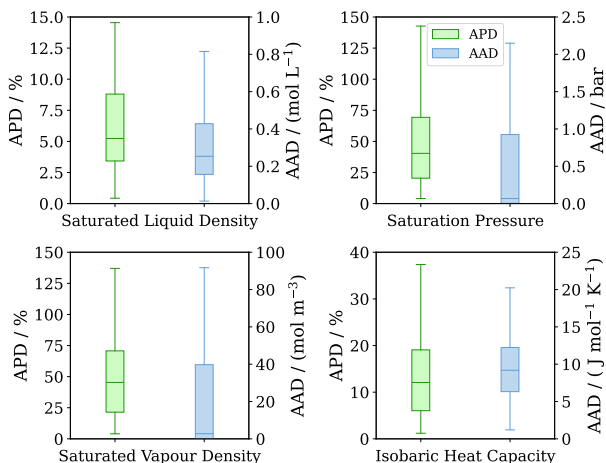


Figure 9: AADs and APDs of some thermodynamic properties using our model. The isobaric heat capacity is computed at 1 bar.

uid density and saturation pressure data again suggests that the model succeeds in predicting the training properties from a molecular fingerprint. However, the benefits of a physics-informed ML model are most apparent in the results for isobaric heat capacity. The model was not trained on any isobaric heat capacity data, but the mean APD is just 14%. Integrating SAFT-VR Mie within the machine learning model increases the flexibility of the model; this suggests that any type of thermodynamic data could be used in training our model, irrespective of the amount of data of that type available. Should real experimental data be used with our workflow, this flexibility would likely be very practical, given that experimental data ranges, amounts, and types are often highly varied.

The results above were obtained using a 512-bit atom-pair fingerprint, though we also considered other lengths (Table 5). We see that the validation batch loss generally decreases

as fingerprint length increases, suggesting that inputting a more detailed representation of a molecule may improve the model performance.

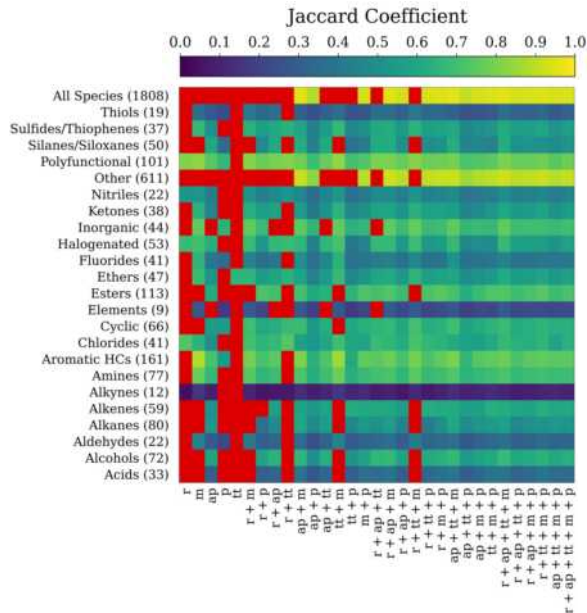


Figure 10: Fingerprint performance for 1808 of the compounds considered by Esper et al. [17], having removed stereoisomers. Red cells indicate combinations where at least one duplicate occurred in the fingerprints generated. The fingerprints were initially 16384 bits long, and a radius of 6 was used, where applicable. r, m, ap, p, and tt refer to rdkit, Morgan, atom-pair, pattern, and topological torsion fingerprint-generation algorithms, respectively.

We also explored other fingerprint-generation methods, including a Morgan algorithm, which enumerates the type of atom within a certain number of bonds ("radius") of every atom in a species. We found this approach generated the

Number Bits (Initial)	Number Bits (Reduced)	Training Batch Loss	Validation Batch Loss	Average Jaccard Coefficient
512	259	0.0287	0.0622	0.7241
1024	327	0.0268	0.0610	0.7486
2048	363	0.0156	0.0433	0.7648
4096	369	0.0126	0.0390	0.7679
8192	370	0.0102	0.0393	0.7684
16384	370	0.0118	0.0414	0.7684

Table 5: ML model performance for different atom-pair molecular fingerprint lengths, indicating that a more-detailed representation of the input molecules increases the accuracy of thermodynamic-property estimates.

same binary vector for multiple alkanes in our training set unless a very large radius was specified encompassing the entirety of each molecule we consider. Specifying ever-larger fingerprint radii is impractical, but the ML model cannot differentiate between species if the fingerprints are identical.

The Morgan algorithm was unsuited to our system of alkanes, but it may be appropriate for other families of compounds. In Figure 10, the red squares indicate where a fingerprint produces any duplicate vectors for different families of compounds. To generate Figure 10 we consider five fingerprint-generation algorithms as well as the twenty-six combinations that are possible from this set of five. We concatenate the bit vector outputted by each algorithm to combine the fingerprints.

Where a fingerprint generates a set of unique bit-vectors for a family of compounds, we illustrate the average Jaccard Coefficient in Figure 10. The Jaccard Coefficient (Equation 21) measures the similarity between two sets, where values of 1 and 0 indicate that the lists are identical or completely different, respectively. The Jaccard Coefficient for each cell in Figure 10 is the average value over all pairs of vectors for each list of fingerprints generated for a chemical family.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (21)$$

The similarity between molecular fingerprints has implications for the success of our workflow. It would be more difficult to differentiate between species with near-identical fingerprints, for example. Equally, it is possible that the predictive power of a model would be hampered if the fingerprints in the training set were very different; inaccuracy may arise in interpolating between highly dissimilar bit vectors. The average Jaccard Coefficient for our system of 80 alkanes only varies from 0.72 to 0.77 (Table 5), so it is difficult to assess the effect of changing fingerprint similarity on our model’s predictive capabilities. The effect of fingerprint similarity may be more prevalent if simultaneously training on multiple families of compounds.

4. Conclusion

In this study, we developed a novel methodology for training ML models to predict EoS parameters. A physics-informed loss function was central to our approach; we incorporated SAFT-VR Mie into our workflow and directly trained on

pseudo-experimental data for 80 alkanes. The SAFT-VR Mie parameters outputted by our ML model are consistent with known linear trends in molecular weight. We also performed a five-fold cross-validation with our model, verifying the accuracy of the output saturation pressures and saturated liquid densities. The benefit of a physics-informed loss function was particularly apparent in our results for isobaric heat capacity. The mean APD between our model outputs and the pseudo-experimental data was just 14%, even though our model was not trained on heat capacity data.

We also explored the characteristics of different fingerprint-generation algorithms, ensuring a unique representation of each molecule in our training set was inputted into the ML model. We extended this analysis to other fingerprint algorithms and chemical families, providing a reference for future work.

This work provides the first physics-informed neural network to use SAFT-VR Mie and is the first open-source ML model to incorporate a SAFT-type EoS.

5. Future Work

Having demonstrated the functionality of an ML model incorporating SAFT-VR Mie using pseudo-experimental data, future work could involve using experimental data with our workflow.

Following that, there are many opportunities for improvement in the ML-model architecture. This could take the form of modifying the number of layers and nodes within the MLP, exploring different molecular representations such as those used in graph neural networks, or adding in more-complex layers like multi-head attention, allowing for a more-expressive model.

Finally, modifications could be made to the physics within the loss function. Primarily, the addition of the association contribution would open up the addition of many new compounds to the training set, and the introduction of a SAFT- γ Mie parameterisation could present an interesting avenue of exploration, enforcing a group-contribution-like approach, while allowing for whole-molecule effects.

Acknowledgements

We would like to thank Pierre Walker and Andrés Riedemann for their invaluable help throughout this project. We additionally thank Daniel McGlinchey.

References

- [1] Nick D. Austin, Nikolaos V. Sahinidis, and Daniel W. Trahan. "Computer-aided molecular design: An introduction and review of tools, applications, and solution techniques". In: *Chemical Engineering Research and Design* 116 (Dec. 2016), pp. 2–26.
- [2] Seimens. *gPROMS*. 2023.
- [3] K.G. Joback and R.C. Reid. "Estimation of Pure-Component Properties from Group-Contributions". In: *Chemical Engineering Communications* 57.1-6 (July 1987), pp. 233–243.
- [4] Leonidas Constantinou and Rafiqul Gani. "New group contribution method for estimating properties of pure compounds". In: *AIChE Journal* 40.10 (Oct. 1994), pp. 1697–1710.
- [5] Vasileios Papaioannou et al. "Group contribution methodology based on the statistical associating fluid theory for heteronuclear molecules formed from Mie segments". In: *The Journal of Chemical Physics* 140.5 (Feb. 2014), p. 054107.
- [6] Pierre J. Walker and Andrew J. Haslam. "A New Predictive Group-Contribution Ideal-Heat-Capacity Model and Its Influence on Second-Derivative Properties Calculated Using a Free-Energy Equation of State". In: *Journal of Chemical & Engineering Data* 65.12 (Dec. 2020), pp. 5809–5829.
- [7] Sayandeep Biswas et al. "Predicting Critical Properties and Acentric Factors of Fluids Using Multitask Machine Learning". In: *Journal of Chemical Information and Modeling* 63.15 (Aug. 2023), pp. 4574–4588.
- [8] Gustavo Chaparro and Erich A. Müller. "Development of thermodynamically consistent machine-learning equations of state: Application to the Mie fluid". In: *The Journal of Chemical Physics* 158.18 (May 2023), p. 184505.
- [9] Kobi Felton et al. *ML-SAFT: A machine learning framework for PCP-SAFT parameter prediction*. preprint. Chemistry, May 2023.
- [10] Jonas Habicht, Gabriele Sadowski, and Christoph Brandenbusch. "Fitting Error vs Parameter Performance-How to Choose Reliable PC-SAFT Pure-Component Parameters by Physics-Informed Machine Learning". In: *Journal of Chemical & Engineering Data* (Oct. 2023), acs.jced.3c00411.
- [11] Benedikt Winter et al. *Understanding the language of molecules: Predicting pure component parameters for the PC-SAFT equation of state from SMILES*. Sept. 2023.
- [12] Raymond E. Carhart, Dennis H. Smith, and R. Venkataraghavan. "Atom pairs as molecular features in structure-activity studies: definition and applications". In: *Journal of Chemical Information and Computer Sciences* 25.2 (May 1985), pp. 64–73.
- [13] Günter Klambauer et al. *Self-Normalizing Neural Networks*. Sept. 2017.
- [14] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. Jan. 2017.
- [15] Thomas Lafitte et al. "Accurate statistical associating fluid theory for chain molecules formed from Mie segments". In: *The Journal of Chemical Physics* 139.15 (Oct. 2013), p. 154504.
- [16] Pierre J. Walker, Hon-Wa Yew, and Andrés Riedemann. "Clapeyron.jl: An Extensible, Open-Source Fluid Thermodynamics Toolkit". In: *Industrial & Engineering Chemistry Research* 61.20 (May 2022), pp. 7130–7153.
- [17] Timm Esper et al. "PCP-SAFT Parameters of Pure Substances Using Large Experimental Databases". In: *Industrial & Engineering Chemistry Research* 62.37 (Sept. 2023), pp. 15300–15310.
- [18] Nicolás Ramírez-Vélez et al. "Parameterization of SAFT Models: Analysis of Different Parameter Estimation Strategies and Application to the Development of a Comprehensive Database of PC-SAFT Molecular Parameters". In: *Journal of Chemical & Engineering Data* 65.12 (Dec. 2020), pp. 5920–5932.
- [19] Pouya Hosseinifar, Mehdi Assareh, and Cyrus Ghotbi. "Developing a new model for the determination of petroleum fraction PC-SAFT parameters to model reservoir fluids". In: *Fluid Phase Equilibria* 412 (Mar. 2016), pp. 145–157.
- [20] Fuzhen Zhuang et al. *A Comprehensive Survey on Transfer Learning*. June 2020.

Author Index

Linked to paper number

Adejumobi, Oluwanifemi E.	60	Huang, Shannan	30
Affrie Shah, Nur Alya Fariesha	40	Ho, Joon Yean	12
Al-Wsaifer, Naser	13	Hong, Weng Hin	31
Alias, Afiqah	49	Izaga, Mencia	14
Ayyar, Pavan	18	Jones, Mia Sophie	33
Bai, Hannah	3	Kanjanabult, Khan	72
Barnes Bush, Finlay	10	Karas, Diketso	11
Barr, Helen	52	Kendell, Osei	26
Bennett-Gant, Tazz	60	Khor, Li Xun	31
Berry, Joseph	59	Kohn, Jeremy	4
Bhugun, Aashna Jaya	40	Kovacs, Alexander	62
Boonnasitha, Serene	72	Kumaran, Sanjay	38
Boschin, Claudia	9	Ke, David	39
Chan, Chee Hung	12	Lai, Samson	53
Cheung, Joyce	47	Lam, Elton	48
Chen, Haochuan	57	Langi, Esha	39
Chen, Qiyun	71	Le Boutillier, Noah	59
Cheong, Bing Zheng	68	Lewis, Elinor	6
Chow, Rouxuan	68	Liang, Shuxuan	54
Christodoulou, Ermis	30	Lim, Amanda Jiayi	37
Costa, João	11	Lukman, Farhad	2
Dita, Alexa	10	Lum, Cheryl	70
Delano, Feyintoluwa Mojinyinoluwa	56	Mabon, Sinclair	46
Demirdesen, Defne	67	Mahzan, Nursyazana	25
Derqui Serrano, Ana	33	Malik, Samreen	51
Dharmasena, Shiwaanan	3	Mansuri, Humaira	25
Diong, Joella	65	Marqueste, Solen	23
Emirzadeoglulari, Eran	15	Marsh, Matthew	64
Evripidou, Christos	36	McGlinchey, Daniel	38
Fadlelmawla, Ali	17	Must, Andreas	32
Fraser, Harrison	44	Mustapha, Babafemi	44
Fung, Andrew	23	Naoki, Fukushima	7
Gadaloff, Michael	73	Nguyen-Phuong, Thao Vy	6
Geller, Jay	55	Nicolaidis, Alkmini	13
Gerard, Nicholas	48	Nwokolo, Victoria Ebubechukwu	56
Ghandour, Hashem	14	Nyayadhish, Radhika	63
Gill, Ajai	42	O'Driscoll, Cahan	43
Glover, Alex	35	Pakradouni, Maria	19
Godfrey, Nicole	21	Paoli, Luc	73
Gosain, Khushali	66	Parias Moreno de los Rios, Begoña	52
Gu, Kexian	8	Parkinson, Tom	18
Hakim, Adnan	69	Pchelintsev, Ivan	32
Han, Ruishan	2	Qiu, Yiyao	51
Harashima, Kenji	20	Reents, Konrad	62
He, Jiahui	50	Roberts, Kadiy'a	26
Huang, John	22	Robinson, Brooklyn	67

Shamash, David	16	Watts, Adam John	34
Shao, Yiheng	47	Wei, Qing Li	21
Shenoy, Rohan	9	Wen, Di	1
Simonetou, Barbara	5	White, James	50
Sivarasan, Kabishan	69	Wong, Amanda	65
Smith, Emma	66	Wong, Zen	53
Solomon, Isabel	42	Woo, Jan Yii	45
Spencer, George	64	Wright, Jonathan	29
Stewart-Tull Joseph John	34	Yang, Yijun	71
Suthagar, Aaron	29	Yang, Yu Hang	70
Tala, Achour	7	Yao, Putian	27
Tandon, Alexander	58	Yee Tong Wah, Jeremy	16
Teegala, Sulekh	45	Yeung, Inny	5
Teow, Denzel Martin	43	Yong, Gordon Wu Shun Yong	37
Thakrar, Sathya	24	Yu, Luen	54
Toh, Ke Jing	49	Yu, Junxin	61
Tomes, Oliver	22	Zafet, Mikaela	19
Tran, Yung	20	Zavate, Antonia Georgiana	28
Triguero Munoz, Mario	15	Zhan, Lucia	55
Trisno, Gabriela	63	Zhang, Tianze	41
Vadukul, Anand	24	Zhang, Zhenran	1
Vaughan, Alexander	58	Zheng, Yunxiang	36
Verkammen, Vincent	4	Zhu, Yingqing	61
Waqar, Shihabuddeen	17	Zhuang, Zhongqi	46
Wang, Barry	41		
Wang, Minqi	27		

Supervisor Index

Linked to paper number

Adjiman, Claire	34, 39, 68
Cabral, João	52
Chachuat, Benoit	10, 11, 48
del Rio Chanona, Antonio	4, 28, 32, 55
Elani, Yuval	5, 26
Eslava, Salvador	1, 18, 47
Fennell, Paul	3, 7, 45
Galindo, Amparo	24, 44
Hallett, Jason	25, 36, 46
Hankin, Anna	31, 62
Hammond, Ceri	40, 54
Hawkes, Adam	59
Hellgardt, Klaus	35, 43, 58
Heng, Jerry	9, 49, 53
Jackson, George	13, 73
Kalliadasis, Serafim	15
Kazarian, Sergei	56
Kontoravdi, Cleo	22, 27, 72
Li, Kang	57, 61
Luckham, Paul	2
Markides, Christos	29, 41, 69
Matar, Omar K.	51
Müller, Erich A.	38
Petit, Camille	12, 19, 37
Pini, Ronny	20, 60, 66
Papathanasiou, Maria	64, 67, 70
Polizzi, Karen	23, 63
Rinaldi, Roberto	33, 42
Shah, Nilay	6, 16, 21
Song, Qilei	65
Tighe, Chris	8
Titirici, Magda	14, 50, 71
Trusler, Martin	30
Yetisen, Ali	17

