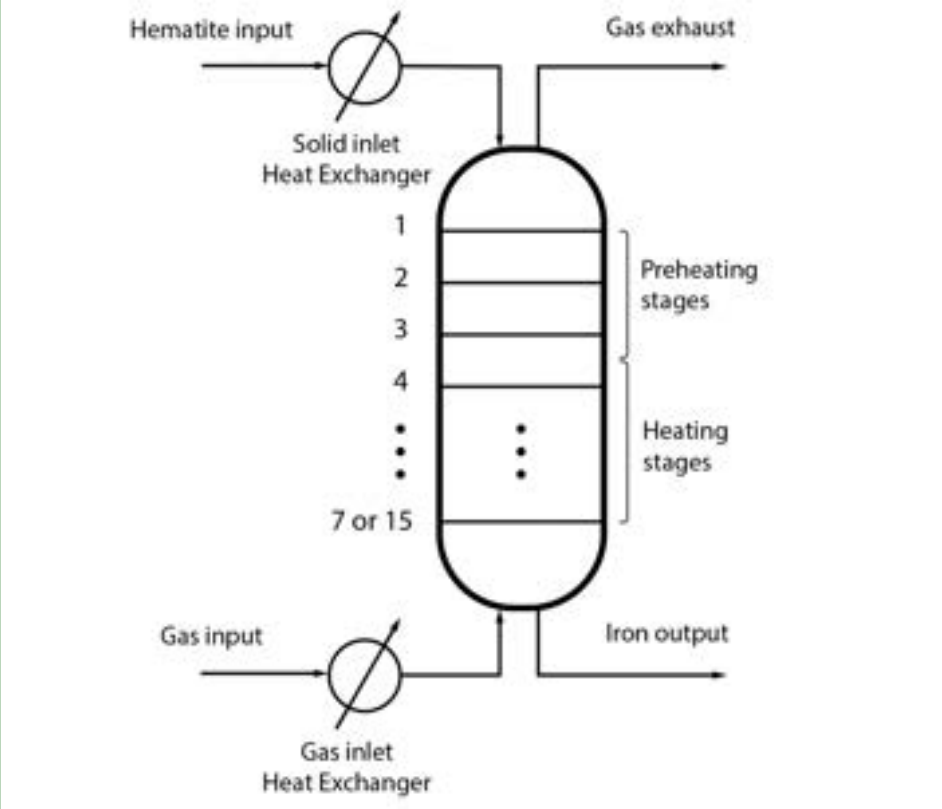
CHEMICAL ENGINEERING RESEARCH

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

VOLUME 5



Edited by Erich A. Müller

February, 2023

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 5

2023

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.

E-ISBN 978-1-9160050-4-4

First published in 2023

Preface

This volume of *Chemical Engineering Research* collects the unedited research project reports written by 4th year undergraduates (Class of 2023) of the M.Eng. course on Chemical Engineering in the Department of Chemical Engineering at Imperial College London. The research project spans for one term (Autumn) during the last year of the career and has an emphasis on independence, ability to plan and pursue original project work for an extended period, to produce a high quality report, and to 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 obviously only touch on a small number of research groups and interests. For a full description of the research at the department, the reader is referred to the departmental website[1].

The papers presented are in no particular order and they are identified by a manuscript number. 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 of the reports are missing, being embargoed, as they contain confidential information. A few of the reports correspond to industrial internships, called LINK projects, performed in collaboration with Shell.

Cover figure corresponds to a diagram of a simplified ZESTY reactor model (taken from the work of Javier Monteliu and Jien Feung Jason Goh, manuscript 12).

London, February 2023

[1] https://www.imperial.ac.uk/chemical-engineering

Title Index

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

Author and Supervisor index at the end of the book.

Hybrid ZIF-8@Porous Graphene Oxide (PGO) Hollow Fibre Membranes with Improved Molecular Sieving Property for Nanofiltration

Jianing Li and Zihao Li
Department of Chemical Engineering, Imperial College London, U.K.

Abstract
Membranes with ultrafast water transport and high molecular rejection are desired for various separation processes. Graphene oxide (GO) based membranes have demonstrated their potential for nanofiltration. Despite the promising water separation performance, the permeance and molecular rejection of the GO-based membranes still require improvement. In this study, high-performance zeolitic imidazolate framework-8 (ZIF-8) hybridised porous graphene oxide (PGO) membranes were fabricated via *in situ* growth of ZIF-8 nanocrystals within the PGO membranes' lamellar structure. The crystallisation of ZIF-8 mitigated the rejection loss due to nonselective defects and slightly enlarged interlayer d-spacing of the PGO membranes. The ZIF-8 hybridised membranes made of PGO nanosheets with 5 hours of mild-etching treatment (ZIF-8@PGO_5h) maintained a pure water permeance of 2.81 $LMHbar^{-1}$. Various dye molecules were used to measure the permselectivity of the membranes on dye-containing wastewater. The ZIF-8@PGO_5h membrane showed a molecular weight cut-off (MWCO) of approximately 314 $g\ mol^{-1}$, indicating the promising future of MOFs hybridised PGO hollow fibre membranes for nanofiltration.

Keywords: Graphene Oxide, Porous Graphene Oxide, Metal-organic Framework, ZIF-8, Hollow Fibre Membranes, Nanofiltration, Dye Removal

1. Introduction

In recent decades, rapid industrialisation and population growth have intensified freshwater scarcity and lack of safe drinking water, even threatening people's lives, especially in developing areas. Membrane technology significantly improves the separation processes efficiency in wastewater treatment and seawater desalination [1]. Compared with conventional separation technologies, nanofiltration is more cost and energy- effective and environmentally friendly [2]. High-performance membranes usually have one or more features: ultrahigh pure water permeance, low molecular weight cut-off, as well as excellent mechanical and chemical stability [3].

Graphene Oxide (GO), a two-dimensional nanomaterial with atomic-scale thickness and high aspect ratio, has been studied extensively. Its tuneable porosity, surface functionalities, and decent chemical and mechanical stability make it an emerging star nano-scale material [4].

GO membranes can effectively retain small molecules and multivalent salts because of the explicit interlayer space between GO flakes [5]. However, inconsistent water permeation observations were reported by various research groups indicating the structural instability of GO membranes [5, 6, 7]. The unstable structure and molecular transport behaviour could be triggered by: (1) the hydrogen bonding between passing water molecules and oxygen functional groups on the edge and the basal plane of the GO membrane and (2) the interlayer interaction between charged solute (e.g., metal ions) and the negatively charged groups on GO membranes [7]. Besides, the water flux could drop tremendously after more than ten hours, indicating that the GO flakes are not packed in the uniform, ordered lamellar structure initially [5].

In this study, PGO-based membranes have been regarded as the starting point. Wu et al. implemented a mild chemical etching method to prepare PGO dispersion and subsequently fabricated PGO membranes on alumina hollow fibre substrates [8]. Ammonia (NH_4OH) and hydrogen peroxide (H_2O_2) were added into GO dispersion, and the dispersion was gently stirred for a specific time. The pore size and density of the PGO nanosheets were well controlled by manipulating the chemical etching time. The water permeation of PGO with 5 hours of chemical etching treatment showed a 23-fold increase compared with the pristine GO membrane. However, compared to pristine GO membranes, its rejection of methyl red (MR) dye molecules declined significantly due to the nonselective voids in the membrane microstructure.

As aforementioned, the water permeation can be improved by increasing the chemical etching time but at the cost of losing molecular permselectivity. Thus, a novel fabrication method and/or material have to be introduced to mitigate the rejection loss while maintaining high water permeance.

Metal-organic frameworks are a new type of microporous crystalline materials formed by self-assembling organic ligands and metal ions through coordination bonds. Distinctive features of MOFs including large specific surface area and pore volume, and tuneable pore size make them promising materials for a wide range of applications in catalysis, adsorption, and separation. However, the formation of defect-free, structurally stable MOFs-based thin-film membranes is a crucial challenge. The fabrication of MOFs@GO

composites is a new concept to take advantage of desired properties of both GO and MOFs to achieve good size-sieving ability and structural integrity. The porous MOFs materials with tailored and uniform pore size can patch the nonselective defects in the GO membranes. The oxygen functional groups on GO nanosheets enable them to form coordination bonds with the metal ions in MOFs. This characteristic not only promotes the long-term stability of the composite membrane even under hydraulic pressure or cross-flow condition but also enables the selective growth of MOFs to boost the separation performance of membranes.

For instance, Ying, Y. et al. intercalated $UiO-66-(COOH)_2$ into GO membranes via pressure-assisted self-assembly (PASA) filtration method on PAN supports [9]. The MOFs@GO membrane with 0.3 mg of wet $UiO-66(Zr)-(COOH)_2$ (corresponding MOF loading, 23.08 wt %) per 1 mg of GO showed a 159% improvement compared with the pristine GO membrane in water permeation and obtained a 99.5 wt% water content in the ethyl acetate-water pervaporation process. The intercalation of $UiO-66(Zr)-(COOH)_2$ successfully reduced the nonselective voids. The PASA technique ensured that the interlayer spacing only changes little even with the presence of MOFs, which gave the membrane a compact structure.

Recently, a highly stable and ultra-permeable hybrid membrane was synthesised by Zhang, W. et al. via *in situ* crystallisation of zeolitic imidazolate framework-8 (ZIF-8) at the edges of freeze-dried GO (f-GO) nanosheets following an ice-templating technique [10]. The ZIF-8 nanocrystals were preferentially grown along the edges of the f-GO nanosheets. This was because Zn^{2+} in ZIF-8 formed a coordination bond with the carboxyl functional group which are predominantly found along the edges of the f-GO nanosheets [4]. The clusters of ZIF-8 nanocrystals filled these major microstructural defects. The resulting water permeation increased 30-fold compared with that of pristine GO membrane and the perm selectivity of the hybrid membrane increased by 6 times. This type of MOFs hybrid GO-based membrane with stable microstructure, high water permeation, and good molecular perm selectivity is the desirable building block for future nanofiltration membranes. However, the neutral dye rejection results showed a molecular weight cut-off (MWCO) of 1300 Da which is considerably high for nanofiltration membranes. Considering the small pore size of ZIF-8 nanoparticles (0.34 nm), this relatively poor size-sieving performance of ZIF-8@GO composite membranes implies that ZIF-8 nanoparticles were not selectively grown on the edges and micro-defects of GO membranes.

In this study, ZIF-8 nanoparticles were grown in PGO membranes on alumina hollow fibre substrates through a modified two-step *in situ* growth approach to mitigate the trade-off between water permeation and molecular permselectivity. During experiments, only GO etched for 5 hours (named PGIO_5h) was used for membrane fabrication. The two-step *in situ* growth method was conducted by soaking PGO membranes in the metal solution (1h /3h /12h) and then in the ligand solution (1h /3h /12h). The two-step in-situ growth allows for the selective growth of ZIF-8 nanoparticles in the lamellar structure of PGO membranes. Through immersing PGO HF membranes in metal solution, metal ions with positive charge coordinated with oxygen functional groups on PGO nanosheets. Immersion in ligand solution leads to ZIF-8 nanoparticle growth. The loading of ZIF-8 nanoparticles in ZIF-8@PGO membranes can be controlled by adjusting metal and ligand immersion time so that water permeation will not be sacrificed.

To evaluate the sieving properties of ZIF-8@PGO HF membranes, two neutral dyes, Methyl Red (MR) and Disperse Red (DR), were used in the rejection test. As the PGO membrane is negatively charged, neutral dyes were used to ensure rejection performance only depends on the size exclusion rather than the electrostatic interaction.

2. Experiments

2.1 Materials

Alumina powder (99.9% metals basis) was obtained from Alpha Aesar. 1-methyl-2-pyrrolidine (NMP) and poly (methyl methacrylate (PMMA)) were selected as the ceramic suspension solvent and binder. Arlacel P135 was supplied by Croda. 99% (metals basis) Zinc nitrate hexahydrate was purchased from ThermoFisher Scientific. 99% 2-Methylimidazole was purchased from Sigma-Aldrich. Methanol (absolute) was provided by VWR. PGO dispersion and aluminium oxide (Al_2O_3) hollow fibre (HF) substrates were made in the lab. Araldite® Epoxy Adhesive was selected to fix hollow fibre substrate on the Swagelok tube fitting. Hydrogen peroxide (H_2O_2) and ammonia hydroxide (NH_4OH) were used in PGO synthesis.

2.2 GO and PGO synthesis

Graphene oxide (GO) was synthesised according to the modified Hummer's method [11, 12]. Briefly, sulfuric acid (H_2SO_4) and oxidising agent potassium permanganate ($KMnO_4$) were gently added to graphite powder in an agitated two-wall glass reactor and thoroughly mixed at 35 °C overnight. Water was added dropwise to dilute the GO suspension and then hydrogen peroxide (H_2O_2) was introduced to remove the excess manganese ions (Mn^{2+}) [13]. The dispersion was then filtered and washed out with diluted hydrochloric acid (HCl) aqueous solution. The resultant GO cake was further dried at room temperature in the vacuum oven for 3 days. The dried GO was deeply washed again using

acetone in bath sonication and vacuum filtered to remove any remaining acids and ensure high purity of GO. After drying GO for at least 3 days under vacuum at room temperature, GO powder is finally ready to be used for homogenous GO dispersion preparation.

Porous graphene oxide (PGO) dispersion was prepared via a mild chemical etching method [8]. NH_4OH and H_2O_2 were added into GO dispersion ($GO/H_2O_2/NH_4OH$:20/1/1 vol) and the dispersion was gently stirred for 5 hours. The PGO dispersion was centrifuged and purified by a dialysis membrane to remove any remaining NH_4OH and H_2O_2.

As how etching time affects the water permeation and nanofiltration performance has been investigated before and it has been confirmed that PGO has shown good performance [8], in this study, 0.1 mg/mL and 0.05mg/mL of PGO dispersion were used in experiments aiming to obtain an optimised recipe of PGO-based membranes.

2.3 Preparation of ceramic suspension

3g of dispersant Arlacel P135, 180 g of NMP and 150 g of alumina powder were transferred into a ceramic jar and mixed with a planetary ball miller at 283 rpm for 48 hours. After that, PMMA was added, and the suspension was further mixed for 48 hours in a roll miller. To evacuate the bubbles before spinning, the suspension was degassed under vacuum for 4 hours.

2.4 Preparation of PGO @ ceramic hollow fibre membranes

The suspension was transferred to a stainless-stell syringe after degassing. The Al_2O_3 hollow fibres were prepared using a combined phase-inversion/sintering process, and sintering was conducted at 1450 °C to improve the mechanical strength [14]. Aluminium Oxide (Al_2O_3) hollow fibre substrates are immersed into acetone and put in the ultrasonic bath for washing. Hollow fibre substrates were put into tube fittings and glued with epoxy to be fixed. The top of the HF substrate was also sealed with epoxy resin. A stronger epoxy was utilised to avoid the softening of epoxy by organic solvent and any possible leakage during membrane fabrication and performance evaluation test. The diluted 0.1 mg/mL PGO dispersion was prepared and sonicated for 1-2 minutes. Hollow fibre substrates, with one end sealed, were dipped into the PGO dispersion solution. Under vacuum filtration, PGO was stacked to the outer surface of substrate and formed a nanosheet with a thickness of 200-400 nm. The thickness of the PGO membranes can be tuned by altering the concentration of the PGO dispersion or the coating time. The accomplished PGO membranes were dried at 40 °C under vacuum for 3 hours to remove any residual water in the nanosheets or substrates. A mild drying temperature is applied to prevent drying-related reduction of PGO membranes.

2.5 Intercalation of ZIF-8 into PGO HF membranes

Zinc nitrate hexahydrate (metal) and 2-Methylimidazole (ligand) are the ingredients for *in situ* ZIF-8 composite growth. DI water and methanol are two solvents that have been used to prepare metal and ligand precursor solutions. For both metal and ligand precursor solutions, concentration and immersion time are two critical parameters influencing the loading of ZIF-8 growing on the pristine PGO membrane.

For metal precursor solution, 3000 ppm was fixed as the concentration of metal precursor solution regardless of solvent type. 1 hour and 3 hours metal soaking time were set in the first step of in-situ growth to allow metal ions to interact with oxygen on PGO membranes.

Regarding the ligand precursor solution, 16000 ppm and 32000 ppm ligand precursor were prepared to test whether ligand concentration affects the number of ZIF-8 crystals and the rejection performance. Similarly, 1 hour, 3 hours and 12 hours ligand immersion time were set up. When the ZIF-8 nanocrystal growth was done, membranes were stored in the dye solution overnight.

2.6 Membrane Characterisation

The pore size and distribution on the basal plane of GO and PGO nanosheets were detected by a high-resolution transmission electron microscope (HR-TEM, JEOLJEM-2100F).

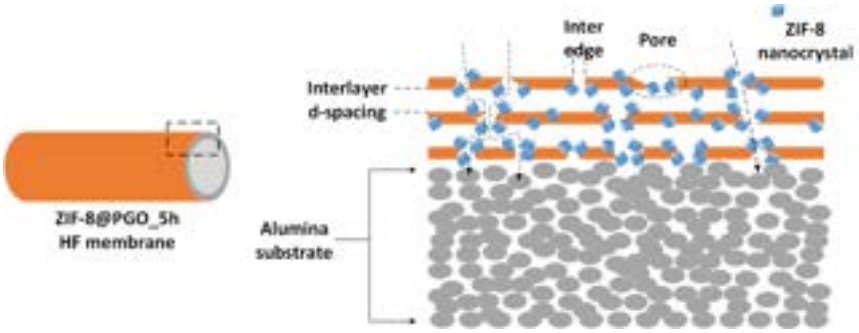


Figure 1 Schematic illustration of ZIF-8@PGO HF membranes

3

The formation of ZIF-8 nanocrystals and the morphology of GO membrane, PGO HF membrane, and ZIF-8@PGO HF membrane with different fabrication methods were observed with a high-resolution field emission gun scanning electron microscope (FEG-SEM, LEO Gemini 1525).

X-ray diffraction (XRD) spectra was collected using X' Pert PANanlytical instrument operated in a 2θ range of 5º - 30º. The voltage and current were set at 40 kV and 20 mA, respectively. The XRD spectra would indicate the successful *in situ* growth of ZIF-8 nanoparticles and could be used to calculate the interlayer d-spacings of GO membranes and PGO membranes based on Bragg's equation:

$$n * \lambda = 2 * d * \sin(\theta)$$

where $n,\ \lambda, d, \theta$ represent a positive integer, the wavelength of the X-ray, the interlayer d-spacing between two nanosheets, and the X-ray incidence angle.

X-ray photoelectron spectroscopy (XPS) was used to determine the binding nature within the ZIF-8@PGO membrane.

2.7 Water Permeation

The pure water permeation and nanofiltration tests were carried out with dead-end filtration system 2. The ZIF-8@PGO HF membrane was mounted into a cell fully filled with water and pressurised using injected nitrogen gas. The mass of collected permeate was recorded every 10 minutes so that the permeance J ($LMHbar^{-1}$) could be calculated:

$$J = \frac{\Delta M}{\rho \cdot 1000 \cdot \Delta t \cdot A \cdot p}$$

where ΔM, ρ, Δt, A, p represent the change in mass (g), the water density (g/cm^3), the change in time (s), the cross-sectional area of membrane (m^2) and the pressure (bar).

Although the permeate contains some dye molecules, as the collected permeate volume and the dye concentration are both small compared to the dye solution in the feed vessel and water permeance was assumed to be equal to the MR or DR permeance in this study.

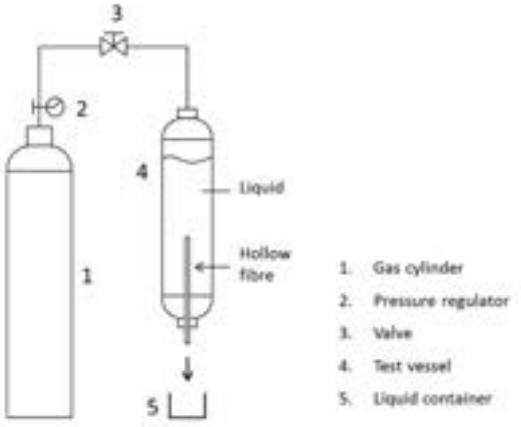


Figure 2 Scheme of the setup for pure water permeation and nanofiltration tests

As the collected permeate volume was rather smaller than that of feed in solution in the test vessel and the dye solution 20 mg/L was very dilute, water permeation tests were conducted with dye solution assuming the density of dye solution is the same as water density.

2.8 Dye rejection test

To wash off the remaining organic ligand groups and prevent the impact of adsorption of the dye molecules, PGO HF membranes were kept in the dye dispersion (Methyl Red (MR) and Disperse Red (DR)) with a concentration of 20 mg/L for 12 hours. Like the water permeation procedure, the membrane connected to the tube fitting was mounted into a cell with dye solution and pressurised by nitrogen. At least 3mL of the permeate was collected, and the permeate concentration was measured by UV-Vis spectrophotometer. Rejection was determined by:

$$R = \left(1 - \frac{C_p}{C_f}\right) \times 100\%$$

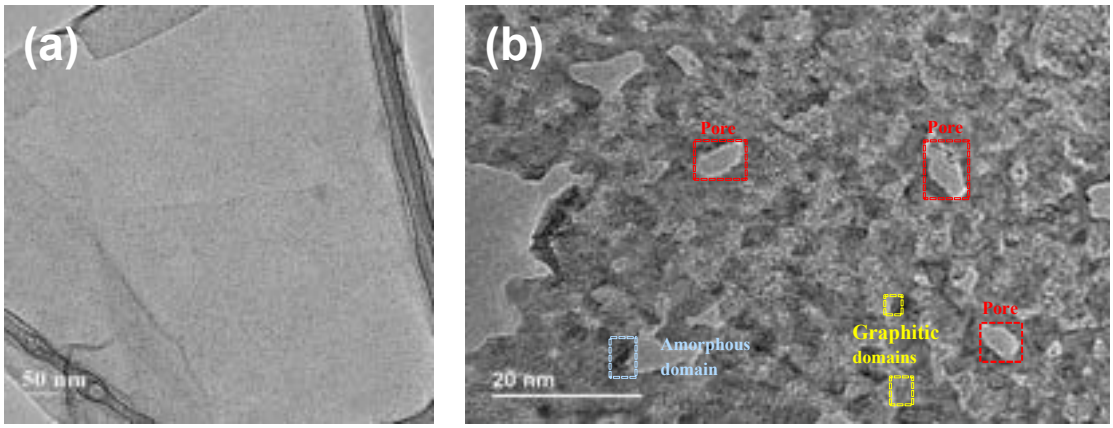


Figure 3 HR-TEM images of (a) GO nanosheet and (b) PGO nanosheet

4

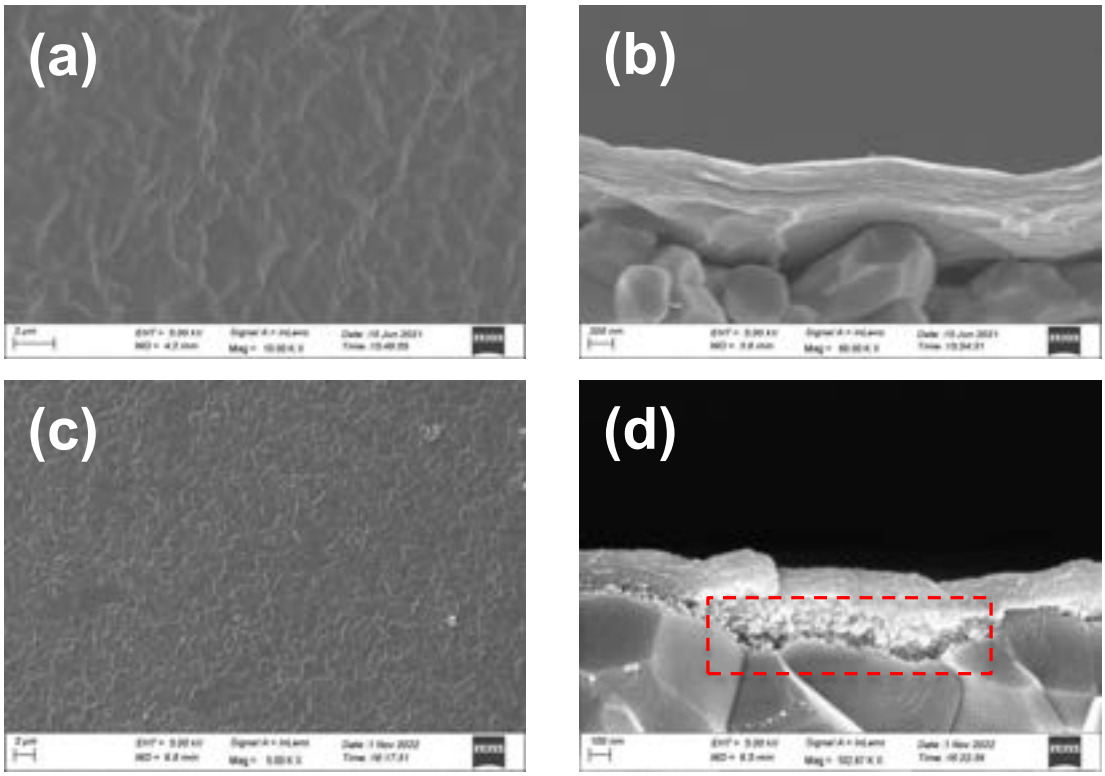


Figure 4 Surface and cross-sectional SEM images of (a, b) PGO HF membrane and (c, d) ZIF-8@PGO HF membrane (M30003h_M1600012h)

where R is the membrane rejection to the dye molecules, C_p and C_f are the concentration of the permeate and the feed solution (mg/L).

3. Results

3.1 Characterisation of ZIF-8@PGO HF membranes

The pores generated through mild chemical etching treatment were observed with HR-TEM. The oxidation of graphene created intrinsic pores on the basal plane of GO nanosheets. Nevertheless, the long and tortuous water transport pathways due to low pore density and the small pore sizes hindered the fast water permeation. Figure 1 shows the HR-TEM images of GO and PGO nanosheets. No pores were observed in Figure 1(a) due to the minimal pore sizes of GO nanosheets that might be lower than the TEM detection range. For PGO nanosheets (Figure 1(b)), both the pores size and pore density increased substantially. The generated pores, the crystalline graphitic domains, and the amorphous domains are labelled. The pore sizes of PGO nanosheets were likely to be smaller than those presented in the HR-TEM images, as the electron beams emitted by the HR-TEM during the scanning might have damaged the samples and enlarged the pores on the nanosheets.

Growing nanocrystals along the edges of the PGO nanosheets selectively has two purposes. First, to increase the permselectivity of the membrane by filling the nonselective voids in the PGO membranes while maintaining high water permeation. Second, to reinforce the membrane structure and increase the mechanical stability of the laminate layers. ZIF-8 nanocrystals were explicitly selected due to their ability to crystallise *in situ* under mild conditions, good water stability, minimal resistance to water permeation, and high selectivity against undesired molecules due to small pore size (0.34 nm).

A uniform distribution of ZIF-8 nanocrystals throughout the ZIF-8@PGO HF membrane was expected to achieve the desired functionalities. Compared to the SEM cross-sectional image of the PGO HF membrane (Figure 4(b)), the SEM cross-sectional image of the ZIF-8@PGO HF membrane (Figure 4(d) shows that many nanocrystals were grown along the edges of the PGO nanosheets and between the membrane and the alumina substrate, which indicates both the metal source and ligand source could fully penetrate the PGO membrane through selective defects. Nevertheless, the composition and the growth position of the nanocrystals required further investigation.

From the X-ray diffraction (XRD) spectra analysis (Figure 5(a)), GO powder exhibited a sharp peak at $2\theta = 9.56^{\circ}$, corresponding to an interlayer d-spacing of 0.836 nm. After 5 hours of mild chemical etching treatment, the PGO membrane showed a peak at $2\theta = 15.11^{\circ}$, corresponding to interlayer d-spacing of 0.808 nm. The decrease in interlayer d-spacing is due to the removal of oxygen functional groups. Furthermore, the ZIF-8@PGO membrane XRD spectra portrayed three peaks at $2\theta = 7.56^{\circ}$, 10.63°, and 13.04°, corresponding to the ZIF-8 characteristic peaks [15], thereby confirming the

5

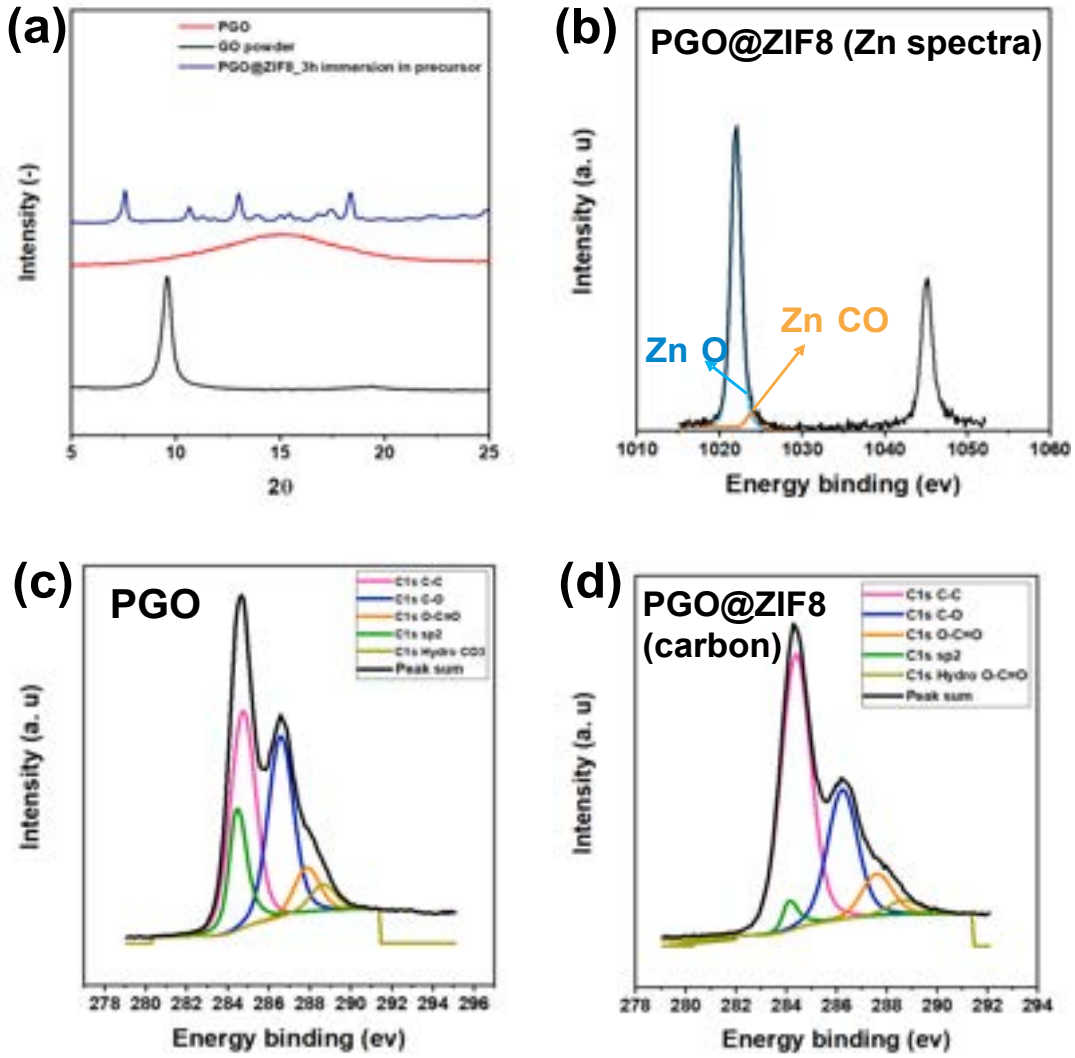


Figure 5 (a) XRD spectra of GO powder, PGO, and ZIF-8@PGO, XPS spectra of (b) ZIF-8@PGO(Zn), (c) PGO (carbon), and (d) ZIF-8@PGO (carbon)

successful growth of ZIF-8 nanocrystals in the lamellar structure of the PGO membrane.

X-ray photoelectron spectroscopy (XPS) zinc spectra of the ZIF-8@PGO membrane (Figure 5(b)) showed two peaks centred at approximately 1021.98 eV and 1023.98 eV, which correspond to the Zn-O and the Zn-CO bonds respectively. The presence of the Zn-CO bond confirms the bond formation between zinc ions and oxygen-containing functional groups on PGO nanosheets. In PGO nanosheets (Figure 5(c)), the dominant oxygen-containing functional group in the pristine PGO membrane is the hydroxyl group (C–OH peak), followed by a much smaller amount of carbonyl group (O=C–O peak) and carboxyl group (hydro O-C=O peak) [16]. After the *in situ* growth of ZIF-8 nanocrystals, no major change took place in oxygen-containing functional groups binding energies and peak intensities, indicating no significant structural change occurred due to ZIF-8 crystallisation. The binding energy of hydroxyl group (C-O peak) shifts from 286.68 eV to 286.28 eV, which is ascribed to the coordination bonds between zinc ions and hydroxyl groups. The above changes in XPS spectra results proved that ZIF-8 successfully crystallised *in situ* and preferably grew along the periphery of PGO nanosheets where the hydroxyl groups locate. Yet a fraction of ZIF-8 nanocrystals also grew at the surface of the PGO nanosheets, which might increase the interlayer d-spacing of the lamellar structure. These findings are consistent with the graphene oxide layer structure model proposed in the literature [17].

3.2 Water permeation and dye rejection performance

To understand how coating time affects the water permeance, 0.05mg/mL PGO dispersion was used to prepare a batch of pristine PGO membranes without any MOF composites. As the coating time decreases, the thickness of PGO nanosheets decreases. From Figure 6, when the coating time is half, water permeation rises with an increasing rate.

To achieve both high water permeance and good dye rejection performance, coating time, PGO dispersion concentration, metal and ligand precursor solvent, ligand precursor concentration, metal and ligand precursor immersion time were altered in the

6

preliminary stage. After coating 0.1mg/mL PGO dispersion onto the hollow fibre substrates for 30s and dried, the PGO HF membranes were immerged in 3000 ppm metal methanol precursor for 3 hours (M_3000_3h) and then 16000 ppm ligand methanol precursor for 12 hours (M_16000_12h). 100% MR rejection was reached with a low permeance (0.585 $LMHbar^{-1}$). Water as a precursor solvent could not achieve high rejection therefore methanol was fixed as both the metal and ligand precursor solvent for the next-step experiments.

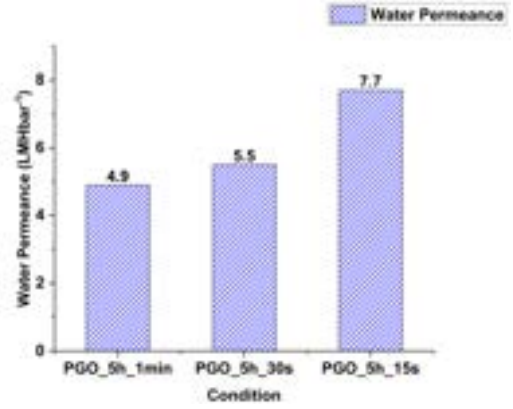


Figure 6 Water permeance for pure PGO HF membranes with different coat times

In the following experiments, 4 factors including the PGO dispersion concentration, ligand concentrations, metal and ligand immersion times were investigated, and it was found that all factors affect the water permeance and rejection. The highest water permeance and 100% dye rejection were found at 0.05mg/mL, M_3000_3h, M_16000_3h, which is as the optimal condition. Further improvement could be explored based on this recipe.

Factor 1: ligand immersion time

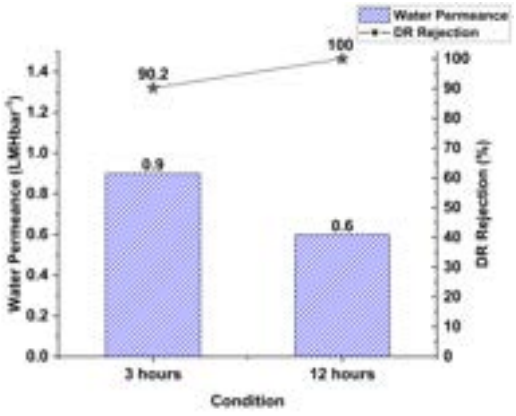


Figure 7 Water permeance and dye rejection for 0.1mg/mL, M3000_3h, M16000_3hours and 12 hours different ligand immersion times

In Figure 7, longer ligand immersion time, in the other word, longer reaction time allowed more ligands to link to the metal nodes and from ZIF-8 nanocrystals. In that way, water was blocked by clusters of ZIF-8 crystals therefore the water permeance declined and 12 hours demonstrated a perfect rejection.

Factor 2: ligand concentrations

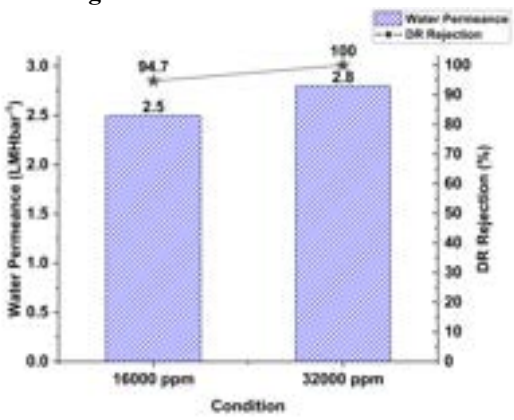


Figure 8 Water permeance and dye rejection for 0.05mg/mL, M3000_3h, M16000ppm, 32000 ppm_1h different ligand concentrations

According to the idea of changing ligand immersion time, higher ligand concentrations also provided more ligands to coordinate with the metal ions which may disadvantage the water permeance. As shown in Figure 8, 32000 ppm induced slightly higher permeance and 100% dye rejection. It is highly likely that the repulsion between negatively charged ligands and oxygen on PGO nanosheets enlarges the interlayer spacing therefore more water molecules are able to pass through. However, the water permeance difference could also be an acceptable experimental error (~10%). It is worth applying HR-TEM to measure the d spacing of the membrane under different ligand concentrations to confirm the microstructure.

Factor 3: metal immersion time

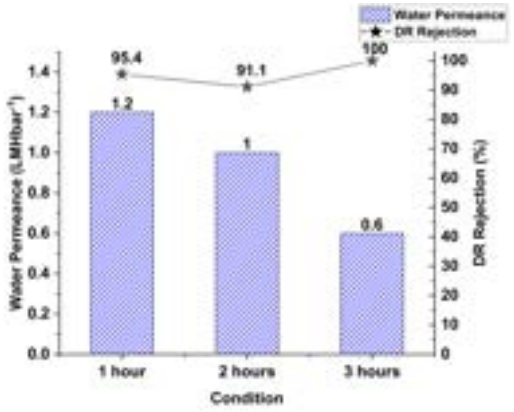


Figure 9 Water permeance and dye rejection for 0.1mg/mL, M3000_1h, 2h 3h, M16000_12h different metal immersion times

From Figure 9, shortening the metal precursor soaking time effectively improved the permeance

7

but lowered the full rejection. That is because positive metal ions have less time to interact with the negative oxygen functional groups on the PGO nanosheets in 1 hour than 3 hours, even if 12 hours of immersion in ligand precursor was not able to coordinate more ligands to the metal nodes. The micro-defects on PGO nanosheets were not fully covered by the limited number of ZIF-8 composites leading to a lower dye rejection. The lower rejection happened at 2 hours than 1 hour could be triggered by the UV-Vis error (5%). Nevertheless, 95.4% is a good rejection performance, hence 1 hour metal immersion time could be repeated in the following experiments.

Factor 4: PGO concentrations

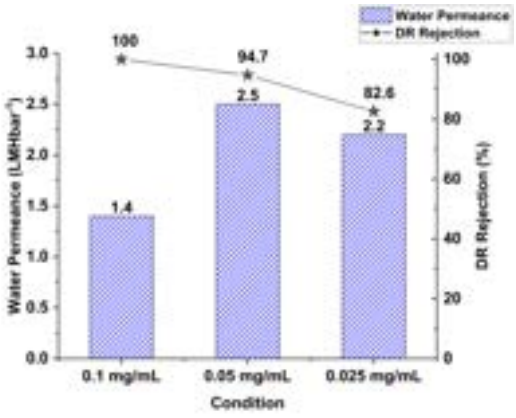


Figure 10 Water permeance and dye rejection for 0.1mg/mL, 0.05mg/mL and 0.025 mg/mL, M3000_3h, M16000_1h different PGO concentrations

Keeping the coating time and other ZIF-8 intercalation conditions identical, PGO dispersion concentrations were altered to tune the thickness of the PGO nanosheet. With increasing the concentration, water permeance decreases because of the increment in thickness and tortuosity. Comparing 0.05 and 0.025 mg/mL conditions in Figure 10, there was a minor water permeance reduction from 2.5 to 2.2 LMHbar-1. It's very possible that a concentration change in relatively dilute PGO dispersion does not dominate the PGO nanosheet thickness. In contrast, focusing on the higher concentrations 0.1 and 0.05 mg/mL, water permeance as well as the PGO nanosheet thickness heavily depends on the PGO concentrations. The rejection dropped with the decreasing PGO concentrations. 82.6% is not an ideal rejection, so it is not suggested to try any concentrations lower than 0.025 mg/mL.

Factor 5: Cross-linking effect (No ligand immersion)

There was a specially designed condition skipping the ligand immersion step. The external insertion of metal ions between the PGO layers crosslinks to the negatively charged oxygen on adjacent PGO flakes decreasing the interlayer spacing. [18] As a result, the permeance is substantially lower than all listed conditions, which indicates crosslinking happened. The dye rejection (86.4%) was not low, which may be attributed to the packed PGO nanosheets with a very small interlayer spacing.

4. Discussion

Methyl Red (MR) is a neutral dye with a small molecular weight (269.3 g/mol) that has been widely used in membrane dye removal tests. [19, 20] However, as MR is widely used as a pH indicator, it is very sensitive to pH change.

In the pressurised cell, it is unpreventable that some of the charged metal nodes and ligand organic groups were washed off and interacted with MR. As a pH indicator, its colour and physical properties greatly depend on the pH values. This leads to significant errors during UV-Vis absorbance measurements and unreliable dye rejection results. Therefore, after the preliminary experiments, MR was replaced by another stable neutral dye, Disperse Red (DR), with a relatively small molecular weight (314.34 g/mol) which facilitated a more accurate molecular weight cut-off determination. Besides, the surface charge of ZIF-8 composite PGO membranes could be checked using zeta potential instrument. After quantitatively understanding the surface charge, charged dyes could be introduced to the performance test which better simulate the seawater desalination in real life.

The growth of metal-organic framework crystal heavily depends on the temperature, growing time, and solvent. From the rejection results, for either metal or ligand precursor, when water was chosen as the precursor solvent, the neutral dye rejection has never exceeded 65%. This indicates that compared with water, methanol provides a preferable environment for ZIF-8 growing. XPS could be carried out to determine the growth behaviour of ZIF-8 nanocrystals with water as the precursor solvent.

As the optimal performance occurred at 0.05 mg/mL PGO, M_3000_3h and M_16000_1h showed 94.7% rejection, 1 hour metal and ligand immersion time could be tried to investigate if a desirable rejection can still be realised.

5. Conclusions

In this study, we report a two-step *in situ* growth of zeolitic imidazolate framework-8 (ZIF-8) into porous GO (PGO) hollow fibre (HF) membranes and the substantial improvement in molecular rejection performance. The pore size and pore density of PGO nanosheets were controlled by adjusting the etching time. The *in situ* crystallisation of ZIF-8 was achieved by immersing PGO HF membrane into the metal ion source and the organic

8

ligand source in turn. The amount of ZIF-8 grown was controlled by adjusting the immersing time. HR-TEM images confirmed the creation of nanopores with tailored pore sizes on the basal plane of GO nanosheets. SEM, XRD, and XPS analyses results demonstrated that the ZIF-8 precursors successfully penetrated the PGO membrane, and ZIF-8 preferably crystallised along the edges of the PGO nanosheets as expected. ZIF-8@PGO HF membrane with a thickness of 200-400 nm maintained a high water permeance up to 2.81 $LMHbar^{-1}$ while achieving an outstanding MWCO of 314 Da. Such performance results exhibit the excellent potential of ZIF-8 nanocrystals in improving the performance of GO-based membranes.

Acknowledgements

The authors gratefully acknowledge the research funding provided by EPSRC in the United Kingdom (SynHiSel project grant no EP/V047078/1) and the steady stream of teaching and help offered by Dr. Farhad Moghadam and Miss Mengjiao Zhai throughout the research project.

References

[1] Obotey Ezugbe, E., & Rathilal, S. (2020). Membrane Technologies in wastewater treatment: A Review. *Membranes*, *10*(5), 89. https://doi.org/10.3390/membranes10050089

[2] Balaji, K. R., Abdellah, M. H., Kumar, V. G. D., Santosh, M. S., Reddy, R., Kumar, S., & Szekely, G. (2023). Nanofiltration membranes composed of carbonized giant cane and Pongamia Meal Binder for ion sieving in water and molecular sieving in organic solvents. *Sustainable Materials and Technologies*, *35*. https://doi.org/10.1016/j.susmat.2022.e00517

[3] Moghadam, F., & Park, H. B. (2018). Two-dimensional materials: An emerging platform for gas separation membranes. *Current Opinion in Chemical Engineering*, *20*, 28–38. https://doi.org/10.1016/j.coche.2018.02.004

[4] Ma, J., Ping, D., & Dong, X. (2017). Recent developments of graphene oxide-based membranes: A Review. *Membranes*, *7*(3), 52. https://doi.org/10.3390/membranes7030052

[5] Chong, J. Y., Wang, B., Mattevi, C., & Li, K. (2018). Dynamic microstructure of graphene oxide membranes and the permeation flux. *Journal of Membrane Science*, *549*, 385–392. https://doi.org/10.1016/j.memsci.2017.12.018

[6] Chong, J. Y., Wang, B., & Li, K. (2018). Water transport through graphene oxide membranes: The roles of driving forces. *Chemical Communications*, *54*(20), 2554–2557. https://doi.org/10.1039/c7cc09120f

[7] Liu, G., Jin, W., & Xu, N. (2016). Two-dimensional-material membranes: A new family of high-performance separation membranes. *Angewandte Chemie International Edition*, *55*(43), 13384–13397. https://doi.org/10.1002/anie.201600438

[8] Wu, T., Moghadam, F., & Li, K. (2022). High-performance porous graphene oxide hollow fiber membranes with tailored pore sizes for water purification. *Journal of Membrane Science*, *645*, 120216. https://doi.org/10.1016/j.memsci.2021.120216

[9] Ying, Y., Liu, D., Zhang, W., Ma, J., Huang, H., Yang, Q., & Zhong, C. (2017). High-flux graphene oxide membranes intercalated by metal–organic framework with highly selective separation of aqueous organic solution. *ACS Applied Materials & Interfaces*, *9*(2), 1710–1718. https://doi.org/10.1021/acsami.6b14371

[10] Zhang, W.-H., Yin, M.-J., Zhao, Q., Jin, C.-G., Wang, N., Ji, S., Ritt, C. L., Elimelech, M., & An, Q.-F. (2021). Graphene oxide membranes with stable porous structure for Ultrafast Water Transport. *Nature Nanotechnology*, *16*(3), 337–343. https://doi.org/10.1038/s41565-020-00833-9

[11] Hummers, W. S., & Offeman, R. E. (1958). Preparation of graphitic oxide. *Journal of the American Chemical Society*, *80*(6), 1339–1339. https://doi.org/10.1021/ja01539a017

[12] Zaaba, N. I., Foo, K. L., Hashim, U., Tan, S. J., Liu, W.-W., & Voon, C. H. (2017). Synthesis of graphene oxide using modified Hummers method: Solvent influence. *Procedia Engineering*, *184*, 469–477. https://doi.org/10.1016/j.proeng.2017.04.118

[13] Li, D., Müller, M. B., Gilje, S., Kaner, R. B., & Wallace, G. G. (2008). Processable aqueous dispersions of graphene nanosheets. *Nature Nanotechnology*, *3*(2), 101–105. https://doi.org/10.1038/nnano.2007.451

[14] Lee, M., Wu, Z., Wang, R., & Li, K. (2014). Micro-structured alumina hollow fibre membranes – potential applications in wastewater treatment.

Journal of Membrane Science, *461*, 39–48. https://doi.org/10.1016/j.memsci.2014.02.044

[15] Nordin, N. A., Ismail, A. F., Misdan, N., & Nazri, N. A. (2017). Modified zif-8 mixed matrix membrane for CO2/CH4 separation. *AIP Conference Proceedings*. https://doi.org/10.1063/1.5005424

[16] Lesiak, B., Kövér, L., Tóth, J., Zemek, J., Jiricek, P., Kromka, A., & Rangam, N. (2018). C SP2/SP3 hybridisations in carbon nanomaterials – XPS and (X)AES study. *Applied Surface Science*, *452*, 223–231. https://doi.org/10.1016/j.apsusc.2018.04.269

[17] Aliyev, E., Filiz, V., Khan, M. M., Lee, Y. J., Abetz, C., & Abetz, V. (2019). Structural characterization of graphene oxide: Surface functional groups and fractionated oxidative debris. *Nanomaterials*, *9*(8), 1180. https://doi.org/10.3390/nano9081180

[18] Safaei, J., Xiong, P., & Wang, G. (2020). Progress and prospects of two-dimensional materials for membrane-based water desalination. *Materials Today Advances*, *8*, 100108. https://doi.org/10.1016/j.mtadv.2020.100108

[19] Ahmadipouya, S., Mousavi, S. A., Shokrgozar, A., & Mousavi, D. V. (2022). Improving dye removal and antifouling performance of polysulfone nanofiltration membranes by incorporation of uio-66 metal-organic framework. *Journal of Environmental Chemical Engineering*, *10*(3), 107535. https://doi.org/10.1016/j.jece.2022.107535

[20] Hu, M., Yang, S., Liu, X., Tao, R., Cui, Z., Matindi, C., Shi, W., Chu, R., Ma, X., Fang, K., Titus, M., Mamba, B. B., & Li, J. (2021). Selective separation of dye and salt by PES/SPSF tight ultrafiltration membrane: Roles of size sieving and charge effect. *Separation and Purification Technology*, *266*, 118587. https://doi.org/10.1016/j.seppur.2021.118587

10

Cell-free Protein Synthesis towards Investigating *in vitro* Glycosylation to Optimize the SUGAR-TARGET Platform

Yixue Dong, Yuyang Ye
Department of Chemical Engineering, Imperial College London

ABSTRACT

Recombinant proteins are a crucial in industry, comprising component ranging from therapeutic agents to industrial enzymes[3]. However, their natural biological properties give rise to a number of problems, which have limited their applications. For instance, glycosylation is a common post-translational modification of proteins. However, it is still challenging to create glycoproteins with targeted glycosylation of high homogeneity because of the promiscuity of enzymes involved[11]. Scientists have attempted to resolve this, but methods for industrial-scale production in glycoengineering and respective native reaction network are still lacking. To address these limitations, *in vitro* glycosylation has recently been considered and a small-scale sequential glycosylation reactions for tailored sugar structures (SUGAR-TARGET) platform has been established[11]. In this study, successful expression of proteins with targeted basal glycosylation was implemented using Chinese hamster ovary (CHO) cell-based cell-free proteins synthesis (CFPS). After that, the SUGAR-TARGET platform was optimized at the enzymatic *in vivo* biotinylation followed by one-step immobilisation/purification step. It was found that the efficiency of biotinylation was proportional to the extent of biotin added to maximize the yield of glycosyltransferases. Furthermore, a larger-scale SUGAR-TARGET platform was built, which acts as a prototype to support its further applications at an industrial scale, and raises the prospect for use of the SUGAR-TARGET platform in combination with CHO cell-based CFPS systems.

Keywords - Recombinant protein, CFPS, Glycosyltransferase, Glycosylation, SUGAR-TARGET

1. INTRODUCTION

Recombinant proteins have become an invaluable biological molecule in a plethora of different industries. Despite their wide applications such as health and biotechnology, proteins are mainly used in pharmacology as therapeutic proteins and industrial fields as enzymes[14]. Due to their large molecular weight, complex composition and structure, proteins have limited solubility, as well as thermal and proteolytic stability, which leads reduced efficacy and greater immunogenetic side effects of therapeutic proteins[3]. In addition, these limitations hinder the development of enzymes and increase the production cost on an industrial scale[3].

Scientists have explored different techniques to enable the manipulation of protein stability, specificity and alteration in its overall function to achieve the desired properties[14]. Among all the methods tested, glycoengineering appeared to be one of the most reliable for future investigation. The principle of glycoengineering is changing glycosylation[7], which is one of the most common and most poorly understood post-translational modifications of proteins[19]. There are two major glycosylation pathways, which are called N-linked and O-linked glycosylation. They respectively indicate two ways glycans could attach to, either the side chain nitrogen (N) atoms of Asn residues or the side chain oxygen (O) atoms of Ser and Thr residues[3].

Glycoproteins are proteins containing glycans attached to amino acid side chains during glycosylation. Since human cells are fundamentally different from those of other species, including microorganisms, fungi, insects and plants, the possibility of creating recombinant glycoproteins for therapeutic purposes in humans using cultivated mammalian cells has generated a lot of interest[19]. In general, Chinese hamster ovary (CHO) is one of the mammalian cells most widely used for the expression and production of recombinant N- or O-linked glycoproteins[9], as it is able to produce glycoproteins at a high rate and could be grown in large-scale bioreactors[3].

Despite the establishment of functional importance from previous work, it is still challenging to produce glycoproteins with targeted glycosylation of high homogeneity[5]. The promiscuity of glycosylating enzymes results in a heterogeneous glycoprofile[2]. Furthermore, methods for large-scale production in glycoengineering and respective native reaction networks are still lacking and the strategies for engineering bespoke functions on proteins using sugar

chemistry is time-consuming[11]. To address these demands, *in vitro* glycosylation has been recently considered. It has the benefit of allowing for specific sugar modifications on recombinant glycoproteins[15], which means therapeutic glycoproteins could be readily generated in a relatively short time, regardless of the production scale.

2. BACKGROUND

In previous work, in order to deal with the limitation of glycoengineering techniques, a platform for Artificial Golgi Reactions sequential glycosylation reactions for tailored sugar structures (SUGAR-TARGET) was developed which allowed bespoke, controlled N-linked glycosylation *in vitro*[11]. Apart from that, an innovative method including *in vivo* biotinylation was also created followed by one-step immobilisation/ purification. The human-like glycosylation pathway was chosen using three enzymes, which were N-Acetylglucosaminyltransferase I (GnTI), α-Mannosidase II (ManII) and β-1,4-Galactosyltransferase (GalT)[11]. The reaction cascade is shown below (**Figure 1**). Therefore, these immobilised enzymes were used to mimic the reaction cascade of the human N-linked glycosylation process, where promiscuity is present naturally. However, the SUGAR-TARGET platform was established for small-scale reactions with sequential incubation of the reaction mixtures with individual enzymes followed by removal using magnetic bead separation.

The purpose of this project was to investigate *in vitro* glycosylation carried out by cell-free protein synthesis (CFPS) and to assess whether there was scope to supplement these products into an optimized SUGAR-TARGET platform. Given this, a larger-scale prototype of the SUGAR-TARGET system would need to be established using packed columns. In order to achieve these goals, we firstly focus on the expression and purification of proteins with targeted glycosylation by conducting CFPS using CHO cells. After that, modification on previous small-scale SUGAR-TARGET platform was done with optimization of the reaction conditions. Lastly, an industrial-scale prototype of this system was designed with the usage of polypropylene columns in lab.

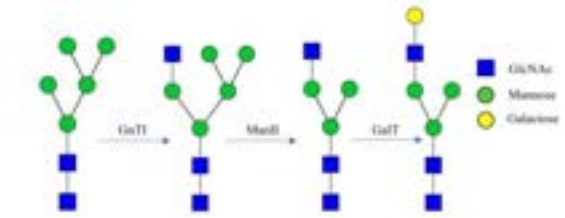


Figure 1. *Reaction Cascade of Immobilised Enzymes*

3. METHODOLOGY

3.1. Expression and purification of protein with targeted glycosylation

3.1.1. Expression of GFP and mFc-GFP

Green fluorescent protein (GFP) and green fluorescent protein with a monomeric crystallizable fragment antibody domain (mFc-GFP) were expressed in Chinese hamster ovary (CHO) cells. Cell-free protein synthesis (CFPS) was chosen for the expression of protein with targeted glycosylation. The main advantage of choosing CFPS is that the metabolic and cytotoxic burdens are abandoned in CFPS. Thus, it can provide an independent, open system to produce specific proteins directly and synthesize difficult-to-express or toxic with a high throughput production[10]. What's more, since all the materials are used to produce the protein, it simplifies purification steps and reduces the loss of targeted protein during purification, which results in a higher yield[8]. Three components are essential for CFPS: a DNA plasmid encoding targeted protein, a cell-free lysate containing cellular components required for DNA transcription and translation, and a reaction mixture providing amino acids, nucleotides, energy sources and crowding reagents[4]. However, the yield of protein expressed by CHO cell-based CFPS is low, which is mainly due to the stress-induced eukaryotic initiation factor 2(eIF2) phosphorylation[10]. To solve this problem, accessory protein tGADD34 is added to the expression reaction to dephosphorylate eIF2α and inhibit stress-induced gene expression to improve yields[13]. The expression of proteins is motivated when the genetic materials combine with the cell-free extract.

The DNA plasmid of GFP and mFc-GFP used for expression were given as a gift from a member of Polizzi Lab. The lysate, tGADD34 and reaction mix used for CFPS reaction was prepared as previous research described[10]. 25μL reactions were made to perform CFPS with the following procedures: 2.5μL purified accessory protein tGADD34 was added to 12.5μL lysate. Following mixing the sample with gentle flicking, it was incubated at room temperature for 10 minutes. 9.5μL reaction mix was then added to the sample and gently mixed with flicking. At the same time, the plate reader was pre-heated to 30 °C to ensure expression could be started immediately once all the components are assembled. 0.5μL DNA of GFP or mFc-GFP was then added to the sample. A negative control was performed to cancel the background fluorescence by repeating the same

procedure described before with a modification: 0.5µL water was used instead of the prepared plasmid. The reaction samples were then transferred into the wells of the Corning 384 well plate (or other multi-well plate). Finally, the plate was sealed with membrane and placed into the plate-reader CLARIOstarPlus (BMG Labtech) with a 300 rpm shaking condition for at least 8 hours for expression.

3.1.2. Purification of the expressed mFc-GFP

After expressing mFc-GFP with targeted glycosylation, it was purified with a one-step purification method using Ni-NTA beads. A native condition was chosen to purify the protein to avoid effects on glycosylation by denatured condition. 40µL (0.5mg) of Ni-NTA Magnetic Beads were first mixed with 160µL of equilibrium buffer (100mM sodium phosphate, 600mM sodium chloride, 0.05% Tween – 20 Surfact – Amps Detergent Solution and 30mM imidazole). After vortexing the mixture for 10 seconds, a magnetic stand was used to collect the beads and the supernatant was discarded. The beads were then mixed with 400µL of equilibration buffer and vortexed for 10 seconds. The supernatant was removed and discarded. 40mL of the pre-expressed mFc-GFP was diluted with an equal volume of equilibration buffer. The prepared protein extract was then mixed with beads using an end-over-end rotator for 30 minutes. After collecting the beads using a magnetic stand, the beads were washed twice with wash buffer (100mM sodium phosphate, 600mM sodium chloride, 0.05% Tween – 20 Surfact – Amps Detergent Solution and 50 mM imidazole). 400µl wash buffer was added to the beads and vortexed for 10 seconds, the supernatant was then removed and discarded. After washing, 25µL of elution buffer (100mM sodium phosphate, 600mM sodium chloride and 250mM imidazole) was added and vortexed until all the beads are submerged in the elution buffer. The mixture was then incubated for 15 minutes on a rotating platform and the supernatant containing His-tagged protein was collected and saved. The elution step was repeated using 25µL of elution buffer. The beads were finally incubated for 10 minutes and saved.

3.1.3. SDS-PAGE analysis for purified protein

Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) was applied to analyse the purification of the expressed protein. It can separate different proteins with different molecular weights and investigate the purity of the glycoprotein sample.
9% gels were first prepared as follows using a Bio-Rad gel casting stand: The glass plates and spacers of the gel casting unit were cleaned, and the plates were assembled. To cast 2 × 9% gel, resolving gel solution was prepared (4.2mL water; 3mL acrylamide; 2.5mL Tris-HCl, pH8.8; 0.150mL SDS; 0.150mL ammonium persulfate ;0.009mL TEMED). The aliquot gel was then poured into the plates and water was added to reach the top edge of the plates to maintain an even and horizontal resolving gel surface. The gel was left for 20 minutes to solidify, and the overlaid water was then discarded. Stacking gel solution was then prepared (3.69mL water; 0.625mL acrylamide; 0.630mL Tris-HCl, pH6.8; 0.05mL SDS; 0.05mL ammonium persulfate; 0.005mL TEMED). The stacking gel was poured until it overlaid the edge of the plates. The 10-well comb was inserted to make sure no air bubbles were trapped in the gel. The gel was left for 20 minutes to complete the set-up.

40µL of the supernatant containing purified mFc-GFP was mixed with 10µL 5 × SDS-PAGE buffer (0.225M Tris·Cl, pH 6.8; 50% glycerol; 5% SDS; 0.05% bromophenol blue; 0.25M DTT). The sample was heated and incubated at 100° C for 10 minutes. 20µL of the incubated sample was loaded on the gel alongside PageRuler Prestained Protein Ladder (ThermoFisher). Detailed information for the ladder can be found in **Appendix A**. The gels were then run at constant electric current (mAmp) using 25mAmp/gel for 50 minutes. After the complete running of the SDS-PAGE, the gels were washed and stained with SimplyBlue™ SafeStain (ThermoFisher) following the microwave stained protocol provided by the manufacturer[6].

3.2. Enzyme expression and immobilisation

In this research, both GnTI and GalT were expressed in a bacterial system. Immobilised enzymes have higher reusability, stability, and controllability compared to mobile enzymes[12]. Therefore, the expressed enzymes were immobilised to build an enzyme cascade using a one-step immobilisation/purification method. Since ManII has to be expressed in insect cells and biotinylated by a chemical biotinylation technique, it was not expressed and purified in this research due to time limitations.

3.2.1. Optimizing the concentration of d-biotin for inducing enzyme expression

From the previous research, GalT and GnTI were biotinylate using 20µM biotin. To improve the extent of biotinylation, the concentration of biotin was optimised by expressing GalT and GnTI with a final concentration of biotin varying from 10µM to 30µM with an increment of 5µM. A modified experiment was done for induction of 20µM and 100µM biotin to express GalT and GnTI to

obtain a more visible result and further investigate the overall change in biotinylation performance with different concentrations of biotin.

3.2.1.1. GalT and GnTI expression

The expression of glycosyltransferase GalT and GnTI using a bacterial system took 3 days. On the first day of the expression process, 900mL of Luria Broth Base (Miller's LB Broth Base) ™ (LB) medium (1% peptone from casein; 0.5% yeast extract; 1% NaCl in water) was prepared. The co-transformed cells of GnTI and GalT were given by a member in the Polizzi Lab. A single colony of the co-transformed GnTI or GalT was added to 5mL of LB, together with 100μg/mL ampicillin and 10 μg/mL chloramphenicol. The whole process was done near a fire to eliminate possible contaminants in the air. The culture samples were incubated in a shaking incubator overnight at 37 °C.

After one night of growth, the pre-induction culture was prepared by diluting 125μL starter culture with 12.5mL LB media containing 0.2% glucose. The same pre-induction culture was prepared for each sample with different biotin concentrations added after. To prevent the elimination, the whole process was done near a fire. The samples were then incubated in a shaking incubator at 37 °C until the optimal density at 600nm (OD600) reached 0.6-0.8 with a 1mL of LB as the blank. 1mL of the pre-induction culture was collected for each sample to trace the experiment. The new volume for each sample was reduced to 11.5mL. The expression of the enzymes was then induced by adding isopropyl β-D-1-thiogalactopyranoside (final concentration 0.1mM) and biotin into the pre-induction culture. The volume of 4mM biotin added to each sample is shown below:

Table 1. *Amount of Biotin Added in Each Sample for the First Experiment*

Sample No.	Concentration of biotin / μM	Amount of 4mM biotin added / μL
GalT		
1	10	28.75
2	15	43.13
3	20	57.50
4	25	71.88
5	30	86.25
GnTI		
6	10	28.75
7	15	43.13
8	20	57.50
9	25	71.88
10	30	86.25

The samples were then incubated in a shaking incubator overnight at 20 °C.

After overnight incubation, 1mL of each post-induction sample was collected for tracing the experiment. The cells were then harvested by centrifugation (4 °C, 4000×g, 30 minutes). The pellet for each sample was resuspended in 1.25mL of lysis buffer (5ml/gr cells; 20mM Tris-HCl, pH 7.5; 200mM NaCl; 5% glycerol; 0.1mM phenylmethylsulfonyl fluoride) and the samples were left for 20 minutes to break the cells. The samples were then sonicated for 6 minutes with 10 seconds on/off pulses (40%). Finally, the lysate was centrifuged (4 °C, 12864 × g, 30 minutes). The supernatant was collected and filtered with a 0.45μm filter. The filtrate was collected for further analysis.

With two modifications, the same experiment was done to express GalT and GnTI with 20μM and 100μM biotin. The volume of 4mM biotin added to each sample in the second experiment was shown as follows:

Table 2. *Amount of Biotin Added in Each Sample for the Second Experiment*

Sample No.	Concentration of biotin / μM	Volume of 4mM biotin added / μL
GnTI		
1	20	57.50
2	100	287.50
GalT		
3	20	28.75
4	100	287.50

Besides, after the same induction and harvesting procedure, the pellet of each sample was resuspended in 350μL of lysis buffer to ensure protein concentration was high enough to show clearer results on the gel.

3.2.1.2. Quantification of the extent of biotinylation by a streptavidin gel-shift

The strategy of *in vivo* biotinylation is illustrated in **Figure 2a**. Specifically, the catalytic domain of recombinant proteins was initially fused to Maltose Binding Protein (MBP) at N-terminus and AviTag at C-terminus. Subsequently, a small two-residue Glycine-Serine linker was inserted before AviTag to ensure functionality. This allowed for the site-specific biotinylation of both enzymes by co-expressing with biotin ligase BirA and supplementation of the medium with biotin[1].

Before quantifying the extent of biotinylation, a gel-shift analysis with streptavidin by SDS-PAGE was used due to the non-covalent biological interactions of the binding between biotin and streptavidin (**Figure 2b**). As a result,

except the original bands indicating the molecular weight of BirA (34kDa) and enzymes (93.8kDa for GnTI and 76.4kDa for GalT), this binding would be shown on the gel in the form of the tint bands of enzymes and additional new upper bands with the control of streptavidin (66kDa), which was called gel shift. The intensity difference of the bands of the expressed enzymes with and without streptavidin would implicate the efficiency of biotinylation.

9% SDS-PAGE gel was prepared as described previously. To prepare the samples used for gel shift assay, 8.25µL of 2 × SDS-PAGE buffer without DTT (0.09M Tris·Cl, pH 6.8; 20% glycerol; 2% SDS; 0.02% bromophenol blue) was added to 7.5µL of each biotinylated enzyme. The mixed samples were then heated at 95°C for 5 minutes using a PCR system (ProFlex). The samples were completely cooled down to room temperature. 0.75µL of streptavidin was then added to each sample. For negative controls, 0.75µL of DI water was added instead. After that, the samples were loaded on the prepared 9% SDS-PAGE gel alongside PageRuler Prestained Protein Ladder (ThermoFisher) and run at a constant electric current (mAmp) using 25mAmp/gel for 50 minutes. The same microwave staining protocol was applied to the gels as mentioned before. Finally, the intensity of the bands on the gel was measured using Image Lab Software to carry out gel shift assays on the biotinylated enzymes.

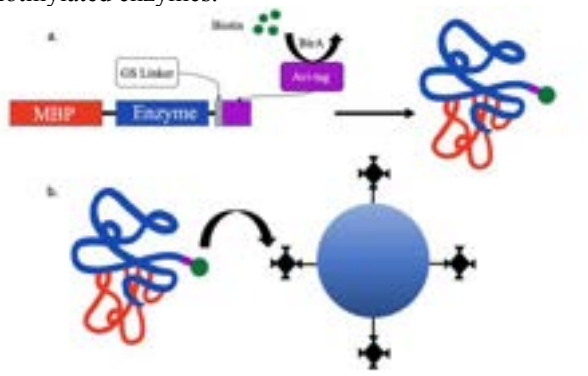


***Figure 2**. Principle of in vivo Biotinylation. a. Process of binding between BirA and AviTag; b. Interaction of binding between biotin and streptavidin.*

3.2.2.1. One-step immobilisation/purification of enzymes

Streptavidin silica particles 1%w/v. 1.0-1.4µm (Spherotech) were used to carry out the one-step immobilisation of the expressed enzymes. For each sample of the expressed enzyme, 50µL of streptavidin silica particles were centrifuged using a microcentrifuge at 6.8rpm for 3 minutes to remove the storage buffer. The beads were then washed 3 times. In one wash cycle, 200µL of 0.1M Tris-HCl (pH 7.5) was added to resuspend the beads and then centrifuged at 6.8 rpm for 3 minutes. 30µL of the expressed enzyme sample was then mixed with the washed beads with a dilution by 970µL of the 0.1M Tris-HCl (pH 7.5). The beads were harvested by centrifugation for 8 minutes at 6.8rpm. Finally, the particles were washed with 2mL of 0.1M Tris-HCl (pH 7.5) 3 times.

3.2.2.2. SDS-PAGE analysis for immobilised enzymes

SDS-PAGE was applied to analyse the immobilised enzyme samples. 9% SDS-PAGE gel was prepared as previously mentioned. The samples for running on the gels were then prepared. To prepare the samples of expressed enzymes, 40µL of each pro-immobilised enzyme sample was mixed with 10µL of 5 × SDS-PAGE buffer (0.225M Tris·Cl, pH 6.8; 50% glycerol; 5% SDS; 0.05% bromophenol blue; 0.25M DTT). The samples were then heated up to 95 °C using a PCR system (ProFlex) for 5 minutes. To prepare the samples of immobilised enzymes, 80µL of beads in DI water was mixed with 20µL 5 × SDS-PAGE buffer (0.225M Tris·Cl, pH 6.8; 50% glycerol; 5% SDS; 0.05% bromophenol blue 0.25M DTT), followed by incubation at 100° C for 10 minutes. The samples were then centrifuged for 1 minute at 13.4rpm to remove the beads.

20µL of each expressed enzyme sample and 20µL of the supernatant of each immobilised bead sample was loaded on the gel aside from the PageRuler Prestained Protein Ladder (ThermoFisher) and run at a constant electric current (mAmp) using 25mAmp/gel for 50 minutes. The same microwave staining protocol was applied to the gels as mentioned before.

3.2.3. Lab-scale SUGAR-TARGET platform design

2mL disposable columns (Thermo Scientific™) and silica particles 5% w/v. 1.26µm (Spherotech) were used to build a prototype for a lab-scale SUGAR-TARGET platform. From the previous study, the conditions for inducing desired glycosylation on the targeted protein in a large-scale SUGAR-TARGET platform were computed theoretically[18]. Two designs were considered: the first design is a continuous packed-beds column containing enzyme cascade, while the second one is three columns linked in series with the packing of different glycosyltransferases respectively. The advantages and disadvantages of the designs were considered from functional, efficient, and economic aspects.

4. RESULTS AND DISCUSSION

4.1. Expression and purification of protein with targeted glycosylation

4.1.1. Expression of GFP and mFc-GFP

Initially, the ability of tGADD34 to increase expression of GFP and mFc-GFP in CHO cells-based CFPS was analysed using the plate-reader CLARIOstarPlus (BMG Labtech). **Figure 3** showed the expression behaviour of newly purified tGADD34 in expression GFP and mFc-GFP overtime by CFPS, as mentioned in **Section 3.1.1**. The plate reader can detect the signal given by GFP and quantify the fluorescent intensity[20]. The more GFP and mFc-GFP were expressed, the higher fluorescence was reported. The trend of the positive control and the expression of GFP using new tGADD34 are similar, and it proves the new tGADD34 can work in the expression of GFP and mFc-GFP. Since the fluorescence detected in the expression of GFP using new tGADD34 is higher than that using old tGADD34 at any time, the new tGADD34 worked better in the CFPS, which is mainly due to the denaturation of old tGADD34 during storage. Furthermore, CFPS performed better in expressing GFP compared to mFc-GFP. This is due to the additional antibody fragment interfered with GFP folding, thus reduced the fluorescence. What's more, since mFc-GFP is greater than GFP in size, by applying the same amount of materials for protein expression, the amount of GFP produced is higher than mFc-GFP.

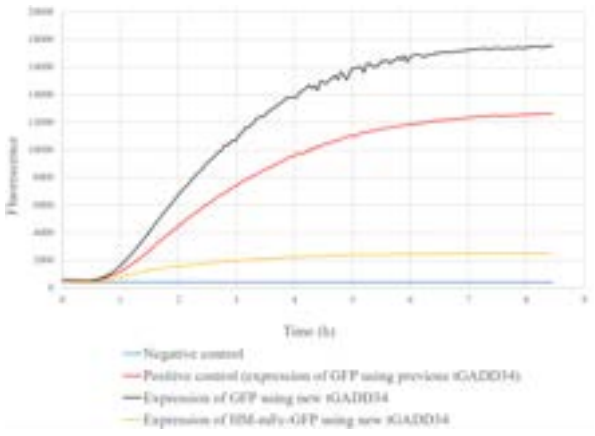


Figure 3. *Testing New tGADD34 and Assessing Expression of GFP and HM-mFc-GFP*

The formal expression of mFc-GFP was conducted using the same CFPS method. From **Figure 4**, the trend of the curve for the expression of mFc-GFP is similar to that shown in **Figure 3**. A sudden reduction in the fluorescence occurred at 2.8 hours. By analyzing the raw data, the fluorescence of the sample of expressed mFc-GFP in the first well reduced, resulting in a sharp decrease in the average fluorescence, which could be due to the machinery error. While the overall trend of the curve still shows the successful expression of mFc-GFP was achieved.

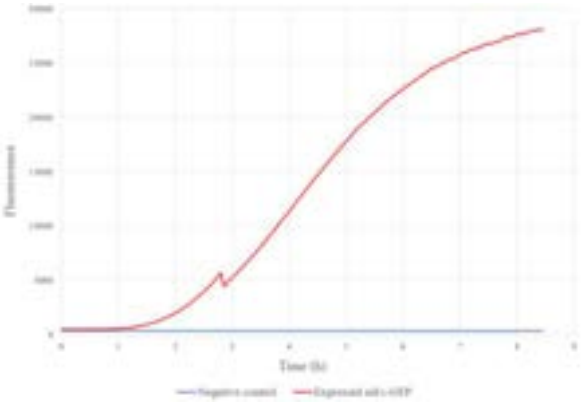


Figure 4. *Expression of mFc-GFP with Targeted Glycosylation using CHO Cell-based CFPS*

4.1.2. Purification of the expressed mFc-GFP and SDS-PAGE analysis for purified protein

The purification of mFc-GFP was carried out using Ni-NTA beads with a one-step purification method, which was described in **Section 3.1.2**. **Figure 5** shows the SDS-PAGE result analysing the purified protein. There is a band occurs at 54.8kDa, which specifically represents mFc-GFP. It further proves the protein expression is successful. Since the band is clear and distinguishable from others, it shows the purification method using Ni-NTA beads works to purify the expressed mFc-GFP with targeted glycosylation. It should be noticed that some impurities are observed in the sample. In that case, a more specific purification method could be investigated to give a better result.

However, the glycan analysis shows glycans were missing in the purified protein. The His-tagged purification using Ni-NTA beads did not work for the glycan purification since the detergents were not sufficient for microsome lysis, or the beads used were not able to specifically capture the His-tagged protein. It can also be because of the impurity issues of the sample, which prevent the expressed mFc-GFP to release glycans. Besides, since the band of the mFc-GFP shown on the gel is thin, the concentration of purified mFc-GFP with targeted glycosylation in the sample is lower than 100μg/mL. Therefore, the glycans cannot be detected. The detailed glycan analysis result is in **Appendix B**.

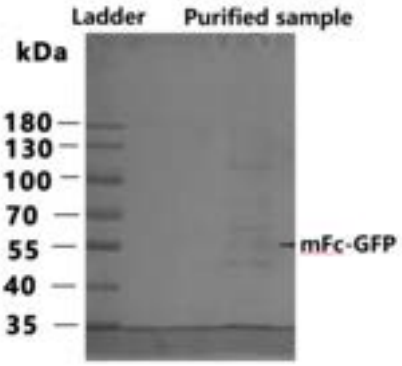


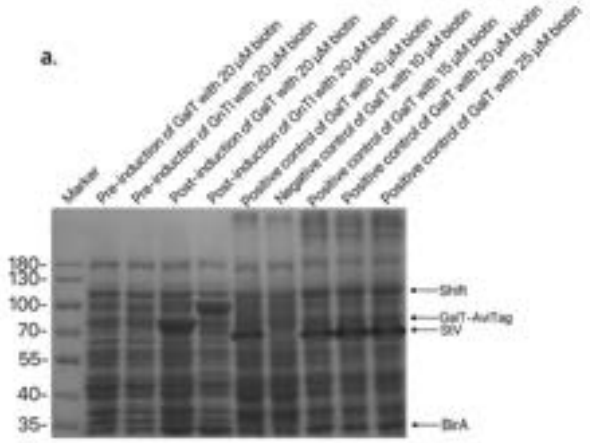
Figure 5. *SDS-PAGE Analysis on Ni-NTA Purified Protein with Targeted Glycosylation*

4. 2. Artificial Golgi reactions

4.2.1 Optimization of enzymatic *in vivo* biotinylation during expression of glycosyltransferases

In order to optimize the yield of proteins (GnTI and GalT) from expression, we intended to vary the biotin concentration during enzymatic *in vivo* biotinylation experiments. The binding between biotin ligase BirA and AviTag provided the foundation of this efficient reaction[1], which has been mentioned in **Section 3.2.1**.

We initially planned to vary it from 10μM to 30μM with the increment of 5μM as mentioned in **Section 3.2.1** to explore the trend of biotinylation efficiency and thus achieve the optimization. However, it was quite difficult to see the bands of both GalT and GnTI from **Figure 6**, which would a challenge in measuring the intensity difference. This result might be due to serval reasons. Firstly, the change of biotin concentration of the original idea was too small to figure out the obvious gel shift, so it seemed to achieve the same biotinylation. Secondly, the lysate was too dilute which resulted in a small concentration of proteins, and therefore, respective tint bands.



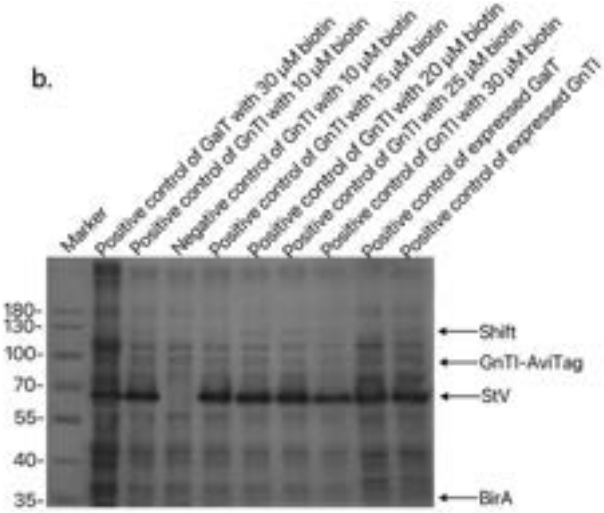


Figure 6. *First Attempt of Biotinylation Confirmation Using Gel-shift Assays; Each lane was loaded with and without BirA and streptavidin in the absence of reducing agent DTT. a. Gel assays of GalT; b. Gel assays of GnTI.*

Having learned the experience of failures before, the protocol for *in vivo* biotinylation was modified (**Section 3.2.1**). The much clearer results of gel-shift assays are shown in **Figure 7**. Since streptavidin had four binding sites, it could bind to more chains of the biotinylated target, which corresponded to multiple bands on the gel[11]. After measuring the intensity of bands for both enzymes with Image J software, biotinylation efficiency with respective biotin concentration was calculated and the results are generated in **Table 3**. It was obvious to see that 100μM of biotin concentration corresponded with higher biotinylation, especially for GalT. Therefore, it could be concluded that the efficiency of enzymatic biotinylation would be proportional to the addition of biotin concentration. Furthermore, the intensity of gel shift was interestingly found to be different in both cases, as can be seen in **Figure 7**. For GnTI, a darker shift band was shown when the concentration of biotin was 100μM. This was confirmed after the measurement of protein concentration, and it was found to be much higher than that of 20μM ones. While in terms of GalT, a darker band at 20μM might be a result of operation error, such as poor performance when loading samples.

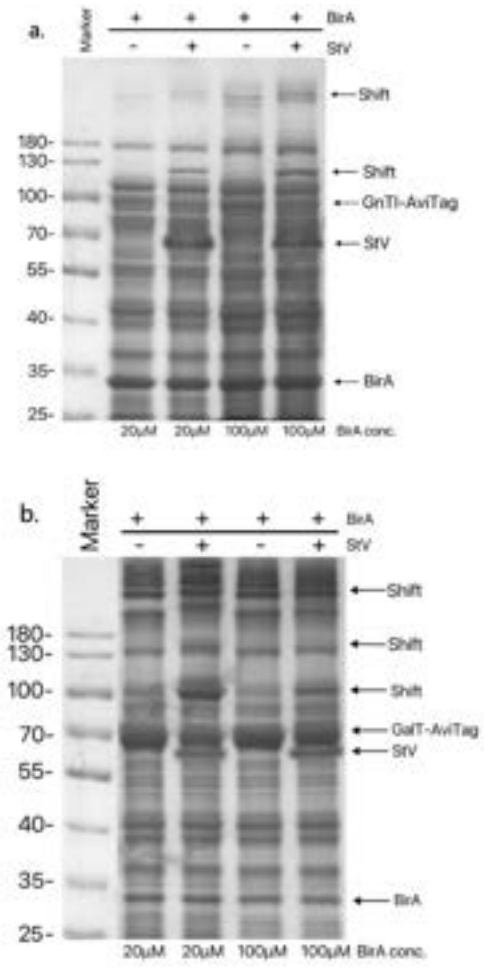


Figure 7. *Second Attempt of Biotinylation Confirmation Using Gel-shift Assays; Each lane was loaded with and without BirA and streptavidin in the absence of reducing agent DTT. a. Gel assays of GnTI; b. Gel assays of GalT.*

Table 3. *Results of biotinylation efficiency with respective enzymes and biotin concentration added*

Biotinylated enzyme with different concentration	% biotinylation
GnTI with 20μM biotin	31.78
GnTI with 100μM biotin	50.14
GalT with 20μM biotin	38.40
GalT with 100μM biotin	65.00

4.2.2 One-step immobilisation/purification

After the successful expression and implementation of *in vivo* biotinylation, one-step immobilisation/ purification of GnTI and GalT was conducted with streptavidin-silica beads. Gel electrophoresis was then used for confirmation, where the results were shown in **Figure 8**. The presence of multiple bands on the lanes of binded beads suggested that the recovery of GnTI was not as good as that of GalT. This could be a human error, resulting in the poor immobilisation of enzymes. In addition, it was clear to see that the bands of BirA were also shown when non-specifically bound to beads, which might be due to the complex formed with its substrate AviTag.

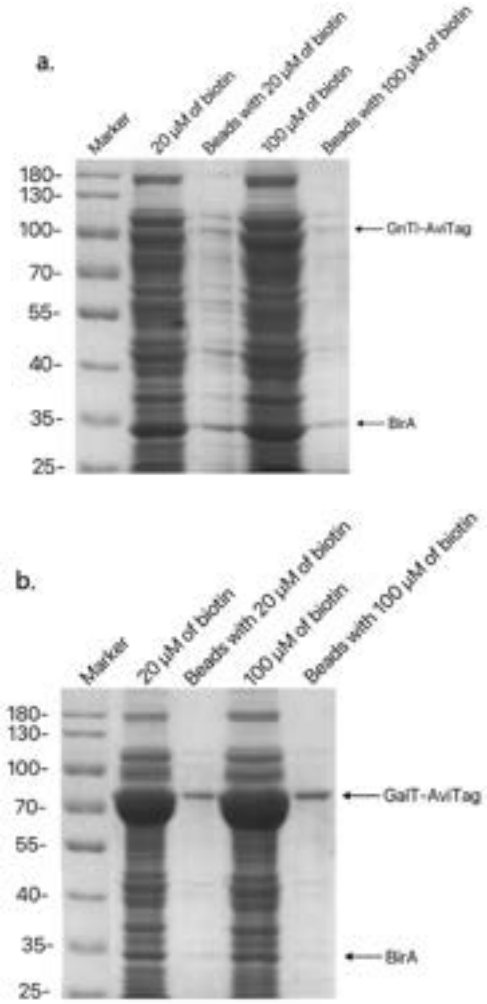


Figure 8. *Confirmation for the Recovery of Enzymes after One-step Immobilisation/ purification; a. Gel assays of GnTI; b. Gel assays of GalT.*

4.2.3 Design of larger-scale SUGAR-TARGET platform prototype

Following the novel methods of the *in vivo* biotinylation and site-specific immobilisation of glycosyltransferases with additional changes to optimize conditions, we considered to set up a larger-scale SUGAR-TARGET platform by using disposable polypropylene columns based on the conceptual and computational prototype mentioned in **Section 3.2.3**. Mainly speaking, there were two types of packed bed reactors designed, following the cascade of GnTI-ManII-GalT.

4.2.3.1 Continuous packed bed reactor

The first ideal setup was the continuous bed reactor. In this kind of system, immobilised enzymes are packed in a single continuous column with filters in between, which could be seen clearly by three distinct colours from **Figure 9**. When considering the implementation in terms of industrial scale, the system had the advantage of automatic control and operation, which would help to reduce both labour and capital costs[21]. However, the flowrate and residence time of each specific glycoprotein could not be controlled separately. Besides, it was necessary to consider the pressure drop across the packed column, as well as the effect of column dimensions like length to achieve the desired reaction rate[21].

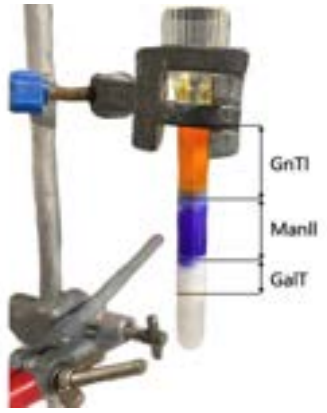


Figure 9. *Establishment of Continuous Packed Bed Reactor with Immobilised Enzyme Cascade*

4.2.3.2 Sequential packed bed reactors

Another idea was to connect three packed columns in series (**Figure 10**) and then control the flow of glycoproteins by manually opening or closing, though it was somewhat time-consuming. Also, variables like temperature and pressure could be controlled easily since glycoproteins were separated. Due to the fact that the reaction buffers for three immobilised enzymes were specific throughout the cascade, it was suggested to replace the buffer system in between. To solve this problem, a desalting column was considered to be added between each packed column. Initially, the heavy glycoproteins took much more time to flow through the bed compared with small salt molecules. Nevertheless, if an inlet stream of buffer was introduced into the desalting column and then suddenly flashed out, glycoproteins would be quickly pushed downwards to flow to the next column and complete buffer exchange. Despite the whole system had many advantages, one point to consider was that the operating costs would be higher with the usage of multiple columns.

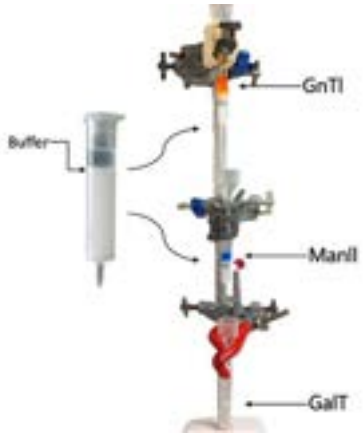


Figure 10. *Establishment of Sequential Packed Bed Reactor with Immobilised Enzyme Cascade*

5. CONCLUSIONS

In this research, mFc-GFP was successfully expressed using a CHO cells-based CFPS method. It was sequentially purified with a one-step Ni-NTA beads purification method. However, the impurity issue caused the failure in detecting glycans in the purified protein, which requires further investigation in the specific purification method. In terms of modification of the previous small-scale SUGAR-TARGET platform, it resulted that the extent of enzymatic biotinylation was proportional to the increase in biotin content by optimizing the biotin concentration. The expressed enzymes were then successfully purified with a one-step immobilisation/purification method using streptavidin silica beads. Lastly, a prototype for the large-scale SURGAR-TARGET platform was built, which allowed to apply this system for industrial uses in the future.

6. FUTURE WORK

The glycan analysis result shows glycans are missing in the purified mFc-GFP. To improve the glycosylation of the proteins, better detergents can be used to release the glycans. A more specific purification technique can also be investigated to purify the expressed protein more efficiently, such as protein A beads, which is an antibody-binding protein that can bind to the Fc region[16]. Besides, the gel shift assay for enzyme expression shows the extent of biotinylation did not reach 100% even for 100μM biotin concentration. Future experiments can be conducted with biotin concentrations greater than 100μM to optimize *in vivo* biotinylation. Last but not the least, based on the lab-scale SUGAR-TARGET platform and the theoretical data from the previous research[18], an industrial-scale SUGAR-TARGET system can be built

with further evaluation of economic potential, practical feasibility, environmental and safety factors.

7. ACKNOWLEDGEMENTS

We would like to first express our gratitude to Prof. Karen M. Polizzi for providing invaluable support and feedback on experimental procedures. We also thank to Oscar Marshall and Elli Makrydaki for their support throughout the entire duration of the project.

8. REFERENCES

1. A. Gautier, M. J. Hinner, M. Fairhead, and M. Howarth, "Site-Specific Biotinylation of Purified Proteins Using BirA," in *Site-specific protein labeling methods and protocols*, 2015.
2. A. M. Sinclair and S. Elliott, "Glycoengineering: The effect of glycosylation on the properties of therapeutic proteins," *Journal of Pharmaceutical Sciences*, vol. 94, no. 8, pp. 1626–1635, 2005.
3. B. Ma, X. Guan, Y. Li, S. Shang, J. Li, and Z. Tan, "Protein glycoengineering: An approach for improving protein properties," *Frontiers in Chemistry*, vol. 8, 2020.
4. B. Melinek, N. Colant, C. Stamatis, C. Lennon, S. S. Farid, K. Polizzi, M. Carver, and D. G. Bracewell, *Toward a Roadmap for Cell-Free Synthesis in Bioprocessing*, 2020. URL: https://www.researchgate.net/publication/350383278_Toward_a_Roadmap_for_Cell-Free_Synthesis_in_Bioprocessing.
5. B. S. Hamilton, J. D. Wilson, M. A. Shumakovich, A. C. Fisher, J. C. Brooks, A. Pontes, R. Naran, C. Heiss, C. Gao, R. Kardish, J. Heimburg-Molinaro, P. Azadi, R. D. Cummings, J. H. Merritt, and M. P. DeLisa, "A library of chemically defined human N-glycans synthesized from microbial oligosaccharide precursors," *Scientific Reports*, vol. 7, no. 1, 2017.
6. Bio-Rad, "A guide to polyacrylamide gel electrophoresis and detection," 2021a. URL: https://www.bio-rad.com/webroot/web/pdf/lsr/literature/Bulletin_6040.pdf.
7. C. F. Goochee, M. J. Gramer, D. C. Andersen, J. B. Bahr, and J. R. Rasmussen, "The oligosaccharides of glycoproteins: Bioprocess factors affecting oligosaccharide structure and their effect on glycoprotein properties," *Bio/Technology*, vol. 9, no. 12, pp. 1347–1355, 1991.
8. C. H. Chiba, M. C. Knirsch, A. R. Azzoni, A. R. Moreira, and M. A. Stephano, "Cell-free protein synthesis: Advances on production process for biopharmaceuticals and immunobiological products," *BioTechniques*, vol. 70, no. 2, pp. 126–133, 2021.
9. E. Grabenhorst, P. Schlenke, S. Pohl, M. Nimtz, and H. S. Conradt, "Genetic engineering of recombinant glycoproteins and the glycosylation pathway in mammalian host cells," *Glycotechnology*, pp. 1–17, 1999.
10. E. Makrydaki, O. Marshall, C. Heide, G. Buldum, C. Kontoravdi, and K. M. Polizzi, "Cell-free protein synthesis using Chinese hamster ovary cells," *Recombinant Protein Expression: Prokaryotic Hosts and Cell-Free Systems*, pp. 411–435, 2021.
11. E. Makrydaki, R. Donini, A. Krueger, K. Royle, I. Moya-Ramirez, D. A. Kuntz, D. R. Rose, S. M. Haslam, K. Polizzi, and C. Kontoravdi, "Immobilised enzyme Cascade for targeted glycosylation," 2022.
12. H. H. Nguyen and M. Kim, "An overview of techniques in enzyme immobilization," *Applied Science and Convergence Technology*, vol. 26, no. 6, pp. 157–163, 2017.
13. I. Novoa, H. Zeng, H. P. Harding, and D. Ron, "Feedback inhibition of the unfolded protein response by GADD34-mediated dephosphorylation of eif2α," *Journal of Cell Biology*, vol. 153, no. 5, pp. 1011–1022, 2001.
14. J. Puetz and F. M. Wurm, "Recombinant proteins for industrial versus pharmaceutical purposes: A review of process and pricing," *Processes*, vol. 7, no. 8, p. 476, 2019.
15. M. Thomann, T. Schlothauer, T. Dashivets, S. Malik, C. Avenal, P. Bulau, P. Rüger, and D. Reusch, "In vitro glycoengineering of IGG1 and its effect on Fc receptor binding and ADCC activity," *PLOS ONE*, vol. 10, no. 8, 2015.
16. M. Zarrineh, I. S. Mashhadi, M. Farhadpour, and A. Ghassempour, "Mechanism of antibodies purification by protein A," *Analytical Biochemistry*, vol. 609, p. 113909, 2020.
17. N. E. Gregorio, M. Z. Levine, and J. P. Oza, "A user's guide to cell-free protein synthesis," *Methods and Protocols*, vol. 2, no. 1, p. 24, 2019.
18. O. V. Klymenko, N. Shah, C. Kontoravdi, K. E. Royle, and K. M. Polizzi, "Designing an artificial golgi reactor to achieve targeted glycosylation of monoclonal antibodies," *AIChE Journal*, vol. 62, no. 9, pp. 2959–2973, 2016.
19. S. A. Brooks, "Appropriate glycosylation of recombinant proteins for human use: Implications of choice of expression system," *Molecular Biotechnology*, vol. 28, no. 3, pp. 241–256, 2004.
20. T. Wei and H. Dai, "Quantification of GFP signals by fluorescent microscopy and flow cytometry," *Methods in Molecular Biology*, pp. 23–31, 2014.
21. X. Kang, *Packed bed bioreactor*. URL: https://userpages.umbc.edu/~xkang/ENCH772/packed.html.

Linear Modelling and Control of a Refrigeration System

Isra Parwaiz and RanHee Kim
Department of Chemical Engineering, Imperial College London, U.K.

Abstract

United Kingdom's net zero goals has led to many businesses working proactively to improve the efficiency of their energy use and reduce their carbon footprint. Supermarkets, in particular, are extremely energy intensive which means there is an added focus on reducing their energy use through a variety of data-driven modelling followed by advance control techniques. This paper presents a possible approach for modelling the compressor work for a UK supermarket and reducing the compressor work through the implementation of a simple bang-bang controller. A linear model as well as a neural network model was investigated. The linear model was chosen because it is simpler and can be physically represented as an equation. The bang-bang controller was simulated with and without temperature setpoint changes with the aim of improving energy efficiency. Overall, energy savings of 1.5% and cost savings of 0.8% were achieved, which were deemed as reasonable.

Keywords: Refrigeration Systems, Data-driven Modelling, Bang-Bang Control

1 Introduction

With the announcement of the United Kingdom (UK) being the first major economy to pass the ambitious 2050 net zero target, following the 26th UN Climate Change Conference of the Parties (COP26), there has been an increase in interest and urge to reduce carbon emissions within all aspects of the economy [1]. Many regulations and policies have been put in place to promote the paradigm shift to renewable energy sources. However, energy demand and consumption continue to increase and the capacity of energy from renewable resources is not sufficient. Moreover, the rise of digitally connected devices has also amplified the electricity demand. Smart energy management aims to monitor energy consumption patterns so that businesses can anticipate amounts of usage to help control and reduce wasted energy and become more energy efficient [2].

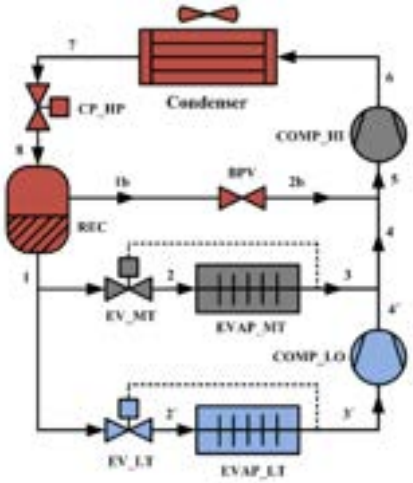


Figure 1: Schematic of the studied refrigeration cycle considered in this paper, taken from [18].

Supermarkets are the most energy-intensive buildings, and, in the UK, these supermarkets take up 3% of the total energy consumption of the economy [3]. Within these supermarkets, refrigeration systems account for 30% - 60% of the total electricity use. Smart energy management techniques have been investigated in refrigeration systems in supermarkets as a method to reduce electricity use. Reducing this electricity use allows for cost reduction and lets supermarkets rely more on renewable sources as the overall demand for energy decreases.

A basic layout of a typical refrigeration system with a booster configuration is shown in Figure 1. A typical refrigeration system includes multiple fridges, one at a lower temperature and another at a medium temperature, and freezer display cases with two compressor banks. The dynamic model of the refrigeration system is a system of energy balance equations. In the cold reservoir, there is a transfer of heat from the foodstuffs to the cooled air, which is then transferred from the cooled air to the circulated refrigerant. The act of being able to store electrical energy as coldness by lowering temperatures of the foodstuffs presents a significant potential in the energy storage capacity of supermarket refrigeration systems by utilising demand side management [4]. Over the course of this paper, demand-side management is explored, and a potential approach is presented.

2 Background

Modelling of refrigeration systems is an exciting field in research and has been investigated in numerous studies in the literature to learn about various aspects of the system's behaviour. Several modelling techniques have been investigated in the literature, one of those techniques is data-driven modelling. Each component in the refrigeration system, such as the compressor or the evaporator can be individually modelled, and their behaviour is studied [5].

[6] explores data-driven modelling through artificial neural networks alongside predictive control. An attempt is made to improve the efficiency of the compressor in the refrigeration system by varying the compressor motor speed. Although significant energy savings are obtained, the shortcoming lies in the fact that training of neural networks is an extensive and expensive process and requires good quality experimental data. [5] further goes to establish a non-linear model based on the thermal dynamics of the cabinet and the evaporator as opposed to the compressor. This is a simpler model and achieves

promising results, however, it does compromise slightly on the accuracy of predictions.

However, as recognised by [7], while neural networks are much more flexible and good for eliminating noise, a physical model cannot be established in the form of an equation. Similar to [5], this paper uses a black box technique identifying time-varying and time-invariant linear models to design and optimise an MPC Controller. The time-varying model in particular yields favourable results, showing sufficient energy savings. Despite the potential of a linear dynamic model, an artificial intelligence (AI) approach using neural networks or fuzzy logics has been generally more popular for research. AI techniques for refrigeration modelling have been extensively discussed in literature, such as in [8], [9], and [10].

Establishing a functional model is only the first step to developing a control strategy that could potentially save energy and reduce costs. A supermarket refrigeration system suitable for supervisory control in the smart grid is explored in [11]. The fully integrated model consisted of three independent models which represented the three different subsystems [11]. The three subsystems were: the display case, suction manifold, and the condenser. Dynamic equations were identified to model heat transfers within the subsystems. Whereafter, parameter estimations were performed via an iterative prediction-error minimisation (PEM) method. Supervisory control was implemented with a PI controller to regulate the power consumption to the reference level received from the grid. The control objective was to ensure the average power consumption of the refrigeration system follows the reference and simultaneously stays within the temperature limits of the display cases. Simulations of the model with these subsystems corresponded well in confirming the effectiveness of the model approach.

In [12] control strategies in the heat recovery of a CO_2 booster system in a supermarket has been simulated and analysed for its energy, environmental and economic benefits. Five different heat recovery strategies were investigated. Each strategy analysis considered the energy consumption, operational cost along with its environmental impact. The models validated that it was able to reproduce the behaviour of the refrigeration system with high precision. The best energy strategy would be able to result in a reduction of total consumption by 32%. Heat recovery systems bring about significant benefits to improve the energy efficiency of energy-intensive buildings such as a supermarket. However, trade-offs will need to be considered carefully. This paper concludes that it is very difficult to model such a system's behaviour with a steady-state model of the installation. Therefore, a more accurate model would consider the system dynamics [13].

3 Motivation

Previous work has been conducted in designing, installing, and costing a model predictive control framework for a Heating, Ventilation and Air Conditioning (HVAC) system. The MPC scheme aims to provide an optimal temperature setpoint that will minimise the overall costs and carbon usage whilst maintaining the thermal comfort range of temperatures for the occupants of the supermarket. The full scope of this project aims to develop a fully functional refrigeration model. However, due to time limitations, this paper aims to simulate the refrigeration dynamics of several cabinets and implement control systems to determine how variations in the temperature setpoints of the cabinets can affect overall energy costs using available telemetry data.

4 Data

4.1 Description

All data used for the project is from a UK based supermarket. All data used is from 25th October 2022 to 22nd November 2022. The cabinet telemetry data is at a resolution of 2 minutes, while the compressor work data is at a resolution of 10 minutes.

4.2 Raw Data Visualisation

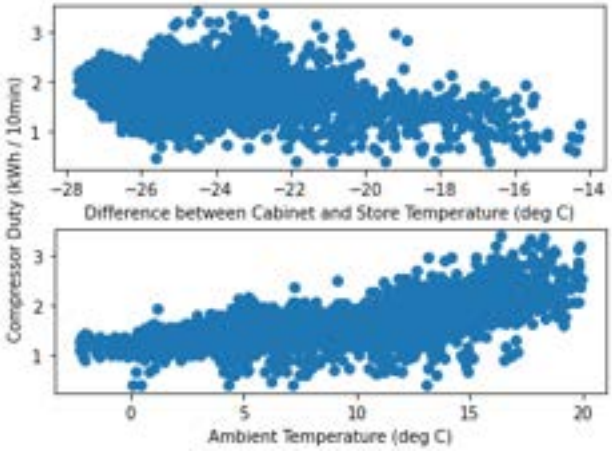


Figure 2: Plots showing raw data for the ambient temperature and the temperature difference between cabinet and store temperatures (ΔT) against the compressor duty.

Figure 2 depicts the general effect of ΔT and ambient temperature on the compressor duty. It is evident that as ambient temperature increases so does the compressor duty and as ΔT decreases (in magnitude), compressor duty decreases.

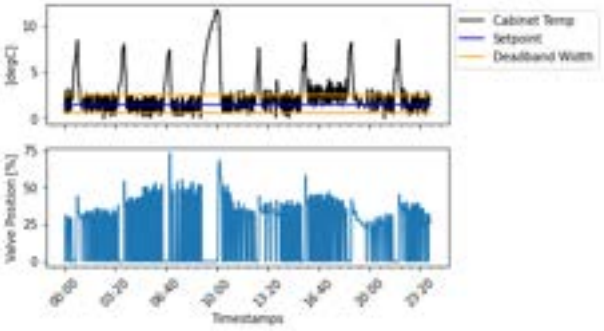


Figure 3: Plots showing the raw temperature data of a fridge cabinet and its corresponding valve positions for a sample day: 26th October 00:00-23:58.

Figure 3 shows the temperature variations and the valve positions of a fridge cabinet in time resolutions of 2mins. The temperature setpoint for this particular cabinet is observed to be 1.5°C with a deadband of 1°C. It is evident that a bang-bang controller is utilised as

there are fluctuations in the temperature between two states: above and below the desired setpoint.

5 Materials and Methods

5.1 Software

Python 3.9 was used throughout this project for all simulation and modelling purposes.

5.2 Model for the Compressor Duty

For the first model, a linear approximation was used. Noting that the driving force for the compressor work is the temperature difference between cabinet temperature and the store temperature, and accounting for the effects of the ambient temperature, a linear approximation is made in the form of the following equation:

$$\dot{W} = a_1(\bar{T}_{cabinet} - T_{store}) + a_2 T_{ambient} + a_3 \quad (1)$$

Where $\bar{T}_{cabinet}$ is the weighted average calculated with respect to each cabinet type mentioned in Table 1.
The equation can be simplified and written as:

$$\dot{W} = a_1 x_1 + a_2 x_2 + a_3 \quad (2)$$

Real-time data for x_1, x_2 and $\dot{W}$ is fed to python and the LinearRegression feature of scikit-learn is used to obtain the constants a_1, a_2 and a_3.

In order to assess if compressor duty is a linear function of the aforementioned temperatures, a neural network was also trained on the data. This was done using the MLPRegressor from scikit-learn.

Lastly, the effects of humidity were also investigated. This was done by simply including another linear term in the equation correlating to the humidity.
For all models, the data was split into test and training sets with a test size of 30% at random. This is kept consistent throughout all the models discussed in the course of this report.

5.3 Model for the Cabinet Temperature

Table 1: Classification of cabinet types based on their temperature range.

Cabinet Type	Temperature Range	Description	Total
1	-23 ≤ T ≤ -18	Freezer	13
2	0 ≤ T ≤ 2	Fridge	4
3	-1 ≤ T ≤ 3	Fridge	8
4	-1 ≤ T ≤ 7	Fridge	5
5	-1 ≤ T ≤ 5	Fridge	5

The cabinet temperature depends on the flow of the coolant which is controlled through a valve. The valve opening can be manipulated to adjust the cabinet temperature which in turn can be used to adjust the compressor duty. In order to devise a model for the cabinet temperature, a simple heat balance is taken over the cycle such that:

$$\frac{dT_{cabinet}}{dt} = \frac{Q_{cabinet}}{mc_p} \quad (3)$$

where $Q_{cabinet}$ is the cabinet duty denoted by $m_c\lambda$ and the mass of the coolant, m_c can be calculated by multiply the valve position by the valve constant, k_V. Substituting these inputs into the main equation yields:

$$\frac{dT_{cabinet}}{dt} = \frac{k_V x \lambda}{mc_p} \quad (4)$$

This equation can be further simplified into a discrete time temperature model, using a constant change in time, Δt:

$$T_{cabinet}(n+1) = \frac{k_V x \lambda}{mc_p}\Delta t + k_C T_{cabinet}(n) + c \quad (5)$$

This equation finally can be written in the form of $Y = mX + C$ such that:

$$T_{cabinet}(n+1) = m_1 x(n) + m_2 T_{cabinet}(n) + c \quad (6)$$

Similar to the compressor duty model, historic data for cabinet temperatures and valve positions are fed to python and the LinearRegression feature of scikit-learn is used to obtain the constants m and c. Within the refrigeration system, there are several cabinets, each operating over a range of temperatures. Based on temperature range of the product, the cabinets were classified into five types as shown in Table 1. A separate model was established for each cabinet type.

5.4 Control Scheme for the Refrigeration Cycle

With the obtained multivariate linear model equation for the cabinet temperatures as a function of valve position and previous cabinet temperatures, bang-bang control was implemented to control and regulate the temperature of the cabinet within a dead band of 1°C of the desired temperature set point. The principal theory of bang-bang control is the switch between two states when the state conditions are met.

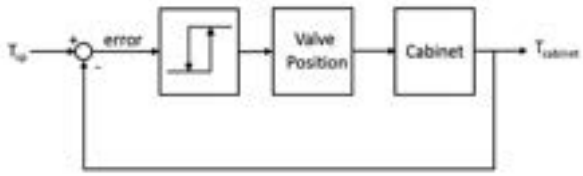


Figure 4: Control loop diagram for the bang-bang controller.

There was a total of five different types of cabinets that were investigated as shown in Table 1. The valve positions were used as the actuators in the control loop. Whereby, when the temperature of the cabinet falls below $T_{sp} - deadband$, the valve would be set at the lower limiting value and when the cabinet temperature reaches above $T_{sp} + deadband$, the valve position would be set to the upper limit value.

Table 2: Manually tuned upper and lower limits for valve positions for each cabinet type with their temperature setpoints.

Cabinet Type	T_{sp} (°C)	Valve Position (%)	
		Lower Limit	Upper Limit
1	-21.0	10	50
2	1.0	10	35
3	1.0	10	90
4	3.0	10	100
5	2.0	10	40

And so, with the implementation of bang-bang control for each type of cabinet, a weighted average temperature of all the cabinets at each timestamp was calculated and stored as a variable to use in the work

compressor model equation. This would then be used to simulate the variation of the work compressor duty as a function of time. Hence, an evaluation of the time periods of the highest and lowest duties was conducted and used to investigate how altering the temperature set points at different periods of the day could affect the overall work compressor duty and electricity tariffs.

A representative electricity price profile of an economy-7 type tariff was used for the cost analysis of the compressor duties of the cabinets. Table 3 presents the variation in electricity prices during the day.

Table 3: Electricity price used for the cost analysis during off-peak and on-peak tariff times.

Hour of the Day	Electricity Price [p/kWh]
00:00 – 16:00	10
16:00 – 20:00	20
20:00 – 00:00	10

The hourly tariff information was used to modify the temperature setpoints of the cabinets to optimise the energy consumption of the compressors in the refrigeration model. Whereby, temperature setpoints were increased during the hours of higher prices.

6 Results and Discussion

6.1 Model for the Compressor Duty

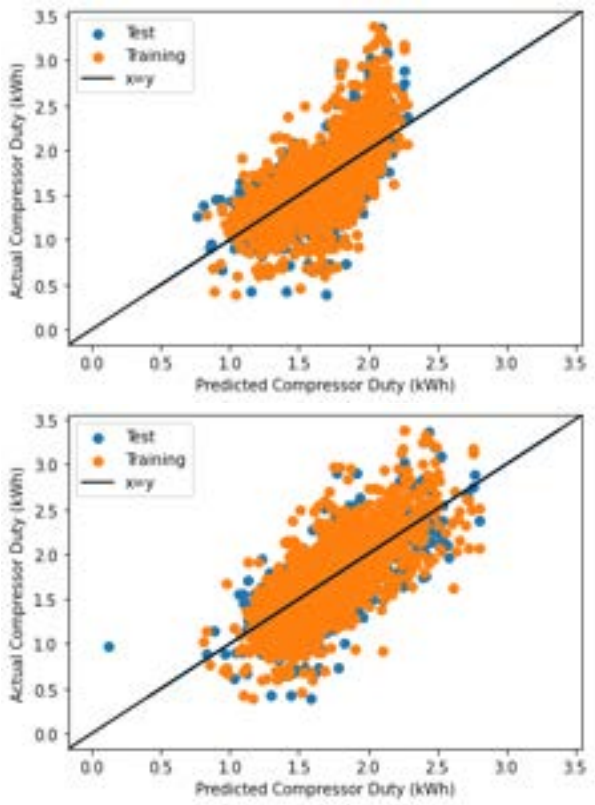


Figure 5: Linear model (top) and Neural network-based model (bottom) for the compressor duty, depicting model predicted values against real-time values for testing data and training data.

Figure 5 shows plots of the model predicted values against the actual values for the compressor duty for the linear regression and the neural net respectively.

The linear model shows favourable results with a decent R^2 score of 0.45 and a MAPE of 14.4%. The non-linear model performs slightly better with a marginally higher R^2 score of 0.53 and a MAPE of 13.4%. Although the neural network performs better, for simplicity, the linear model was utilised for further analysis. The coefficients a_1, a_2 and a_3 are summarised in Table 4.

Table 4: Results from the multivariate linear regression for the compressor duty, the coefficients correlate to Equation (1).

Coefficients	Value
a_1	0.057
a_2	-0.032
a_3	0.32

As explained in section 5.2, the effect of humidity was also assessed. The inclusion of humidity did not improve the results, in either case, yielding approximately the same R^2 score and MAPE. To assess the validity of this conclusion a p-value test was implemented on the coefficients. The results of the test are summarised in Table 8 in Appendix A. The results of the p-value test reiterated the conclusion that humidity does not significantly affect the compressor duty and only the temperature gradient as well as the ambient temperature have a measurable impact on the compressor duty. It is noted that the model is based on temperature conditions during the winter months thus the model validity is limited and is not accurate for the warmer summer months.

6.2 Model for the Cabinet Temperature

The model for each cabinet type performs very well, with an R^2 score greater than 0.9 for each cabinet type. Table 5 summarises the results from the multivariate regression with the constants for Equation (6). Figure 6 depicts the regression results for each cabinet type.

Table 5: Results from the multivariate regression for the cabinet temperature, the coefficients correlate to Equation (6).

Cabinet Type	m_1	m_2	c
1	-0.024	0.988	0.433
2	-0.041	0.880	1.113
3	-0.019	0.890	0.890
4	-0.031	0.946	0.744
5	-0.003	0.989	0.161

The R^2 score is very close to 1 for all models, denoting a near perfect fit. This makes sense as the cabinet temperature is mostly dependent on the coolant flow which is manipulated through the valve position.

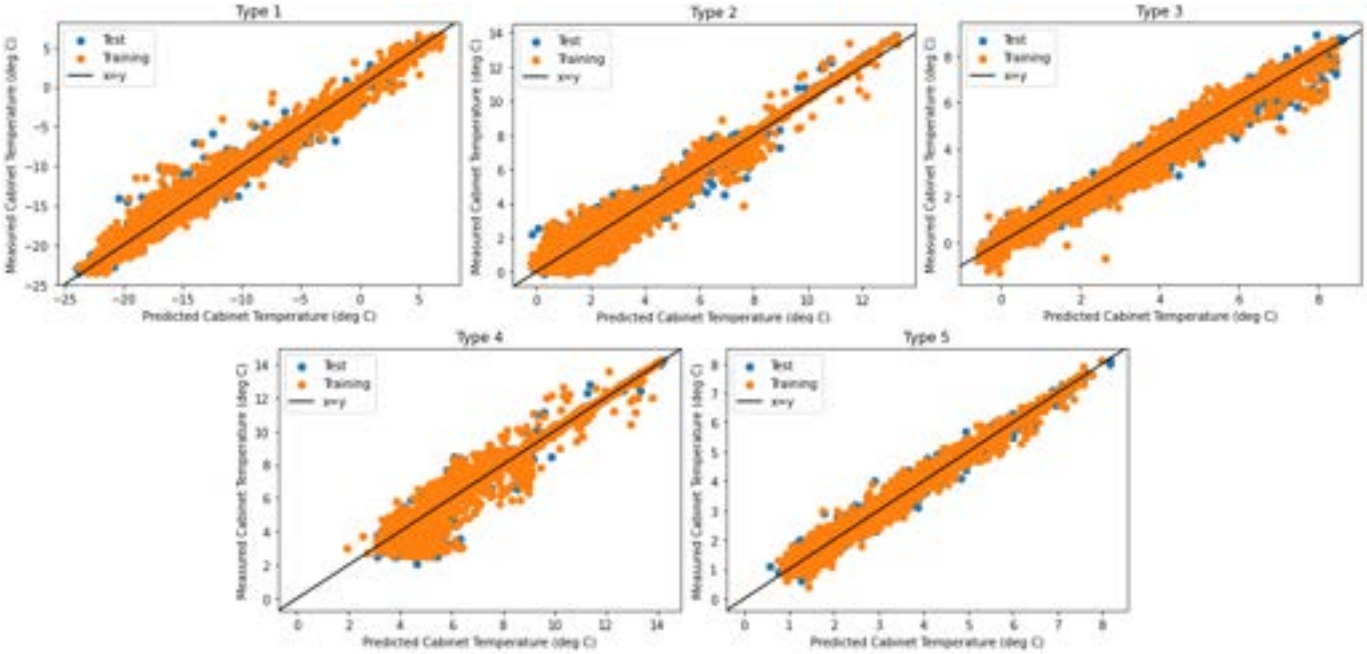


Figure 6: Linear model for the cabinet temperature, depicting model predicted values against real-time values for testing data and training data.

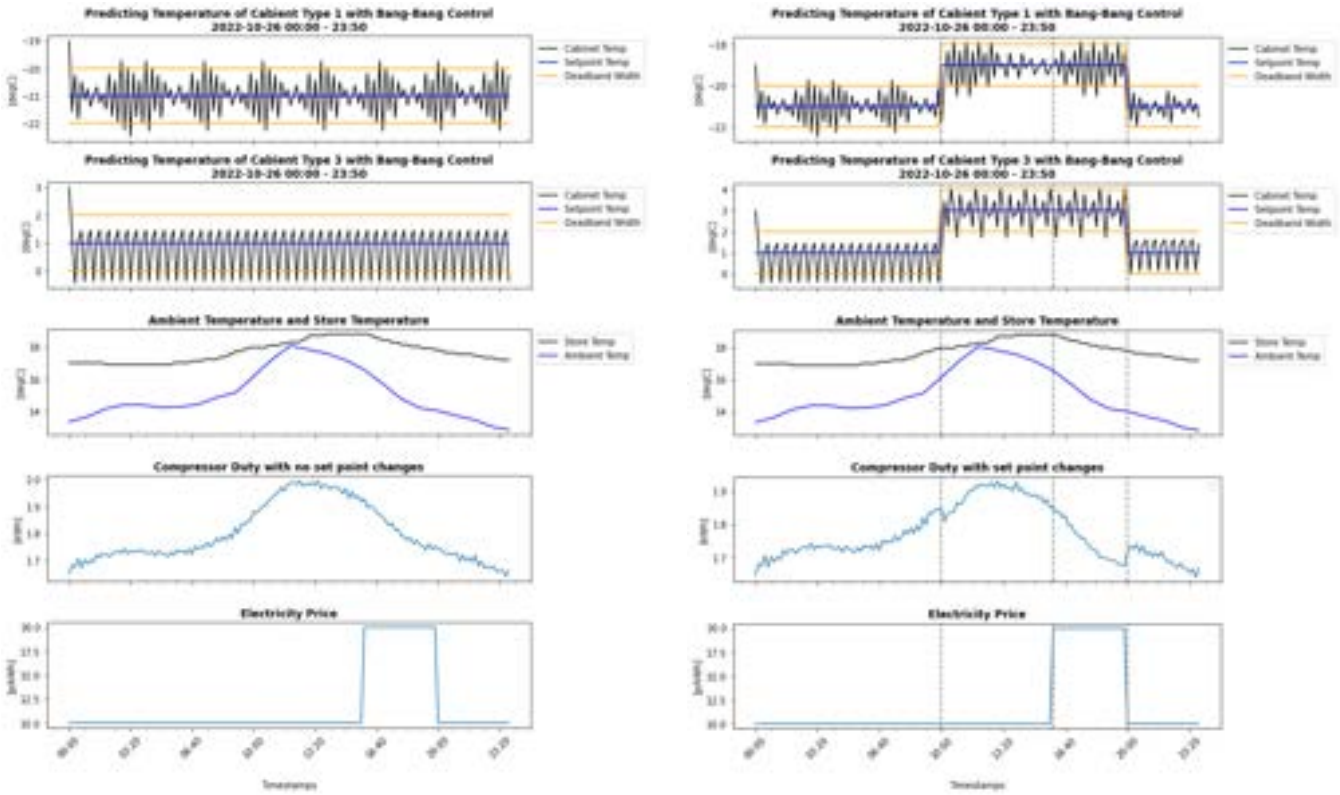


Figure 7: Compiled plots of the key results in the two simulations cases: no temperature setpoint changes (left) and with temperature setpoint changes (right).

6.3 Control Scheme for the Refrigeration Cycle

6.3.1 Controller Simulation

Upon simulation of the cabinet model equations with bang-bang control, Figure 10 in Appendix A presents the regulation of the temperatures of each of the cabinets to their corresponding setpoints along with how the valve positions change to accommodate the regulation for a sample day, 26th October 2022 00:00 – 23:50. Bang-bang control works well to control the cabinet temperatures as the figure shows the abrupt switches between the two states of the valve positions when the pre-determined high and low set points of the cabinet temperatures are reached.

As shown in Figure 10 in Appendix A, there are variations in how quickly the temperatures of the cabinets respond to the changes in the valve positions and this is due to the different types of cabinets having different resulting coefficients in their corresponding linear model equations. Cabinet type 4 seems to have a much slower response in changing its temperature with changes in the valve position as compared to the other types of cabinets. Therefore, it was important to manually adjust the upper and lower limits of the valve positions for each type of cabinet separately to ensure that the bang-bang controller was operating in the desired manner.

6.3.2 Compressor Duty

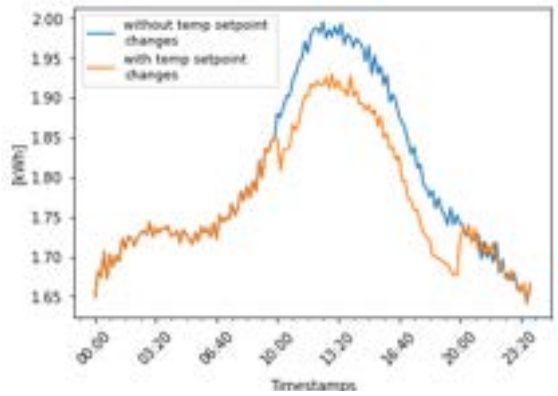


Figure 8: Variation of Compressor duty without and without temperature setpoint changes on 26th October 2022 00:00 – 23:50.

Using Equation (1) and the coefficients from Table 4, Figure 8 shows the resulting plot of how the calculated compressor duty varies throughout the day for the two cases: no temperature setpoint change and with temperature setpoint changes. The total duty was calculated to be 259kWh/day for the case with no setpoint change.

The compressor duty was the highest during the hours of 10:00 – 16:00. These were the periods of the day where the ambient temperature was the highest as depicted in the subplots of Figure 7. The combined subplots of Figure 7 present a useful insight into how the key parameters are affected in the two cases.

As the compressor duty model linear equation is a function of ambient temperature, it confirms that there is an increase in the compressor duty as ambient temperature increases. From this analysis, it can be deduced that this compressor duty will be much higher during the warmer seasons as generally, the ambient temperatures can get 15°C higher than those that have been used in the model dataset for this project. [14] confirms the relationship that compressor duty is expected to rise during the summer months and reduce during the winter months as a result of the higher ambient temperatures. They found that there was a 15% increase in the compressor duty in the summer months relative to the winter months. Hence, this confirms that as ambient temperatures increase, the compressor duty is also expected to increase.

Using the economy-7 electricity tariff as shown in Table 3 of section 5.4, the total daily energy cost from the compressor duty was calculated to be £30.25.

Further analysis was conducted to evaluate the relationship between the total daily compressor duty and overall temperature setpoint changes for all cabinets. The temperature setpoints were varied by 1°C intervals between +3°C and -3°C from the initial setpoint temperatures.

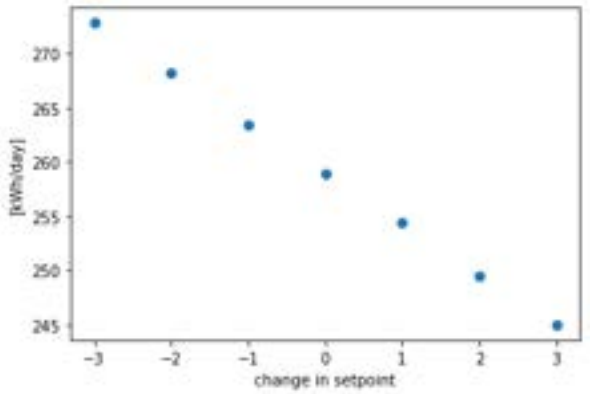


Figure 9: Daily compressor duty against 1°C interval changes in the temperature setpoint of all the cabinets.

Figure 9 shows a linear relationship between a change in setpoint across all cabinets and their compressor duty. With every 1°C increase in the setpoint, there is a 1.8% decrease in the compressor duty per day. This is expected as raising the temperature setpoint of the cabinets reduces the temperature difference between the store and the cabinet and hence will require less compressor work to maintain the lower temperatures.

Considering the daily electricity prices and periods of the day where the compressor duty is the highest, temperature setpoint changes were implemented to all the cabinets as presented in Table 6.

Table 6: Chosen variations in the temperature setpoints across all cabinets

Hour of the Day	Change in temperature setpoint
00:00 – 10:00	0
10:00 – 16:00	+2
16:00 – 20:00	+2
20:00 – 00:00	0

According to [15] and [16], it is good practice to store and handle frozen foods in temperatures of -18°C or lower and cold foods at 5°C. The increases in the temperature setpoint of all the cabinets outlined in Table 6 have accounted for these food safety standards.

With the temperature setpoints changes as shown in Table 6, the daily compressor duty was recalculated to be 255kWh/day. This is a 1.5% reduction in the compressor duty as a result of varying the temperature setpoints which has considered the higher compressor duty periods and periods of higher electricity rates.

The total daily energy cost from the compressor duty with setpoint changes was calculated to be £30.02. This is a savings of only 23p per day. The small savings can be due to the lower ambient temperatures and the store temperatures in the colder seasons. Following equation 1, there is a reduced driving force as the temperature difference between the cabinet and the store would be less. Hence, the compressor duty would not be significantly large to incur large savings. Overall daily compressor duties and energy costs have been summarised in Table 7.

Table 7: Summary of compressor duty and energy costs.

	Base case	**Case 1**	**Difference (%)**
Total compressor duty [kWh/day]	259	255	-1.5
Total Cost [£/day]	30.25	30.02	-0.8

7 Conclusion

An attempt was made to establish a simplistic model for compressor duty and alter it indirectly by controlling cabinet temperatures by changing valve position. A simple bang-bang control model was used for this purpose. With the implementation of the control scheme, savings of 23p per day were obtained, which translates to saving £83.95 a year. The small amount of savings seems to suggest that the implementation of the control scheme is not worthwhile, especially because of the costs associated with installing and maintaining a cloud-based energy management system.

Additionally, with respect to the costs before implementation, a 0.8% decrease in costs is observed as well as 1.5% savings in energy. The percentages very well suggest that the cost and energy savings are significant enough to consider cloud-based solutions, primarily due to the reduction in the supermarket's carbon footprint.

8 Outlook

A more in-depth study of the compressor work duty correlation is required to obtain a more accurate model to improve the quality of the fit. Additionally, the effects of various weather conditions should be studied to establish a more rigorous model. With a better prediction of compressor duty, the control model will work qualitatively better, yielding higher savings.

It is also important to note that the model equations obtained from simulations have been based on the air temperatures of the cabinets as opposed to the temperatures of the foodstuffs. However, it has been assumed that the heat transfer coefficients of the foodstuffs are very low that it will take a long time for the temperature of the foodstuffs to vary for a change in the cabinet air temperature, so a model based on foodstuff temperatures is worth exploring in future analysis.

Further investigations into more advanced control theories, such as proportional-only control (P control) and proportional and integral control (PI control) as well as feedback control are recommended. These controllers would work to reduce the error between the measured cabinet temperatures and the desired setpoint, hence it would help to eliminate the fluctuations in the cabinet temperatures within the deadbands and optimise control to reach desired setpoint.

While control strategies explored in this paper do reduce costs, an investigation into alternative approaches to reduce the overall compressor duty and hence the energy consumption of refrigeration systems within supermarkets is a simpler solution. Currently, supermarkets do not have doors installed on the cabinets and this can have opposing effects in the energy balance whereby unnecessary energy can be lost due to the heat transfer from the cooler cabinets to the warmer stores. This would mean more compressor duty is required to maintain the lower temperatures of the cabinets and also to maintain the comfortable temperature ranges of the stores due to the additional cooling from the heat transfer from the cabinets to the store. Simple solutions such as installing doors on the cabinets could be more cost effective and sustainable.

Refrigeration systems typically use regular Joule-Thomson valves. Since these systems in supermarkets are large enough, it might be worth investing in turbo-expanders. Turbo-expanders can be used to recover wasted energy and therefore can have beneficial impacts in increasing the energy efficiency of the system [17].

9 Acknowledgement

We would like to thank Max Bird for his continued support and advice over the course of this project.

References

[1] "What did the UK Presidency aim to achieve at COP26?," UN Climate Change Conference UK 2021, [Online]. Available: https://ukcop26.org/cop26-goals/. [Accessed 01 12 2022].

[2] C. Read, "SMART ENERGY MANAGEMENT: HOW WILL THE GRID COPE WITH OUR CHANGING ENERGY HABITS?," Thales, 06 12 2021. [Online]. Available: https://www.thalesgroup.com/en/united-kingdom/news/smart-energy-management-how-will-grid-cope-our-changing-energy-habits. [Accessed 01 12 2022].

[3] S. Acha, Y. Du and N. Shah, "Enhancing energy efficiency in supermarket refrigeration systems through a robust energy performance indicator," *International Journal of Refrigeration,* vol. 64, 2016.

[4] S. E. Shafiei, H. Rasmussen and J. Stoustrup, "Modeling Supermarket Refrigeration Systems for Demand-Side Management," *Energies,* vol. 6, pp. 900-920, 2013.

[5] K. Leerbeck, P. Bacher, C. Heerup and H. Madsen, "Grey box modeling of supermarket refrigeration cabinets," *Energy and AI,* vol. 11, 2023.

[6] Ö. Kizilkan, "Thermodynamic analysis of variable speed refrigeration system using artificial neural networks," *Expert Systems with Applications,* vol. 38, no. 9, pp. 11686-11692, 2011.

[7] T. Siqueira Dantas, I. Franco, A. Fileti and F. Silva, "Dynamic linear modeling of a refrigeration process with electronic expansion valve actuator," *International Journal of Refrigeration,* vol. 75, pp. 311-321, 2017.

[8] X. Cao, Z.-Y. Li, L.-L. Shao and C.-L. Zhang, "Refrigerant flow through electronic expansion valve: Experiment and neural network modeling," *Applied Thermal Engineering,* vol. 92, pp. 210-218, 2016.

[9] M. Hosoz, H. Ertunc and H. Bulgurcu, "An adaptive neuro-fuzzy inference system model for predicting the performance of a refrigeration system with a cooling tower," *Expert System Applications,* vol. 38, pp. 14148-14155, 2011.

[10] R. Ahmed, S. Mahadzir, N. E. Mohammad Rozali, K. Biswas, F. Matovu and K. Ahmed, "Artificial intelligence techniques in refrigeration system modelling and optimization: A multi-disciplinary review," *Sustainable Energy Technologies and Assessments,* vol. 7, 2021.

[11] E. Shafiei, H. Rasmussen and J. Stoustrup, "Modeling supermarket refrigeration systems for supervisory control in smart grid," *Proceedings of the American Control Conference,* pp. 5660-5665, 2013.

[12] E. J. Escriva, S. Acha, N. LeBrun, V. Francés, J. Ojer, C. Markides and N. Shah, "Modelling of a real CO2 booster installation and evaluation of control strategies for heat recovery applications in supermarkets," *International Journal of Refrigeration,* vol. 107, 2019.

[13] A. Beghi, M. Rampazzo and S. Zorzi, "Reinforcement Learning Control of Transcritical Carbon Dioxide Supermarket Refrigeration Systems.," *IFAC-PapersOnLine,* vol. 50, pp. 13754-13759, 2017.

[14] Y. Ge and S. Tassou, "Performance evaluation and optimal design of supermarket refrigeration systems with supermarket model "SuperSim", Part I: Model description and validation," *International Journal of Refrigeration,* vol. 34, no. 2, pp. 527-539, 2011.

[15] Food Standards Agency, "Safe Method: FROZEN STORAGE AND DISPLAY.," [Online]. Available: https://www.food.gov.uk/sites/default/files/media/document/sfbb-retailers-frozen-storage.pdf. [Accessed 10 12 2022].

[16] Food Standards Agency, "Chilling food correctly in your business," 05 12 2018. [Online]. Available: https://www.food.gov.uk/business-guidance/chilling-food-correctly-in-your-business#freezing. [Accessed 12 12 2022].

[17] Ipieca, "Turbo-expanders," [Online]. Available: https://www.ipieca.org/resources/energy-efficiency-solutions/power-and-heat-generation/turbo-expanders/. [Accessed 14 12 2022].

[18] E. Shafiei, J. Stoustrup and H. Rasmussen, "A supervisory control approach in economic MPC design for refrigeration systems.," *2013 European Control Conference,* pp. 1567-1530, 2013.

Development of a Catalytic System for Furfural Oxidation

Linqi Chen and Nicholas Lau
Department of Chemical Engineering, Imperial College London, U.K.

Abstract Furfural oxidation to furoic acid is a crucial step in producing a new biobased surfactant, SAF. Two schemes, stoichiometric and heterogenous oxidation, are investigated to find the optimal conditions and yield of furoic acid synthesis. It was found that the yield can approach 73% for stoichiometric oxidation using H_2O_2 as the oxidizer, methanol as solvent, and at a temperature of 60 ℃. For aerobic oxidation, the maximum yield is 75%, which can be achieved at a temperature of 120 ℃ and 5% Ru/C catalyst loading. The catalyst shows good potential for recyclability, whose activity only declines about 1.2% per round of experiments. The catalyst was then subject to physiochemical studies, suggesting that the oxide overlayers' leaching on the ruthenium surface is the main reason for the loss in ruthenium contents. Despite the comparable yields observed in both oxidation schemes, we concluded that aerobic oxidation is a more desirable scheme for SAF commercialization as it does not require the purification of furfural and uses air as the oxidant. Future work on a continuous system can be applied to study the deactivation and regeneration of catalysts in more detail.

Introduction and Background

Surfactants are surface active agent substances that can modify the interfacial tension liquid-gas or liquid-liquid interfaces because of their unique molecular structures. Thanks to this propriety, surfactants have become one of the most used commodity chemicals, with applications in diverse fields (household, industry, agriculture, personal care, oil and gas, food, and pharmaceuticals)[1]. The global surfactant market size was 39.42 billion in 2020 and was expected to increase to 57.81 billion by 2028.[2] The increase is driven by the growing demand for consumer products such as detergents. Due to the large volumes commercialized, biodegradability and toxicity are essential factors to be considered in combination with the low cost and sustainability of the raw material.

Currently, the most widely used surfactants are anionic surfactants, which account for more than 20% of a typical household product.[3] In particular, linear alkyl benzene sulfonates (LAS) and sodium dodecyl sulfates (SDS) are among the most widely used surfactants to meet household laundry requirements. LAS, a typical petroleum-based surfactant, can cause various detrimental effects on aquatic/terrestrial ecosystems besides contributing to positive net CO2 emissions.[4] While biobased surfactants, such as SDS, often perform lower due to lower resistance in hard water and relatively high critical micelle concentration (CMC). Therefore, an environmental-friendly surfactant with performance and costs comparable to those of LAS is needed.

Recently, the concept of sugar-based surfactants has received more and more attention. With well-established industrial technologies, a glucose-based surfactant, alkyl polyglycolide, reached a production capacity of approximately 80000 tons annually in the early 21st century.[5] However, it is a non-ionic surfactant with several inherent disadvantages compared to anionic surfactants since less effective.

Few catalytic pathways have been reported to transform sugars into surfactants through furanic intermediates to produce sugar-based anionic surfactants with high performances. However, many of them suffer from a complex and expensive pathway, which significantly limits the scalability of the process.

In 2022, Al Ghatta and coworkers introduced a new biobased family of anionic surfactants called sulfonated alkyl furoates (SAFs) based on ester linkages, which has excellent potential for up-scaling. Compared with other sugar-based surfactants, SAFs have superior performances because of their high resistance to hard water with low CMC and high flammability with improving Krafft point. Compared to other furan surfactants, the synthesis of SAFs is highly scalable because of its well-established catalyst selection and purification techniques. Besides, SAFs also have a more favorable atom economy than SDSs, and their production from waste resources, such as corn cob, dramatically benefits the future circular economy. As a result, SAFs have shown a promising prospect as a rising sustainable surfactant with good performance and reasonable costs.[6]

The synthetic route of SAFs is illustrated in **Figure 1**. Li and coworkers extracted furfural from corn cob under mild reaction temperature (170-190℃): γ-valerolactone as the solvent, H-ZSM-5 as the catalyst, with a maximum furfural yield of 71.68%.[7] Hidayat and coworkers evaluated corn cob's economic value for furfural production and designed a preliminary corncob-based furfural manufacturing plant.[8] Furoate ester can be synthesized by mixing furoic acid (FA) and dodecanol (DOD) under 150 ℃ and acidic conditions with a Dean-Stark apparatus. The SAFs can then be subsequently produced by sulfonation, adding chlorosulfonic acid at stoichiometric conditions. [6] In contrast, the chemical step from furfural to furoic acid has been considered the weakest link in the synthetic chain because of its complex mechanism with multiple competing pathways.

Two schemes can be considered for producing FA from furfural: stoichiometric oxidation and heterogeneous catalytic aerobic oxidation. Stoichiometric oxidation systems always require expensive oxidants, such as H_2O_2, with the advantage of not using a catalyst, while heterogeneous reaction systems can use more economical O_2 as an oxidizer, but a noble metal catalyst is generally required. As a result, both schemes have the potential to be implemented in SAF production.

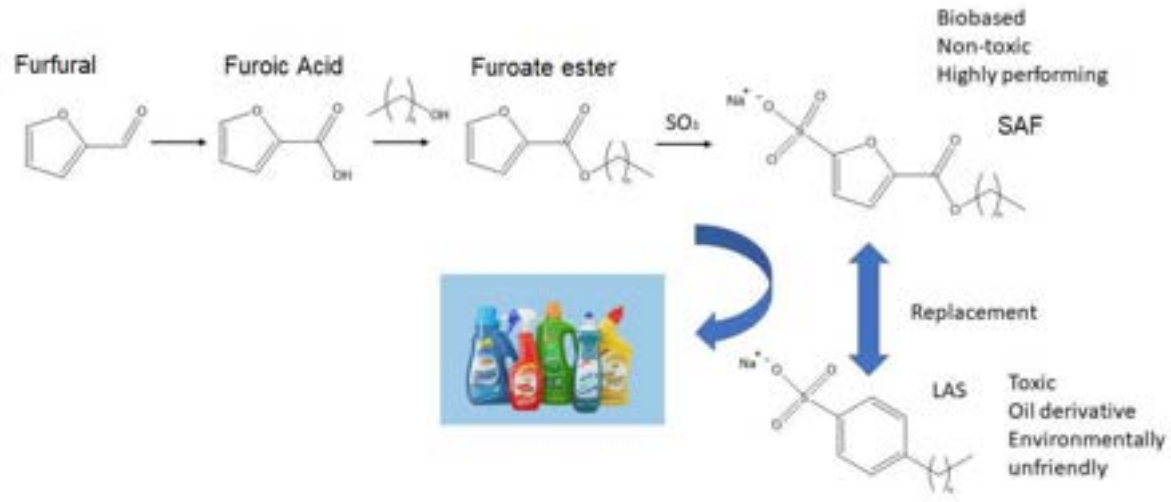


Figure 1. Schematic of SAF synthetic route

The most common route for large-scale production of furoic acid is through the Cannizzaro reaction, which uses sodium hydroxide to convert furfural into equimolar amounts of furoic acid and furfuryl alcohol, as shown in **Figure 2**. The major drawback of the Cannizzaro reaction is that the maximum theoretical yield is only 50%.[9] Although furfural can be directly oxidized to furoic acid using oxygen gas with a metal catalyst; such a reaction scheme still needs to be tested at a large scale due to the great difficulty of recovering heterogeneous catalysts after the reaction.[10]

Figure 2. Schematic of the Cannizzaro reaction

Kazuhiko and coworkers demonstrated homogenous catalyst-free oxidation of non-furan aldehydes to carboxylic acids using H_2O_2 in an acidic condition. They investigated the effects of reaction temperature and substrate- H_2O_2 ratio on the yield of desired carboxylic acid, achieving yields as high as 93%, proving the potential of homogenous oxidation.[11] The use of H_2O_2 as an oxidizing agent has also attracted attention from researchers in green chemistry due to its cleanliness.[12] Within this research, we present a reaction scheme for furfural similar to Kazuhiko's work, using H_2O_2 as the oxidizing agent and NaOH as the base, as shown in **Figure 3**. Due to basic conditions, however, the Cannizzaro reaction competes with the primary oxidation reaction, causing the overall selectivity to decrease. Furthermore, the decomposition of H_2O_2 to water and oxygen is catalyzed at high pH. At high H_2O_2 concentrations, the oxidation of furan compounds also tends to be uncontrolled, leading to over-oxidation of the desired product and the formation of by-products.[13] These aspects of the reaction had not been investigated in published work and are studied within the project.

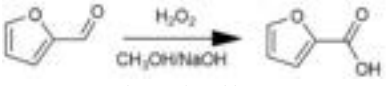


Figure 3. Schematic of homogeneous oxidation using H_2O_2

Another scheme of FA production is through aerobic oxidation, as shown in **Figure 4**. Sadier and coworkers evaluated the oxidation of furfural to furoic acid in an alkaline solution under 15 bar of air and room temperatures in a batch reactor with a TiO_2-supported Ag catalyst. They investigated the effect of temperature, air pressure, base equivalent, and the nature of the inorganic base used separately and achieved a maximum yield of 96% FA under optimum conditions. However, the high yield of FA is only possible with NaOH and a very high pH. Furthermore, the catalyst recyclability is limited, where the catalytic performance declines significantly after three runs. The decline in catalyst activity is associated with an increase in the size and the re-oxidation of the silver particles in a combination of side products poisonous to the active site.[14] Nocito and coworkers described the use of MnO2@CeO2 core−shell oxide in the selective aerobic oxidation of furfural. They investigated the role of the morphology of the catalyst on the reaction yield in detail and claimed that up to 96% yield of FA was observed. The catalyst also has excellent recyclability, where no significant loss was seen after ten cycles and 50 hours of operation. However, to achieve a high yield of FA, a high catalyst loading (20%-30% mass ratio) and long operating time (5h) are required. Moreover, the reaction mechanism is not fully understood and proven, and the high loading of the catalyst doesn't justify the regeneration risk leading to deactivation in the long term. [15]

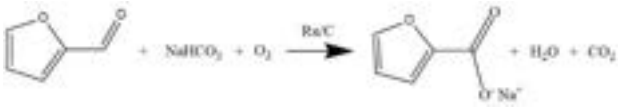


Figure 4. Schematic of aerobic oxidation on Ru/C

Ru/C is a commercial catalyst with vast applications in green and sustainable chemistry. Yi and coworkers researched the catalytic conversion of 5-hydroxymethyl furfural (HMF) to 2,5-furan dicarboxylic acid (FDCA) under basic conditions. With commercial Ru/C catalysts, they managed to achieve an FDCA yield of 88%.[16] The Ru/C catalyst requires a weaker base to achieve maximum yield and a relatively short reaction time, which opens the prospect of using more concentrated furfural and increases

the potential of scalability since furfural is unstable under strongly basic conditions. However, schematic determination of the optimal reaction conditions and investigation of the catalyst recyclability for Ru/C has yet to be extensively reported, which is studied within the project.

Aim of The Project

The project aims to study different approaches for furfural oxidation through the usage of a stoichiometric oxidant and aerobic approach, performing an optimization study on the critical parameters which affect the selectivity. For the stoichiometric oxidation, temperature, solvent, and the furfural-NaOH-H_2O_2 ratio will be varied, while temperature, air pressure, catalyst loading, and the base ratio will be assessed for the aerobic oxidation through a design of experiment approach. In addition, the recyclability of Ru/C for the aerobic oxidation of furfural is studied.

Ultimately, we aim to compare the maximum yield of stoichiometric with aerobic oxidation, along with their advantages and limitations. This should give an overview regarding which system is more beneficial for a scale-up. The research outcome should fill the gap in the synthetic route of SAF and contribute to the commercialization of SAF in the future.

Methods

Materials

All reagents, including pure furfural, furoic acid crystals, absolute dry methanol, 50% hydrogen peroxide, ruthenium on carbon catalyst, and sodium hydrogen carbonate, were purchased from Sigma-Aldrich and used without purification. 20% and 50% sodium hydroxide solutions were prepared by dissolving sodium hydroxide pellets in deionized water, weighted accordingly.

Analytical Method

Standard furfural and furoic acid solutions were prepared by dissolving respective compounds in deionized water and diluted accordingly. The standard solutions were placed in high-performance liquid chromatography (HPLC) to obtain calibration curves for each compound, enabling measurement of their concentration in samples produced by experiments.

Study of furfural stability

The reaction was conducted in a round flask with 20g, 3% furfural dissolved in deionized water in a heating plate and stirred with a magnetic bar. 20% sodium hydroxide solution was added in one portion, with the amount depending on the furfural-NaOH ratio to be analyzed. Samples were withdrawn from the mixture every 5 to 10 minutes for 1 hour, diluted, and analyzed with HPLC. The conversion of furfural is calculated with a calibration curve. Such reaction was repeated at five different furfural-NaOH ratios (0.01 to 1) at three different temperatures (20, 40, and 60 degrees C).

Stoichiometric oxidation of furfural

5g pure furfural was diluted with methanol to reach the target furfural concentration, employing a magnetic stirrer in a heating plate (**Figure 5**). 50% hydrogen peroxide and 50% sodium hydroxide were added dropwise to the mixture. To avoid thermal runaway, the addition of reagents was stopped when the temperature of the mixture exceeded 60 ºC and resumed when the temperature dropped below the same threshold. Samples were withdrawn from the mixture and analyzed through HPLC. The reaction was repeated at different furfural-H2O2-NaOH ratios and at 15 ºC, and an ice bath was used to absorb heat from the reaction. Following the addition of reagents, the mixture was heated at 55 ºC for 1 hour. A condenser was also connected to reduce the evaporation of methanol. Precipitates formed from the reaction were filtered and analyzed through HPLC.

Aerobic Oxidation

3% furfural was prepared by dissolving furfural in distilled water. Before performing aerobic oxidation, 3% furfural was filtered to remove any precipitation, which could potentially behave as poison on the ruthenium catalyst.

The high-pressure Parr reactor (**Figure 6**) is preheated to 100 ℃ to reduce the heating time. 10g of 3% furfural was prepared, and the catalyst and base were measured based on the equivalent ratio specified in the DOE. The reaction mixture was put in a vial and placed inside the preheated reactor. The reactor was quickly sealed before the vial temperature reached 100℃. Upon proper sealing, the reactor was pressurized using an air cylinder. The reaction time is counted when the temperature reaches the set point. 5℃ fluctuation was considered acceptable.

At the end of the reaction, the heater and mixer were turned off, and the reactor was cooled inside an ice bath. After the temperature decreased below 50℃, the reactor could be depressurized and unsealed. The vial was removed from the reactor, and the reaction mixture mass was measured. A sample (30-50 mg) was withdrawn and diluted with water (1 mL). The sample was then analyzed in HPLC, and the FA concentration could be calculated accordingly.

Figure 5. Experimental set-up of stoichiometric oxidation

Figure 6. Parr Series 5500 HPCL Reactor w/4848 Controller & Opt. Expansion Modules [17]

Catalyst Recyclability Test

20g of 3% furfural was prepared and reacted through aerobic oxidation following the methodology above. After each reaction, the catalyst was separated through vacuum filtration. Water and acetone were used to wash the catalyst (6*20 ml). The catalyst was then dried, and its mass was remeasured upon drying. The pH of the liquor was measured.

The dried catalyst was then used to perform a second experiment, and the amounts of base and furfural were adjusted accordingly to keep the same initial ratio. Similar procedures were performed until the catalyst was used six times. All samples were analyzed through HPLC. Catalyst characterization was done through Transmission Electron Microscopy (TEM), Inductively Coupled Plasma (ICP), and Brunauer–Emmett–Teller (BET) analysis to describe the particle size, ruthenium content, and catalyst surface area, respectively.

Results and Discussion

Stability Test of Furfural in NaOH

Under the presence of NaOH, furfural undergoes a Cannizzaro reaction to produce equimolar amounts of furoic acid and furfuryl alcohol. The conversion of furfural increased with the quantity of NaOH (**Table 1**). During the reaction, the pH of the reaction mixture decreased, indicating the consumption of hydroxide ions. Although the conversion of furfural was not expected to be higher than twice the base ratio due to reaction stoichiometry, such phenomena were observed. This suggests that since the furfural ring is prone to attack by nucleophiles in the aromatic ring (C1 carbon), aldol condensation of furfural could have resulted through ring opening forming side polymerized products. [18]

The conversion was also found to increase with temperature, reaching 64% at 60°C, contributed by the increase in reaction rate. It was observed that the

Base ratio	Temperature (°C)	Conversion (%)
0.01	20	13
0.05	20	20
0.5	20	25
1	20	34
0.05	40	29
0.05	60	64

Table 1. The effect of NaOH quantity on the stability of furfural in water

temperature of the mixture didn't change at all conditions, suggesting that the Cannizzaro reaction has a low enthalpy change of reaction.

Stoichiometric Oxidation of Furfural

The effects of molar ratios of NaOH to furfural, H_2O_2 to furfural, temperature, and solvent on the conversion of furfural and yield of furoic acid were investigated, with the aim of finding the optimal reaction condition (**Table 2**).

In experiment A1, a conventional Cannizzaro reaction pathway for furoic acid production was attempted by adding only NaOH. Such a reaction converted 40% of furfural achieving a 20% yield of furoic acid, which is an exact match to the theoretical selectivity of Cannizzaro reaction. The temperature of the reaction mixture also remained unchanged despite the continuous addition of furfural into NaOH. Due to the unreacted furfural present, the mixture was light orange in color, however, contained no precipitate.

When an equimolar amount of base and peroxide were added, an unsatisfactory yield and selectivity of 13% and 12% resulted respectively (A2). Despite the continued addition of excess base and peroxide until 4 and 8 molar equivalents (A3), the yield of furoic acid only increased by 4%. Cong and coworkers investigated the stoichiometric oxidation of substituted benzaldehydes at the same substrate-base-peroxide ratio, solvent, and temperature, yet achieving yields from 81% to 97%. [19] However, as the furan ring is prone to uncontrolled oxidation in excess peroxide, side reactions such as ring-opening might have decomposed furoic acid produced. This phenomenon may impose a maximum limit to the ratio of peroxide in order for yield to be optimal, such that over-oxidation is not dominant.

The reaction also produced 6.7g of white precipitate soluble in water under acidic conditions, generating gas bubbles during dissolution. The solid could not be successfully characterized through NMR or HPLC.

In experiments A4-6, the peroxide ratio was varied from 1 to 4, and a higher base ratio of 2 was selected. The yield increased from 51% to 73%, indicating that the yield and selectivity limit of the Cannizzaro reaction had been exceeded. A fast rise in temperature was also observed, suggesting the oxidation by peroxide dominated. 1.84g of precipitated was filtered from the mixture.

Experiment	Furfural mass (%)	Base ratio	Peroxide ratio	Temperature (°C)	Solvent	Yield (%)	Conversion (%)	Selectivity (%)
A1	10	2	0	60	Methanol	20	40	50
A2	10	1	1	60	Methanol	13	82	12
A3	10	4	8	60	Methanol	17	82	17
A4	10	2	1	60	Methanol	51	68	75
A5	10	2	2	60	Methanol	63	100	63
A6	10	2	4	60	Methanol	73	100	73
A7	10	2	1	60	Toluene	91	92	99
A8	10	2	2	60	Toluene	100	100	100
A9	100	1	2	60	-	39	83	47
A10	100	1	1	10	-	26	87	30
A11	100	2	2	10	-	29	95	31
A12	10	2	2	10	Methanol	19	29	66
A13	5	2	2	10	Methanol	19	25	76
A14	20	2	1	10	Methanol	33	100	33
A15	20	2	2	10	Methanol	40	100	40

Table 2. The effect of reaction parameters on stoichiometric oxidation of furfural

HPLC analysis indicated only 37mg of furoic acid was present in the solid, corresponding to a purity of 2%, while furfural was undetected.

The reaction was repeated at the same condition (A7, A8), but with toluene as solvent. With one molar equivalent of peroxide, a yield of 91% was achieved with selectivity approaching 100%. However, the phase separation between the toluene and ionic phase posed a challenge to the accurate calculation of yield and conversion. As residual furfural likely remained in toluene while furoic acid remained in the aqueous phase, after normalizing with the total volume of the mixture, the total quantity of furfural was underestimated, and that of furoic acid was overestimated. Therefore, the actual yield and conversion are smaller than the ones reported. Rodrigues and co-workers reported the formation of 2-methyl furoate in a similar stoichiometric oxidation using methanol as solvent and H_2O_2 as oxidant.[20] Due to the absence of esterification, the change to toluene as a solvent can potentially improve yield and selectivity.

On the other hand, when pure furfural was oxidized without any solvent, the yield and selectivity dropped to 39% and 47% respectively. Using a lower temperature of 10°C in experiments A10-11 resulted in even lower yields and selectivity. The chromatogram from HPLC analysis of the reaction mixture and precipitate showed several peaks, indicating the abundance of side products. The reaction mixture also appeared dark, suggesting the polymerization of furfural had occurred.[21] As there was no solvent to dilute H_2O_2, the rate of oxidation increased significantly, causing stronger competition from side oxidations and ring-opening reactions as mentioned previously.

The hypothesis is reaffirmed by observing the trend of selectivity when the concentration of furfural in methanol was varied. From experiments A12-15, a higher selectivity towards furoic acid was achieved at a lower concentration of furfural. Nevertheless, a longer reaction time may be needed to obtain the same conversion and yield due to the slow reaction rate.

In general, the conversion at 60°C exceeded that of 10°C due to an increase in reaction rate. The selectivity towards furoic acid also followed the same trend. An analysis of reaction kinetics showed that at a higher temperature, the rate of generation of furoic acid from the Cannizzaro reaction increased by a larger extent than oxidation by peroxide. The activation energy of the former is 69.3kJ/mol and follows the 2nd-order rate law in furfural, while the latter has a smaller activation energy of 44.6kJ/mol and follows the 1st-order rate law in furfural.[22,23] Therefore, the high selectivity of reaction (>50%) was contributed by both the Cannizzaro reaction and oxidation, with oxidation being favored at low temperature and low furfural concentration.

We conclude that the optimal reaction condition is the use of a furfural-base of 2 and the furfural-peroxide ratio of 4, methanol as solvent, and at a temperature of 60°C. Under such conditions, furoic acid can be obtained at a 73% yield. In systems where the reaction time of stoichiometric oxidation of furfural to furoic acid needs to be minimized, the concentration of furfural and peroxide ratio ought not to be increased beyond optimum since doing so will result in the prevalence of over-oxidation. Instead, the temperature and solvent type should be modified.

Aerobic Oxidation of Furfural

Since the aerobic oxidation needs to be carried at high pressure, the addition of NaOH as done in the previous set-up is not possible. Therefore, it was decided to use $NaHCO_3$ as a weaker base to limit the side reaction of furfural. To understand the stability of furfural in the presence of sodium hydrogen carbonate, a stability test has been performed at different temperatures. It was observed that at 90°C, the conversion of furfural was 27% after 10 minutes and achieved 60% in one hour, suggesting that furfural degradation is significant even with a weak base.

Experiment	Pattern	Catalyst Loading	Temperature (℃)	Air Pressure (bar)	Base Ratio	Yield	Complete Conversion
B1	+-+-	7	90	35	1	64	Y
B2	-+-+	3	120	10	1.5	35	Y
B3	++--	7	120	10	1	75	Y
B4	--++	3	90	35	1.5	56	N
B5	-++-	3	120	35	1	41	Y
B6	----	3	90	10	1	35	N
B7	+--+	7	90	10	1.5	52	N
B8	++++	7	120	35	1.5	73	Y

Table 3. Results of 1st DOE

Experiment	Pattern	Catalyst Loading	Temperature (℃)	Base Ratio	Yield
C1	0-+	7	115	1.5	65
C2	+0-	9	127.5	1	61
C3	000	7	127.5	1.25	97
C4	0+-	7	140	1	64
C5	-0-	5	127.5	1	63
C6	+-0	9	115	1.25	62
C7	0--	7	115	1	74
C8	0++	7	140	1.5	61
C9	000	7	127.5	1.25	63
C10	++0	9	140	1.25	53
C11	-0+	5	127.5	1.5	65
C12	000	7	127.5	1.25	60
C13	-+0	5	140	1.25	59
C14	--0	5	115	1.25	72
C15	+0+	9	127.5	1.5	62
C16	000	7	127.5	1.25	58
C17	000	7	127.5	1.25	60

Table 4. Results of 2nd DOE

The stability of furfural with the catalyst at reaction temperature was also investigated. Surprisingly, furfural concentration fluctuates, where the furfural conversion (49%) after 30 minutes is lower than that after 15 minutes (57%). Ultimately, the conversion reached approximately 74% after 1 hour. We inferred that this might arise from a reversible side reaction or furfural adsorption on the catalyst support (carbon black).

Design of Experiments (DOE) with JMP was used to investigate the effect of temperature, air pressure, catalyst loading, and base ratio on the FA yield and determine the optimal reaction conditions.

The 1st DOE was applied to screen the effect of these parameters, and it provides a basic understanding of the reaction (**Table 3**). The maximum yield achieved was 75%, with high catalyst loading, high temperature, low air pressure, and low base ratio (B3). Experiments at low catalyst loading (B2, B4, B5, and B6) were characterized by lower yield. By increasing the temperature and air pressure, the yield was expected to increase by a small amount according to the DOE prediction tool. However, more experiments are needed to increase the level of statistical significance. The fact that the air pressure has a negligible influence on the FA yield implies that the reaction is not oxygen-transfer limited. The DOE also predicted that there is no influence on the base ratio. Nevertheless, as there are four parameters and only eight experiments, our confidence level is insufficient to draw discrete conclusions.

It was also found that the conversion of furfural reached 100% at 120 ℃, while at 90 ℃, the conversion is incomplete. The FA yield remains low even if the conversion of furfural reaches 100% at high temperatures and low catalyst loading, which suggests the presence of side reactions or furfural adsorption on the carbon matrix

To obtain more accurate results, a more detailed 2nd DOE was performed (**Table 4**). Since the reaction is not oxygen-transfer limited, air pressure is removed from the DOE variables. The range of catalyst loading and temperature was adjusted toward the more favorable intervals.

It was found that the FA yield decreases with increasing temperature at the range of 115-140 ℃. The catalyst loading (from 5% to 9%) did not have a significant impact on the FA yield, suggesting that 5% catalyst loading is enough for aerobic oxidation. The results also confirmed that the base ratio does not influence the FA yield.

The central point condition (catalyst loading=9%, temperature=127.5℃, base ratio=1.25) is tested with five experiments (C3, C9, C12, C16, and C17) to understand the repeatability of experiments. It was found that, except for experiment C3, the FA yield is in the range of 58% to

63%, suggesting the high replicability of the experiments. The high conversion of 97% is likely due to the error in FA sampling. We also investigated the effect of reaction time on the FA yield. We adjusted the reaction time from 20 minutes to 1.5 hours through separate experiments (**Figure 7**). At optimal reaction conditions, it was found that the furfural conversion is 100% even with a reaction time of 20 minutes. However, the FA yield decreases to about 55% for a 30-minute reaction and further to as low as 20% for a 20-minute reaction. For a 90-minute reaction, the FA yield increases to 76%, which suggests that 1 hour is the optimal reaction time for aerobic oxidation.

We concluded from the results of both DOEs that the optimal condition for aerobic oxidation is 5% catalyst loading and 120 °C for a reaction time of 1 hour. At the optimal condition, the conversion of furfural is complete, and the FA yield is about 75%. Yi and coworkers studied the base-free conversion of 5-hydroxymethylfurfural (HMF) to 2,5-furan dicarboxylic acid (FDCA) over a Ru/C catalyst on a similar setup and were able to achieve an FDCA yield of 88% with oxygen (2 bars) and longer reaction time (10 hours).[16] Indeed, we achieved a FA yield close to 80% with 3-5 bar oxygen pressure and a reaction time of 2 hours. However, such conditions are not ideal for future commercializing of the process despite the higher yield. Longer processing time can result in a considerably higher cost for a continuous reactor. To reach a high temperature without evaporating the reaction mixture, the oxygen pressure must be carefully chosen, making the reaction system very difficult to operate on a larger scale. Moreover, a high oxygen pressure risks poisoning the Ru/C catalyst by oxidizing ruthenium to ruthenium (IV) oxide.[24]

Six recycling experiments were conducted to assess the catalyst stability (**Figure 8**). The FA yield decreased by 1.2 % after each cycle (75% to 69%), suggesting reasonably good catalyst recyclability. The low yield after the 2nd round can be explained by an operational problem during the set-up which led to air leakage in the system during the reaction. After the 4th round, we observed a FA yield of about 85%, which is higher than the result of the 1st run and all the experiments from both DOEs. We inferred that the carbon is constantly adsorbing furfural. After furfural accumulation reaches a certain amount, it de-adsorbs from the carbon and participates in aerobic oxidation again.

It was found that the ruthenium content of the fresh catalyst was 6.3%, while that of the used catalyst after 6 cycles decreased to 4.6%. Analysis through BET showed that the surface area of the catalyst decreased from 840 to 717 m^2/g, which was primarily due to a decrease in micropores area. The average pore volume decreased from 0.68 to 0.58 cm^3/g, with the average pore diameter decreasing from 7.0 to 6.5 nm and the average pore width remaining at around 3.2 nm.

TEM images from the fresh (**Figure 9a**) and used catalysts (**Figure 9b**) at 120k magnification were used to observe the morphology and particle sizes. Using Image J software, it was roughly estimated that there was a 16% decrease (from 2.9nm to 2.5nm) in unagglomerated particle size, which partially explains the 27% loss of ruthenium contents observed through ICP analysis. It is clear that the ruthenium distribution on the carbon base is not uniform and therefore further investigation using TEM is required to get a statical distribution of particle size and to understand better the ruthenium contents.

It is evident that some agglomerate of ruthenium particles is present in both fresh and used catalysts, as illustrated by the dark clusters of particles. The sizes of those agglomerates range from 5nm to around 20 nm. This agrees with the result of Hitrik's, which investigated the agglomeration of ruthenium through a six-step mechanism.[34] With increasing particle size, the surface-to-volume decreases sharply, and so does the catalytic activity of the particles. This effect can be reduced by using nitrogen-doped mesoporous carbons (NMCs) as supports and tuning the nitrogen content, which could effectively decrease the average particle size to around 2nm. [35]

It is surprising that the FA yield only declines of only 6% upon 27 % of Ru leaching. Catalyst deactivation studies on ruthenium reveal that the formation of ruthenium oxide can happen at room temperatures, which is often associated with a significant loss in catalyst activity.[32] Aßmann and coworkers studied the microscopic process of catalyst deactivation and found that oxygen absorbs on the ruthenium surface, forming some chemical-inactive distinct ordered oxide overlayers.[33] It is likely that, during the preservation of the catalyst, such a layer is formed on the ruthenium surface, which is being leached out during aerobic oxidation. This explains the relatively high ruthenium content loss, moderate reduction in particle size, and the relatively small decline in catalytic activity.

We have focused on the physiochemical studies of ruthenium particles so far. However, as stated by Lin and coworkers, the carbon support loss through oxidation could also have an essential effect on catalyst activity through the change in active sites, which is worth future investigation.[36]

Understanding Furfural Adsorption and Decomposition

A kinetic and thermodynamic study of furfural adsorption onto commercial-grade activated carbon (ACC) from an aqueous solution was performed by Sahu and coworkers. It was suggested that for 0.05 wt% furfural, the furfural removal could achieve 12mg/g. The initial concentration of furfural provides a driving force to overcome the

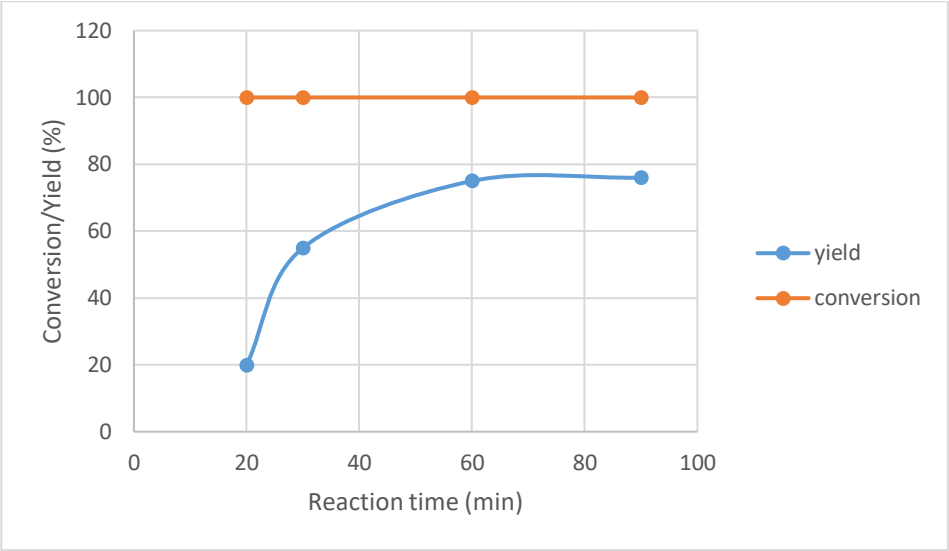


Figure 7. The effect of reaction time on conversion/yield

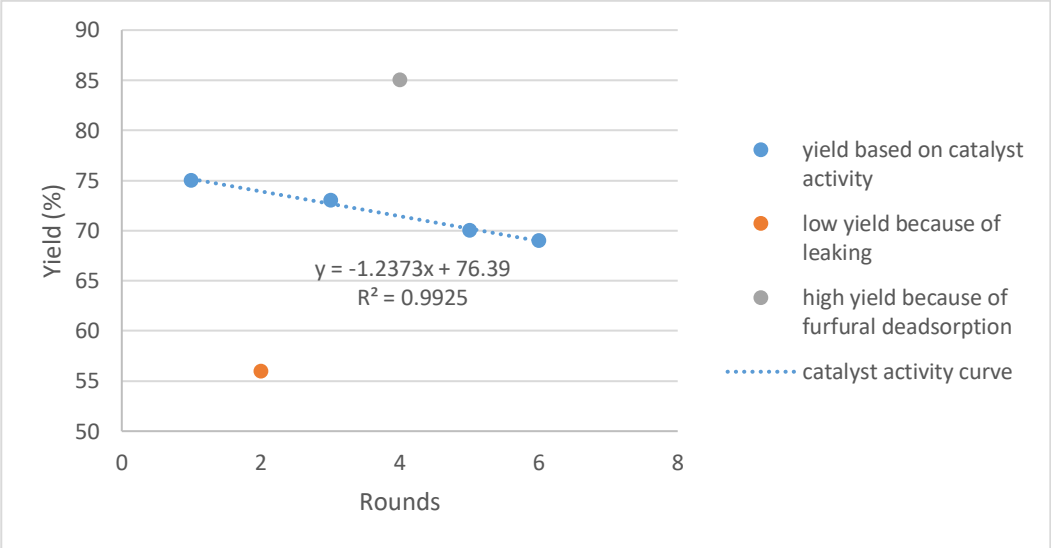


Figure 8. Catalyst Recyclability Test of Ru/C

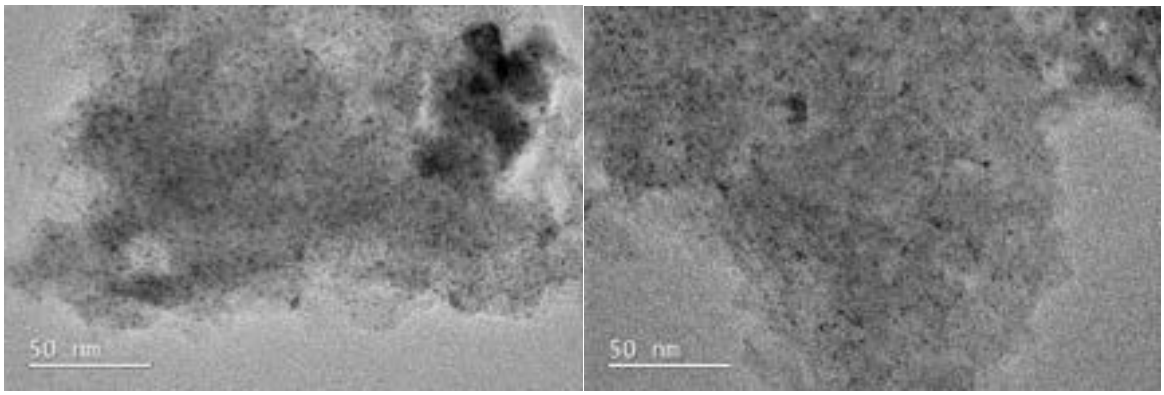


Figure 9. TEM images of catalyst (a) fresh on the left (b) used on the right

mass transfer resistance, and hence, the amount of furfural adsorbed per unit mass increases with increasing initial concentration. With increasing temperatures, the adsorptive removal of furfural decreases as adsorption is in general an exothermic process.[25] It was also found that, with air oxidation in activated carbon (ACAOx), the furfural removal rate was much faster.[37] The Weber-Morris intra-particular diffusion model can be used to determine the rate-determining step of the adsorption process, as shown in **Equation 1**:

$$q_t = k_{pi} t^{1/2} + C$$

where $\boldsymbol{k_{pi}}$ is the intra-particle diffusion rate constant for adsorption at stage i and C is the intercept that represents the boundary layer thickness (mg/g).[26] It was found that the Weber-Morris plot of the furfural adsorption process is not a straight line over the whole-time range, suggesting

the presence of more than one adsorption mechanism. Further analysis suggests that the adsorption proceeds via a complex mechanism, with the rate-determining step being the intra-particle diffusion of furfural into micropores. Furthermore, it was found that the furfural adsorption onto ACC can be best fitted into the Redlich-Peterson isotherm, as illustrated in **Equation 2**.[25]

$$q_e = \frac{K_R C_e}{1 + \alpha_R C_e^{\beta}}$$

The aerobic oxidation of furfural was fitted using a pseudo-first-order model, as illustrated in **Equation 3**.[27]

$$R = k[C_5H_4O_2][O_2] \approx k'[C_5H_4O_2]$$

In contrast, the adsorption of furfural is usually modeled by a pseudo-second-order model, as shown in **Equation 4**:[25,28]

$$\frac{dQ}{dt} = k_2(Q_e - Q)^2$$

where Q is the adsorption quantity at time t and Q_e is the adsorption quantity at equilibrium. Since the adsorption kinetic is of higher order, a high concentration of furfural favors the adsorption compared with aerobic oxidation. This explains the low FA yield observed when the reaction time is reduced to 20 or 30 minutes. The concentration fluctuation observed in the stability test can also be interpreted by the competition between adsorption and oxidation.

However, this does not explain the low yield observed at high temperatures and low catalyst loading, which should have a relatively low adsorption rate. Therefore, the exceptionally low yield is more likely to be explained by the presence of undesirable side reactions. This is confirmed by the yellowish color of the product, while the color for FA is white. One possibility is the formation of humins through the intermediate of α-carbonyl aldehyde. After 5 hours of reaction at 120 ℃, the carbon yield of humins can achieve 23.4%.[29] It is also confirmed that water is essential for furfural derivatives to form humins and hydrolytic open-ring products, where no humins formation is observed in experiments carrying with ethyl acetate. Indeed, Jin and coworkers conducted a detailed experimental study of furfural oxidation, where they concluded that furfural decomposition is mainly initialized through a ring-opening isomerization reaction to form formyl vinyl ketene.[30] Furoin is also a potential side product for the aerobic oxidation on Ru-based catalyst, where Gupta and coworkers observed a furoin yield of 14% at 140 ℃ for a 20-minute reaction.[31] However, furoin formation is favored by the use of a strong base. With sodium hydrogen carbonate, we expect a much lower selectivity to furoin.

In general, it is very difficult to characterize the side products accurately, and it is likely that more than one type of side product exists. However, through the tunning of the reaction temperatures, reaction time, and catalyst loading, the side reactions can be effectively avoided, and a high FA yield can be achieved.

Conclusion

A catalyst-free stoichiometric approach to the oxidation of furfural to furoic acid has been investigated, with yield and selectivity exceeding that of a conventional Cannizzaro reaction. Using methanol as solvent and a furfural-peroxide ratio of 4, the rates of unwanted reactions such as over-oxidation, esterification, and ring-opening can be minimized, achieving a 73% yield. In contrast, the maximum yield with aerobic oxidation on Ru/C is 75%, with the optimal condition of 120℃ and 5% catalyst loading.

However, with stoichiometric oxidation, purification of furfural (pre-distillation to obtain concentrated furfural) is required, which accounts for around 70% of the operating cost of furfural production. Aerobic oxidation uses 3% furfural with a more economical oxidizer (air), and the catalyst shows good recyclability. Overall, we considered that aerobic oxidation of furfural is a better scheme for future commercialization.

Understanding of furfural adsorption and decomposition is still limited. To the best of our knowledge, there is no evidence in the literature investigating the mechanism of the competitive pathways of furfural adsorption and decomposition with aerobic oxidation. The method to reduce such undesirable side reactions is still to be explored. We are confident that, with a better understanding of the reaction mechanism, the yield can be further increased.

Even though the catalyst shows good recyclability within the six rounds of experiments, batch operation results in a significant loss in catalyst weight. Therefore, for the commercialization of SAF, the transformation into continuous operation is necessary. With continuous operation, the catalyst can be recycled with more rounds, and the deactivation and regeneration of the catalyst can be studied in more detail.

Acknowledgment

The author appreciates all the help and guidance provided by Prof. Jason P. Hallett and Dr. Amir Al Ghatta during the research term. Dr. Amir Al Ghatta provided valuable information in the early stages in determining the direction of the research and assisted greatly in the later stages of the operation of the experiments.

References

1. Mohamed Naceur Belgacem; Gandini, A. *Monomers, Polymers and Composites from Renewable Resources*; Elsevier, 2011; pp 153−178.
2. Surfactants Market Size, Share, Growth | Global Report [2028] https://www.fortunebusinessinsights.com/surfactants-market-102385.
3. YU, Y.; Jin ZHAO; Bayly, A. E. Development of Surfactants and Builders in Detergent Formulations. *Chinese Journal of Chemical Engineering* **2008**, *16* (4), 517–527. https://doi.org/10.1016/s1004-9541(08)60115-9.
4. Badmus, S. O.; Amusa, H. K.; Oyehan, T. A.; Saleh, T. A. Environmental Risks and Toxicity of Surfactants: Overview of Analysis, Assessment, and Remediation Techniques. *Environmental Science and Pollution Research* **2021**. https://doi.org/10.1007/s11356-021-16483-w.
5. Hill, K.; Rhode, O. Sugar-Based Surfactants for Consumer Products and Technical Applications. *Lipid - Fett* **1999**, *101* (1),

25–33. https://doi.org/3.0.co;2-n">10.1002/(sici)1521-4133(19991)101:1<25::aid-lipi25>3.0.co;2-n.
6. Al Ghatta, A.; Aravenas, R. C.; Wu, Y.; Perry, J. M.; Lemus, J.; Hallett, J. P. New Biobased Sulfonated Anionic Surfactants Based on the Esterification of Furoic Acid and Fatty Alcohols: A Green Solution for the Replacement of Oil Derivative Surfactants with Superior Proprieties. *ACS Sustainable Chemistry & Engineering* **2022**, *10* (27), 8846–8855. https://doi.org/10.1021/acssuschemeng.2c01766.
7. Li, X.; Liu, Q.; Si, C.; Lu, L.; Luo, C.; Gu, X.; Liu, W.; Lu, X. Green and Efficient Production of Furfural from Corn Cob over H-ZSM-5 Using γ-Valerolactone as Solvent. *Industrial Crops and Products* **2018**, *120*, 343–350. https://doi.org/10.1016/j.indcrop.2018.04.065.
8. Hidayat, N.; Hidayat, A. N.; Gozan, M. Preliminary Design of Corncob Based Furfural Plant. *AIP Conference Proceedings* **2019**. https://doi.org/10.1063/1.5086595.
9. Hoydonckx, H. E.; Van Rhijn, W. M.; Van Rhijn, W.; De Vos, D. E.; Jacobs, P. A. Furfural and Derivatives. Ullmann's Encyclopedia of Industrial Chemistry 2007. https://doi.org/10.1002/14356007.a12_119.pub2.
10. Isenhour, L. L. Method for the production of furoic acid https://patents.google.com/patent/US2041184A/en (accessed 2022 -11 -29).
11. Sato, K.; Hyodo, M.; Takagi, J.; Aoki, M.; Noyori, R. Hydrogen Peroxide Oxidation of Aldehydes to Carboxylic Acids: An Organic Solvent-, Halide- and Metal-Free Procedure. Tetrahedron Letters 2000, 41 (9), 1439–1442. https://doi.org/10.1016/s0040-4039(99)02310-2.
12. Noyori, R.; Aoki, M.; Sato, K. Green Oxidation with Aqueous Hydrogen Peroxide. Chemical Communications 2003, No. 16, 1977. https://doi.org/10.1039/b303160h.
13. Franco, A.; Negi, A.; Luque, R.; Carrillo-Carrión, C. Selectivity Control in the Oxidative Ring-Opening of Dimethylfuran Mediated by Zeolitic-Imidazolate Framework-8 Nanoparticles. ACS Sustainable Chemistry & Engineering 2021, 9 (24), 8090–8096. https://doi.org/10.1021/acssuschemeng.1c00708.
14. Sadier, A.; Paul, S.; Wojcieszak, R. Selective Oxidation of Furfural at Room Temperature on a TiO2-Supported Ag Catalyst. *Catalysts* **2022**, *12* (8), 805. https://doi.org/10.3390/catal12080805.
15. Nocito, F.; Ditaranto, N.; Linsalata, D.; Naschetti, M.; Comparelli, R.; Aresta, M.; Dibenedetto, A. Selective Aerobic Oxidation of Furfural into Furoic Acid over a Highly Recyclable MnO_2@CeO_2 Core–Shell Oxide: The Role of the Morphology of the Catalyst. *ACS Sustainable Chemistry & Engineering* **2022**, *10* (26), 8615–8623. https://doi.org/10.1021/acssuschemeng.2c02341.
16. Yi, G.; Teong, S. P.; Zhang, Y. Base-Free Conversion of 5-Hydroxymethylfurfural to 2,5-Furandicarboxylic Acid over a Ru/c Catalyst. *Green Chemistry* **2016**, *18* (4), 979–983. https://doi.org/10.1039/c5gc01584g.
17. Stirred Reactors https://www.parrinst.com/products/stirred-reactors/ (accessed 2022 -12 -01).
18. Hu, X.; Nango, K.; Bao, L.; Li, T.; Hasan, M. D. M.; Li, C.-Z. High Yields of Solid Carbonaceous Materials from Biomass. *Green Chemistry* **2019**, *21* (5), 1128–1140. https://doi.org/10.1039/C8GC03153C.
19. Cong, Z.; Wang, C.; Chen, T.; Yin, B. Efficient and Rapid Method for the Oxidation of Electron-Rich Aromatic Aldehydes to Carboxylic Acids Using Improved Basic Hydrogen Peroxide. *Synthetic Communications* **2006**, *36* (5), 679–683. https://doi.org/10.1080/00397910500408456.
20. da Silva, M. J.; Rodrigues, A. A. Metal Silicotungstate Salts as Catalysts in Furfural Oxidation Reactions with Hydrogen Peroxide. *Molecular Catalysis* **2020**, *493*, 111104. https://doi.org/10.1016/j.mcat.2020.111104.
21. Hermundsgård, D. H.; Ghoreishi, S.; Tanase-Opedal, M.; Brusletto, R.; Barth, T. Investigating Solids Present in the Aqueous Stream during STEX Condensate Upgrading—a Case Study. *Biomass Conversion and Biorefinery* **2022**. https://doi.org/10.1007/s13399-022-03593-9.
22. Douthwaite, M.; Huang, X.; Iqbal, S.; Miedziak, P. J.; Brett, G. L.; Kondrat, S. A.; Edwards, J. K.; Sankar, M.; Knight, D. W.; Bethell, D.; Hutchings, G. J. The Controlled Catalytic Oxidation of Furfural to Furoic Acid Using AuPd/Mg(OH)2. *Catalysis Science & Technology* **2017**, *7* (22), 5284–5293. https://doi.org/10.1039/C7CY01025G.
23. Murzin, D. Yu.; Bertrand, E.; Tolvanen, P.; Devyatkov, S.; Rahkila, J.; Eränen, K.; Wärnå, J.; Salmi, T. Heterogeneous Catalytic Oxidation of Furfural with Hydrogen Peroxide over Sulfated Zirconia. *Industrial & Engineering Chemistry Research* **2020**, *59* (30), 13516–13527. https://doi.org/10.1021/acs.iecr.0c02566.
24. Gallezot, P.; Chaumet, S.; Perrard, A.; Isnard, P. Catalytic Wet Air Oxidation of Acetic Acid on Carbon-Supported Ruthenium Catalysts. *Journal of Catalysis* **1997**, *168* (1), 104–109. https://doi.org/10.1006/jcat.1997.1633.
25. Sahu, A. K.; Srivastava, V. C.; Mall, I. D.; Lataye, D. H. Adsorption of Furfural from Aqueous Solution onto Activated Carbon: Kinetic, Equilibrium and Thermodynamic Study. *Separation Science and Technology* **2008**, *43* (5), 1239–1259. https://doi.org/10.1080/01496390701885711.
26. Hwang, K.-J.; Shim, W.-G.; Kim, Y.; Kim, G.; Choi, C.; Kang, S. O.; Cho, D. W. Dye Adsorption Mechanisms in TiO_2 Films, and Their Effects on the Photodynamic and Photovoltaic Properties in Dye-Sensitized Solar Cells. *Physical Chemistry Chemical Physics* **2015**, *17* (34), 21974–21981. https://doi.org/10.1039/c5cp03416g.
27. Kim, M.; Su, Y.; Fukuoka, A.; Hensen, E. J. M.; Nakajima, K. Aerobic Oxidation of 5-(Hydroxymethyl)Furfural Cyclic Acetal Enables Selective Furan-2,5-Dicarboxylic Acid Formation with CeO 2 -Supported Gold Catalyst. *Angewandte Chemie* **2018**, *130* (27), 8367–8371. https://doi.org/10.1002/ange.201805457.
28. Ghosh, S.; Falyouna, O.; Malloum, A.; Othmani, A.; Bornman, C.; Bedair, H.; Onyeaka, H.; Al-Sharify, Z. T.; Jacob, A. O.; Miri, T.; Osagie, C.; Ahmadi, S. A General Review on the Use of Advance Oxidation and Adsorption Processes for the Removal of Furfural from Industrial Effluents. *Microporous and Mesoporous Materials* **2022**, *331*, 111638. .
29. Shi, N.; Liu, Q.; Cen, H.; Ju, R.; He, X.; Ma, L. Formation of Humins during Degradation of Carbohydrates and Furfural Derivatives in Various Solvents. *Biomass Conversion and Biorefinery* **2019**, *10* (2), 277–287. https://doi.org/10.1007/s13399-019-00414-4.
30. Jin, Z.-H.; Yu, D.; Liu, Y.-X.; Tian, Z.-Y.; Richter, S.; Braun-Unkhoff, M.; Naumann, C.; Yang, J.-Z. An Experimental Investigation of Furfural Oxidation and the Development of a Comprehensive Combustion Model. *Combustion and Flame* **2021**, *226*, 200–210. https://doi.org/10.1016/j.combustflame.2020.12.015.
31. Gupta, N. K.; Fukuoka, A.; Nakajima, K. Metal-Free and Selective Oxidation of Furfural to Furoic Acid with an N-Heterocyclic Carbene Catalyst. *ACS Sustainable Chemistry & Engineering* **2018**, *6* (3), 3434–3442. https://doi.org/10.1021/acssuschemeng.7b03681.
32. Argyle, M.; Bartholomew, C. Heterogeneous Catalyst Deactivation and Regeneration: A Review. *Catalysts* **2015**, *5* (1), 145–269. https://doi.org/10.3390/catal5010145.
33. Aßmann, J.; Crihan, D.; Knapp, M.; Lundgren, E.; Löffler, E.; Muhler, M.; Narkhede, V.; Over, H.; Schmid, M.; Seitsonen, A. P.; Varga, P. Understanding the Structural Deactivation of Ruthenium Catalysts on an Atomic Scale under Both Oxidizing and Reducing Conditions. *Angewandte Chemie* **2004**, *117* (6), 939–942. https://doi.org/10.1002/ange.200461805.
34. Hitrik, M.; Sasson, Y. Aggregation of Catalytically Active Ru Nanoparticles to Inactive Bulk, Monitored *in Situ* during an Allylic Isomerization Reaction. Influence of Solvent, Surfactant and Stirring. *RSC Advances* **2018**, *8* (3), 1481–1492. https://doi.org/10.1039/c7ra11133a.
35. Gogoi, P.; Kanna, N.; Begum, P.; Deka, R. C.; C. V. V, S.; Raja, T. Controlling and Stabilization of Ru Nanoparticles by Tuning the Nitrogen Content of the Support for Enhanced H_2 Production through Aqueous-Phase Reforming of Glycerol. *ACS Catalysis* **2019**, *10* (4), 2489–2507. https://doi.org/10.1021/acscatal.9b04063.
36. Lin, B.; Guo, Y.; Lin, J.; Ni, J.; Lin, J.; Jiang, L.; Wang, Y. Deactivation Study of Carbon-Supported Ruthenium Catalyst with Potassium Promoter. *Applied Catalysis A: General* **2017**, *541*, 1–7. https://doi.org/10.1016/j.apcata.2017.04.020.
37. Fang, K.; Yang, R. Modified Activated Carbon by Air Oxidation as a Potential Adsorbent for Furfural Removal. *Alexandria Engineering Journal* **2021**, *60* (2), 2325–2333. https://doi.org/10.1016/j.aej.2020.12.032

Modelling of optical components for hydrogen production in modular photoelectrochemical reactors

Jinjie Zhu, Sid Halder

Department of Chemical Engineering, Imperial College London, United Kingdom

Photoelectrochemical (PEC) water splitting reactors are a developing technology for a sustainable production of hydrogen – a fuel, platform chemical and an energy storage vector. This paper describes a new COMSOL Multiphysics 6.1 model of a linear Fresnel lens array coupled with a stepped thickness waveguide. Fresnel lenses are used as solar concentrators because of their low cost and high optical efficiency. The stepped thickness waveguide redirects the light that is concentrated by the lens using embedded mirrors and the principle of total internal reflection. As a result, concentrated light is efficiently directed to the waveguide edge, which will potentially be coupled with the optical window of a PEC reactor. The combined system acts as a solar concentrator, providing a reasonable way of overcoming the optical challenges of PEC systems. The polymethyl methacrylate (PMMA) material chosen for modelling the Fresnel lenses and waveguide has the advantage of low cost and light weight. Overall, an optical efficiency of 53%, a steady state geometrical concentration ratio of 591 and a steady state optical concentration ratio of 43 were determined by the proposed solar concentrator model based on the ray tracing simulation. Different mirror inclination angles within the waveguide were also studied, with the optimum determined to be 20°. Further investigation would include coupling the optical model with the water splitting reactor, introducing a solar tracking system and modeling the heat transfer effect.

Keywords: water splitting, photoelectrochemical reactors, hydrogen, Fresnel lens, waveguide, concentration ratio

1 Introduction

Climate change is the defining problem of this generation. Consistent use of fossil fuels and polluting manufacturing practices have dramatically increased the levels of greenhouse gases, particularly carbon dioxide, in the atmosphere, accelerating the rate of global warming [1]. Our infrastructure relies too heavily on fossil fuels - to prevent catastrophic disruption, there much be a rapid global transition to a more sustainable society, predominantly powered by renewable energy. The drawback of renewable energy is the inherent intermittent nature of the resource; there is a need to develop energy storage carriers that can be paired with renewable electricity and hydrogen is a key candidate to fulfil this role.

Key sectors such as energy generation, transportation, heating and chemical manufacturing can be decarbonised by employing the use of green hydrogen, which can be generated using various electrolyser technologies, all of which split water into hydrogen and oxygen. Equations 1 - 3 detail the reactions occurring in alkaline electrolysers, showing that the water splitting reaction has no associated emissions with it. The hydrogen evolution reaction (HER) occurs at the cathode of the electrolyser, with the oxygen evolution reaction (OER) occuring simultaneously at the anode:

$$\mathrm{HER}: 2\,H_2O + 2\,e^- \rightleftharpoons 2\,OH^- + H_2 \quad (1)$$

$$\mathrm{OER}: 4\,OH^- \rightleftharpoons 2\,H_2O + 4\,e^- + O_2 \quad (2)$$

$$\mathrm{Overall}: 2\,H_2O \rightleftharpoons 2\,H_2 + O_2 \quad (3)$$

Theoretically, an equilibrium potential (ΔV^{ϕ}) of 1.23 V must be applied to start the reaction. However, larger overpotentials are required to run the electrolyser, increasing with scale-up. This is due to inefficiencies involving charge and mass transfer, even with the use of noble metal-based catalysts such as iridium oxide at the anode and platinum at the cathode. Optimisation of electrolysers is an ongoing field of research, the aim being to reduce the electrical energy consumption and costs of all water electrolyser units. The oxygen evolution reaction (OER) that occurs at the anode is the kinetically limiting reaction because of its higher electron stoichiometry ($4\,e^-$) when compared to reduction ($2\,e^-$), leading to a higher overall activation barrier [2].

Photoelectrochemical (PEC) reactors integrate photovoltaics and electrolysers into a single device, essentially harvesting solar energy and converting it to green hydrogen through photoelectrochemical water splitting. This combination of devices reduces the need of power electronics, minimising the power loss and material usage. Additional advantages include *in situ* catalysis and no requirement for critical platinum group metal (PGM) catalysts. Although PEC reactors typically use low-cost semiconducting and catalyst materials, efficiency is still a major issue. Two photoelectrodes are often used (photoanode and photocathode), driving the water splitting half reactions separately. The benefit of this is that the system is able to utilise a greater portion of the solar spectrum. Principal phenomena that adversely affect the performance of upscaled PEC systems are electrode orientation (optical challenges), photoelectrode substrate resistivity and bubble evolution at the electrodes [2]. This paper attempts to address the optical challenges associated with PEC reactors.

The research objectives of this paper were to:

- Create an optical system consisting of a linear Fresnel lens array and stepped thickness waveguide on COMSOL Multiphysics 6.1;
- Quantify the performance of the optical system by running the model simulation and determining the optical efficiency, geometric and optical concentration ratio;
- Investigate how the stepped thickness waveguide parameters affect the performance of the optical system through a parameter sweep analysis;
- Build upon existing COMSOL PEC reactor models by coupling the optical system with the aim of determining the relationship between concentration ratio and hydrogen production flux.

2 Background

A basic theoretical summary of the process of photoelectrochemical water splitting in PEC reactors is given in this section, along with an insight into optical components that could be integrated with the reactor to increase the hydrogen flux produced.

2.1 Photoelectrochemical water splitting

PEC reactors use semiconducting photoelectrodes, which consist of a substrate coated with a layer of semiconducting material such as titanium oxide. The semiconducting material can absorb solar photons with energies greater than the material band gap, which is the energy gap between the valence band and conducting band. This generates electron-hole pairs known as excitons in the following reactions:

$$\mathrm{SC} + \mathrm{hv} \underset{\text{recombination}}{\overset{\text{absorption}}{\rightleftharpoons}} \mathrm{SC}(\mathrm{e^-_{CB}}, \mathrm{h^+_{VB}}) \tag{4}$$

$$2\,\mathrm{H_2O} + 2\,\mathrm{e^-_{CB}} \rightleftharpoons 2\,\mathrm{OH^-} + \mathrm{H_2} \tag{5}$$

$$4\,\mathrm{OH^-} + 4\,\mathrm{h^+_{VB}} \rightleftharpoons 2\,\mathrm{H_2O} + \mathrm{O_2} \tag{6}$$

In a PEC reactor, one of both electrodes can be synthesised from semiconducting materials, performing both photon absorption and catalysis simultaneously. Photoelectrodes utilise an electric field that exists in the depletion layer at the semiconductor | electrolyte interface, providing the driving force for the separation of the negatively charged electron and positively charged hole, generating a photocurrent. These holes then migrate to the photoanode | aqueous solution interface, where they catalyse the OER reaction, producing oxygen. The electrons are transferred across the photocathode electrolyte | interface, driving the HER reaction. The overall reaction is the same as a traditional electrolyser (Equation 3) [3].

Photoanodes tend to be more chemically robust that photocathodes and so they have been researched to a greater extent. The PEC reactor being modelled in this study comprises of a photoanode and a metal cathode and requires an external voltage input to generate hydrogen – this is a stepping stone to the development of a device that will split $\mathrm{H_2O}$ spontaneously, with no input other than solar photons.

A PEC reactor undergoing photo-assisted water electrolysis can be modelled quantitatively. Figure 1 shows a typical schematic of a PECR that highlights the electrodes (note dimensions not to scale). For testing purposes, the cathode is fixed at a potential of 0 V (grounded) and the anode potential is varied relative to the reference electrode (standard hydrogen electrode).

The output current consists of three major components - the ionic current, the cathodic current and the anodic current (consisting of photocurrent and dark current). We refer the readers to a more rigorous theoretical modelling of the system by Hankin et al [4].

Conventional electrolysers are compatible with large electrode geometrical areas due primarily to the fact that uniform electric field distributions are possible between the anode and the cathode. This is difficult to recreate in PEC reactors due to the nature of the photoelectrode – the semiconductor is coated on the surface of an inert substrate that does not take part in the reaction. This results in a non-uniform electric field distribution, causing electron transfer deficiencies. For this reason, perforated cathodes have been used to facilitate a more uniform electric field distribution as shown in Figure 1.

PECR's will be predominantly modular devices but this has multiple drawbacks – each reactor will need its own balance of plant, including electrolyte circuitry and gas manifolds [2].

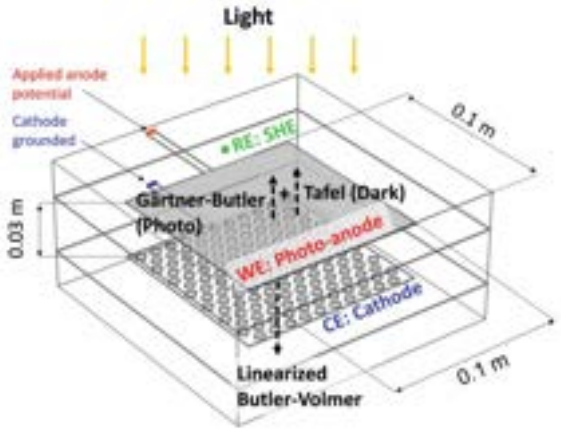


Figure 1: Prototype photoelectrochemical reactor schematic, also highlighting ionic, photo and cathodic current approximation frameworks. Courtesy of Dr Anna Hankin.

2.2 Integrating optical components with PEC reactors

Coupling optical components with PEC reactors can enhance the amount of light reaching the reactor, thus increasing the produced $\mathrm{H_2}$ flux. The issues of modular reactors can be circumvented by concentrating the incident light, increasing the photocurrent density and subsequently increasing the hydrogen flux output. Fig-

ure 2 displays a potential PEC reactor coupled with a linear Fresnel lens array and waveguide.

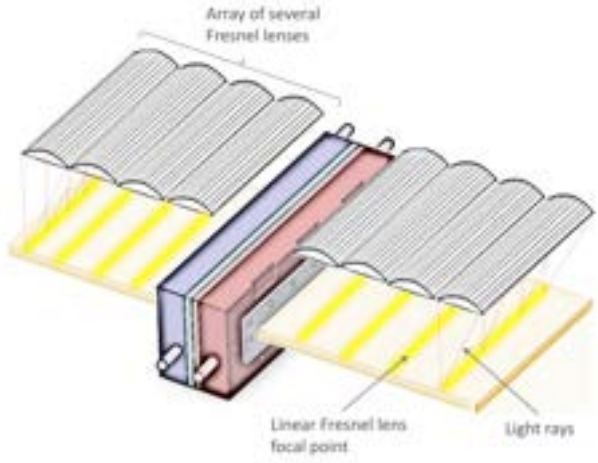


Figure 2: Speculated PEC reactor coupled with a linear Fresnel lens array and waveguide. Courtesy of Dr Anna Hankin.

2.3 Concentrating solar radiation

The intensity of incident solar radiation is dependent on a number of variables, such as the time of year, time of day, latitude and extent of cloud cover. The solar radiation that reaches the Earth's surface without being scattered by molecules or particles in the Earth's atmosphere is called direct solar radiation. Non-direct, isotropic radiation delivered by scattered photons is called diffuse radiation. The sum of the direct and diffuse solar radiation components is called global solar radiation. Latitudes around the equator will receive more sunlight because the solar incidence angle is closer to the normal. Therefore, equatorial regions gain more direct solar radiation than other regions at a given day of the year. Direct solar radiation of a given location will vary throughout the year because the earth revolves around the Sun.

Solar concentration is the most general way to use solar radiation, providing a way to fulfil electrical and thermal energy demands. Reflectors with parabolic surfaces and lenses with convex-shaped surfaces (hyperbolic surfaces) both have the function of converging light into a point, thus increasing the concentration and subsequently the intensity of light. Surfaces used in concentrated solar power technologies are highly susceptible to rapid degradation due to adverse environmental conditions and manufacturing defects, resulting in a substantial drop in efficiency. Traditional concentrators are relatively expensive – there is therefore a need to develop cheaper concentrators that are also maintenance free, light weight and resistant to degradation. A potential viable solution are Fresnel lenses, which are able to focus or collimate light [5].

2.3.1 Fresnel lenses

Fresnel lenses operate based on refraction, which is the phenomenon that occurs when light rays pass through mediums with different densities, causing a path deviation at the boundary interface. More specifically, if light travelling through a medium hits a boundary consisting of a denser material, the light will bend towards the normal. This is primarily due to a change in the velocity of light as different mediums will have different refractive indexes. Fresnel lenses are characterised by their focal length, which is the distance between the plane of the lens and the plane of the focal point, denoted the focal plane. Another important parameter is the f-number, which is defined as the focal length divided by the diameter of the lens. For a linear Fresnel lens, the diameter becomes the width.

Optimal selection of the slope facet, facet spacing, draft facet, and slope angle ensure that incident light is directed towards the focal region. A schematic of a Fresnel lens is shown in Figure 4. The main rationale for using Fresnel lenses over their plano-convex counterparts, as depicted by Figure 3, is that a Fresnel lens reduces material use and increases the compactness of the component while still retaining the same optical performance. These factors make it ideal in applications where size is a limitation, such as in the case of PEC reactors. Not all light rays will be refracted - a small fraction will be reflected due to inherent surface irregularities, decreasing the performance of the component. Other limitations include losses due to geometry and absorption.

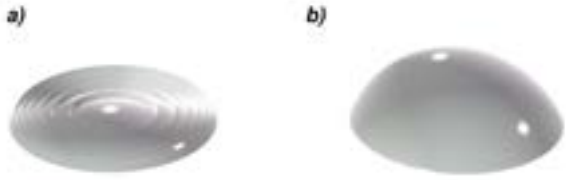


Figure 3: a) Circular Fresnel lens, b) Spherical plano-convex lens, as depicted by the COMSOL Fresnel Lens tutorial [6]

As for the material, PMMA (polymethyl methacrylate) is advantageous when compared to glass as it can be manufactured at scale with low costs. Given the geometry of PEC reactors as depicted in Figure 2, a linear Fresnel lens is suited better than a circular Fresnel lens. Linear Fresnel lenses operate in the same way but instead of focusing light onto a focal point they focus light onto a linear region in the focal plane. This concentrated light can then be directed to the photoelectrodes inside the PECR reactor using another optical component known as a waveguide. Linear Fresnel lenses can also be integrated with the PECR by simple attachment, aligning the Fresnel lens and waveguide with the top and bottom surfaces of the PECR respectively.

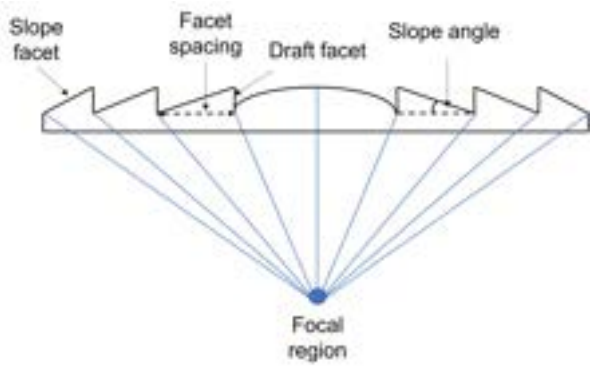


Figure 4: Schematic of a Fresnel lens, pointing out the design parameters that define the geometry [5].

2.3.2 Waveguides

Optical waveguides are devices that guide electromagnetic waves in the visible spectrum. They allow light to be directed to a specific location. Vu et al [7] discusses the implementation of a large-scale daylighting system based on a coupled linear Fresnel lens array and stepped thickness waveguide system. Daylighting refers to the use of natural lighting indoors as opposed to artificial lighting, producing a more comfortable indoor environment and reducing the impact of illnesses such as seasonal affective disorder.

There are two broad categories of waveguide, referred to colloquially as 'lossy' and 'lossless' systems, as depicted by Figure 5. Both enable light to propagate through the waveguide by total internal reflection, the phenomenon that occurs when light rays arriving at the boundary between one medium and another are not refracted into the second external medium, but completely reflected back into the first medium. 'Lossy' structures propagate the primary concentrated light through a flat waveguide via coupling prisms that help guide the light horizontally. However, this leads to geometrical decoupling losses and undesired ray interactions at light entry points, diminishing the efficiency of the system. Planar waveguides direct light rays in both horizontal direction due to the geometry of the coupling losses, meaning that they cannot be used in applications that require light propagation in one horizontal direction only. 'Lossless' systems avoid these issues by instead using a stepped thickness waveguide, preventing additional ray leakages. Concentrated light from the Fresnel lens is focused into a linear region directly below the midpoint of the lens. In a stepped waveguide, there is a midpoint step aligned with each Fresnel lens that re-directs the incoming light laterally into the waveguide. Each step adds thickness to the waveguide, which poses an issue as the number of the linear Fresnel lens in the array increases. In summary, 'lossless' systems are constrained by thickness and 'lossy' systems are constrained by efficiency. It is important to realise that the term 'lossless' is not technically correct - a proportion of light will not be reflected and will simply leave the waveguides, in addition to absorption losses as the light rays meet each waveguide boundary. A typical schematic of a stepped thickness waveguide is depicted by Figure 6.

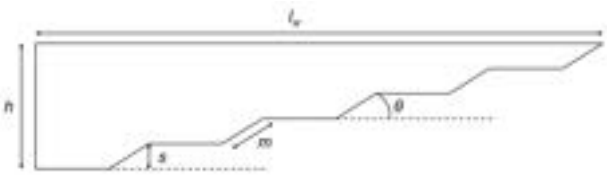


Figure 6: 2D schematic of a stepped thickness waveguide, highlighting key design parameter.

3 Methodology

The model construction and simulation were carried out using the Ray Optics module in the finite element modelling software, COMSOL Multiphysics 6.1. The Geometrical Optics physics interface in COMSOL was ideal for our study as it is designed for the analysis of ray optic simulations of cameras, telescopes, spectrometers, solar collectors and so on. There are some basic requirements that must be specified before the investigation can be carried out. A light source needs to be defined to carry out a ray trajectory simulation study. Each component of the model should have a defined material with some known physical properties such as the refractive index. The boundary conditions of the model should also be specified. The optical system created and defined by Vu et al. [7] was used as a foundation for parameter values of our model, given that the sizing of the system is comparable to speculated PEC reactor sizes. A predefined 'extremely fine' mesh with a maximum element size of 0.02 mm and a minimum of 0.0002 m was used for the optical model.

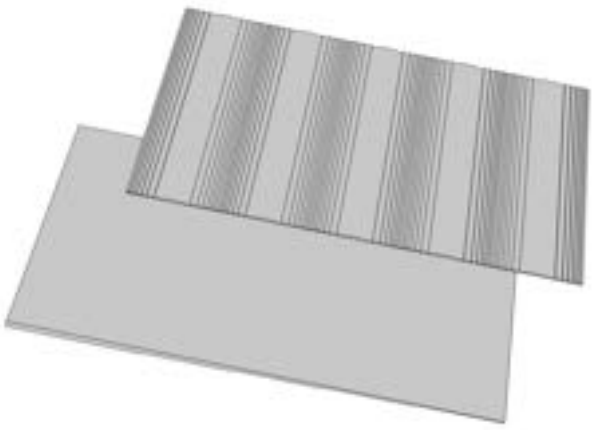


Figure 7: Waveguide and Fresnel lens model on COMSOL

The final geometric model of the system proposed is given in Figure 7. A simulated solar radiation passes through the Fresnel lens array, which directs and focuses the sunlight onto the waveguide. As for the simulation of a light ray source, a 'Release from Grid' node was used. The node produces a user-defined grid of points

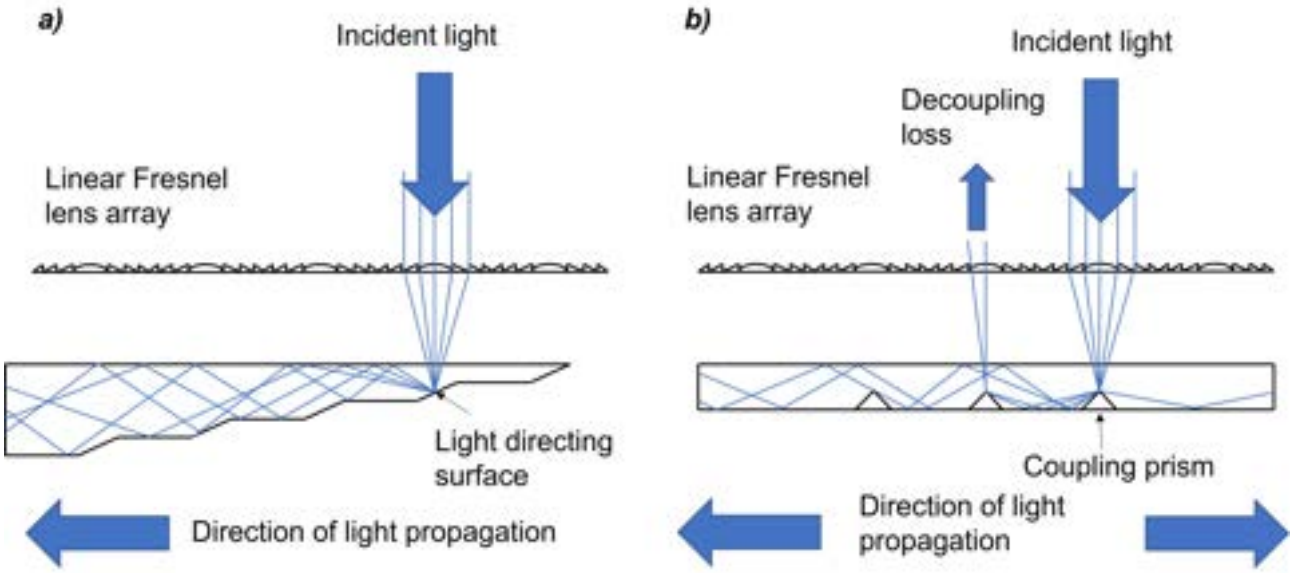


Figure 5: a) 'Lossless' system without decoupling losses, consisting of a stepped thickness waveguide b) Lossy system with decoupling losses consisting of a planar waveguide with coupling prisms [7]

where each point releases a ray. In our study, a rectangular light source with a planar wavefront is chosen. The shape of it is the same as the Fresnel lens. The total number of rays released is calculated as the number of points along the width times number of points along length. The initial coordinates of the source as well as the light vector direction were also defined. The size of the light source aims to cover the whole Fresnel lens array exactly. For the ray tracing study, the initial light intensity and total source power are chosen as 1000 W/m^2 and 812.6 W respectively, simulating a typical solar intensity. Both the Fresnel lens array and stepped thickness waveguide are made of PMMA because of its excellent light transmission properties. A monochromatic light source with wavelength of 660 nm was chosen. Different wavelength inputs will lead to different results.

3.1 Construction of linear Fresnel lens

The Fresnel lens array construction started with a spherical plano-convex lens from a predefined COMSOL parts library. Fresnel lenses replace the curved surface of the lens with a series of concentric grooves, minimising the footprint of the lens. These contours act as individual refracting surfaces and lead to the same focal length as the original lens. Since the aim was to build a linear Fresnel lens array, a work plane was introduced at the cross section of the lens to extract the 2D axisymmetric geometry. This cross section was further extruded to 1000 mm to obtain the linear Fresnel lens. The final geometry was built by duplicating the lens five times using the array feature, finally creating a linear Fresnel lens array. After building the model, a simple test was conducted to see if the planar Fresnel lens had the expected function of focusing light into linear regions. The light source used here has the same parameters (intensity and size of the grid matrix) as the simulated solar radiation mentioned earlier. The direction of the light rays was defined as vertically down towards the Fresnel lens array. The simulation output below (Figure 8) confirmed that the Fresnel lens was operating as expected.

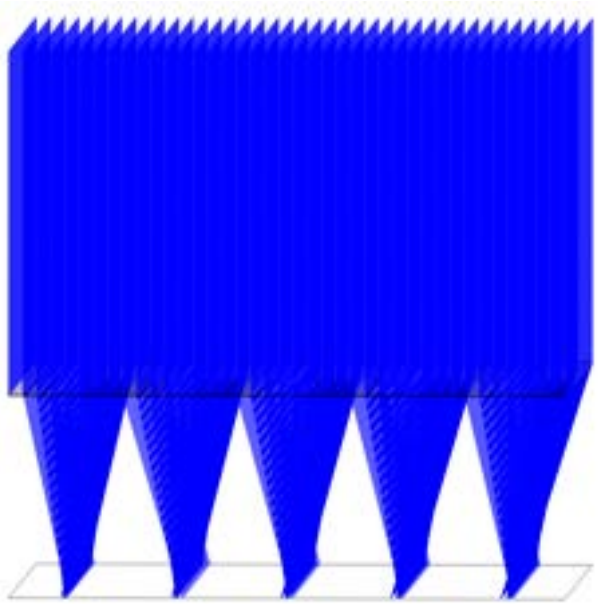


Figure 8: Simulating incident light onto Fresnel lens to observe the concentrating effect

The midpoint of each Fresnel lens in the array was aligned with the inclined surfaces, the mirror surfaces, of the waveguide located below. The mirror condition was defined based on the reflection coefficient only, which is a parameter that describes the proportion of a wave that is reflected by an impedance discontinuity in the transmission medium. The impedance discontinuity is anything that affects the ratio between the inductance of the trace and its capacitance. The coefficient was chosen as 1 to represent a perfect mirror with no energy loss. The governing equations are shown below:

$$\mathbf{n_r} = \mathbf{n_i} - 2\cos(\theta_i)\mathbf{n_s} \quad (7)$$

$$\mathbf{k_{u,r}} = \mathbf{k_{u,i}} - 2\mathbf{k_{u,s}}\cos(\theta_i) \quad (8)$$

$$\mathbf{k_{uv,r}} = -\mathbf{k_{uv,i}} + 2\mathbf{k_{uv,s}} \quad (9)$$

$$\mathbf{k_{v,r}} = \mathbf{k_{v,i}} - \frac{2}{\cos(\theta_i)}\mathbf{k_{uv,s}} \quad (10)$$

$$\Psi_r = \Psi_i + arg(r) \quad (11)$$

Where $\mathbf{n_i}$ is a unit vector in the direction of the incident ray, $\mathbf{n_s}$ is a unit vector normal to the material discontinuity and $\mathbf{k}$ is the wave vector. Equations 7 to 9 represent a reflected curvature tensor in the uu, uv and vv components. Ψ_i and Ψ_r are the refracted ray phase of the incident ray and reflected ray respectively. The difference between these two terms is the phase shift $arg(r)$. Table summarises the Fresnel lens dimensions:

Fresnel Lens Specification	
Thickness (t)	3.5 mm
Width of singular linear Fresnel lens (w_f)	165 mm
Length (l_f)	1000 mm
Number of Fresnel lenses in array (N)	5
Focal length (f)	300 mm
Material	PMMA

Table 1: Summary of Fresnel lens design parameters

3.2 Construction of stepped thickness waveguide

The stepped thickness waveguide was constructed by first creating a 2D axisymmetric cross section, then specifying the coordinates of the edge point in a 2D coordinate system and finally extruding by 1000 mm, converting it to a 3D model. At the material discontinuity where there are changes of the medium from the material of model to air, the refracted wave vector is controlled by Snell's law based on the refractive index on either side. If the incident ray undergoes total internal reflection, no refracted ray is produced. The angle of incidence θ_i is computed:

$$\theta_i = \arccos\left(\frac{\mathbf{n}_i \cdot \mathbf{n}_s}{|\mathbf{n}_i|\,|\mathbf{n}_s|}\right) \quad (12)$$

In an isotropic medium, the electromagnetic properties such as the refractive index are the same in all directions. At a boundary between two isotropic, non-absorbing media, the refracted ray propagates in the direction $\mathbf{n_t}$ is given by the following relations:

$$\mathbf{n}_t = \eta\mathbf{n}_i + \gamma\mathbf{n}_s \quad (13)$$

$$\gamma = -\eta\cos\theta_i + \cos\theta_t \quad (14)$$

$$\eta = \frac{n_1}{n_2} \quad (15)$$

$$\theta_t = \arcsin(\eta\sin\theta_i) \quad (16)$$

Where the ray propagates from the medium with refractive index n_1 into the medium with refractive index n_2. θ_t is the refractive angle. η and γ are two defined variables. For medias that are non-absorbing, the quantities n_1, n_2, θ_i, and θ_t are real-valued.

Two wall conditions were introduced to the waveguide model. The first is at the bottom surface of the waveguide. In order to achieve a maximum performance with minimum power losses of light, the waveguide is modelled to have specular reflection if various conditions are met. This interface is set such that there was specular reflection if the incident angle is greater than the critical angle, and rays would be removed if the angle is less than the critical angle, allowing only total internal reflection to be seen. A pair of variables were detailed in the parameters setting to calculate the refractive index of the waveguide material, and the critical angle on its surfaces given that the free-space region has refractive index of the material. Specular reflection primary behaviour is used, giving access to a variable describing the angle of incidence of the ray on the boundary. The primary ray condition option can then be used to compare this angle to the critical angle. Another wall condition is used to describe the exit of the waveguide. The condition was set to 'Freeze', which stops any rays from further propagating and therefore the wave vector of the ray remains at the same value as when the ray initially strikes the wall. This boundary condition helps to study the ray power or intensity at the instant contact was made with the wall. The key design parameters of the waveguide are shown in Table 2.

Waveguide Specification	
Total height (h)	10 mm
Inclined angle (θ)	30°
Height of each step (s)	2 mm
Width (w)	1000 mm
Length (l_w)	900 mm
Light directing surface length (m)	4 mm
Material	PMMA

Table 2: Summary of stepped thickness waveguide design parameters

After the geometry is created and the model physics is defined, the next step is to build the mesh. COMSOL uses the finite element analysis (FEA) method to solve time-dependent problems. The partial differential equations (PDEs) used to define these problems can be approximated as numerical model equations. The solution to these equations act as an approximation to the real PDEs. A mesh is doing exactly what has been described above and plays an important role in how the model is solved and directly affects the accuracy of the solution. User defined elements such as prism and tetrahedron divide the 3D geometry into small finite parts. These parts are studied separately and together will give the final solution.

The ray tracing study is carried out by defining the time steps and maximum time span. To be specific, the ray propagation within the time it takes for the waves

(which represent photons traveling at the speed of light in a vacuum, 3×10^8 m/s, to propagate through the system was studied.

4 Results

The ray tracing simulation was carried out as specified. Starting from the initial coordinates of the simulated light source, rays were sent out in the direction defined earlier, which was vertically downw towards the linear Fresnel lens array. As time goes on, the light rays pass through the Fresnel lens, concentrating the light into five linear regions on the waveguide. The ray tracing simulation will stop if the maximum time span defined is reached. The propagation of light through the system is shown in Figure 9.

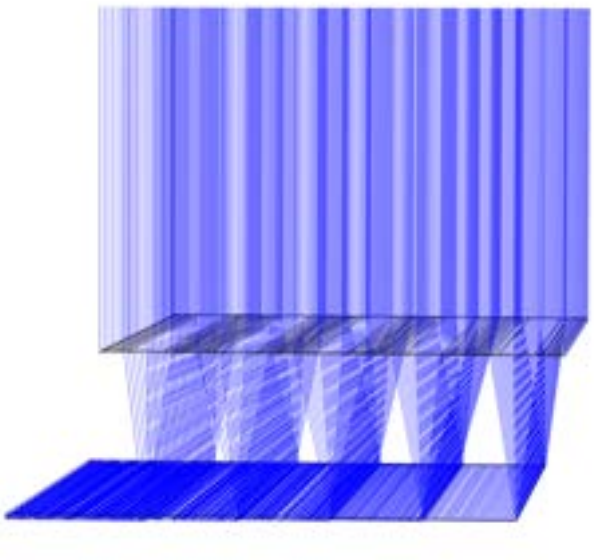


Figure 9: Simulating the ray trajectories of the light propogation through the optical system. A GIF of the process can be found [8]

4.1 Intensity profile on focal plane

By introducing a physical plane at the focal length on COMSOL, a 'Ray Accumulator' node can be used to calculate the light intensity profile of the plane.

Figure 10 shows the intensity distribution on the focal plane, clearly displaying the five linear regions of concentrated light as expected. The maximum intensity reached is 2.05×10^5 W/m^2 and can be found at the centre of the spots which are shown in white. The blue parts represent a lower intensity, and the black parts indicate regions where the light did not penetrate the plane.

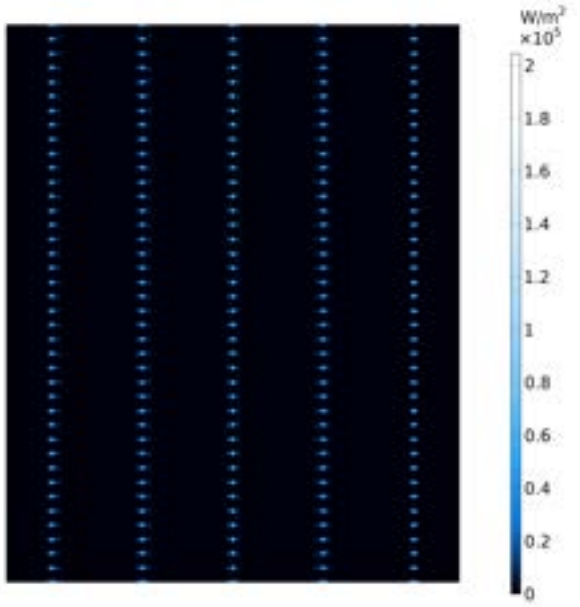


Figure 10: Intensity plot at the focal plane of the Fresnel lens

4.2 Concentration ratio profile at waveguide exit

A ray detector was introduced at the exit of the waveguide, its role being to compute information about rays that arrive at a set of selected boundaries. In our case, one selected boundary is the surface of the waveguide exit. The ray detector can provide the accumulated power at the surface. The concentration ratio was then studied by taking the ratio of input light intensity and exit light intensity.

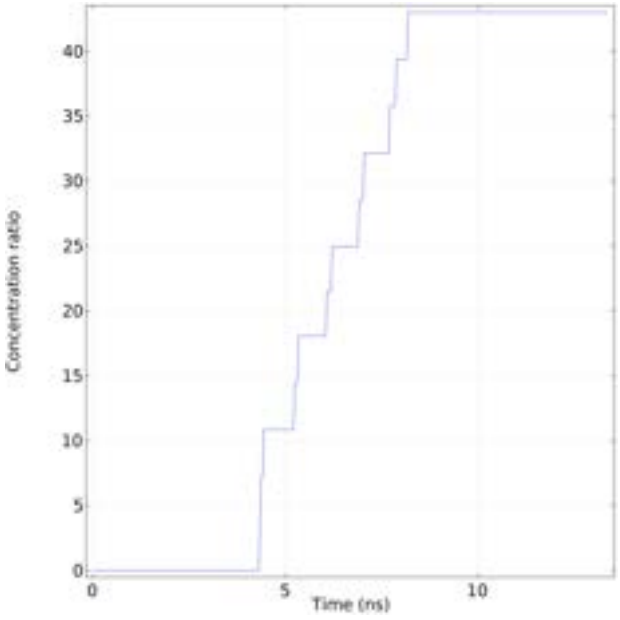


Figure 11: Optical concentration ratio at waveguide exit against simulation time

We chose to plot the time as the independent variable against the concentration ratio, the dependent variable in our analysis, as this reveals how light propagation varies with time. A maximum concentration ratio of 43

is reached at approximately 8.2 ns as highlighted. The results are shown in Figure 11.

4.3 Ray power profile at waveguide exit

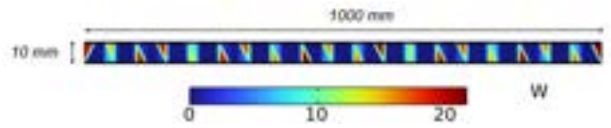


Figure 12: Ray power profile at the waveguide exit

The power distribution diagram at the exit port of waveguide is shown in Figure 12. The shape of the light focus points and distribution come from the shape of the light source, which is rectangular. The red colour represents locations with the highest power, around 23 W. Summing up the power represented by these shapes will give the total accumulated power at the exit surface. The intensity at the exit can be calculated by dividing the power by the cross sectional area.

5 Discussion

The optical efficiency of the coupled Fresnel lens and waveguide can be calculated using the following expression:

$$Opticaleffiency = \frac{P_{exit}}{P_i} \tag{17}$$

Where P_{exit} is the steady state power at the waveguide exit in W and P_i is the incident light power in W.

Using data output by the COMSOL simulation, the optical efficiency of the system was found to be 53%.

As for optical losses, the mirror is modelled as ideal hence there is no energy loss stemming from this. However, in reality the reflection ratio will not be 1 due to a number of limitations, such as manufacturing defects and material degradation. Losses come from the absorption of light within the Fresnel lens array and light absorbed from the waveguides. Also, there are losses due to Fresnel reflection where the two media having different refractive indexes. To be specific, the amount of light lost depends on the properties of material.

5.1 Concentration ratios

The geometric concentration ratio C_0 which is used to evaluate the quality of solar collectors is defined as:

$$C_0 = \frac{A_{Fresnel}}{A_{exit}} \tag{18}$$

Where $A_{Fresnel}$ is the area of Fresnel lens array and A_{exit} is the area of focus spots at the waveguide exit port. A geometric concentration of 591 is obtained from the model. Typically, in concentrator photovoltaic (CPV) system, the concentrator needs to have a concentration ratio in the range of 400 – 1000 times to effectively use high-efficiency solar cells [9]. Therefore, the proposed model is considered to be acceptable. However, geometric concentration does not take power loss into account as it assumes uniform radiation flux. For modelling purposes, a more representative way of deducing the efficiency of an optical system is to use the optical concentration ratio, C_{opt}, which is defined as the intensity ratio at the lens and at the receiver.

$$C_{opt} = \frac{I_{Fresnel}}{I_{exit}} \tag{19}$$

A maximum optical concentration ratio of 43 is reached for our model.

The $f - number$ of a linear Fresnel lens is defined as:

$$f - number = \frac{f}{w_f} \tag{20}$$

According to Davis' simulation results [10], for a circular Fresnel lens, a f-number of 1.85 gives an optical concentration ratio of 17. This indicates the implement of waveguide gives rise to an improvement in concentration ratio.

5.2 Investigating the incline angle of the stepped thickness waveguide

The optimum incline angle of the waveguide mirrors was determined by performing a parametric sweep in COMSOL, whereby the model solutions are found iteratively for each different inclined angle. Since this requires significant computational power, we chose a range of angles in increments of 5°, relative to the initial value of 30°that was chosen by Vu et al. [7], who offered no explanation for why they chose this specific value.

Figure 13 depicts the results of the parametric sweep, revealing an interesting interplay between the inclined angle and optical concentration ratio. The light rays take a specific time to i) pass through the Fresnel lens and ii) propagate through the waveguide, finally reaching the ray detector at the exit port of the waveguide. This is the reason that there is a stepped increase in the optical concentration ratio with time, until all the simulated rays have propagated through the waveguide and a steady state optical concentration ratio is reached. A general observation that can be made is that decreasing the inclined angle increases the concentration ratio. However, in practice this could have drawbacks. Decreasing the angle slowly reverts the stepped thickness waveguide back to a planar waveguide, which as discussed before is a lossy system, decreasing the efficiency of the system and also exacerbating further ray leakage. Although the general trend is that a decreased incline angle increases the concentration ratio, it was found that an angle of 20°achieves the optimum concentration ratio. This suggests that there is a degree of optimality in the system which could be investigated further through reactor parameter optimisation.

Analysis of the 'staircase' structure also gives an insight into the non-steady state light propagation through the system. Five distinct 'steps' are observed, corresponding to the number of Fresnel lens in the array.

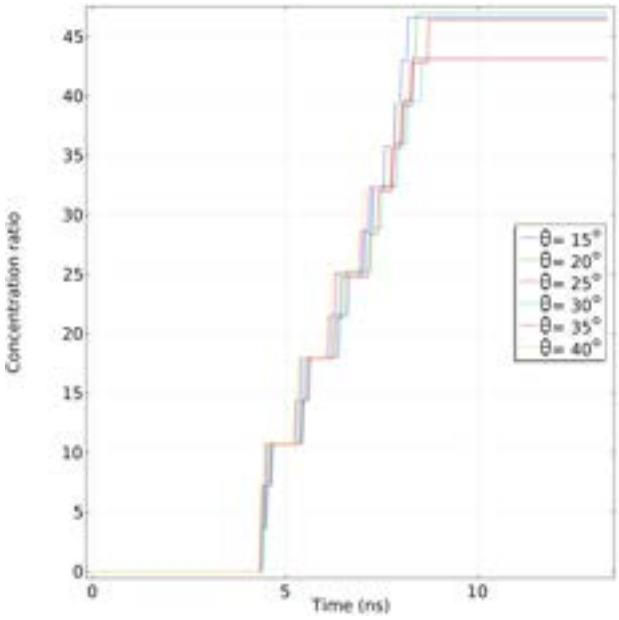


Figure 13: Parametric sweep results, investigating the relationship between the waveguide inclined angle and optical concentration ratio

The waveguide is a 3D component and the light rays that are being focused by the Fresnel lens hit the waveguide from all lateral angles. Clearly the incident light rays from the Fresnel lens that hit the light directing surface closest to the waveguide exit will reach the ray detector first, followed by light rays hitting the second closes light directing surface and so on. This explains the number of steps on the plot. A time delay is also observed during the concentration ratio step increase. We speculate that this occurs because of ray interactions with waveguide boundaries and also with other rays. This time delay increases fractionally after each step. Intuitively this makes sense as the rays that hit the light directing surface furthest away from the ray detector will interact with the most boundary surfaces and rays within the waveguide on average, causing the largest time delay. The inclined angle parameter sweep also surprisingly reveals that the time delay varies in a peculiar non-linear way, varying significantly with different inclined angles. Perhaps running the COMSOL model with a higher resolution would smooth out this disparity in the time delay, or it could be that a change in the incline angle alters the ray distribution of the system in a nontrivial manner. Nevertheless, the results show that the inclined angle of the waveguide is a parameter that can be optimised to maximise the concentration ratio and minimise the time delay of the system. The concept of steady state and non-steady state when referring to ray propagation through the system could be described as redundant because of the almost instantaneous reaching of the steady state.

5.3 Global solar radiation intensity

The concept of global solar radiation intensity was introduced in the background section. Considering the global intensity at London, as shown in Figure 14, the solar intensity that reaches the Earth's surface varies with each day of the year and the maximum global intensity is reached on the 168th day of the year (midyear). Therefore, the intensity input to the Fresnel lens surface will not be exactly 1000 W/m^2 as specified in our model at all times – this was more of an average value. This means recording the exact solar intensity from time to time becomes important as it changes model results significantly. In this case, the optical concentration ratio will vary and geometric concentration ratio becomes applicable as it depends on the system geometry only. In order to develop a more accurate model, proper solar tracking and measurement systems could be investigated. A dual-axis tracker designed by Nguyen [11] claimed to provide an additional 40% of solar energy over the year, relative to a system in a fixed position. Thus, a solar tracking system guarantees a maximum intensity of light can be captured and a good measurement gives an accurate intensity input to the model. Another reason why a solar tracking system is important is because the light source in the model is considered to be ideal, which means the incident angle of light is always 0°and all the incident light rays hit the Fresnel lens array. In reality, the incident light will come in all directions.

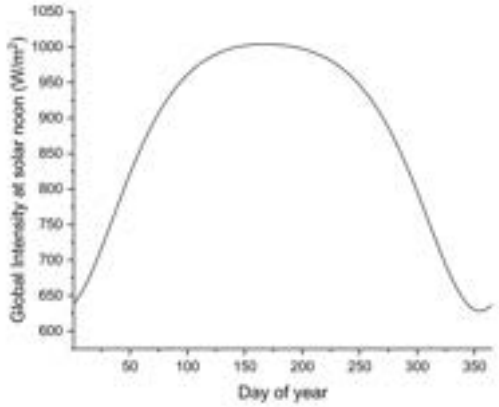


Figure 14: Global radiation intensity in London versus day of year

5.4 Thermal management

A consequence of concentrating light into small focal regions is that this greatly increases the heat transfer of the system, leading to extremely high temperatures at the focal regions. These temperatures cause adverse degradation of the materials involved. In the application that this study describes, the light rays are concentrated by a linear Fresnel lens array onto a stepped thickness waveguide made of PMMA, which has a melting point of approximately 160 °C. Heat dissipation systems must be implemented to ensure that the waveguide material does not melt - in other solar concentration applications passive and active cooling water systems have been implemented [2].

6 Conclusion

This paper reports model-based predictions of the performance of an optical system for harvesting, concentrating and directing solar radiation into a photoelectrochemical reactor. A model of the proposed PMMA linear Fresnel lens array and stepped thickness waveguide was built using COMSOL Multiphysics 6.1. The Geometrical Optics module was chosen, which took into account refraction according to Snell's law at a material interface. The stepped thickness waveguide caused light rays to propagate via total internal reflection. The bottom waveguide surfaces have reflective coating to minimize the energy loss. Given an input light source of 1000 W/m^2, the geometric concentration ratio and optical concentration ratio obtained were 591 and 43 respectively in an ideal case. The optical efficiency of the system was found to be 53%. The maximum optical concentration ratio at the waveguide exit could be improved by changing the inclined angle of the waveguide to the optimum value of 20°. For a more realistic study, a solar tracking system could have been added to the COMSOL model to increase the accuracy of the simulation. Further work may include simulating the heat transfer of the optical system on COMSOL to determine the viability of certain waveguide materials, as well as discovering new waveguide geometries such as a double layer planar waveguides proposed by Vu et al. [12] used in concentrator photovoltaic systems.

Acknowledgements

The authors would like to express their gratitude to Dr Anna Hankin for her continual support and guidance.

References

(1) Parmesan, C.; Yohe, G. A globally coherent fingerprint of climate change impacts across natural systems. *Nature* **2003**, *421*, 37–42.

(2) Moss, B.; Babacan, O.; Kafizas, A.; Hankin, A. A Review of Inorganic Photoelectrode Developments and Reactor Scale-Up Challenges for Solar Hydrogen Production. *Advanced Energy Materials* **2021**, *11*, 2003286.

(3) Carver, C.; Ulissi, Z.; Ong, C.; Dennison, S.; Kelsall, G.; Hellgardt, K. Modelling and development of photoelectrochemical reactor for H2 production. *International Journal of Hydrogen Energy* **2012**, *37*, 2010 AIChE Annual Meeting Topical Conference on Hydrogen Production and Storage Special Issue, 2911–2923.

(4) Hankin, A.; Bedoya-Lora, F. E.; Ong, C. K.; Alexander, J. C.; Petter, F.; Kelsall, G. H. From millimetres to metres: the critical role of current density distributions in photo-electrochemical reactor design. *Energy Environ. Sci.* **2017**, *10*, 346–360.

(5) Kumar, V.; Shrivastava, R.; Untawale, S. Fresnel lens: A promising alternative of reflectors in concentrated solar power. *Renewable and Sustainable Energy Reviews* **2015**, *44*, 376–390.

(6) Ray Tracing Simulation of a Fresnel Lens, COMSOL.

(7) Ngoc-Hai, V.; Shin, S. A Large Scale Daylighting System Based on a Stepped Thickness Waveguide. *Energies* **2016**, *9*, 71.

(8) Halder, S.; Zhu, J. Solar Rays propagating through a linear Fresnel lens array and waveguide with embedded mirrors, Electrochemical Systems Laboratory, Available at: https://www.youtube.com/watch?v=ojuk8rbVFX0, 2022.

(9) Xie, W.; Dai, Y.; Wang, R.; Sumathy, K. Concentrated solar energy applications using Fresnel lenses: A review. *Renewable and Sustainable Energy Reviews* **2011**, *15*, 2588–2606.

(10) Davis, A. Fresnel lens solar concentrator derivations and simulations. *Proceedings of SPIE - The International Society for Optical Engineering* **2011**, *8129*, DOI: 10.1117/12.892818.

(11) Nguyen, N. H. In 2016.

(12) Vu, N. H.; Pham, T. T.; Shin, S. Large Scale Spectral Splitting Concentrator Photovoltaic System Based on Double Flat Waveguides. *Energies* **2020**, *13*, DOI: 10.3390/en13092360.

Data-driven modelling of twin-column chromatographic multi-column counter-current solvent gradient purification (MCSGP) process

Haonan Wang and Yanlong Zhao
Department of Chemical Engineering, Imperial College London, U.K.

Abstract Chromatography is often involved in the downstream bioprocess for the separation of monoclonal antibodies (mAbs). A semi-continuous chromatographic process named multicolumn counter-current solvent gradient purification (MCSGP) process had been developed and proved feasible for the separations of mAbs. High-fidelity simulations of such processes often demand computing a large set of partial differential and algebraic equations (PADES) rendering great computational efforts. In contrast, a black box model based on the mapping of inputs and outputs using purely mathematical relationships could save a lot of computational effort. This paper explores the feasibility of developing a data-driven model using feed-forward artificial neural networks (ANNs) for a twin-column MCSGP set-up. Two models were developed with saving computational time as a priority. The two models were designed based on the best mean square error (MSE) of predictions and optimised through exploring different learning rates to the 'Adam' optimiser. The results proved the two models can be either used in combination or individually to give a preliminary screening of varying inputting operating parameters that would be considered feasible given the requirements such as purity, yield, and product collection window. The developed models were also proved to demand negligible computational time compared to the mechanistic model simulations.

Keywords: Chromatography, Twin-column MCSGP, Black box model, Machine learning, ANN

1 Introduction

Monoclonal antibodies (mAbs) have contributed vitally to advancement in the treatment of infectious diseases, cancer, and autoimmune diseases (Torphy, 2002). However, due to their targeted treatment of chronic diseases such as cancer and their relatively low potency which results in the need for high cumulative doses, mAbs are usually amongst the most expensive of drugs (Farid, 2007). Driven by the increasing pressure of mAbs market demand, there exists significant research interest related to lowering the cost of producing mAbs. It has been shown that the cell culture titre is one of the most prominent cost drivers (Werner, 2004). Increasing the titres in turn drives the search for novel approaches or alternatives for the downstream purification processes to ensure there is a net gain in lowering the production cost (Farid, 2007).

Chromatography is often employed in the downstream bio-process for the separation of mAbs. Traditionally, it was performed using gradient batch processes. In the past few decades, significant efforts have been made to the development of continuous chromatographic processes, which have been proven to be much more economical than batch processes for large-scale productions of the biomolecules (Aumann & Morbidelli, 2007). Classical continuous chromatography processes such as simulated moving bed (SMB) and recycling chromatography have significant advantages over the batch types attributed to their ability to conserve partial separation thereby resulting in better productivity (Müller-Späth et al., 2008). Therefore, these processes were often deployed in the binary separations of bio-molecules. However, mAbs separations generally requires splitting the feed stream into three fractions with the product having intermediate adsorptive properties and the weak and strong components as impurities. Although (Kim et al., 2003) proved the applicability of SMBs for ternary separations by cascading two SMBs in series, it was limited to the case where there is little of the most strongly or weakly adsorbed impurities. In addition, the characterised drawbacks for SMB and recycling chromatography as their inability to perform linear solvent gradients and strong dilution effects of product sample prior to reinjection respectively urge the development of a continuous chromatographic process specifically targeted for the effective separation of three component biomolecular mixtures.

MCSGP, first introduced by (Ströhlein et al., 2006), combined the advantages of both a solvent gradient batch and a continuous SMB. An MCSGP process can perform solvent gradient elution as in batch units but using a continuous countercurrent unit. The countercurrent nature of an MCSGP process refers to the solid phase resins switching in position opposite to the flow direction. The readers are referred to (Aumann & Morbidelli, 2007) for a detailed explanation of the working principle and a preliminary design based on a batch gradient and the corresponding chromatogram for the original 6-column set-up, provided by the inventors. (Aumann & Morbidelli, 2008) later demonstrated the practicality of reducing the columns to 3 and provided a design of the process based on the same experimental method. However, these designs generally lack relationships describing the physiochemical phenomenon happening in the process, a more robust design approach will be based on the relationships describing the adsorp-

tion equilibria, and the mass transfer properties of all components. (Müller-Späth et al., 2008) developed such a mathematical model with Langmuir adsorption isotherms and a lumped kinetic model. The performance parameters were defined as purity and productivity. The issue is that to describe the periodic nature and complicated adsorptive phenomena underlaid, the mathematical models generally comprise large sets of partial differential and algebraic equations (PADEs) rendering great computational efforts.

There exists, however, another approach based on the mapping of input and output data using purely mathematical correlations while involving minimum physiochemical process description. The data-driven model can be called a black box model, which, although lacks physiochemical knowledge, generally saves computational time and effort (Tabora, 2012). The data-driven modelling is particularly applicable to the MCSGP process since not only the process requires great computational power to simulate but also the feed compositions will vary simultaneously subject to the disturbances from the upstream bioprocess. This renders the optimisation of the operating parameters very computationally expensive. The black box model saves the computational time and effort of computing every time point for infeasible sets of initial conditions until the process reaches the cyclic steady state (CSS). This paper explores the approach for the modelling of the MCSGP process at CSS that is based on feedforward artificial neural networks (ANNs) following four main steps: (1) Data generation from high-fidelity simulations of the MCSGP process. (2) Determining the ANN structure including the activation functions, hidden layer size and neuron size. (3) Data-driven model optimisation via tuning learning rate. (4) Cross-validation of the data-driven model using additional data generated from the mechanistic simulations. Two data-driven models were developed following the above approach, one static model aimed at screening infeasible combinations of inputs and one pseudo-dynamic model designed to approximate the product concentration profile. The results showed that the static model can give a decent preliminary screening of the combinations of operating parameters that could potentially satisfy the purity and yield requirements, whereas the pseudo-dynamic model was able to approximate the concentration profile through an operating cycle given the input conditions. Both models presented require negligible time to run compared to the simulation based on the mathematical relationships.

2 Theoretical Background

2.1 Twin-column MCSGP

The focus of the paper was on the twin-column MCSGP process setup. As explained by (Krättli et al., 2013), figure 1 showed the principle of the process operation, where the middle of the two columns is connected by flow streams indicated by the arrows. The figure depicts an entire cycle of the MCSGP that is the complete repeating element of the process.

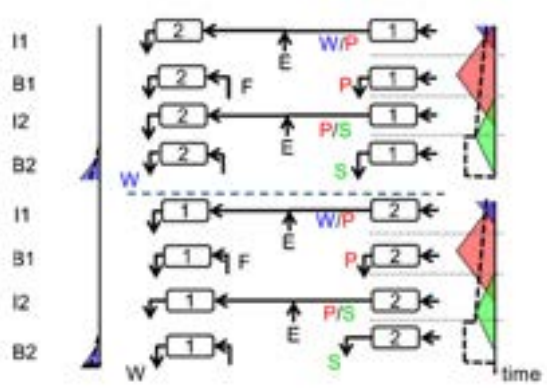


Figure 1: Schematic overview of a complete cycle of the twin-column MCSGP process, adapted from (Krättli et al., 2013)

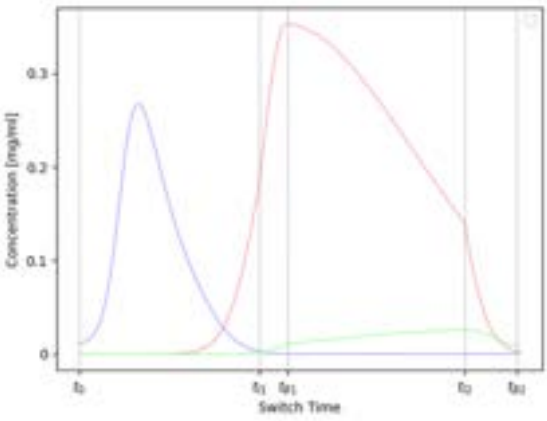


Figure 2: Plot of concentration against switch time, the blue curve represents weak impurity, the red curve represents the product and the green curve represents the strong impurity

To understand the process, it is convenient to notice the general observation that throughout all cycles, the column on the right is carrying out the gradient elution task, whereas the column on the right is performing recycling and feeding tasks.

The process starts with column 2 being equilibrated and emptied. During phase I1, overlapping regions of weak (W) and product (P) are eluted from column 1. Before the stream enters column 2, it is mixed with an additional stream of pure eluent (E) to enhance the absorption. Figure 2 shows that phase I1 ends when negligible impurity W concentration is found to be eluted from column 1. During phase I1, all components being fed into column 2 are expected to be retained in column 2. Phase B1 operates in batch mode. During this phase, the stream eluted from column 1 is designed to have the highest P concentration with minimal impurities contamination throughout the whole switch. Therefore, phase B1 is also known as the product collection window. Fresh feed is introduced into and retained in column 2 during this phase. Phase I2 starts when the overlapping regions of P and S components are eluted from column 1. The modifier concentration in column 1 contin-

ued to increase throughout the switch to counteract the increasing adsorptive interactions in the order of component (W, P, S) with the solid phase resin. Since impurity W interacts the least strongly with the resin, it travels the quickest inside the column. Therefore, when phase I2 finishes with the recycling of the P/S stream. Phase B2 starts with mostly S being eluted from column 1 and with W being eluted from column 2. Phase B2 continues until column 1 is empty, meaning all components (W, P, S) were eluted out. After this point, the switch occurs, meaning that column 1 swaps position with column 2. The gradient elution task is now performed by column 2, which is the column on the right at this point. The cycle continues through all phases until column 2 has eluted out all components. Then, both columns will be switched back to the initial positions, and the process will repeat exactly the path described above.

2.2 Cyclic Steady State Operation

In contrary to continuous steady-state, CSS is more relevant to the MCSGP process due to its nature where both the continuous sector and batch sector are involved in the process. CSS exhibits uniform operating parameters and produces similar output from batch to batch given the batch time being constant throughout each cycle. (Minceva et al., 2003) This means that both the liquid and solid phase concentration will, at a certain time after the process starts, follow an almost identical profile axially along the bed. The cyclic concentration profile of each component will therefore be indistinguishable from one another between each cycle. From this time onwards, it can be identified as that the process has reached CSS. In the MCSGP process, other than the start-up and shut-down procedure, CSS will be applicable most of the time. Therefore, the process performance at CSS will be the focus of this study.

2.3 Mathematical Formulations

A mathematical model describing the detailed physiochemical behaviour of the process comprised of a lumped kinetic model and Bi-Langmuir adsorption isotherm was developed and validated against experiments by (Müller-Späth et al., 2008), where detailed mathematical relations and nomenclature can be found. Such mathematical models developed based on the first-principal approach can be classified as mechanistic models. A mechanistic model is also known as a white box model since its structure is well-defined and transparent. Although mechanistic models provide precise and detailed descriptions of the underlying physiochemical system, it generally comprises a large set of complex PADEs. For the twin-column MCSGP process, the simulated model involves 50 space discretisation points, resulting in 4119 equations and 3309 variables (Papathanasiou et al., 2016). The main model complexities arise from equations 1 and 2 since both partial differential equations are a function of time and space. As listed below, the two equations describe the liquid and solid phase concentrations of the four components (i.e., the modifier, the weak impurities, the product, and the strong impurities).

Liquid Phase Concentration

$$\frac{\partial \mathrm{c_{i,h}}}{\partial \mathrm{t}} = \mathrm{D_{ax}} \frac{\partial^2 \mathrm{c_{i,h}}}{\partial \mathrm{t}^2} - \frac{\mathrm{Q_h}}{\mathrm{A_{col}} \varepsilon_\mathrm{i}} \frac{\partial \mathrm{c_{i,h}}}{\partial \mathrm{z}} - \frac{(1-\varepsilon_\mathrm{i})}{\varepsilon_\mathrm{i}} \frac{\partial \mathrm{q_{i,h}}}{\partial \mathrm{t}} \quad (1)$$

Solid Phase Concentration

$$\frac{\partial \mathrm{q_{i,h}}}{\partial \mathrm{t}} = k_i \left(q^*_{i,h} - q_{i,h}\right) \quad (2)$$

where $t \in [0, t_{\mathrm{end}}]$ represents the time, $z \in [0, L_{\mathrm{col}}]$ the column length, $i = 1, \ldots, n_{\mathrm{comp}}$ the component, and $h = 1, \ldots, n_{col}$ the column index.

The competitive bi-Langmuir isotherm

$$q^*\left(\mathrm{c_{i,h}}\right) = \frac{\mathrm{c_{i,h}} \cdot \mathrm{H^I_{i,h}}}{1 + \sum_{i=2}^{n_{comp}} \frac{\mathrm{c_{i,h}} \cdot \mathrm{H^I_{i,h}}}{q^{\mathrm{I}}_{i,h}}} + \frac{\mathrm{c_{i,h}} \cdot \mathrm{H^{II}_{i,h}}}{1 + \sum_{i=2}^{n_{comp}} \frac{\mathrm{c_{i,h}} \cdot \mathrm{H^{II}_{i,h}}}{q^{\mathrm{II}}_{i,h}}} \quad (3)$$

Computing both the liquid and solid phase concentration values at every time and every space point would demand large computational power. Therefore, when carrying out a sensitivity analysis or later implementation of a control and optimisation strategy, the computational time could usually be a major concern for the mechanistic model.

3 Methodology

The framework as illustrated by figure 3 was followed through this study.

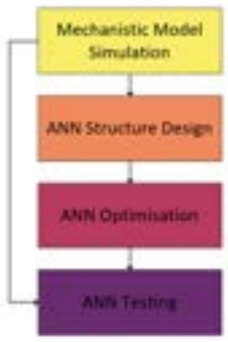


Figure 3: The framework of black box modelling

3.1 Data Generation and Analysis

The ANNs were trained by the data generated from the mechanistic model built by Papathanasiou's research group at Imperial College London on gPROMS® ModelBuilder v7.1.1. The data was generated using the gPROMS's Global Sensitivity analysis (GSA) feature. Although a GSA analysis is particularly applicable to this process as described above, it is not the main concern of this work. The focus instead was to use this feature as a convenient way of

generating the data since the gPROMS's Global Sensitivity analysis feature could automatically compute a large set of combinations of inputs (initial conditions), thereby generating the corresponding output responses. The initial conditions used as inputs for the data generation are listed in the table below:

First, 500 samples each of varying combinations of inputs were generated from the mechanistic model simulation on gPROMS® ModelBuilder. A subsequent data handling was carried out to determine the CSS. It was decided to record the maximum concentration of the product at the outlet of the column executing gradient elution during each cycle. The CSS had been reached once the maximum product concentration fluctuates within 0.1% from adjacent cycles.

The purity and yield at CSS were then computed for each sample. The performance indicators were decided to be the average purity and yield during CSS, which were computed by the following equations.

Purity

$$\mathrm{Pur}_{av,j} = \frac{C_{avPs,j}}{C_{avWs,j} + C_{avPs,j} + C_{avSs,j}} \tag{4}$$

Yield

$$Y_j = \frac{C_{avPs,j}}{C_P^{\mathrm{feed}}} \tag{5}$$

Where, $j = 1, \ldots, n_{\mathrm{cycle}}$ the cycle index and $s = 1, \ldots, n_{\mathrm{outlet}}$ the outlet stream

Given the nature of the product, it was decided that at least 98% purity was required. The yield requirement was decided to be above 80% to ensure a satisfactory profit. The average concentrations were computed using numerical integration with a time interval of 0.2s.

3.2 ANN Structure Determination

Python V3.10 was used to build all models. Numpy V1.23 and Pandas V1.5.2 libraries were used for data processing. TensorFlow V2.11 and Keras V2.3 libraries were used for building, training, and validation of ANNs.

Two models were decided to build—-one static and one pseudo-dynamic model. The first one consisting of 1 ANN predicting the purity and yield was called the static model. The second one utilised 2 ANNs where each ANN predicted the average product concentrations at the outlet of one column through one switch. The pseudo-dynamic model's predictions consisted of the average product concentration at 8 time points corresponding to 8 stages through 2 switches. In this way, by tracking the product concentration from the outlet of one column through all stages in a whole cycle, the concentration profile along the other column would be identical except for a time difference. Since the model only predicted the average concentration at eight instants of the process during a CSS cycle, it could only approximate the concentration profile resulting in the model not being truly dynamic. To distinguish it from the static model, the latter model was defined as pseudo-dynamic.

The inputs for both models were decided to be the same as the inputs for the data generation as shown in table 1 to incorporate as many details as possible. Since the static model predicts yield and purity, to restrict the output between 0 to 1, a sigmoid activation function was added to the output layer; whereas to ensure the average concentration output from the pseudo-dynamic model to be positive, the output layer for each of the 2 ANN was implemented with softplus activation function. The two activation functions are defined below.

Sigmoid

$$f(x) = \frac{x}{1 + e^{-x}} \tag{6}$$

Softplus

$$f(x) = \ln\left(1 + e^{x}\right) \tag{7}$$

To decide the ANNs structure such as the number of hidden layers, activation function, and neuron size on each layer. Mean squared errors, computed using the following formula, of the predictions to the actual data were calculated based on the dataset comprised of 500 samples. 1000 epochs were decided to train the ANNs of both models, whereas 25% percent of the 500 samples were split into the validation set. An average mean square error (MSE) was computed for the last 100 epochs for each combination of hyperparameters. This is to avoid the chance of training stopping at an abnormality.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 \tag{8}$$

The following values of the ANN hyperparameters were explored to decide the final structure of each ANN in the model.

Table 2: Table summarising ANN features explored in the structure determination

Features	Values
Activation functions	sigmoid(6), tanh(9), hard sigmoid(10), ReLu(11)
Hidden layer size	0,1,2,3,4,5,6,7
Neuron size	4,8,16,32,64,128,256

Activation functions:

$$f(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{9}$$

$$f(x) = \begin{cases} 0, & \text{if } x \leq -2.5 \\ 0.2x + 0.5, & \text{if } -2.5 < x < 2.5 \\ 1, & \text{if } x \geq 2.5 \end{cases} \tag{10}$$

$$f(x) = \max(0, x) \tag{11}$$

Table 1: Table summarising inputs conditions for the Training and Validation of ANNs.

	Boundary for 500 sets	Boundary for 100 sets	Unit	Distribution for 100 sets
Feed concentration of modifier	[2.0 3.0]	[2.1 2.8]	mg/mL	Normal distribution with mean 2.6, variance 0.6
Feed concentration of weak impurities	[0.05 0.08]	[0.04 0.07]	mg/mL	Uniform distribution
Feed concentration of product	[0.3 0.4]	[0.25 0.45]	mg/mL	Normal distribution with mean 0.35, variance 0.6
Feed concentration of strong impurities	[0.03 0.05]	[0.03 0.05]	mg/mL	Normal distribution with mean 0.04, variance 0.2
Inlet flowrate of the column executing the gradient elution during phase I1	[0.1 1.0]	[0.2 1.1]	mL/min	Normal distribution with mean 0.4, variance 0.6
Inlet flowrate of the column executing the gradient elution during phase I2	[0.1 1.0]	[0.1 0.9]	mL/min	Uniform distribution
Initial modifier concentration for the column executing the gradient elution	[2.0 3.0]	[2.2 2.8]	mg/mL	Normal distribution with mean 2.6, variance 0.4
Initial modifier concentration for the column executing the recycling and feeding tasks	[0.8 1.2]	[0.9 1.1]	mg/mL	Uniform distribution

3.3 Model optimisation

After the ANNs structure has been determined, 30 logarithmically (with base 10) spaced learning rates ranged from 1×10^{-5} to 1×10^{-3} were analysed as input for the optimiser 'Adam'. 'Adam' is a stochastic gradient descent method that is based on adaptive estimation of first-order and second-order moments. The advantage of 'Adam' over other optimisation algorithms arises from its well-suited behaviour for problems with a large set of data or parameters. (Kingma & Ba, 2014)

Sensitivity analysis based on varying learning rates was conducted by training both models with 1000 epochs and with the data split into 3:1 training and validation datasets. An average MSE loss was computed for the last 100 epochs for both the training and validation set at each learning rate for the same reason as described above. The variance of the MSE loss for the last 100 epochs was also calculated to quantify the degree of overtraining or fluctuation of MSE.

3.4 Model Validation

To further validate the accuracy of the trained models, an additional set of 100 samples comprised of varying input conditions was generated. In table 1, Column 'Boundaries Train Set' refers to the 500 samples used for the training of the ANNs of both data-driven models, whereas column 'Boundary Validation Set' refers to the 100 additional samples used for the validation. Column 'Distribution' refers to the distribution of each input in the validation set, whereas all inputs in the Train Set were uniformly distributed. An additional data set was generated using different distribution functions and bounds of the input parameters. To simulate the disturbance in the feed stream resulting from the upstream processes, up to 20% variation was applied randomly to the upper and lower bounds of feed concentrations in both columns. To further eliminate the effect of monotonic increment, uniform distribution was changed to normal distribution with the mean shifted up to 15% from the middle value between the upper and lower bound of each variable. The generated data was inspected and was completely different from the data set used for training.

4 Results and Discussion

4.1 Simulated Data Analysis

CSS Determination

As depicted by figure 4, CSS was reached at different cycle numbers depending on the initial conditions. For some combinations of inputs, CSS was reached from 7^{th} cycle onwards. However, the CSS was far yet reached even at the 10^{th} cycle for other cases. The previously mentioned combinations of inputs could potentially satisfy the purity and yield requirements, it takes too long to reach the CSS resulting in an undesired loss since it was decided that the simulation only runs for 20 cycles.

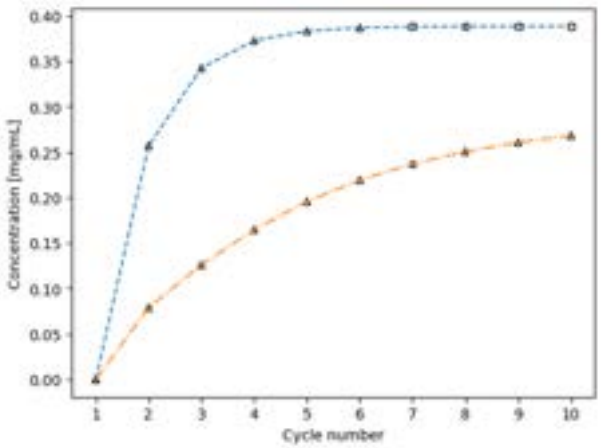


Figure 4: Figure illustrating cyclic steady state behaviour of the MCSGP process. The blue curve with '- -' line shape indicates a case that CSS has reached before 10^{th} cycle is reached whereas the orange curve with '- .' indicates a case that CSS is not reached yet at 10^{th} cycle. Each empty triangle locates at maximum cyclic product concentration where CSS has not been reached, whereas the empty squares indicate maximum CSS product concentration.

Non-linear Interactions in Inputs

The periodic nature and complex adsorptive reactions rendered the non-linear dependency of the process performance to the operating conditions. For example, a change in the impurities concentration will also affect the adsorptive behaviour of the product, rendering sometimes the modifier concentrations less effective. In addition, difference in flowrates inside the column during the interconnected states will also

have an impact on the degree of partial separations of eluting components, resulting in a different concentration pattern at the outlet stream during each stage. Varying each or some of the initial conditions such as feed concentrations or flowrates could result in a totally different process performance and may sometimes render the process infeasible. The data generation with 500 samples was also proved to be very time-consuming since to compute the process response, it requires the simulation to run through every time point until CSS has been reached. For this reason, a quick preliminary screening of the infeasible combinations of process operating conditions will be extremely useful for the later process optimisation subject to different set-ups or feed compositions.

As shown by figure 5, only a very small portion (4.4%) of the input combinations satisfied the purity and yield requirment at the 20th cyclic cycle. In addition, a wide spread of purities was observed through all yield values which demonstrates non-linear behaviour of the process outputs subject to different input conditions. For the high density of observed purities at yield approaching 0 is, for example, due to for certain flowrates (of the interconnected states) combinations, product may not be actually eluted out during the B1 stage. Moreover, the switch times and solvent gradients being kept the same for all inputs therefore remained unoptimised for each set of other input conditions explains why there is only a small fraction of the observed purities satisfy both requirements.

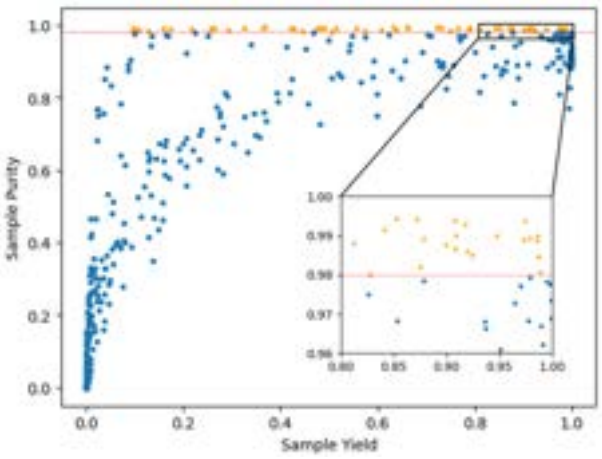


Figure 5: Plot of sample product purity against product yield, the red dotted line is a plot of $y = 0.98$, the orange triangle scatters represent samples which has reached 0.98 and the blue circle scatters represent samples which has lower purity than 0.98

This demonstrates the significance of developing a data driven model for quick screening of infeasible initial conditions.

4.2 The Static Model

The result from the determination of ANN structure for the static model in table 3 had shown that amongst all the tested activation functions, 'tanh' and 'ReLu' activation functions performed better than 'Sigmoid' and 'hard Sigmoid' activation functions with regards to the average MSE loss.

Table 3: Table summarising Best MSE Loss against ANN features for the static model. The 'Best MSE Loss' refers to the average MSE of the last 100 epochs out of the 1000 epochs used for training the ANN

Activation function	Best average MSE loss	Hidden Layer Size	Neuron Size
ReLu	0.0583	4	128
Tanh	0.0249	4	128
Sigmoid	0.1619	3	128
Hard Sigmoid	0.1613	4	4

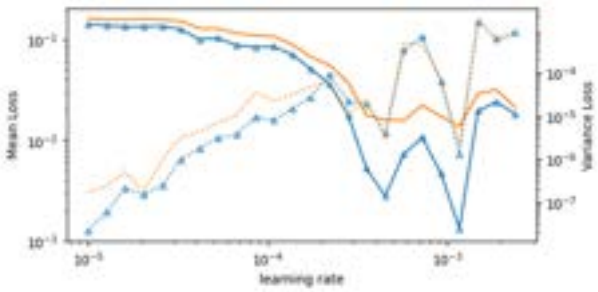
The static model composed of 5 feedforward layers, which include 4 hidden layers and 1 output layer, with 128 neurons on each hidden layer was decided. Further optimisation with regards to the learning rate was conducted to decide whether 'ReLu' or 'tanh' activation function performed better for the static model.

In general, the optimal learning rate was between 1×10^{-3} and 1×10^{-4}. With lower learning rates, the neural network was converging slowly and was likely to stop at the local optima, this will result in MSE not decreasing further to where the global optima are located. However, larger learning rates may also result in overtraining given 1000 epochs. It also lowered the possibility of locating an optimum. The training would then mispredict the global optima location when optimising the neural network. This was confirmed as higher variances were observed in the plot for learning rates that were higher than 2×10^{-3}. Therefore, a learning rate that results in both low mean and variance of MSE was chosen (figure 6).

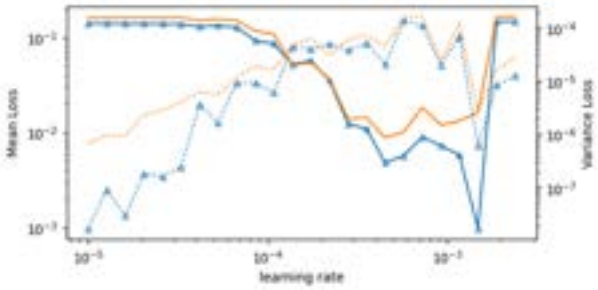
It was also noticed that 'Relu' activation function had a less stable MSE than the 'tanh' function as demonstrated by larger fluctuations in variance towards larger learning rates. This implied that there were potentially multiple optima for a short range of learning rates for the 'Relu' activation function. Since both activation functions had similar MSE at a comparative range of learning rates, to ensure the stability of the ANN, 'tanh' was chosen to avoid locating at different optima.

The final ANN structure had been designed as consisting of 5 layers with 4 hidden layers using 'tanh' and the output layer deployed the sigmoid activation function. For each hidden layer, 128 neurons were used. A learning rate of 1.05×10^{-3} was decided for the 'Adam' optimiser. Figure 7 showed the prediction result of this ANN.

After completion of model training, an MSE against epochs plot (figure 8) was generated for the ANN. Although fluctuations of MSE could be seen throughout the model training which indicated that the optimisation encountered various optima, a general decreasing trend was observed with increasing epoch number. Although fluctuations could be seen for the MSE value towards the end of the training, which indicated a certain degree of overtraining, it was considered acceptable given the MSE value was still decreasing.



(a) ReLu



(b) Tanh

Figure 6: Figure illustrating the average MSE loss and variance against learning rates for the 'Adam' optimiser of two activation functions (Upper plots: 'Relu'; Lower plots: 'tanh'). The solid lines represent the mean MSE and the dashed lines plot the variance of MSE. In both plots, the blues line with empty triangle illustate the training set data where the orange line shows the validation set data.

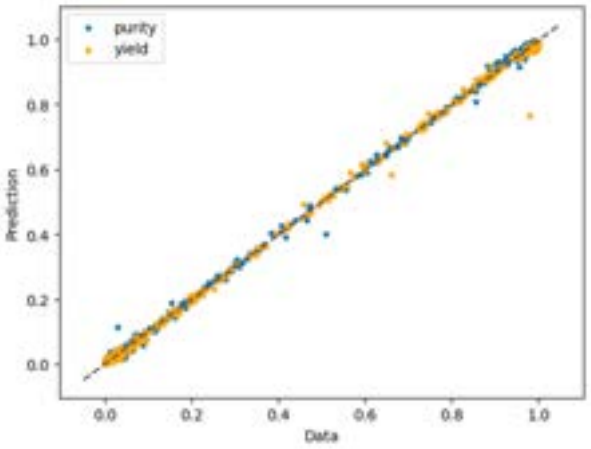


Figure 7: Figure illustrating the training result of the static model. The vertical axis represents the predicted result of purity and yield in range of 0 to 1 and the horizontal axis plots the sample purity and yield generated from the mechanistic model. The dashed line is a plot of $prediction = data$. The predictions would therefore be more accurate when the scattered points approach towards the dashed line.

The final MSE for the training and validation set were 3.6×10^{-4} and 1.8×10^{-3} respectively as shown in figure 8. This resulted in an average error of 0.02 for the training set and 0.04 for the validation set with respect to the simulation data. The percentage errors of either purity or yield prediction were therefore 2% for the train set and 4% for the validation set compared to the simulation data.

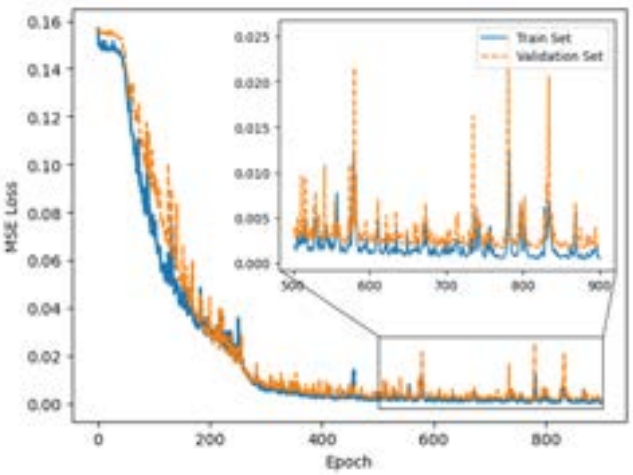


Figure 8: MSE loss against training epoch for the static model. The dashed orange line represents the 25% validation set whereas the blue line is plotted for the 75% training set out of the 500 training samples.

Table 4 below depicted the accuracy of prediction on the training dataset. Three target purity were chosen: 0.98, 0.95 and 0.90. The accuracy was not exactly the ratio between the number that reached target purity in the data and the number that met target purity in the prediction. This was due to the model's prediction error so that for certain samples the purities were predicted to be higher than the simulated data. Although 98% purity was the strict requirement for the targeted product, given there was only a very small portion of the samples satisfied the requirements as discussed above, a relatively minor error could result in a large deviation of the prediction accuracy calculated in such a way. Therefore, a larger allowance for the purity requirement such as 0.95 or 0.90 was decided to compute for prediction accuracy. Given the model served only as a preliminary screening of input parameters, it was justifiable that a relatively less strict purity requirement could be decided for predictions so that samples that achieved a relatively higher purity such as 0.90 could be optimised later via fine-tuning input parameters.

Table 4: Summary table for the static model accuracy against training dataset

Target Purity	Sample Number	Predicted Sample Number	Accuracy
0.98	60	45	66.7%
0.95	98	93	89.8%
0.90	134	132	95.5%

The accuracy of the static model's prediction was measured with the 100 samples (generated previously as a test set consisting of different input combinations) for different purity requirements. Table 5 below depicted the accuracy in predicting the number of cases that met the target purity.

The test samples were suggesting that the static model could predict the cases of input that would meet the target purity. This could be further im-

Table 5: Summary table for the static model accuracy against testing dataset

Target Purity	Sample Number	Predicted Sample Number	Accuracy
0.98	17	10	52.9%
0.95	23	19	73.9%
0.90	31	30	90.3%
0.80	37	37	94.6%

proved by optimisation of the ANN. Due to the non-reproducible nature of model training, it would be extremely difficult to locate the exact optima such that the ANN could predict the exact output from the simulated data.

4.3 The Pseudo-dynamic Model

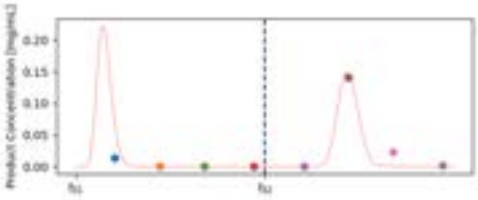
Following the same methodology as for the static model, the ANNs structure was also decided for the pseudo-dynamic model. Each ANN composing the pseudo-dynamic model consisted of 3 feedforward layers. The 2 dense layers consisted of 256 neurons deployed with 'ReLu' activation function, and the output layer used 'softplus' activation function to avoid negative average concentration outputs.

Table 6 illustrates the MSE for both ANNs of the pseudo-dynamic model. In switch 1, the average error present in the average product concentration for the training and validation set were 8.79×10^{-3} mg/mL and 0.046 mg/mL respectively, whereas for switch 2, the errors were 5.90×10^{-3} mg/mL and 7.61×10^{-3} mg/mL respectively. This implied that the general trend from stage to stage would not be drastically affected and the model would still provide sensible information to eliminate sub-optimal input conditions.

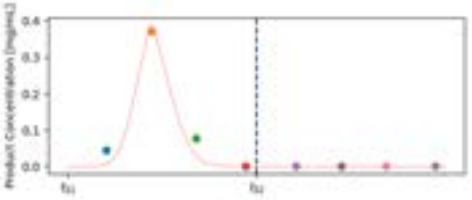
Figure 9 are the plots of samples for which the static model had predicted to have a product purity of above 0.98. The blue curve shows the product concentration profile from the mechanistic model. The scattered points are the predicted average product concentration profile of each stage from the dynamic model. The dashed line indicates the shift from switch 1 to switch 2 corresponding to each ANN. The predicted average product concentrations for each stage were considered accurate in providing an approximate concentration profile for each input condition in the dataset. The pseudo-dynamic model could therefore be used to provide insights for the product profile at the end of each stage for one column, thereby eliminating cases not having the noticeable or highest concentration of product during the collection window.

Table 6: Summary table for the pseudo dynamic model MSE and error

	MSE Train	MSE Validation	Error Train mg/mL	Error Validation mg/mL
ANN-S1	7.73×10^{-5}	1.97×10^{-3}	8.79×10^{-3}	4.44×10^{-2}
ANN-S2	3.48×10^{-5}	5.78×10^{-5}	5.90×10^{-3}	7.61×10^{-3}



(a) Sample 1



(b) Sample 2

Figure 9: Plots of Sample product concentration against time points. The pink curve refers to the data set and the scattered points are the predicted average product concentration for each stage

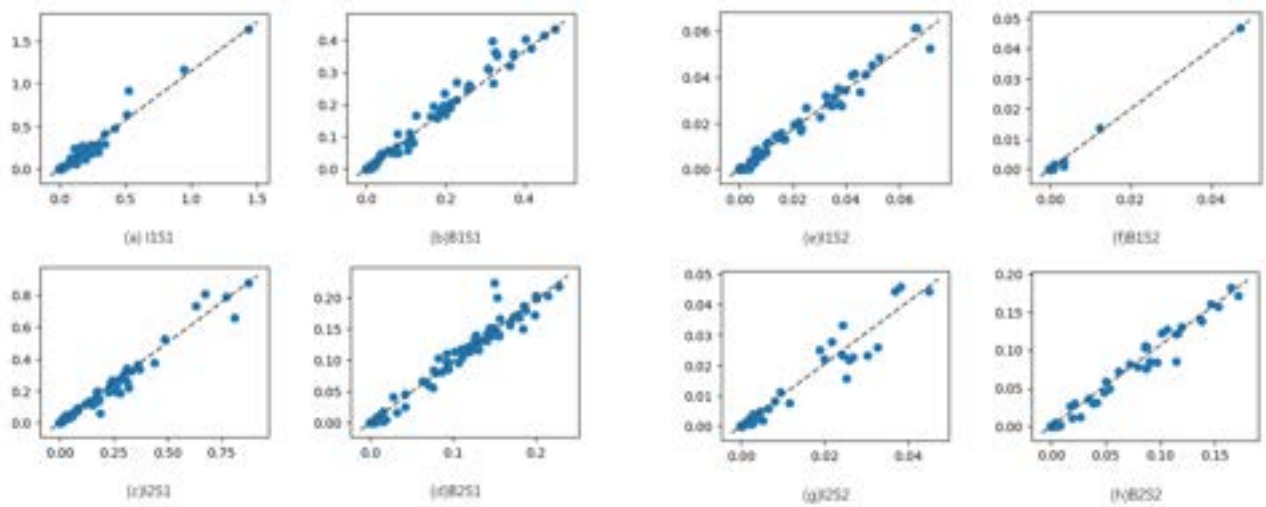


Figure 10: Product concentration plot. The vertical axis represents the predicted product concentration, and the horizontal axis represents the data. The dashed lines are the indicator where the predictions were more accurate when approaching the line.

To further validate the performance, the pseudo-dynamic model was tested against the input parameters from the 100 sample test data to predict the product concentration profile. Figure 10(a)-(h) are plots of the predicted result against the simulated data. The prediction from the pseudo-dynamic model being worse than that of the static model was due to a wider spread of average concentration compared to either the purity or yield that both ranged from 0 to 1. In addition, the compromise made, which implied as only 2 ANNs were designed for the pseudo-dynamic model other than designing in total 8 ANN for each stage, was to save the computational efforts. The results suggested that larger errors could be seen for the ANN predicting switch 2 where the targeted column is performing feeding and recycling tasks. This was considered acceptable since for most samples that met the purity and yield requirement, the product component was not expected to be eluted out during switch 2.

5 Conclusion

The purpose of this work was to create a black box model that could predict the purity, yield, and product concentration at CSS given the input conditions listed in table 1 above with saving computational time as a priority. Although the static model and the dynamic model required cross-validation from the mechanistic model simulations, both models would still provide valuable preliminary insights into the process given the input conditions that were not generated with the mechanistic model simulations. Since the responses were highly non-linearly related, simple interpolations between data points were thus invalid while the black box model could predict more accurately. This would help to eliminate most operating parameters that would yield low purity at first using the static model. The pseudo-dynamic model will then help to eliminate the samples where the concentration profile is not considered optimal.

There are still some limitations to the black box model. The model is only applicable to the current operating conditions such as concentration gradients of the modifier, maximum flow rate during the batch states, and switch times. It is also only applicable to the current design parameters such as the column length-to-diameter ratio and void ratio of the column. Arguably still, the model can be retrained to other setups that may be commonly in use in the industry.

The greatest advantage of the black box model compared to the mechanistic model is the computing time. To generate a single sample data set, about 60s or more will be required. In contrast, the black box model would take less than 0.1s for estimating outputs for 100 samples.

Other than the black box model, the hybrid model combing features from both the mechanistic and data-driven models has also been explored by some researchers. (H. Narayanan, 2021) However, the best modelling approach that guarantees satisfactory predicting accuracy with less computation as a priority remained unclear and challenging that requires further research.

Acknowledgements

The authors of this report would like to thank Michalopoulou Foteini for her continued support and guidance throughout the duration of this project, especially in providing the requested experimental data for the project. Special thanks to Sachio Steven for his valuable insight on this project.

References

Aumann, L. & Morbidelli, M. (2007) A continuous multicolumn countercurrent solvent gradient purification (MCSGP) process. Biotechnology and Bioengineering. 98 (5), 1043–1055. doi:10.1002/bit.21527.

Aumann, L. & Morbidelli, M. (2008) A semicontinuous 3-column countercurrent solvent gradient purification (MCSGP) process. Biotechnology and Bioengineering. 99 (3), 728–733. doi:10.1002/bit.21585.

Farid, S.S. (2007) Process economics of industrial monoclonal antibody manufacture. Journal of Chromatography B. 848 (1), 8–18. doi:10.1016/j.jchromb.2006.07.037.

Kim, J.K., Zang, Y. & Wankat, P.C. (2003) Single-Cascade Simulated Moving Bed Systems for the Separation of Ternary Mixtures. Industrial & Engineering Chemistry Research. 42 (20), 4849–4860. doi:10.1021/ie030373j.

Kingma, D.P. & Ba, J. (2014) Adam: A Method for Stochastic Optimization. doi:10.48550/ARXIV.1412.6980.

Krättli, M., Steinebach, F. & Morbidelli, M. (2013) Online control of the twin-column countercurrent solvent gradient process for biochromatography. Journal of Chromatography A. 1293, 51–59. doi:10.1016/j.chroma.2013.03.069.

Minceva, M., Pais, L.S. & Rodrigues, A.E. (2003) Cyclic steady state of simulated moving bed processes for enantiomers separation. Chemical Engineering and Processing: Process Intensification. 42 (2), 93–104. doi:10.1016/S0255-2701(02)00038-7.

Müller-Späth, T., Aumann, L., Melter, L., Ströhlein, G. & Morbidelli, M. (2008) Chromatographic separation of three monoclonal antibody variants using multicolumn countercurrent solvent gradient purification (MCSGP). Biotechnology and Bioengineering. 100 (6), 1166–1177. doi:10.1002/bit.21843.

Narayanan, H., Seidler, T., Luna, M.F., Sokolov, M., Morbidelli, M. & Butté, A. (2021) Hybrid Models for the simulation and prediction of chromatographic processes for protein capture. Journal of Chromatography A. 1650, 462248. doi:10.1016/j.chroma.2021.462248.

Papathanasiou, M.M., Avraamidou, S., Oberdieck, R., Mantalaris, A., Steinebach, F., Morbidelli, M., Mueller-Spaeth, T. & Pistikopoulos, E.N. (2016) Advanced control strategies for the multicolumn countercurrent solvent gradient purification process. AIChE Journal. 62 (7), 2341–2357. doi:10.1002/aic.15203.

Ströhlein, G., Aumann, L., Mazzotti, M. & Morbidelli, M. (2006) A continuous, counter-current multi-column chromatographic process incorporating modifier gradients for ternary separations. Journal of Chromatography A. 1126 (1), 338–346. doi:10.1016/j.chroma.2006.05.011.

Tabora, J.E. (2012) Data Driven Modelling and Control of Batch Processes in the Pharmaceutical Industry. IFAC Proceedings Volumes. 45 (15), 708–714. doi:10.3182/20120710-4-SG-2026.00204.

Torphy, T.J. (2002) Monoclonal antibodies: boundless potential, daunting challenges. Current Opinion in Biotechnology. 13 (6), 589–591. doi:10.1016/S0958-1669(02)00385-3.

Werner, R.G. (2004) Economic aspects of commercial manufacture of biopharmaceuticals. Journal of Biotechnology. 113 (1), 171–182. doi:10.1016/j.jbiotec.2004.04.036.

Imperial College London

A New MIQCQP Approach to Symbolic Polynomial Regression

William Baker and Mohammod Halim
Department of Chemical Engineering, Imperial College London, U.K.

Key Words:
Symbolic Regression
MIQCQP
Global Optimisation
Model Identification
Polynomial regression
Machine Learning

Abstract:
Algorithmic searches for physical models have become increasingly popular in recent years, not least in chemical engineering where recent advances have supported the wider adoption of machine learning techniques. Existing parametric and non-parametric approaches such as sparse and symbolic regression respectively, often suffer from fundamental issues however, with overcomplexity and misspecification affecting the former and high computational complexity the latter. This paper presents a new mixed integer quadratically constrained quadratic programme (MIQCQP), which facilitates symbolic multivariate polynomial regression, seeking to strike a balance between these two methodologies. To this end, a new formulation is outlined and the effects of implementing a series of additional symmetry cuts, to help reduce computational times, are explored. The performance of this formulation is then assessed and compared against a tailored form of an existing symbolic regression formulation. Applications and possible expansions of the formulation are also briefly discussed laying the foundations for possible future work. Performance analysis revealed the new formulation to be effective at producing surrogate models to accurately describe non-linear behaviour, outperforming the tailored symbolic regression regarding both computational times and accuracy. The new formulation was also successfully applied to various chemical engineering examples surrounding non-linear dynamical systems and thermodynamic modelling, showing great potential for expansion and improvement. A major limitation of the formulation however, is its high degree of computational complexity, compared to alternative parametric techniques, currently limiting its application to smaller, less complex data sets.

1. Introduction

Machine learning (ML) algorithms based on surrogate modelling approaches have become increasingly popular across chemical engineering in recent years [1]. Advances in both computational processing power and process automation have fostered the union between machine learning and automated systems in many areas relating to chemical engineering including process development [2], reaction modelling [3] and process optimization [4]. Such applications have highlighted the increasing need for new ML algorithms to provide accurate and robust models for data prediction with such models forming an integral part of system automation [1]. As such, the past decade has seen heightened interest in areas relating to algorithmic searches for physical models, with important and relatively recent advances made surrounding both parametric [5] and non-parametric regression techniques [1,6].

For instance, concerning parametric regression, where potentially non-linear behaviour is described using either a linear combination of specified basis functions or with the help of existing knowledge of the behaviour's functional form [1], the ALAMO approach has emerged as a promising methodology [5,7,8]. Examples found in literature, demonstrate the ability of ALAMO to produce accurate surrogate models in as few terms as possible [7], with more recent developments enabling physical knowledge of a system to be transferred to ALAMO to improve its modelling performance [8]. Wider applications of such sparse techniques are abundant in the literature surrounding parametric regression, seeing successful applications in areas such as model selection for hybrid dynamical systems [9], data driven identification of Navier-Stokes equations [10] and activity estimation in spectrometry [11]. Despite recent developments however, drawbacks associated with parametric approaches often come from their high dependency on existing knowledge of the system being analysed. This exposes the methodology to issues such as misspecification alongside other underlying issues such as overfitting and overcomplexity.

Alternatively, non-parametric regression techniques, which require no knowledge of an underlying functional form or specification of basis functions [1], allow for the flexible formulation of free-form equations, removing some of the limitations imposed by parametric techniques. One method for achieving this is symbolic regression (SR), a technique first proposed as a ML method in 2007 by Bongard and Lipson [12] to find valid and useful free-form models capable of accurately making data predictions. SR is commonly facilitated utilising Genetic Programming (GP). This approach is so common in fact, that the term 'symbolic regression' is often used synonymously with 'genetic programming' [6]. Genetic algorithms can often obtain good solutions although there is never a guarantee of global optimality [6]. Additional identified issues with this approach include poor accuracy [13] and restricted application to the discovery of simple functions and to data sets with few input variables and data points [6]. This reality has seen SR lose popularity compared to alternative deterministic parametric techniques which are subsequently used much more extensively in practise [9-11].

In response to some of these issues, an alternative approach towards SR was later proposed in 2018 by

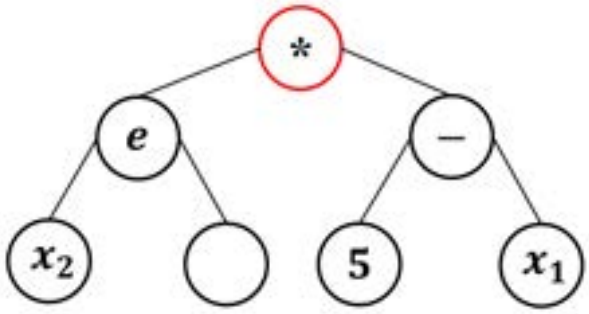
Cozad [6], where mixed integer nonlinear programming (MINLP) was utilised to reformulate SR as a nonlinear, nonconvex, disjunctive program that could be solved to global optimality. Such an approach helped to improve the reliability of SR for inferring behavioural information about a given system. Further improvements to the formulation presented in [6] have been explored in subsequent years. In 2020 for example, improvements were proposed by Neumann [1], where an alternative globally optimal SR formulation was successfully implemented and applied to several chemical engineering examples to accurately identify physical models which correctly represent physical phenomena. More recently, formulations leveraging derivative information to propose suitable functional forms was developed and its advantages through application to determining thermodynamic equations of state were demonstrated [14]. In many of these cases however, the computationally intensive nature of SR, due its high degree of combinatorial complexity, is identified as a limiting factor for its application, limiting its use to the analysis of, at most, a few hundred data points [1,6]. Tackling this issue and expanding the applicability of SR to more complex systems remains a challenge for current research.

This paper discusses the formulation of a mixed integer quadratically constrained quadratic programme (MIQCQP) to facilitate regression via a multivariate polynomial. We consider the problem of regressing a set of input-output data, $(X_{1,d}, \dots, X_{N,d}, Y_d)_{1 \le d \le D}$ by a multi-variate polynomial $y = a_0 + \sum_{m=1}^{M} a_m T_m(x_1, \dots, x_N)$ with $T_m(x_1, \dots, x_N) \coloneqq x_1^{\alpha_{m,1}} \dots x_N^{\alpha_{m,N}}$. In addition to determining the values of the regression coefficients $a_0, \dots, a_M$, the aim is to determine the structure of the regression model in terms of its monomial terms $T_1, \dots, T_M$ through symbolic regression. A polynomial functional form was selected due to its robust ability to describe a wide range of non-linear behaviours accurately [15], a fact that has seen polynomial regression used extensively across science and engineering [16-18]. The resulting formulation hopes to strike a balance between sparse and SR techniques to tackle many of the issues associated with each, namely through a reduction in computational complexity compared to existing SR formulations and an improvement in versatility over existing sparse regression techniques.

The remainder of this paper is organised as follows: An initial background surrounding SR is presented in section 2 before details of the newly proposed MIQCQP formulation are provided in section 3.1. Information concerning the formulation of a tailored MINLP SR, used for comparison, is then presented in section 3.2. Section 3.3 details the test instances considered as well as the test setup. The results of the computational testing are then outlined and discussed in section 4, before efforts to expand and apply the proposed formulation to several chemical engineering examples are presented in section 5. Lastly, final conclusions and opportunities for future research are presented in section 6.

2. Symbolic Regression Overview

Symbolic regression works under the premise that any mathematical expression can be represented by its own expression tree (figure 1). The order of the operations is encoded by the expression tree using operands at leaf nodes and operators everywhere else. The expression tree can then be evaluated by recursively applying a set of selected operators starting at the root node. Under this methodology changes that occur higher up the expression tree i.e., at the root node can often have a significant impact on the final model. The formulated free-form equations can provide insight into the underlying behaviour of physical systems [6] and can subsequently be used to make accurate data predictions.



$$f(x_1, x_2) = (5 - x_1)e^{x_2}$$

Figure 1: An example of a typical SR expression tree

As previously mentioned however, a significant drawback of such an approach lies with its high combinatorial complexity. The large number of feasible expressions obtainable through this method, thanks to the numerous available functional forms, translates into significantly long computational times subsequently hindering the applicability of SR.

3. Methodology

3.1. MIQCQP Approach

3.1.1. Formulation

The proposed MIQCQP formulation is based off the recursive formulation of monomial terms (m) in which higher order terms can be decomposed into a product of terms found earlier on in the solution (figure 2). The complexity of the model is controlled by varying the maximum number of monomials to be included in the final model (M) along with the total number of quadratic terms for decomposing these monomials (Q). M and Q are selected such that $M \le N + Q$ where N is equal to the total number of input variables. The aim of this approach is to utilise symbolic regression to construct the relevant monomials before linearly combining them to produce the final surrogate model. Such a method hopes to be less computationally complex compared to previous SR techniques, as presented in [6], by restricting the functional form to solely polynomial. The formulation also aims to simultaneously be more robust at describing complex non-linear behaviour than existing sparse regression techniques, as seen in [5], by removing the need to specify fixed basis functions.

Given some set of data, for data points $d = 1, \dots, D$ and input variables $n = 1, \dots, N$, $(X_{1,d}, \dots, X_{N,d}, Y_d)_{1 \le d \le D}$ we

propose the following MIQCQP formulation to perform symbolic regression in the least squares sense Eq. (1):

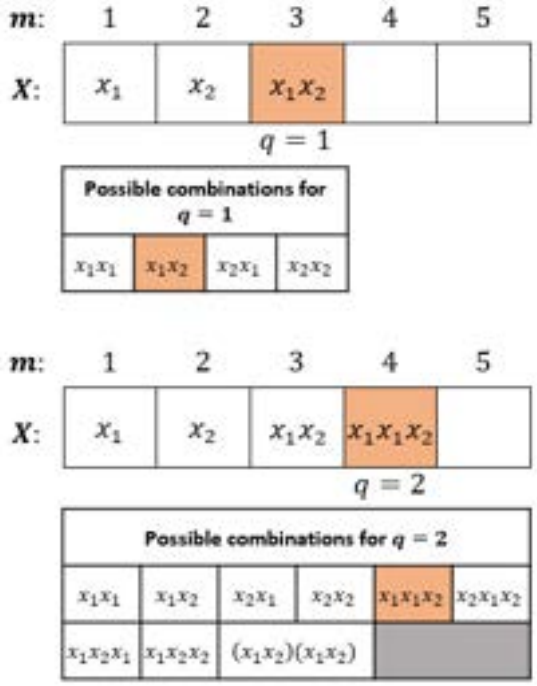


Figure 2: Schematic overview of MIQCQP formulation for input variables x_1, x_2.

Table 1: MIQCQP notation: indices

Description	*Index*	*Range*
Left hand operand position for monomial decomposition	i	$1, \dots, N+Q-1$
Right hand operand position for monomial decomposition	j	$1, \dots, N+Q-1$
Data points	d	$1, \dots, D$
Monomials	m	$1, \dots, M$
Quadratic terms	q	$1, \dots, Q$
Input variables	n	$1, \dots, N$

Table 2: MIQCQP notation: parameters

Description	*Parameter*
Maximum number of monomials	M
Bound magnitude	B
Response values at data point d	Y_d
Value of input variable n at data point d	$X_{n,d}$

Table 3: MIQCQP notation: variables

Description	*Type*	*Variable*
Initial regression constant	Continuous	a_0
Monomial coefficients	Continuous	a_m
Quadratic term values	Continuous	$X_{N+q,d}$
Monomial selection	Binary	z_m
Left-hand term selection	Binary	$\omega_{q,i}^L$
Right-hand term selection	Binary	$\omega_{q,j}^R$

$$\min \sum_{d=1}^{D} \left(Y_d - a_0 - \sum_{m=1}^{N+Q} a_m X_{m,d} \right)^2 \qquad (1)$$

s.t.

$$\sum_{m=1}^{N+Q} z_m \leq M \qquad (2)$$

$$-z_m B \leq a_m \leq z_m B, \qquad \forall m = 1 \dots N+Q \qquad (3)$$

$$\sum_{i=1}^{N+q-1} \omega_{q,i}^L = \sum_{j=1}^{N+q-1} \omega_{q,j}^R = 1, \qquad \forall q = 1 \dots Q \qquad (4)$$

$$| X_{N+q,d} - X_{i,d} \cdot X_{j,d} | \leq \left(2 - \omega_{q,i}^L - \omega_{q,j}^R\right) B, \qquad \forall i,j = 1 \dots N+q-1, \; \forall q = 1 \dots Q, \forall d = 1 \dots D \qquad (5)$$

$$z_m, \omega_{q,i}^L, \omega_{q,j}^R \in \{0,1\} \qquad \forall m = 1 \dots N+Q, \; \forall q = 1 \dots Q, \; \forall i,j = 1 \dots N+q-1 \qquad (6)$$

$$a_m, X_{N+q,d} \in \mathbb{R} \qquad \forall m = 1 \dots N+Q, \; \forall q = 1 \dots Q, \; \forall d = 1 \dots D \qquad (7)$$

In the above formulation binary variables $z_m, m = 1, \dots, N+Q$ control which monomials are to be included within the final regression model. These variables are restricted by a cardinality constraint Eq. (2), which limits the total number of selected monomials to upper bound M. Similarly, continuous variables $a_m, m = 1, \dots, N+Q$ are controlled by Eq. (3) in which regression coefficients for non-selected monomials are forced to zero or bound between some sufficiently large constant ± B otherwise. Binary variables $\omega_{q,i}^L, \omega_{q,j}^R$ encode the decomposition of auxiliary variables $x_{N+q}, q = 1, \dots, Q$ into the product of two terms x_i, x_j for $i,j = 1, \dots, N+q-1$. For each term to be decomposed exactly two terms are needed, as enforced by Eq. (4). Finally, Eq. (5) enforces the identity of term X_{N+q} as the product of selected terms X_i, X_j as defined by binary variables $\omega_{q,i}^L, \omega_{q,j}^R$. Eq. (5) can equivalently be reformulated as two separate constraints in which the LHS is defined as ≤ B and ≥ -B respectively although for conciseness the absolute value of the LHS has been used above.

The proposed algorithm was initialised as a sparse linear regression model by forcing binary variables $z_{N+1}, \dots, z_{N+Q}$ to zero at the beginning.

3.1.2. Illustrative example

Consider input data $x_{n,d}, n = 1,2, d = 1, \dots, 10$, for a series of randomly selected data points $1 \leq x_{n,d} \leq 10$, and response data given by the function:

$$f(x_1, x_2) = 7 + x_1 + 1.2 x_1^2 x_2 \qquad (8)$$

The above MIQCQP formulation can be used to predict this function exactly due its multivariate polynomial nature. For the following example Q and M were selected to be 2 and 3 respectively with B set arbitrarily to 10,000.

Solution:

$$f = 7 + x_1 + (0)x_2 + (0)x_1^2 + 1.2x_1^2x_2 \qquad (9)$$

The resulting regression model Eq. (9) was obtained in 1.55 seconds using the global solver BARON [19]. It can be seen that to produce term three from Eq. (8) the regression must first produce an intermediate term x_1^2 which does not appear in the resulting regression model thanks to $z_3, a_3 = 0$.

Alternatively, if Q = 3 and M = 5 we obtain the following solution after 2.61 seconds.

$$f = 7 + x_1 + (0)x_2 + 9236x_1^2 - 9236x_1^2 + 1.2x_1^2x_2 \qquad (10)$$

In Eq. (10) the allowance for the use of superfluous terms results in redundant terms m = 3 and m = 4. The effects of this may be realised in the slightly longer computational time that is required.

3.1.3. Strengthening the formulation

To avoid instances such as Eq. (10). It becomes necessary to include additional constraints to remove symmetry from the problem and in theory reduce computational times by reducing the size of the search space. Below we present several possible instances in which a single solution can be presented in a multitude of equivalent forms before presenting a series of additional constraints to prevent such superfluous instances from occurring.

Consider the function:

$$f(x_1, x_2) = 7 + x_1 + x_2x_1 + 1.2x_1^2x_2 \qquad (11)$$

Eq. (11) can equally be expressed as:

$$f(x_1, x_2) = 7 + x_1 + x_1x_2 + 1.2x_2x_1^2 \qquad (12)$$

$$f(x_1, x_2) = 7 + x_1 + 4x_2x_1 - 3x_2x_1 + 1.2x_1^2x_2 \qquad (13)$$

$$f(x_1, x_2) = 7 + x_1 + 1.2x_1^2x_2 + x_2x_1 \qquad (14)$$

Eq. (12) presents a case in which *intra-term* symmetry provides an equivalent solution as $x_1x_2 = x_2x_1$. Eq. (13) presents an instance in which the use of *repeated terms* can be used to provide an equivalent solution as $x_2x_1 = 4x_2x_1 - 3x_2x_1$. Lastly, Eq. (14) demonstrates an instance in which *inter-term* symmetry provides an equivalent solution as $x_2x_1 + 1.2x_1^2x_2 = 1.2x_1^2x_2 + x_2x_1$. In an attempt to remove Eq. (12-14) as feasible solutions we enforce the following additional constraints:

$$\sum_{q=1}^{Q} \omega_{q,i}^{L}\omega_{q,j}^{R} \leq 1\,, \qquad \forall i, j = 1 \dots N + q - 1 \qquad (15)$$

$$\omega_{q,i}^{L} \leq 1 - \sum_{j=i+1}^{N+q-1} \omega_{q,j}^{R}\,, \qquad \begin{matrix}\forall i \\ = 1 \dots N + q - 2 \\ \forall q = 1 \dots Q\end{matrix} \qquad (16)$$

$$\left|X_{N+q+1,d'}\right| - \left|X_{N+q,d'}\right| \geq 0, \qquad \begin{matrix}\forall q \\ = 1 \dots Q - 1\end{matrix} \qquad (17)$$

Eq. (15) performs a sum across all quadratic terms $q = 1, \dots, Q$ and stipulates that any combination of active binary variables can occur at most at one value of q. This removes symmetry from the problem by preventing the programme from producing multiple terms with the same configuration of $\omega_{q,i}^{L}, \omega_{q,j}^{R}$. Eq. (16) stipulates that for each quadratic term, any active left-hand binary variable can only ever be paired with an active right hand binary variable whose index j is less than or equal to the selected index i. For example, x_1x_2 would be invalid as $j > i$ but x_2x_1 and x_1x_1 would be valid as $j \leq i$ in both cases. This removes symmetry by preventing the programme from producing multiple equivalent terms. Eq. (17) stipulates that no term $N + q$ can be larger in magnitude than its subsequent term $N + q + 1$ at some data point d' with magnitude $d' \geq 1$. This removes symmetry from the problem by preventing the programme from coming up with multiple equivalent solutions in which only the order of terms is changed. It should be noted that for magnitudes of d' between 0-1, Eq. (17) would not be suitable. This is an important caveat that needs to be considered when selecting an appropriate d' and could necessitate data modification/scaling. For an alternative approach towards tackling both intra and inter term symmetry, not implemented in this study, see appendix B.

Adding these additional constraints could in fact increase computational times, despite the removal of symmetry from the problem, due to the addition of many constraints making the problem harder to solve. As such, the effects of adding each constraint were explored (section 3.3) to select the optimal configuration. The results of this exploration are presented in section 4.1.

3.2. Tailored SR Approach

For the purposes of assessing the relative performance of the MIQCQP formulation outlined above, a tailored MINLP symbolic regression, based off the formulation presented in [6], was constructed for comparison. For details concerning the full formulation see Appendix C.

As previously stated, the newly constructed SR has been tailored for the purposes of facilitating a better comparison with the MIQCQP formulation outlined in section 3.1. To this end, the binary operators made available to the model have been restricted to $\mathcal{B} = \{+, *\}$ and Unary operators restricted to $\mathcal{U} = \{cube\}$. Such restrictions enable the SR to construct solely multi-variate polynomial models, providing a consistent basis for comparison.

3.3. Testing

Each of the discussed formulations were implemented in GAMS version 36.2.0 and tested using an INTEL quad core i5-7500 CPU @ 3.4GHz.

An initial exploration of the effects of adding extra constraints Eq. (15-17) was conducted. To this end, every possible combination of the constraints was

implemented and used to determine an optimum surrogate model to a series of test functions utilising the global solver BARON [19]. Functions included a multivariate non-polynomial function: $f(x_1, x_2) = x_1 e^{0.1x_2}$ **(1)**, a univariate non-polynomial function: $f(x_1) = x_1 \ln x_1$ **(2)** and a multivariate polynomial function: $f(x_1, x_2) = 4 + 3x_1 + 2.1x_2 + 5x_2x_1 - x_1^2 - 7x_2x_1^2$ **(3)**. The CPU time taken to achieve this was recorded for a total of five runs per configuration with the intention of selecting the configuration averaging the fastest for future testing. Each initial run was carried out with Q = 3, N = 2, D = 10 and M = 5. Further testing was then carried out with D = 30 for some test functions to verify the results for larger data sets.

Following the successful determination of the optimal formulation configuration, a series of tests were conducted to assess the relative performance of the optimal MIQCQP formulation compared to the tailored SR discussed in section 3.2. To this end, a multivariate, non-linear, non-polynomial function: $f(x_1, x_2) = 2x_1 e^{-0.1x_2}$ was selected and both formulations were tasked to produce an optimum surrogate. The MIQCQP formulation was tested at various combinations of M, Q, N and D, whilst the tailored SR was tested at varying tree depths, n_{data} and n_{pred}. The performance of the two formulations was then analysed and compared. BARON [19] was selected as the global solver for both models to ensure a fair comparison of performance, although it should be noted that the MIQCQP formulation can also make use of state-of-the-art global solvers such as GUROBI version 10 [20] which may deliver further performance improvements compared to when using BARON [19]. For all cases, data points $x_{n,d}$ were randomly generated such that $1 \leq x_{n,d} \leq 10$.

4. Results and Discussion

4.1. Constraint Analysis

Results from figure 3 reveal that for both the univariate and multivariate non-polynomial functions the fastest configuration included all three additional constraints. When looking at the results for the multivariate polynomial however, the fastest configuration included only one constraint Eq. (17), followed closely by no constraints. From these results we see that when trying to determine an inexact surrogate model, including all three additional constraints can produce the fastest formulation whilst when trying to determine multi-variate polynomials (which can be determined exactly) the reverse appears to be true. It should be noted that the differences in compute times between configurations when trying to determine the multivariate polynomial often remained small for D=10, therefore the analysis was repeated for D=30 yielding comparable conclusions. More broadly such results demonstrate the effects that varying the function to be determined can have on the impact of added symmetry cuts, making it difficult to conclude which configuration would be universally optimum with no single configuration proving the best in all cases.

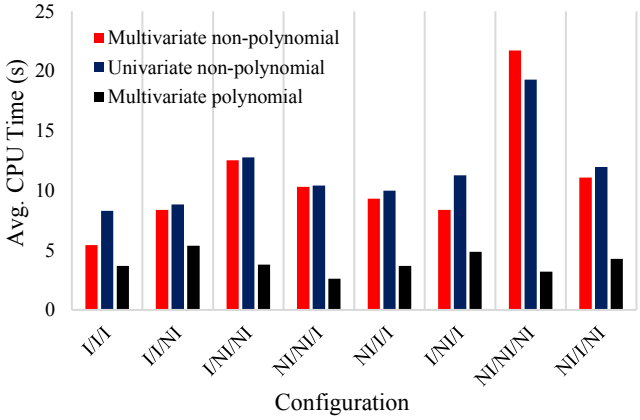
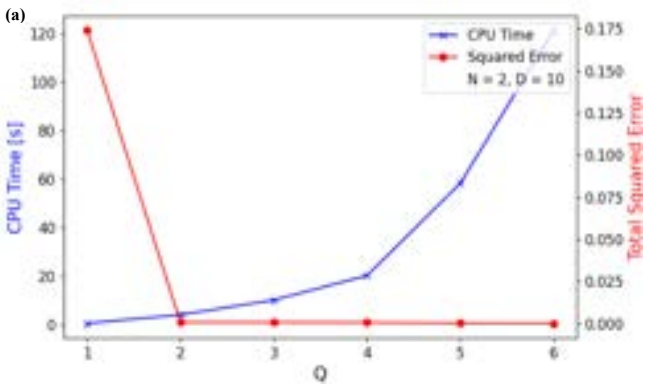
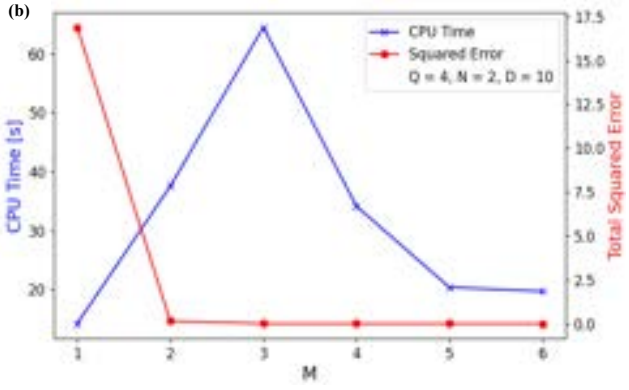


Figure 3: Constraint analysis for D = 10, Eq. (15)/ (16)/ (17): I = Included, NI = Not Included

The analysis for the multivariate non-polynomial function was also repeated for a larger data set (D=30) to explore if the effects of implementing the additional constraints would change as additional complexity was added into the problem. The results yielded similar conclusions with the fastest case being the one in which all additional constraints were included (20.70s). Including none resulted in the second slowest configuration (59.61s) after the case in which only Eq. (15,16) were included (72.02s). Here the effects of including various constraints become more pronounced with compute times varying more significantly. This subsequently enables more reliable conclusions to be obtained. For the purposes of further testing, a configuration in which all three additional constraints are included was selected primarily due to the observed reduced compute times when attempting to produce surrogate models for both D = 10 and D = 30. For fully tabulated results surrounding the constraint analysis and details concerning both the modelled functions and associated error see appendix D.

4.2. MIQCQP Performance Analysis





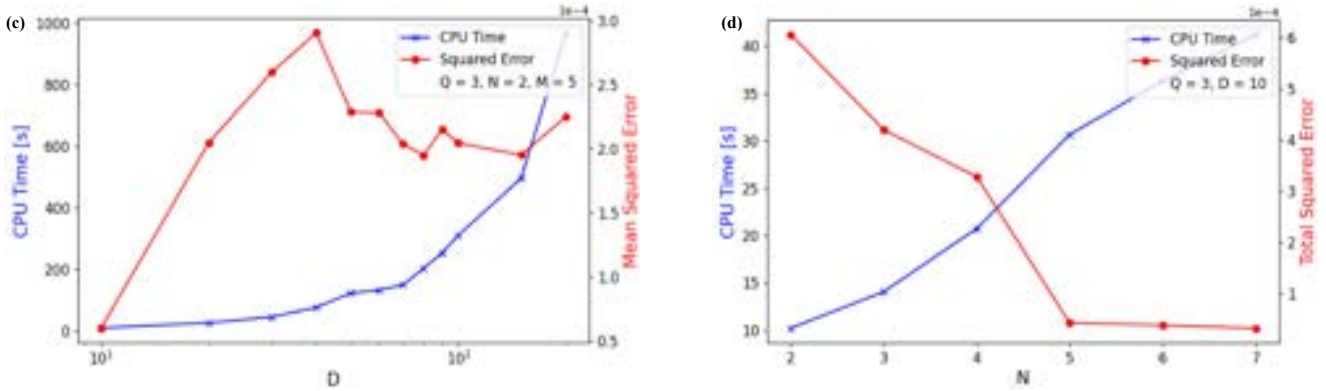


Figure 4: Scalability of proposed MIQCQP formulation as Q varies (a), M varies (b), D varies (c) and N varies (d).

Performance analysis of the MIQCQP formualtion reveals it to possess a high degree of computational complexity whilst also demonstrating its ability to produce accurate surrogate models to complex non-linear functions. For instance, figure 4a shows an exponential increase in compute time as the number of quadratic terms increases. Despite this however, the total squared error can be seen to reduce rapidly to near zero after Q=2 subsequently reducing the need to make use of additional terms and allowing CPU times to be kept well below 30 seconds. Selecting the optimum number of quadratic terms such to optimise error against compute times is therefore an important step when attempting to use this formualtion in practise.

Interestingly, figure 4b shows an initial increase in required compute times as M is increased (for fixed values of Q and N) before revealing a decrease in compute times for higher values of M. This could be due to the increased combinatorial complexity that optimal term selection introduces into the problem when M is low due to the numerous possibilities avaliable for which combination of terms to select. As M is increased to higher values, a greater proportion of N+Q terms are selected (to reduce error) thus reducing the amount of variation that term selection introduces into the problem and therefore resulting in lower compute times.

Figure 4c demonstrates how as the number of data points increases, the complexity increases exponentially, resulting in significangtly longer compute times, reaching nearly 1000 seconds by 200 data points. Furthermore, the mean square error (MSE) can be seen to be increasing with the number of data points, reaching a maximum of around 3.0E-4 at D=40 before decreasing and broadly leveling off. This makes sense as the MSE is an unbaised estimator of variance and as such one would expect to see the MSE converge towards the error variance. The low value of the MSE at D=200 highlights the ability of the formualtion to produce accurate models at larger data sets, even if compute times do increase to relatively high values.

Lastly, figure 4d reveals a steady increase in compute times as the total number of input variables increases, however even at N = 7, compute times are below 60 seconds. Figure 4d also shows how error decreases with the number of input variables. This can be explained by the fact that M was allowed to increase as N increased. By introducing additional input variables, the total number of terms avaliable to the model also increased. More specifically, the model can utilise a greater number of non-quadratic terms consisting solely of input varaibles $X_{n,d}, n = 1, \dots, N, d = 1, \dots, D$. For tabulated resultes of data presented in figure 4, see appendix E.

4.3. Tailored SR Performance Analysis

***Table 9: Tailored SR peformance analysis with varying tree depth*: n_{data}=10, n_{pred}=2.**

Tree Depth	*Nodes*	*Modelled Function*	*CPU Time (s)*	*Square Error*
2	3	$f(x_1) = 1.321x_1$	0.31	17.73
3	7	$f(x_1, x_2) = 5.363 + x_1 - 0.752x_2$	7.48	4.10
4	15	$f(x_1, x_2) = -0.013 + 1.895x_1 + 0.025x_2 - 0.137x_2x_1$	377.34	0.17
5	31	N/A	Time out	N/A

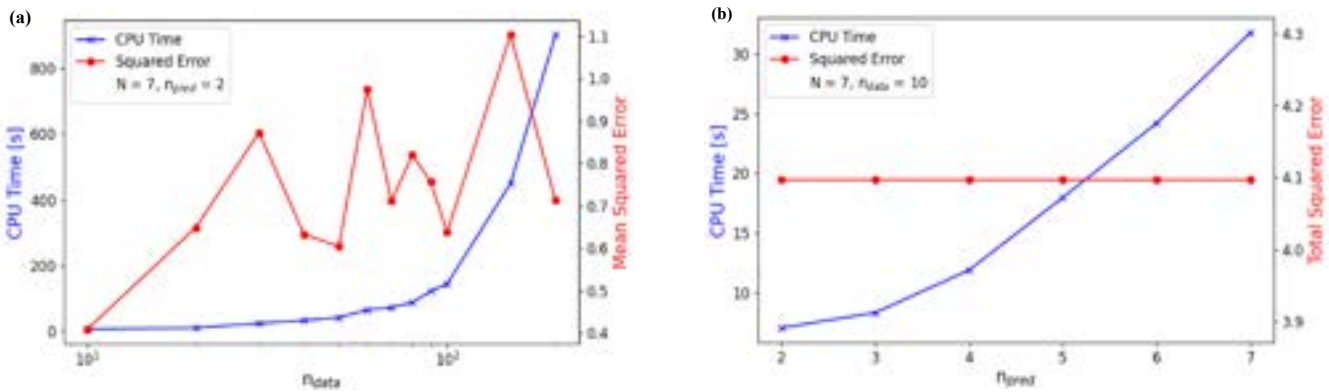


Figure 5:Scalability of tailored SR as n_{data} varies (a) and n_{pred} varies (b).

Table 9 demonstrates the expected significant computational complexity associated with previous SR formualtions as commented on in the literature [1,6]. As tree depth increases the compute time increases by several orders of magnitude until at a tree depth of five, the time out threshhold of 10,000s is reached yielding no final globally optimum solution. This is undesirable as at a tree depth of 4, the squared error still remains relatively high at 0.17 when restricting the SR to produce multivariate polynomails only.

Figure 5a demonstrates how as the number of data points increases the compute time increases signifcantly, especially after 100 data points. Furthermore, the MSE increases significantly as the number of data points increases highlighting the difficulty of producing an accurate surrogate model for large data sets at a tree depth of only 3. It can be predicted however (utilising observations from both table 9 and figure 5a), that if a larger tree depth was used, significantly large compute times and likley time outs would soon become an issue as the number of data points increased (as verified by futher testing).

Figure 5b shows a steady increase in compute time as the total number of input variables increases. For each number of input variables tested, identical models were produced resulting in the same error for all runs. Compared to figure 5a we also see how increasing the number of input variables has a less signifcant effect up to n_{pred} = 7 than when increasing the number of data points. For tabulated results of data presented in figure 5, see appendix F.

4.4. Performance Comparison

When increasing the total number of data points, both models scale similarly regarding compute times with the MIQCQP formulation scaling slightly better for larger data sets. The MIQCQP formulation with Q = 3 and M = 5 predictably runs slower than the less complex tailored SR with a tree depth of only 3, however it is able to achieve a significantly lower MSE for all data sets and, unlike the tailored SR, is successful at producing a relatively accurate surrogate model for D=200. Table 9 highlights that if the tailored SR were to be provided the required layers to yield comparable levels of accuracy to the MIQCQP for each data set (i.e. depth > 4), it would run significantly slower, a trend verified in [1,6]. Therefore, such a comparison reveals the improved capability of the MIQCQP to produce surrogate models of improved accuracy within more reasonable timeframes and thus we can conclude that the MIQCQP is superior to the tailored SR in this regard.

Concerning complexity, the MIQCQP formulation also provides increased flexibility over the tailored SR, with variations in Q/M enabling a finer balance of error with compute times compared to varying the tree depth for the tailored SR. In practise this would support quicker identification of optimal configurations in which error and compute times are sufficiently low (see figure 4a, Q =2).

As the number of input variables increases both formulations scale similarly regarding compute times with neither model requiring more than 60 seconds when 7 input variables are present within the data set. This is significant as the MIQCQP must also handle an increased number of terms as the number of input variables increases, facilitating a lower total squared error.

When comparing the relative performance of the two models however, it becomes apparent that both formulations suffer from high degrees of computational complexity when compared to alternative parametric techniques. In practise, such complexity limits both approaches to dealing with no more than a few hundred data points.

To summarise, comparing the relative performance of the two formulations reveals the MIQCQP to deliver improved performance compared to the tailored SR. The formulation can obtain more complex surrogate models of improved accuracy over the tailored SR and do so within more reasonable timeframes. It also provides increased flexibility when it comes to balancing error with compute times. Compute times for the MIQCQP do however scale similarly to the tailored SR with regards to both the number of data points and input variables. Consequently, the proposed MIQCQP formulation can produce usable surrogate models for data sets larger, and more complex than those suitable for tailored SR, however this will still incur long compute times.

5. Application and Expansion

This section explores the applicability and expandability of the proposed formulation. To this end, various attempts to apply the proposed formulation to practical examples found throughout chemical engineering and further expand upon its capabilities are discussed.

5.1. Application to Dynamical Systems

The proposed MIQCQP formulation could be used in practise across many different areas throughout chemical engineering. For example, it could be used to discover underlying equations of non-linear dynamical systems from data, a task recognised to have enabled the rapid development of knowledge and technology across many disciplines [21]. An existing framework where governing equations of non-linear dynamical systems are determined through sparse identification was presented in 2016 by Brunton, Proctor and Kutz [21]. In this section we attempt to use the proposed formulation as an alternative approach towards tackling this issue.

5.1.1. Height of liquid in a tank

Consider the simple scenario of a tank, cross-sectional area $A = 1m^2$, filling with water at a constant inlet flowrate $F_{in} = 20kgs^{-1}$and an outlet flowrate given by $F_{out} = \alpha\sqrt{h}$ where h is the height of liquid in the tank and α is the outlet flowrate coefficient. For this example, $\alpha = 10\ kgs^{-1}m^{-0.5}$. From first principles we know the differential of the height of liquid in the tank with respect to time is given by:

$$\rho A \frac{dh}{dt} = F_{in} - \alpha\sqrt{h} \tag{18}$$

However, let us assume that this was unknown and only data for the height of the liquid in the tank, h, at various times, t, was available (t = 0-200s, D = 20). Such data can easily be obtained in practise. Using this data, the central difference method can be applied to obtain an estimate for the differential at each point in time. By providing the MIQCQP formulation with all time-dependant variable data (h) and response data given by the approximated gradient, it can produce a surrogate model to Eq. (18). For example, at M=3, Q=2, N=1 and B =10,000 the formulation produces a globally optimum surrogate in 0.50 seconds:

$$dh/dt = A + Bh + Ch^2 + Dh^3 \quad (19)$$

A = 0.0151, B = -0.00596, C=0.000801, D=-0.0000642

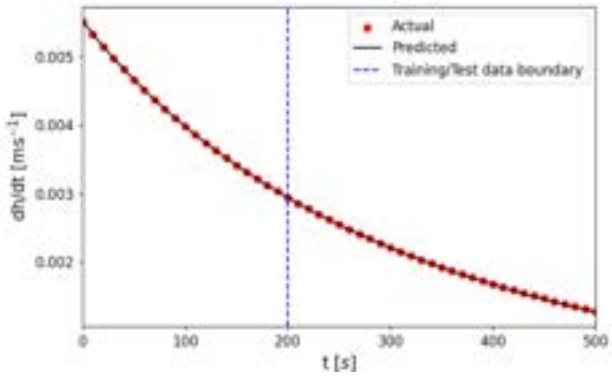


Figure 6: Actual and predicted dh/dt Vs time

Figure 6 demonstrates the ability of the surrogate model to accurately predict the value of the differential at any time t, both within the range of training data (t=0-200s) and beyond (t=200-500s).

5.1.2. Molar holdup in a reacting system

This methodology can easily be expanded to more complex systems in which there are multiple differential equations which are each a function of various time dependant variables. This can be achieved by repeating the process above for each differential in parallel or by modifying the formulation to perform multi-input/multi-output regression. For instance, consider the case of a well-stirred, isothermal reactor in which the following reversible reaction is taking place:

$$CH_3OH + CH_3COOH \leftrightarrow CH_3COOCH_3 + H_2O \quad (20)$$

In this example we want to determine the differential with respect to time of the molar holdup of each component in the reactor. From first principles we know this to be given by:

$$\frac{dN_i}{dt} = F_{in}x_{in,i} - F_{out}x_i + N_T \sum_{j \in REAC} \nu_{ij} r_j, \quad i \in COMP \quad (21)$$

Where $F_{out}, x_i, N_T, r_1, r_2$ all vary in time. Once again, assuming this was unknown and that all that was available was information regarding the molar holdup of each component at time t (t=0-50,000s, D=10, N=4), the same method as used previously can be applied for each component in parallel. The MIQCQP formulation can then once again be utilised to produce a surrogate to Eq. (21). For example, at M = 5, Q=1, N =4 and B = 100,000 the formulation produces a surrogate for each component in 3.81 seconds, 2.57 seconds, 2.16 seconds and 2.15 seconds respectively (figure 7).

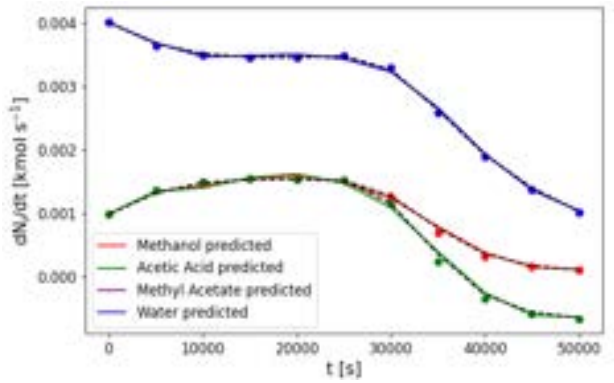


Figure 7: Actual (points), central difference (dotted lines) and predicted (solid lines) dN_i/dt Vs time for methanol, acetic acid, methyl-acetate and water (NB. water and methyl acetate overlap exactly)

Figure 7 demonstrates the applicability of the proposed formulation to more complex dynamical systems, being capable of producing suitably accurate surrogates for all components. For the same reactor, time differentials for outlet flowrates, pressure, total molar holdup etc, could also be determined using this method. It should be noted that some of the discrepancy seen between the actual and predicted values comes from the difference between the actual gradient and the gradient fed to the MIQCQP formulation determined via the central difference method. This can be seen by observing the improved fit against the central difference determined gradient. As such altering this element of the methodology could yield improved results when making data predictions.

5.2. Expansion to Produce Rational Functions

By building upon the same basic methodology (figure 2), the proposed MIQCQP formulation can be expanded such that it is able to produce a more diverse set of models capable of capturing a wider range of behaviours. One possible approach towards achieving this is presented below, as a proof of concept, in which the formulation is modified to produce rational functions. An alternative approach is presented in appendix G where division is introduced for constructing each monomial.

5.2.1. Modified Formulation

The formulation can be expanded by modifying it to approximate behaviour through a multivariate generalised rational function. Previous studies have explored the use of this functional form for approximation and note its increased suitability over standard multivariate polynomials for approximating non-smooth and non-Lipschitz functions [22].

$$\min \sum_{d=1}^{D} (Y_d - R_d)^2 \quad (22)$$

$$\sum_{m=1}^{N+Q} z_{b,m} \leq M \quad (23)$$

$$-z_{b,m}B \le b_m \le z_{b,m}B, \qquad \forall m = 1 \dots N+Q \quad (24)$$

$$\sum_{i=1}^{N+q-1} \theta_{q,i}^{L} = \sum_{j=1}^{N+q-1} \theta_{q,j}^{R} = 1, \qquad \forall q = 1 \dots Q \quad (25)$$

$$| K_{N+q,d} - K_{i,d} \cdot K_{j,d} | \le \left(2 - \theta_{q,i}^{L} - \theta_{q,j}^{R}\right)B, \quad (26)$$
$$\forall i,j = 1 \dots N+q-1,$$
$$\forall q = 1 \dots Q, \forall d = 1 \dots D$$

$$R_d = a_0 + \sum_{m=1}^{N+Q} a_m X_{m,d} \Big/ b_0 + \sum_{m=1}^{N+Q} b_m K_{m,d} \quad (27)$$
$$\forall d = 1, \dots, D$$

$$z_{b,m}, \theta_{q,i}^{L}, \theta_{q,j}^{R} \in \{0,1\} \qquad \forall m = 1 \dots N+Q \quad (28)$$
$$\forall q = 1 \dots Q$$
$$\forall i,j = 1 \dots N+q-1$$

$$b_m, K_{N+q,d} \in \mathbb{R} \qquad \forall m = 1 \dots N+Q \quad (29)$$
$$\forall q = 1 \dots Q$$
$$\forall d = 1 \dots D$$

We modify the formulation presented in section 3.1 by including Eq. (23-29) alongside Eq. (2-7) as well as modifying Eq. (1) to Eq. (22). Eq. (2-7) constructs the original polynomial $(a_m, X_{m,d})$ whilst Eq. (23-29) constructs a second polynomial $(b_m, K_{m,d})$ in parallel. The final rational function used for approximation $R(x_1, x_2)$ is then constructed through Eq. (27). Both polynomials are initialised as sparse linear regressions. Furthermore, to remove symmetry from the problem, b_0 is fixed to 1 to not only avoid division by zero but to prevent equivalent solutions from being formulated in which the numerator and denominator are both multiplied by some scale factor. The complexity of both polynomials is regulated using the same M, Q and B.

5.2.2. Illustrative example

Consider a set of input temperatures (T_2/K) and corresponding vapour pressures (P_2/atm) for water as defined by the Clausius Clapeyron Equation:

$$P_2 = P_1 e^{-\frac{\Delta H_{vap}}{R}\left(\frac{1}{T_2} - \frac{1}{T_1}\right)} \quad (30)$$

For water $\Delta H_{vap} = 40.8\, Kjmol^{-1}$. We also know that at 1atm the vapour pressure of water is 373K, enabling us to define P_1, T_1. Eq. (30) can thus be modelled with the expanded formulation (M=4, Q = 3) using BARON [19] in 15.94 seconds as:

$$P_2 = \frac{A + BT_2 + CT_2^2}{1 + DT_2 + ET_2^2 + FT_2^3} \quad (31)$$

For completeness, and to provide a broad comparison between the proposed formulation and alternative sparse techniques, the resulting surrogate Eq. (31) was compared against a sparse regression in which the basis functions $[1, T_2, T_2^2, T_2^3]$ were provided with coefficients [G,H,I,J] respectively. The resulting regression completed in 0.33 seconds using the BARON solver [19]. Due to the magnitude of the resulting coefficients (A-J) it is important that they are implemented with sufficient accuracy to ensure both models remain good surrogates to Eq. (30). For detailed values of each coefficient, see figure 8.

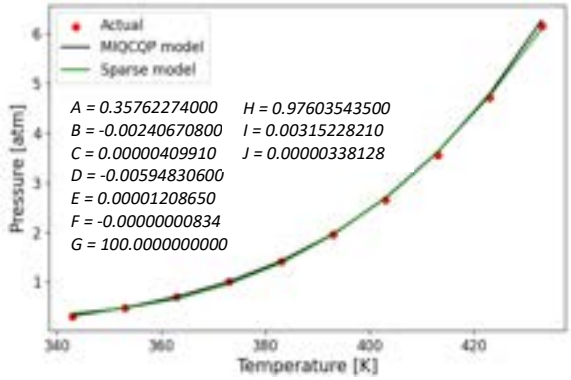


Figure 8: Modelled vapour pressure for water at varying temperatures compared with vapour pressure determined via the Clausius Clapeyron equation.

Figure 8 highlights the inherent strengths of the sparse regression with it being able to produce a surrogate of comparable accuracy to the expanded rational SR in significantly less time. Both approaches however are able to produce accurate surrogates the Clausius Clapeyron equation Eq. (30).

6. Conclusions

This report has outlined a new MIQCQP formulation, to perform symbolic multivariate polynomial regression, with the aim of reducing computational complexity compared to current SR techniques [1,6] whilst also increasing versatility compared to existing parametric techniques [5,7]. The performance of the new formulation (with included symmetry cuts) was assessed using BARON [19] and compared to a tailored version of an existing symbolic regression formulation [6], capable of producing solely multivariate polynomials. Examples where the formulation was then applied to various chemical engineering examples and further expanded upon were then briefly discussed laying the foundation for possible future work.

The new formulation was able to produce surrogate models to accurately describe non-linear behaviour. It was successfully applied to predict thermodynamic properties and as an alternative technique (to existing parametric frameworks [21]) to determine governing equations underlying non-linear dynamical systems. The MIQCQP formulation was also able to outperform the tailored SR, producing more complex and accurate surrogate models in more reasonable timeframes. Despite the MIQCQP formulation's improved performance over the traditional SR formulation [6] however, it does continue to suffer from a high degree of computational complexity compared to alternative parametric techniques. Such complexity resulted in compute times which rise exponentially with both the number of data points and the number of additional quadratic terms. This limits the formulation's ability to describe complex non-linear behaviour for larger data sets, where many terms are required, both accurately and more notably, within reasonable timeframes.

The scope of this paper has encompassed primarily the formulation, testing and comparison of a new MIQCQP formulation to carry out symbolic polynomial regression. Future work could continue to test this formulation by making use of alternative solvers not utilised in this paper (such as GUROBI version 10 [20]) and assessing whether further performance improvements can be obtained. Furthermore, testing the formulation against existing open access benchmarks could provide an alternative avenue to assess and compare its performance. Subsequent work could also explore additional means by which to expand upon the formulation further, for example, by considering multi-output regression and/or incorporating information criterion to control sparsity. Lastly, investigations could be carried out into the effectiveness of the formulation when applied to more complex real-world examples found throughout science and engineering with particular emphasis on how it handles noisy data sets.

References

[1] P Neumann, L Cao, D Russo, V.S Vassiliadis, A.A Lapkin, 2020, *A new formulation for symbolic regression to identify physico-chemical laws from experimental data*, Chemical Engineering Journal, vol. 387, https://doi.org/10.1016/j.cej.2019.123412

[2] T Bikmukhametov, J Jäschke, 2020, *Combining machine learning and process engineering physics towards enhanced accuracy and explainability of data-driven models*, Computers & Chemical Engineering, vol. 138, https://doi.org/10.1016/j.compchemeng.2020.106834

[3] T.I Madzhidov, et al., 2021, *Machine learning modelling of chemical reaction characteristics: yesterday, today, tomorrow*, Mendeleev Communications, vol. 31, https://doi.org/10.1016/j.mencom.2021.11.003

[4] A.D Clayton, et al., 2020, *Automated self-optimisation of multi-step reaction and separation processes using machine learning*, Chemical Engineering Journal, vol. 384, https://doi.org/10.1016/j.cej.2019.123340

[5] Z.T Wilson, N.V Sahinidis, 2017, *The ALAMO approach to machine learning*, Computers & Chemical Engineering, vol. 106, https://doi.org/10.1016/j.compchemeng.2017.02.010

[6] A Cozad, N.V Sahinidis, 2018, *A global MINLP approach to symbolic regression*, Mathematical Programming, vol. 170, https://doi.org/10.1007/s10107-018-1289-x

[7] A Cozad, N.V Sahinidis, D.C Miller, 2014, *Learning surrogate models for simulation-based optimisation, AIChE Journal*, vol. 60, https://doi.org/10.1002/aic.14418

[8] A Cozad, N.V Sahinidis, D.C Miller, 2015, *A combined first-principles and data-driven approach to model building*, Computers & Chemical Engineering, vol. 73, https://doi.org/10.1016/j.compchemeng.2014.11.010

[9] N.M Mangan, T. Askham, S.L Brunton, J.N Kutz, J.L Proctor, 2019, *Model selection for hybrid dynamical systems via sparse regression*, Proc. R. Soc. A, vol. 475, http://dx.doi.org/10.1098/rspa.2018.0534

[10] S.H Rudy, S.L Brunton, J.L Proctor, J.N Kutz, 2017, *Data-driven discovery of partial differential equations*, Science Advances, vol. 3, https://doi.org/10.1126/sciadv.1602614

[11] Y Sepulcre, T Trigano, Y Ritov, 2013, *Sparse Regression Algorithm for Activity Estimation in γ Spectrometry*, IEEE Transactions of Signal Processing, vol. 61, https://doi.org/10.1109/TSP.2013.2264811

[12] J Bongard, H Lipson, 2007, Automated reverse engineering of nonlinear dynamical systems, PNAS, vol. 104, https://doi.org/10.1073/pnas.0609476104

[13] M.F Korns, 2011, *Accuracy in Symbolic Regression*, Genetic Programming Theory and Practise IX, https://doi.org/10.1007/978-1-4614-1770-5_8

[14] M.R Engle, N.V Sahinidis, 2021, *Deterministic symbolic regression with derivative information: General methodology and application to equations of state,* AIChE Journal, vol. 68, https://doi.org/10.1002/aic.17457

[15] R.M Heiberger, E Neuwirth, 2009, *Polynomial Regression*, R Through Excel. Use R. https://doi.org/10.1007/978-1-4419-0052-4_11

[16] G Wang, J Cao, H Wang, M Guo, 2007, *Polynomial Regression for Data Gathering in Environmental Monitoring Applications*, IEEE Global Telecommunications Conference, https://doi.org/10.1109/GLOCOM.2007.251

[17] M Ekum, A Ogunsanya, 2020, *Application of Hierarchical Polynomial Regression Models to Predict Transmission of COVID-19 at Global Level*, Clinical Biostatistics and Biometrics, vol. 6, https://doi.org/10.23937/2469-5831/1510027

[18] M Najafzadeh, D.B Laucelli, A Zahiri, 2016, *Application of model tree and Evolutionary Polynomial Regression for evaluation of sediment transport in pipes,* KSCE Journal of Civil Engineering, vol. 21, https://doi.org/10.1007/s12205-016-1784-7

[19] A Khajavirad, N.V Sahinidis, 2018, *A hybrid LP/NLP paradigm for global optimization relaxations*, Mathematical Programming Computation, vol. 10, https://doi.org/10.1007/s12532-018-0138-5

[20] Gurobi Optimization LLC, 2022. *Gurobi optimizer reference manual, version 10*. https://www.gurobi.com/documentation/10.0/refman/index.html

[21] S.L Brunton, J.L Proctor, J.N Kutz, 2016, *Discovering governing equations from data by sparse identification of non-linear dynamical systems*, PNAS, vol. 113, https://doi.org/10.1073/pnas.1517384113

[22] R.D Millán, V Peiris, N Sukhorukova, J Ugon, 2021, *Multivariate approximation by polynomial and generalised rational functions*, Optimisation and Control, https://doi.org/10.48550/arXiv.2101.11786

Electrochemical Biosensor for Detection of Base Pair Mismatches in DNA Mutations

Bryan Tan[1] and Louis Chew[1]
[1]Department of Chemical Engineering, Imperial College London, U.K.

Abstract

Base pair mismatches present in DNA mutations can have adverse ramifications on human health, product quality, and experimental research analysis, amongst other fields. DNA biosensors can provide a low cost and rapid option for the detection of such base pair mismatches. The concept of an electrochemical biosensor utilising streptavidin-biotin interactions to bind DNA samples onto a modified screen-printed carbon electrode was explored in this study. The biosensor demonstrated successful immobilisation of biotinylated DNA strand onto the streptavidin-modified working electrode surface. Cyclic voltammetry revealed detectable differences in peak currents across double-stranded DNA of varying number of base pair mismatches. Detection of target DNA concentrations down to 1 μM, or 1 % of the immobilised biotinylated cDNA concentration, was achieved. Electrical readouts from the biosensor were easily replicable across biosensor chips. This study hence serves as a proof of concept for an innovative electrochemical biosensor for the detection of DNA base pair mismatches. The electrical biosensor signal output has the potential to be easily integrated into processes across industries, such as quality control of DNA-based products.

Keywords: *Biosensors, Cyclic voltammetry, DNA, Electrochemistry*

1. Introduction and Background

Beyond the Watson-Crick model for deoxyribonucleic acid (DNA), which gives complementary base pairs of guanine-cytosine and adenine-thymine, base pair mismatching in double-stranded DNA (dsDNA) may occur. In DNA replication, such mismatches may occur at a frequency of 1 in 10^7 base pairs per mitosis cycle [1]. These DNA mismatches can have adverse effects on human health and biotechnology, such as in gene mutations [2], pathogen infections [3], cancer [4], and product quality [5]. The detection of DNA mismatches continues to be of significant challenge. Current detection techniques include PCR-based methods (e.g. RT-PCR [6], ASB-PCR [7]) and gel electrophoresis methods (e.g. DGGE [8], TTGE [9]). However, these methods each have inherent limitations in terms of costs, technologies required, or accessibility for industrial uses. PCR-based methods generally have much higher associated costs and technological requirements, while the simpler electrophoresis methods offer much less information. There is hence a technological gap that can be addressed.

Recent developments in DNA biosensors based on nucleic acid hybridisation have received considerable attention due to their low cost and low technological barriers to adoption, while also providing rapid detection. Biosensors are analytical devices that combine the specificity of a biological sensing element with a transducer and converts them into signals [10]. Such DNA biosensors can be extremely useful in various industries, such as clinical diagnostics [11], biological research [12], food safety [13], and environmental monitoring [14].

Specifically, electrochemical biosensors will be investigated in this study due to the high sensitivity of electrochemical transducers. They have the potential to create a simpler and inexpensive detection method compared to

other assays such as spectrophotometry [15] and fluorescence based biosensors [16]. The electrochemical biosensor relies on differences in electrochemical characteristics arising from the different states of hybridisation between complementary single-stranded DNA (ssDNA) versus those with base pair mismatches [17]. The corresponding electrical signals will then be amplified, processed, quantified, and visualised on the display unit of the biosensor.

This proof-of-concept study aims to investigate the feasibility of developing an electrochemical biosensor to detect base pair mismatches in DNA. The electrochemical biosensor designed is based on a screen-printed carbon electrode (SPCE). The SPCE utilises a three-electrode configuration consisting of a working electrode, a counter electrode, and a reference electrode. The reference electrode controls the potential of the working electrode and minimises potential and current loss in the circuit [18]. Advantages of SPCEs include the low sample volume required, and the potential for modification of the electrode surface for specific analytes depending on analytical requirements [19]. Problems with classical solid electrodes, such as memory effects and tedious cleaning processes, can also be avoided.

Selective immobilisation of the ssDNA probe is crucial as direct adsorption of ssDNA onto the electrode surface will lead to poor hybridisation efficiency [20]. This would compromise the reproducibility and sensitivity of the biosensor. The proposed biosensor will leverage on streptavidin-biotin interactions to attach biotinylated probes on the SPCE for improved detection sensitivity. The streptavidin-biotin binding is one of the strongest known non-covalent bonds occurring in nature [21], with one streptavidin protein having the ability to bind with four biotin molecules with high affinity and selectivity [22]. Streptavidin-modified SPCEs would thus be used against biotinylated DNA sequences in this study.

The SPCE biosensor chip is evaluated using cyclic voltammetry (CV) to investigate the electrochemical behaviour on the electrode surface. CV is one of the most common, straightforward, and efficient method for obtaining qualitative and quantitative information on biological and redox reactions [23]. The working principle of CV can be explained with reference to the sample cyclic voltammogram output given in Figure 1.

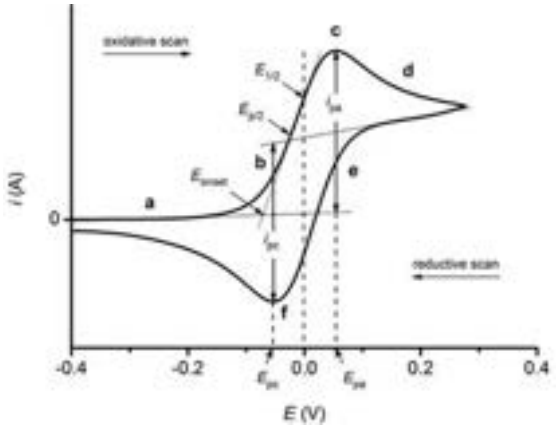


Figure 1. *Example of a cyclic voltammogram [24].*

The CV involves scanning across a range of potentials, with the forward direction indicating an oxidative scan while the backward direction represents a reductive scan [18]. Using the oxidative scan as an example, the current increases exponentially as the analyte is oxidised at the working electrode surface. The current is dictated by the rate of diffusion of oxidant to the electrode. As the scan continues, more oxidant is depleted. The oxidised ions will form a diffusion layer until the current reaches a peak, at which point the current is limited by mass transport of analyte to the electrode. Further increase in potential causes a decrease in current until a steady state is achieved. The reverse happens during the reductive scan. The anodic and cathodic peak currents should be of equal magnitude but with opposite sign, implying that a fully reversible process. Hence, only analysis of a single peak is required.

The electrochemical biosensor tested presents a low cost and simple analytical method for detecting base pair mismatches in DNA. A successful proof of concept can be used for further developments of the method for applications in various fields.

2. Materials and Methods

2.1 Chemicals

Common use chemicals such as phosphate buffered saline (PBS) and salts were supplied by Sigma Aldrich (UK). Oligonucleotide sequences were synthesised by Invitrogen (UK).

Table 1. *Oligonucleotide sequences with underlined mismatches.*

Oligonucleotide	Sequence	Base Pair Mismatches	MW (g/mol)
Biotin-cDNA	5'-CCG GGC GTC GCT GGT GGG-3'-biotin	-	6160.0
tDNA	5'-CCC ACC AGC GAC GCC CGG-3'	0	5116.6
Non-cDNA 1	5'-CCC ACC AGC **CTG** GCC CGG-3'	3	5407.6
Non-cDNA 2	5'-CCC ACC **TCG** **CTG** GCC CGG-3'	6	5398.6
Non-cDNA 3	5'-**GGG** ACC AGC GAC GCC **GCC**-3'	6	5196.6

2.2 Oligonucleotide Design

ssDNA sequences with length of 18 base pairs were designed with a pair of complementary strands, as well as three non-complementary with mismatches of varying number of base pairs and/or locations. The complementary DNA was biotinylated (see biotin-cDNA) for immobilisation on the streptavidin modified SPCE. The full list of oligonucleotide sequences is provided in Table 1.

2.3 Biosensor Preparation

The electrochemical biosensor was constructed using DropSens streptavidin modified SPCE purchased from Metrohm (UK). Preparation of the biosensor comprised of two key steps: immobilisation of sample onto the working electrode and hybridisation of complementary DNA strands. Hybridisation of ssDNA was conducted prior to sample immobilisation in order to achieve greater consistency and control over reaction conditions at this current proof-of-concept stage. It was thought that the hybridisation of DNA strands at elevated temperatures may damage the streptavidin protein attached on the working electrode surface were the ssDNA first immobilised then heated for hybridisation.

DNA Hybridisation. Working solutions of the sequenced ssDNA were diluted in 1x PBS solution to the desired working concentrations. The desired pairings of ssDNA were hybridised together for 5 min at 73 °C, over an electric heating bath. Hybridised aliquots were left at room temperature for 10 min for cooling to minimise material loss due to evaporation.

DNA Immobilisation. 10 µL aliquots of the hybridised solutions comprising biotin-cDNA and relevant ssDNA were pipetted onto the streptavidin coated working electrode surface of the SPCE. The SPCE was left at room temperature for 15 min over a water bath, which provided humidity to maintain the functionality of streptavidin. The working electrode was washed with 1x PBS solution for 5 times to rinse off any biotin-cDNA not fully immobilised onto the electrode surface.

2.4 Cyclic Voltammetry

CV was performed using a Metrohm AutoLab potentiostat on the Nova 2.1.4 interface. The prepared SPCEs had their sensing area fully immersed in an aqueous solution of 5 mM $K_3Fe(CN)_6/K_4Fe(CN)_6$ in 0.1 M KCl. The scan rate was set at 100 mV/s across an experimentally-determined scanning range of -0.2 V to +0.5 V. CV was initialised with 3 pre-treatment cycles to stabilise the working electrode before the data was recorded for a further 5 cycles.

2.5 SDS PAGE

Sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS PAGE) was performed using a 20 % polyacrylamide gel to separate the different ssDNA and dsDNA samples. A high concentration of 20% polyacrylamide was required to achieve separation due to the extremely small differences in molecular weights between the DNA strands measuring only 18 base pairs in length. The gel was prepared with 8 ml of 30 % acrylamide, 3.55 ml of water, and 0.24 ml of 50x tris-acetate EDTA (TAE) buffer. 100 ng samples of the DNA were mixed with 5x loading dye and made up with water to prepare 20 µL aliquots. SDS PAGE was conducted at 100 V for 1 hour. The detached gel was then stained with SYBR gold for 3 hours before images were captured under UV light with NuGenius gel imaging system.

3. Results and Discussion

3.1 SDS PAGE

Successful hybridisation of the ssDNA strands following the designed conditions were validated using SDS PAGE. As shown in Figure 2, visible separation of bands was observed between the ssDNA and dsDNA samples.

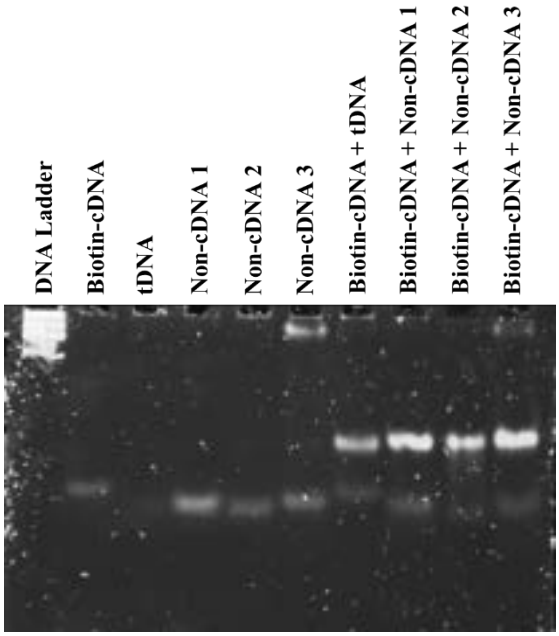


Figure 2. *20 % polyacrylamide gel with ssDNA and hybridised dsDNA samples captured under UV light.*

Since a standard DNA ladder comprising oligonucleotides much longer than the tested samples was used, bunching of the DNA ladder bands was to be expected considering the fact that the samples tested were much shorter and would progress much faster. Nonetheless, some progression of the DNA ladder down the gel showed that the SDS PAGE setup was working as intended. The ssDNA samples gave a single band while the dsDNA samples displayed two bands. The higher band corresponded to the hybridised dsDNA due to their larger molecular weights causing them to move slower than ssDNA. The presence of the second band for dsDNA samples indicated that not all of the ssDNA were completely hybridised. Nonetheless, a simple visual comparison of the brightness of the two bands showed that the amount of unhybridised ssDNA left was much lower than that of hybridised dsDNA present. The small amount of remaining ssDNA following the hybridisation step would be unlikely to significantly affect the electrochemical biosensor performance.

Based on the SDS PAGE gel image, it was not possible to distinguish between different dsDNA strands despite the presence of varying base pair mismatches. This further proved the utility of an electrochemical biosensor in providing complementary data to identify base pair mismatches in DNA mutations.

3.2 CV Parameter Validation

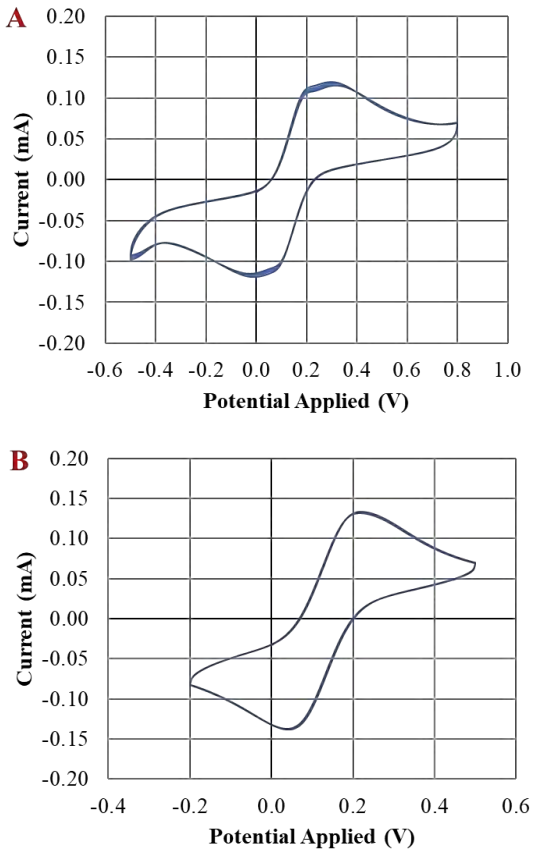


Figure 3. *Cyclic voltammograms recorded for streptavidin modified SPCE with 100 µM hybridised biotin-cDNA and tDNA. Figures A and B correspond to scanning ranges of -0.5 V to +0.8 V and -0.2 V to +0.5 V respectively. Scan rate of 100 mV/s.*

Preliminary CV testing conducted on the SPCEs allowed for confirmation of the key parameters for running CV tests. Figure 3 shows that the smaller scanning range of -0.2 V to +0.5 V was more suitable for the particular SPCE used as minimal degradation of the peak current was observed across consecutive cycles. Reduction in peak current for the larger scanning range in Figure 3(A) could be caused

by the degradation of biological materials due to an excessive potential difference being applied. The smooth curves recorded showed that the scanning rate of 100 mV/s used provided adequate time for the redox reactions to proceed. Good overlap of the 5 consecutive cycles observed in Figure 3(B) also showed that the 3 pre-treatment cycles used were sufficient for initial stabilisation of the working electrode.

The cyclic voltammogram in Figure 3(B) showed well-defined peaks. This indicated that electron transfer kinetics of $Fe(CN)_6^{4-}$ ions on the working electrode was fast, and that the rinsing process of the working electrode with 1x PBS was satisfactory.

3.3 Initial CV Testing

Initial CV testing was performed to ascertain the reliability of the CV setup, as well as provide a baseline comparison case for subsequent tests with dsDNA. The bare carbon electrode was tested against standard cyclic voltammograms available from the manufacturer. Based on the shape of the cyclic voltammogram, peak current location and values, the bare carbon electrode results had good agreement with cyclic voltammograms provided by the manufacturer [25]. Standard cyclic voltammograms were not available for the bare streptavidin modified electrode.

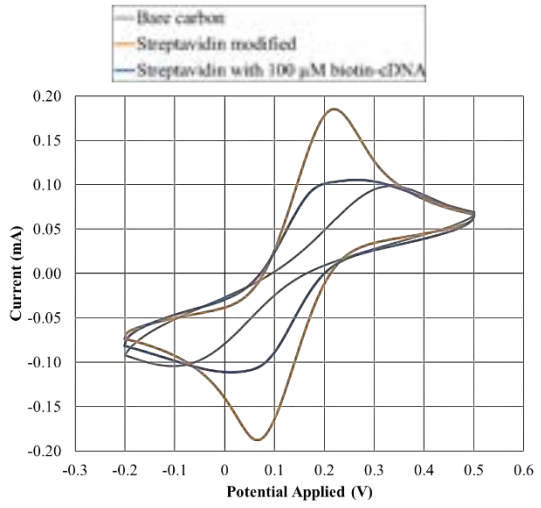


Figure 4. *Cyclic voltammograms recorded for electrodes with bare carbon, bare streptavidin modified, and 100 µM biotin-cDNA. Scan rate of 100 mV/s.*

As shown in Figure 4, both streptavidin modified SPCEs with and without immobilised material recorded peaks at a lower potential as compared to that of the bare carbon electrode. By modifying the surface of the working electrode with streptavidin, the electrochemical characteristics of the working electrode was altered. Hence, the shift in peak location was to be expected.

Of importance would be the magnitude of the peak currents recorded. The bare streptavidin modified electrode had the highest peak current peak current of around 0.18 mA. Peak current decreased to 0.11 mA following immobilisation of the single-stranded biotin-cDNA. This showed that the immobilisation of DNA onto the streptavidin coated working electrode surface was successful and brought about detectable changes in the electrochemical activity of the electrode.

The presence of phosphate groups in nucleotides conferred DNA its negative charge. The repulsive electrostatic interactions of the negatively charged ssDNA and ferrocyanide ions impeded the transfer of ferrocyanide ions to the working electrode surface. This implied a higher electron transfer resistance and thus lower peak current [20]. The recorded peak currents for bare streptavidin and that of immobilised biotin-cDNA provided useful reference points for subsequent comparisons involving different dsDNA combinations.

3.4 Base Pair Mismatch Detection

CV tests conducted with the four pairs of hybridised dsDNA, previously given in Table 1, yielded quantifiable differences in electrochemical readout from the CV. As shown in Figure 5, the magnitude of peak currents of the electrodes with hybridised dsDNA fell between the range given by the peak current of bare streptavidin modified electrode and that of the electrode with only single-stranded biotin-cDNA. Amongst the dsDNA samples, peak current recorded in decreasing order for hybridised pairs had tDNA having the highest peak current, followed by non-cDNA 3, non-cDNA 1, and then non-cDNA 2.

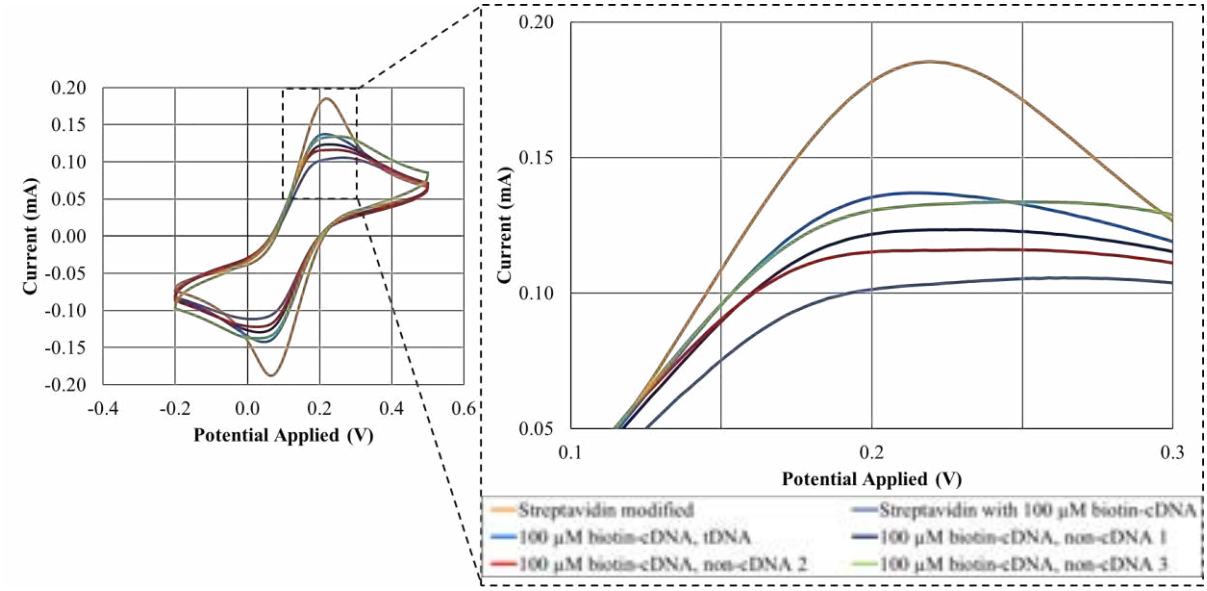


Figure 5. *Cyclic voltammogram of hybridised dsDNA with tDNA and non-cDNA 1 to 3, enlarged over positive current peaks between the range of potential applied from 0.10 V to 0.30 V. Scan rate of 100 mV/s.*

While the cyclic voltammogram for dsDNA was initially expected to give lower peak currents than ssDNA due to an increase in resistance from dsDNA, the CV tests conducted showed otherwise. Peaks from dsDNA were higher than that of ssDNA. Although dsDNA might have inherently higher electron transfer resistance than ssDNA, other factors such as the orientation of immobilised DNA on the working electrode surface may also affect the overall electrochemical characteristic of the electrode [20]. Further investigation beyond this study may be conducted to better understand the actual electron transfer mechanism. Nonetheless, the results obtained were reproducible across multiple chips and repeats. Trends in the current peaks between hybridised pairs of DNA observed were also in order as mentioned previously.

To better quantify the effect of the detection of DNA base pair mismatches on the cyclic voltammogram, the drop in peak current from that of the bare streptavidin modified SPCE was calculated and plotted in Figure 6.

From Figure 6, the single-stranded biotin-cDNA recorded the largest drop in peak current of 0.08 mA. Samples 2 through 4 are in order of increasing number of base pair mismatches. A corresponding increase in the drop in peak current was calculated, ranging from 0.055 mA to 0.065 mA. It could be hypothesised that with increasing number of base pair mismatches, the drop in peak current would eventually trend closer towards that of single-stranded biotin-cDNA as fewer non-cDNA strands would be able to successfully hybridise with the biotin-cDNA.

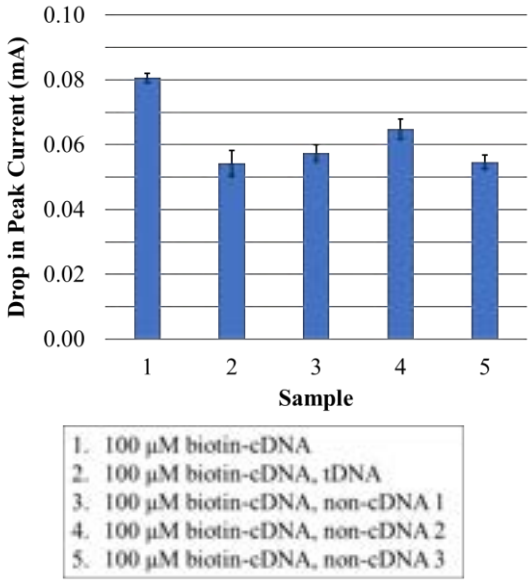


Figure 6. *Drop in peak currents from bare streptavidin modified SPCE for various immobilised DNA strands. Average readings from three SPCEs for each sample recorded and corresponding error bars plotted.*

For sample 5, with non-cDNA 3, the peak current drop was however noted to be lower than sample 4, despite both having 6 base pair mismatches. Hybridisation of dsDNA would be affected by both the number of base pair

mismatches, as well as the location of the mismatches on the DNA sequence. Comparison between the drop in peak currents of just samples 4 and 5 indicate that the difference in electrochemical activity between dsDNA with the same number of base pair mismatches at varying locations could potentially be picked up by the electrochemical biosensor.

Nonetheless, it should be noted that while a trend of drop in peak current with number of base pair mismatches was shown, the differences could not yet be said to be statistically significant between certain samples based on the current small sample size of data. This would be further discussed later in section 3.6.

3.5 Identifying Detection Limits

A brief study into the effect of target DNA concentrations on electrochemical output was conducted with hybridised dsDNA between biotin-cDNA and tDNA. tDNA concentrations of down to 1 µM, or 1 % of the 100 µM biotin-cDNA concentration, were tested.

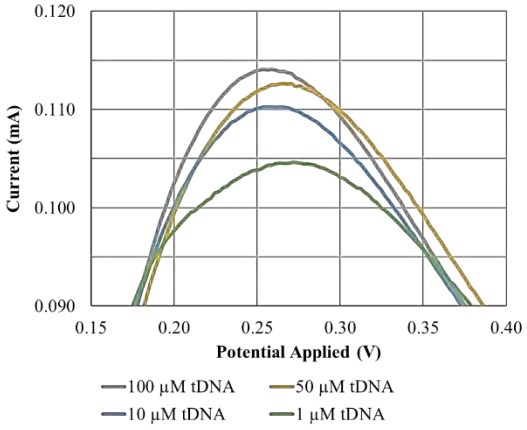


Figure 7. *Cyclic voltammogram for streptavidin modified SPCE with different concentrations of tDNA hybridised with 100 µM biotin-cDNA at 73 °C, enlarged over positive current peaks between the range of potential applied from 0.15 V to 0.40 V. Scan rate of 100 mV/s.*

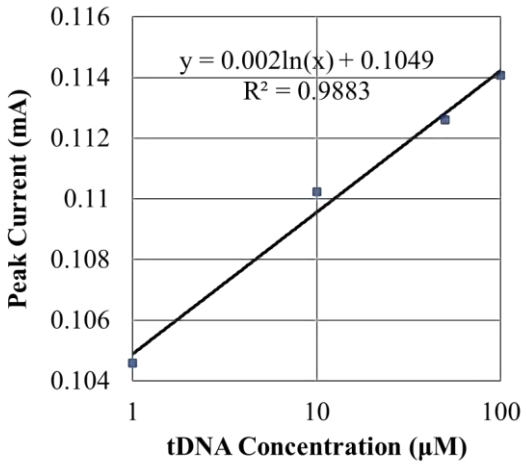


Figure 8. *Correlation between peak current recorded and logarithmic of tDNA concentration hybridised with biotin-cDNA.*

As shown in Figure 7, the cyclic voltammograms showed a decrease in peak current recorded with decreasing tDNA concentrations. From Figure 8, peak current recorded also showed good corelation with tDNA concentrations in logarithmic terms. The fitting equation obtained was $y = 0.002\ln(x) + 0.1049$ with a goodness-of-fit value of $R^2 = 0.9883$. This showed that the constructed electrochemical biosensor performed reasonably well even at reasonably low sample detection limits of 1 % of the biotin-cDNA concentrations.

It would certainly be desirable to conduct further experiments to better determine if lower detection limits could be achieved with the electrochemical biosensor. Nonetheless, the current results showed that the biosensor was sufficiently sensitive to for detection limits of at least 1 % and would be sufficiently robust at this current stage of development.

3.6 Critical Evaluation

As this study is still at an early proof-of-concept stage in the development of the electrochemical biosensor, there was a conscious need to balance the allocation of finite resources available. Hence, this study sought to explore a breadth of various aspects crucial to understanding the biosensor, rather than focusing too in depth on any one aspect in particular. This included the mechanism of immobilisation, effect of base pair mismatches on electrical readouts, and an approximate limit of detection. Certain trade-offs were thus made

in the process in order to determine the overall feasibility of the concept.

Oligonucleotide length. The DNA sequences used in this study were very short strands only 18 base pairs in length. This afforded certain benefits such as minimising the formation of undesired secondary structures which could hinder the hybridisation processes. By keeping the oligonucleotide lengths short, the ability to hybridise would be dependent only on the presence of base pair mismatches in the non-cDNA. The number of mismatched nucleotides as a percentage of the total nucleotide length would also be higher, thus emphasising the difference in hybridisation even for just 3 or 6 base pair mismatches. However, the use of short oligonucleotides would also likely confer a much smaller level of impedance to the working electrode surface. Smaller magnitudes of changes to peak current would thus be expected. Cyclic voltammetry readouts could thus be more greatly limited by the sensitivity of the equipment available. The sensitivity of the equipment in measuring short oligonucleotides should thus be taken into account when evaluating the lack of statistically significant differences in the drop in peak currents mentioned in Section 3.4. The use of longer DNA strands could potentially lead to more pronounced differences between the drops in peak current.

Sample preparation. In this study, the hybridisation event was conducted first at 73 °C before immobilisation to avoid any potential degradation of the streptavidin protein. However, it was noted that there could be a need for immobilisation biotin-cDNA first, which then hybridises and captures complementary target DNA strands. Most samples collected would also likely face contamination by other substances, the effect of which on the SPCE would have to account for. Separately, additional pre-treatment of samples could be necessary for real world samples containing a wide range of impurities. Hence, while the study provides a strong basis for further development of an electrochemical biosensor, much more work would have to be done to arrive at a fully implementable commercial product.

The above two factors raised should be taken into consideration to provide a more holistic view of the results from this study. These were conscious decisions made surrounding experimental design. It is recommended that the above design considerations be accounted for in any further work to be done surrounding the electrochemical biosensor of interest.

4. Conclusions

4.1 Biosensor Proof of Concept

The feasibility of developing an electrochemical biosensor to detect base pair mismatches in DNA mutations, centred around the streptavidin-biotin binding interactions, was proven in this study. Hybridised dsDNA with the attached biotin protein was successfully immobilised onto the streptavidin modified working electrode surface. Proof of successful immobilisation is a key finding in itself as the mechanism of immobilisation could potentially be easily replicated for other forms of biological molecules such as mRNA. The ready availability of commercial SPCEs with the streptavidin modification also ensures lower barriers to implementation of such an electrochemical biosensor.

Detectable changes in recorded peak current outputs depending on the number and location of base pair mismatches provided a desirable outcome for further development of an electrochemical biosensor for detection purposes. The reproducibility of results across multiple chips is also very promising. Given the greater ease of processing, this electrochemical biosensor could be easily integrated with other automated and digital process systems into real world applications.

The output of an electrochemical biosensor could be used to complement experimental data from current established methods such as SDS PAGE. This provides users with a simple alternative method at lower cost to analyse biological samples of interest.

Overall, this study opens up plenty of possibilities for applications in various scientific fields such as molecular diagnostics, therapeutic development, biotechnology, and environmental studies. The ease of use and scalability of an electrochemical biosensor only adds to the attractiveness of such an option in real world applications.

4.2 Future Work

This proof of concept study can be used as a basis for more in-depth studies to better understand the effect of particular parameters. These include, but are not limited to, oligonucleotide length, number and location of base pair mismatches, concentration of immobilised biotinylated cDNA, and hybridisation conditions. Practical considerations of target sample preparation and determining the absolute lower detection limits can also be explored.

5. Acknowledgements

We would like to express our gratitude to Dr Ying Tu, Mr Chileab A B Redwood-Sawyerr, as well as all members of the Polizzi laboratory group for their professional guidance and support. SPCEs were purchased with funding from Wellcome Leap.

6. References

1. Kingsland, A. & Maibaum, L. (2018) DNA Base Pair Mismatches Induce Structural Changes and Alter the Free-Energy Landscape of Base Flip. *The Journal of Physical Chemistry B*. 122 (51), 12251–12259. doi:10.1021/acs.jpcb.8b06007.
2. Mahdieh, N. & Rabbani, B. (2013) An Overview of Mutation Detection Methods in Genetic Disorders. *Iranian Journal of Paediatrics*. 23 (4), 375–388.
3. Boyce, K.J., Wang, Y., Verma, S., Shakya, V.P.S., Xue, C. & Idnurm, A. (2017) Mismatch Repair of DNA Replication Errors Contributes to Microevolution in the Pathogenic Fungus *Cryptococcus neoformans* J.A. Alspaugh (ed.). *mBio*. 8 (3). doi:10.1128/mbio.00595-17.
4. Aarnio, M., Sankila, R., Pukkala, E., Salovaara, R., Aaltonen, L.A., de la Chapelle, A., Peltomaki, P., Mecklin, J.-P. & Jarvinen, H.J. (1999) Cancer risk in mutation carriers of DNA-mismatch-repair genes. *International Journal of Cancer*. 81 (2), 214–218. doi:3.0.co;2-l">10.1002/(sici)1097-0215(19990412)81:2<214::aid-ijc8>3.0.co;2-l.
5. Hlongwane, G.N., Dodoo-Arhin, D., Wamwangi, D., Daramola, M.O., Moothi, K. & Iyuke, S.E. (2019) DNA hybridisation sensors for product authentication and tracing: State of the art and challenges. *South African Journal of Chemical Engineering*. 27, 16–34. doi:10.1016/j.sajce.2018.11.002.
6. Jiang, H., Xi, H., Juhas, M. & Zhang, Y. (2021) Biosensors for Point Mutation Detection. *Frontiers in Bioengineering and Biotechnology*. 9. doi:10.3389/fbioe.2021.797831.
7. Morlan, J., Baker, J. & Sinicropi, D. (2009) Mutation Detection by Real-Time PCR: A Simple, Robust and Highly Selective Method I. Schrijver (ed.). *PLoS ONE*. 4 (2), e4584. doi:10.1371/journal.pone.0004584.
8. Strathdee, F. & Free, A. (2013) Denaturing Gradient Gel Electrophoresis (DGGE). *Methods in Molecular Biology*. 1054, 145–157. doi:10.1007/978-1-62703-565-1_9.
9. Viglasky, V. (2013) Polyacrylamide temperature gradient gel electrophoresis. *Methods in Molecular Biology (Clifton, N.J.)*. 1054, 159–171. doi:10.1007/978-1-62703-565-1_10.
10. Bhalla, N., Jolly, P., Formisano, N. & Estrela, P. (2016) Introduction to biosensors. *Essays In Biochemistry*. 60 (1), 1–8. doi:10.1042/ebc20150001.
11. Moore, B., Hu, H., Singleton, M., De La Vega, F.M., Reese, M.G. & Yandell, M. (2011) Global analysis of disease-related DNA sequence variation in 10 healthy individuals: Implications for whole genome-based clinical diagnostics. *Genetics in Medicine*. 13 (3), 210–217. doi:10.1097/gim.0b013e31820ed321.
12. Mansouri, S., Cuie, Y., Winnik, F., Shi, Q., Lavigne, P., Benderdour, M., Beaumont, E. & Fernandes, J.C. (2006) Characterization of folate-chitosan-DNA nanoparticles for gene therapy. *Biomaterials*. 27 (9), 2060–2065. doi:10.1016/j.biomaterials.2005.09.020.
13. Yasmin, J., Ahmed, M.R. & Cho, B.-K. (2016) Biosensors and their Applications in Food Safety: A Review. *Journal of Biosystems Engineering*. 41 (3), 240–254. doi:10.5307/jbe.2016.41.3.240.
14. Palchetti, I. & Mascini, M. (2008) Nucleic acid biosensors for environmental pollution

monitoring. *The Analyst*. 133 (7), 846–854. doi:10.1039/b802920m.

15. Adiguzel, Y. (2017) Biosensor with UV Spectrophotometric Detection. *Proceedings*. 2 (3), 113. doi:10.3390/ecsa-4-04928.

16. Zeglis, B.M. & Barton, J.K. (2006) A Mismatch-Selective Bifunctional Rhodium−Oregon Green Conjugate: A Fluorescent Probe for Mismatched DNA. *Journal of the American Chemical Society*. 128 (17), 5654–5655. doi:10.1021/ja061409c.

17. Asrat, T.M., Cho, W., Liu, F.A., Shapiro, S.M., Bracht, J.R. & Zestos, A.G. (2021) Direct Detection of DNA and RNA on Carbon Fiber Microelectrodes Using Fast-Scan Cyclic Voltammetry. *ACS Omega*. 6 (10), 6571–6581. doi:10.1021/acsomega.0c04845.

18. Elgrishi, N., Rountree, K.J., McCarthy, B.D., Rountree, E.S., Eisenhart, T.T. & Dempsey, J.L. (2017) A Practical Beginner's Guide to Cyclic Voltammetry. *Journal of Chemical Education*. 95 (2), 197–206. doi:10.1021/acs.jchemed.7b00361.

19. Taleat, Z., Khoshroo, A. & Mazloum-Ardakani, M. (2014) Screen-printed electrodes for biosensing: a review (2008–2013). *Microchimica Acta*. 181 (9-10), 865–891. doi:10.1007/s00604-014-1181-1.

20. Hernández-Santos, D., Díaz-González, M., González-García, M.B. & Costa-García, A. (2004) Enzymatic Genosensor on Streptavidin-Modified Screen-Printed Carbon Electrodes. *Analytical Chemistry*. 76 (23), 6887–6893. doi:10.1021/ac048892z.

21. González, M., Bagatolli, L.A., Echabe, I., Arrondo, J.L.R., Argaraña, C.E., Cantor, C.R. & Fidelio, G.D. (1997) Interaction of Biotin with Streptavidin. *Journal of Biological Chemistry*. 272 (17), 11288–11294. doi:10.1074/jbc.272.17.11288.

22. Lian, H., Jiang, J., Wang, Y., Yu, X., Zheng, R., Long, J., Zhou, M., Zhou, S., Wei, C., Zhao, A. & Gao, J. (2022) A novel multimeric sCD19-streptavidin fusion protein for functional detection and selective expansion of CD19-targeted CAR-T cells. *Cancer Medicine*. 11 (15), 2978–2989. doi:10.1002/cam4.4657.

23. Magar, H.S., Hassan, R.Y.A. & Mulchandani, A. (2021) Electrochemical Impedance Spectroscopy (EIS): Principles, Construction, and Biosensing Applications. *Sensors (Basel, Switzerland)*. 21 (19), 6578. doi:10.3390/s21196578.

24. Robson, H., Reinhardt, M. & Bracher, C. (n.d.) *Cyclic Voltammetry Explained: Basic Principles & Set Up*. Ossila. https://www.ossila.com/pages/cyclic-voltammetry.

25. Metrohm (n.d.) *Metrohm DropSens Screen-Printed electrodes*. Dropsens. https://www.dropsens.com/en/screen_printed_electrodes_pag.html.

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

Jeremy Kim, Heather Page

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

Abstract

The world is undergoing sustained efforts towards decarbonising the energy industry to address the current climate crisis. This report investigates the feasibility of retrofitting an MEA-based post-combustion carbon capture plant to a small-scale waste-to-energy power plant utilising Refuse Derived Fuel. The power plant is modelled and simulated through Aspen Plus V11, alongside the carbon capture process operating with a capture efficiency of 95%. The feasibility is investigated in terms of the energy requirements and implications, as well as the specific post-combustion capture plant area requirement per unit power generation. At the given capture efficiency, an associated energy penalty of 45.7% is imposed, leading to a decrease in net output by 4.3 MW_e, primarily due to high energy consumption at the reboiler and CO_2 compressor. A specific plot sizing of 232.6m^2MW^{-1} is obtained for the physical requirement of the carbon capture plant layout at this capacity and operation. The high energy penalty and large specific plot area lead to complications in the feasibility of this simulated waste-to-energy plant.

1 Introduction

Since the introduction of the Paris Agreement at COP21, there has been an increased combined effort across the world for governments to combat climate change. Signed by 194 parties, including the UK, the Agreement established long-term goals to guide nations to limit the global temperature increase to well below 2°C, while pursuing efforts to further limit this increase to 1.5°C (UNFCCC, 2015). This was reinforced at the most recent COP27, where a 2°C increase was demonstrated to be unsafe, thus focusing on permitting a maximum rise of 1.5°C. To this end, COP27 saw the launch of new programmes to promote climate technology solutions in developing countries and scaling up mitigation ambition and implementation ('COP 27', 2022). Yet, the UN Framework Convention on Climate Change currently places the world on track to reach 2.5°C by the turn of the century (UNFCC, 2022), indicating a need for a 45% reduction in greenhouse gas (GHG) emissions (IPCC, 2018). As such, it is clear that carbon capture holds its place within this industry to help achieve these aims, with increasing importance as the urgency increases day by day. Carbon capture can particularly aid in the transition to renewable energy, decarbonising current and existing emissions sources, which will be needed to provide secure and predictable energy whilst the transition to a low carbon energy system is ongoing. Moreover, the European Energy Roadmap forecasts that electricity produced from biomasses and waste will account for between 7.3 and 10.8% of total production in 2050 (European Commission, 2011), highlighting the importance of utilising biomass and waste as energy sources for developing an energy system independent from fossil fuels. Refuse Derived Fuel (RDF) appears as a promising option, being a sustainable fuel source, with the main drawbacks stemming from direct emissions due to its combustion. Successfully combining an RDF-to-energy plant with carbon capture technology may therefore present a cumulative benefit and encourage the adoption of utilising RDF. Investigating primarily the energy penalty and physical plant footprint will allow for an analysis into the carbon capture feasibility – to include both commercial considerations alongside technical feasibility – of such RDF-to-energy power plants in conjunction with a carbon capture retrofit.

2 Background

2.1 Post-Combustion Capture (PCC)

Fossil fuels play a major role in power generation in the UK, hence it is crucial that technologies which reduce the carbon footprint of fossil fuel-dependent industries are developed, in addition to advancing green energy production. Carbon capture and storage (CCS) enables continuing operation of fossil fuel-based power plants whilst eliminating carbon emissions. The main methods for carbon capture fall into three categories: oxy-fuel combustion systems, pre-combustion carbon capture and post-combustion carbon capture (PCC) – the latter of which is implemented here. The installation of PCC enables the reduction in carbon emissions without significantly changing existing infrastucture making this technique more economically feasible; a CCS process is retrofitted to a fossil fuel power plant, capturing carbon dioxide from the exhaust gas stream before release to the atmosphere.

2.2 CCS Solvent Selection

Monoethanolamine (MEA) solution, an amine-based absorption solvent for CO_2, is selected as the CCS technology for this study, being the most documented method for CCS. Despite MEA being a mature choice of solvent, this technology requires a large solvent regeneration energy which has been seen to cause a high energy penalty when retrofitted to power stations (Goto, Yogo and Higashii, 2013). Research into lowering the energy penalty of CCS processes through novel solvents (Mumford *et al.*, 2015) and by means of other gas separation technologies – such as pressure swing adsorption or membranes (Xie *et al.*, 2019) – is currently being conducted. However, such processes require electricity rather than heat. Thus, assessing the feasibility of operating a MEA-based CCS process, of high regeneration energy, retrofitted to an RDF-to-energy plant is of particular interest here.

2.3 Refuse Derived Fuel

Non-hazardous municipal solid waste (MSW) streams, which cannot be reused or recycled, can be utilised as a source of energy generation whilst helping to overcome problems regarding the depletion of fossil fuels, meeting increasing global energy demand and managing waste generation. RDF is the energy-contained portion of MSW, displaying a wide range of chemical composition due to varying blends of organic mixtures and thus the properties cannot be easily predicted. RDF is prepared from MSW via removal of inert matter, size reduction methods – such as milling – and drying (Jannelli and Minutillo, 2007). Once processed, RDF can be used as a solid fuel source, like coal, to produce energy by combustion. This energy can then be converted within a Combined Heat and Power (CHP) plant; steam turbines coupled with a generator produce electricity whilst thermal energy is rejected in the cooling water (Environ Consultants Ltd., n.d.). In this study, the chemical compositions of several different RDFs from European sources have been obtained from literature to be able to adequately model an RDF-to-energy plant. Despite the extensive research into PCC retrofitted to larger-scale coal fired power plants, the full-scale implementation of retrofitting PCC to smaller-scale, waste fuel-fired power stations is still minimal (Bisinella *et al.*, 2021).

3 Methodology

In this study, a PCC plant has been retrofitted to a CHP RDF plant based on a UK plant processing 4,500 $kghr^{-1}$ of RDF, observed in literature (Environ Consultants Ltd., n.d.). Both processes have been modelled by means of Aspen Plus V11 code (Aspen Plus, 2000).

3.1 Power and Heat Generation Plant

The Aspen Plus flowsheet for this process is displayed in Figure 3.1.1. The RDF power plant consists of various sections:

- RDF drying section: moist RDF is dried via direct hot air drying through a rotary drier to 1% moisture content;
- RDF combustion section: RDF is combusted through a moving grate combustion system at temperatures exceeding 1500°C;
- gas recovery section: product gases are separated from ash produced during the combustion of RDF in a cyclone;
- power generation section: a high pressure (HP) turbine and low pressure (LP) turbine configuration is coupled directly with a generator.

3.1.1 Characteristics of RDF

Table 3.1.1 summarises the proximate analysis, ultimate analysis, and Lower Heating Values (LHV) of five different RDFs sourced throughout Europe. RDF 4 is a UK based source, providing the highest LHV and is thus chosen as the fuel to simulate, being applicable for this study. A fuel with a higher LHV releases more heat when combusted, therefore increasing power generation. These given characteristics are input requirements to model RDF in Aspen Plus since RDF is a non-conventional fuel.

Table 3.1.1: Chemical characteristics of 5 different European-sourced RDF fuels

RDF Fuel	Proximate Analysis (%wt)				Ultimate Analysis (%wt)					Calorific value ($MJkg^{-1}$)
	Moisture content	Volatile content	Fixed carbon	Ash content	C	H	O	N	S	
1	8.50	70.40	3.60	26.00	46.80	5.40	20.40	1.10	0.30	11.40
2	0.99	79.81	9.69	10.50	51.30	7.50	29.72	0.77	0.21	23.09
3	1.30	72.85	10.84	16.31	56.75	4.74	20.40	1.67	0.13	22.35
4	5.90	74.39	11.90	13.71	59.80	8.58	16.47	1.03	0.41	24.92
5	4.20	76.20	10.44	13.36	43.05	5.91	37.07	0.60	0.00	18.45

1 - (García *et al.*, 2021); 2,3 - (Grammelis *et al.*, 2009); 4 - (Materazzi *et al.*, 2015); 5 - (Efika, Onwudili and Williams, 2015)

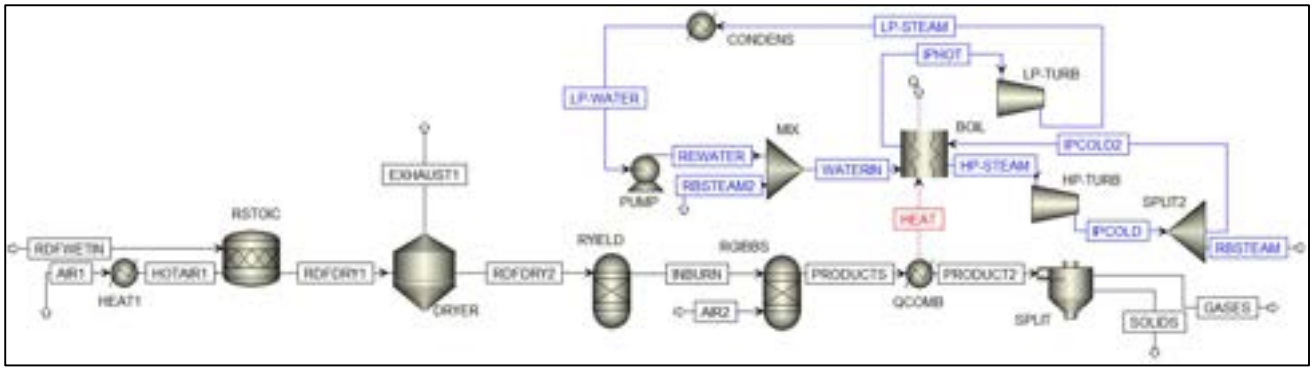


Figure 3.1.1: Aspen Plus flowsheet of RDF-to-energy plant

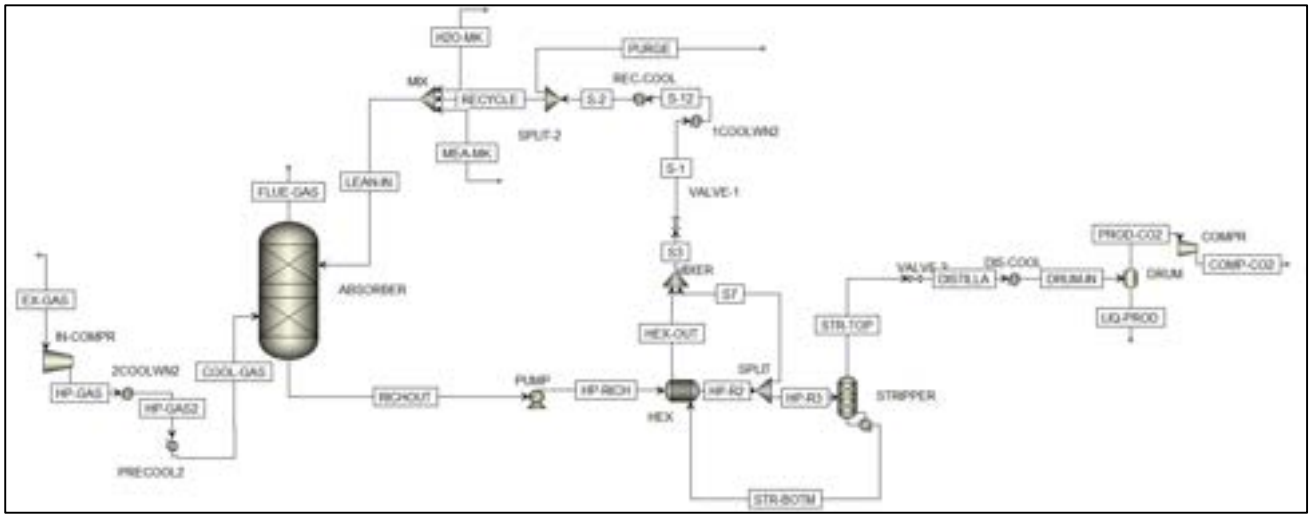


Figure 3.1.2: Aspen Plus flowsheet of PCC plant

3.1.2 Drying

As RDFs are commonly manufactured and received with a non-negligible moisture content, implementing a drying stage is necessary to remove excess moisture and maximise the heat generated and efficiency of combustion. In Aspen Plus, this is modelled by contacting a hot air stream – heated by use of excess heat from the CCS process – with the wet RDF in the RSTOIC block to react a portion of RDF to form water, before passing the single outlet stream through the DRYER model which separates the dried RDF from the moist air exhaust. In practice, this is done within a direct dryer such as a belt or rotary dryer, the most commercially observed dryers for RDF treatment (Kiranoudis, Maroulis and Marinos-Kouris, 1996). An energy balance around this process yields equation 3.1.2.1:

$$\begin{aligned} Q_{min,drying} = {} & \dot{m}_s c_{ps}(T_{sf} - T_{si}) \\ & + \dot{m}_{RDF}\{X_i c_{pl}(T_v - T_{si}) \\ & + X_f c_{pl}(T_{sf} - T_v)\} \\ & + \dot{m}_{moisture}\{\lambda \\ & + c_{pv}(T_{vf} - T_v)\} \end{aligned} \qquad (3.1.2.1)$$

where $\dot{m}_s$, $\dot{m}_{RDF}$, $\dot{m}_{moisture}$ denotes the mass flowrate of dried RDF, initial RDF and evaporated moisture (obtained via a mass balance of the water present) respectively; c_{ps}, c_{pl}, c_{pv} denotes the specific heat capacity of the RDF (which is assumed constant over the temperature range), water and steam respectively and λ the latent heat of water; X_i, X_f denotes the inlet and outlet moisture content of the RDF respectively and T_{sf}, T_{si}, T_v, T_{vf} denotes the temperature of the RDF at the outlet, at the inlet, the saturated water and produced vapor respectively. Incorporating a 5% heat loss inefficiency and for a drying process achieving 1% moisture content, a total heat duty requirement of $Q_t = 0.18\text{MW}$ is identified, equating to an air flowrate of 6kgs^{-1}.

3.1.3 Combustion

The combustion of RDF is quite complex and is therefore only modelled by two major steps. Since RDF is a non-conventional fuel, the Gibbs free energy of RDF cannot be calculated in Aspen and therefore the RDF is first to be decomposed into its constituent elements through the RYIELD reactor (Aspen Plus, 2000). The yield distribution is obtained from the ultimate analysis of RDF. The heat of reaction associated with this decomposition of

RDF is considered then in the RGIBBS reactor, which simulates the combustion process. This block models chemical equilibrium by minimising Gibbs free energy and does not require details of reaction stoichiometry, reacting RDF and air to produce the exhaust gases. A sensitivity analysis was conducted to find the amount of air required for complete combustion, yielding an air flowrate of 55,000kghr^{-1}.

3.1.4 Exhaust Gas-Solid Separation

The product exhaust gases are separated from the ashes produced during combustion via a cyclone, modelled in Aspen Plus by an SSPLIT block, before entering the CCS process. The mass fractions of the main exhaust gases are displayed in Table 3.1.2. The remaining gases includes SO_X, NO_X, CO and H_2 which equate to less than 1% of the exhaust gas stream.

Table 3.1.2: Composition of exhaust gas stream

Gas	Mass Fraction
CO_2	0.157
N_2	0.717
O_2	0.064
H_2O	0.056
Other gases	< 0.01

3.1.5 Power Generation

To generate power, the Rankine steam cycle has been modelled. The hot combustion gases provide heat energy to the boiler, raising the temperature of a recycled water stream to create high-pressure steam at 600°C and 100 bar. The high-pressure steam is then expanded in the HP turbine to a lower pressure where a portion of the heat is converted to work. The stream exiting the HP turbine is again passed back through the boiler and reheated to 600°C, and then expanded in the LP turbine to 0.1bar. The isentropic efficiencies of both turbines are assumed to be 90%. The low-pressure steam is completely condensed in the condenser where the latent heat of condensation is rejected to the cooling water. A pump then pumps condensate back to the boiler at 10bar.

A portion of the steam flow is bled from between the HP and LP turbines and is used to supply the reboiler duty in the CCS process. This stream is later reintroduced to the steam cycle once cooled by mixing with the water stream first entering the boiler: the detailed reasoning for this is explained in Section 3.2.6.

3.2 Carbon Capture Plant

The Aspen Plus flowsheet for the CCS process is displayed in Figure 3.1.2.

The CCS process, designed to yield a 95% CO_2 capture efficiency, is as follows:

- Exhaust gases exiting the power plant are first pressurised and cooled, then enter the absorption column;
- In the absorber, CO_2 in the exhaust gas stream is absorbed by MEA solvent exiting at the bottom of the column, whilst the remaining flue gases exit at the top;
- The CO_2-rich MEA stream is then pumped through a heat exchanger before entering the stripping column;
- The stripper removes CO_2 from the MEA stream which exits out the top of the column, whilst the now-lean MEA stream is recycled back through the process;
- The CO_2 produced enters a condenser, followed by a knock-out drum where remaining liquid is separated, before being compressed to reduce volume and allow transportation to storage sites.

3.2.1 Reaction Mechanism

In order for Aspen Plus to simulate the necessary reactions, the vapor phase is modelled via the Soave-Redlich-Kwong equation of state, while the liquid phase is described through the activity coefficient 'Electrolyte-NRTL' model. This model is defined by the assumptions of like-ion repulsion and local electroneutrality (Moioli and Pellegrini, 2013).

The CO_2-MEA-H_2O reaction mechanism relies on a zwitterion mechanism, and is defined by an equilibrium of the 5 equations as laid out below (Aboudheir *et al.*, 2003), with all species in aqueous solution.

$$2H_2O \leftrightarrow H_3O^+ + OH^- \quad (1)$$
$$CO_2 + 2H_2O \leftrightarrow H_3O^+ + HCO_3^- \quad (2)$$
$$HCO_3^- + H_2O \leftrightarrow H_3O^+ + CO_3^{2-} \quad (3)$$
$$MEACOO^- + H_2O \leftrightarrow MEA + HCO_3^- \quad (4)$$
$$MEAH^+ + H_2O \leftrightarrow MEA + H_3O^+ \quad (5)$$

In addition to these, 5 kinetically controlled reactions complete the mechanism, as shown in equations 6 through 10 below.

$$CO_2 + OH^- \leftrightarrow HCO_3^- \quad (6)$$
$$MEA + CO_2 \leftrightarrow MEA^+COO^- \quad (7)$$
$$MEA^+COO^- + MEA \leftrightarrow MEAH^+ + MEACOO^- \quad (8)$$
$$MEA^+COO^- + H_2O \leftrightarrow H_3O^+ + MEACOO^- \quad (9)$$
$$MEA^+COO^- + OH^- \leftrightarrow MEACOO^- + H_2O \quad (10)$$

3.2.2 Blower

The pressure of the exhaust gas stream is increased by 0.10bar by the blower. This ensures any frictional pressure drops along the gas pipeline are overcome.

3.2.3 Cooler

The exhaust gases are cooled from 160°C to 40°C before entering the absorber. This involves two cooling units: the first unit is contacted with the air stream required for drying RDF in the power plant, rejecting heat to warm the air stream. This is explained in more detail in Section 3.3. To further cool the stream to the required temperature of 40°C, the second heat exchanger utilises cooling water at 20°C.

3.2.4 Absorber

The absorber is modelled by a RADFRAC unit, with an operating pressure of 0.75bar. The cooled exhaust gas stream enters at the bottom of the column, contacting a lean solvent stream which enters at the top of the column. The concentration of MEA solvent and lean loading are set at 30wt% and 22% respectively, observed as suitable valves used in literature (Puxty *et al.*, 2009) (Alie *et al.*, 2005). Optimising such values acts to minimise the total regeneration energy of the process, with most of this energy being consumed by the stripper reboiler. Sulzer MELLAPAK^{TM} 250Y (Chemtech, 2005) is chosen as the internal structured packing, giving a large surface area of packing material and thus good absorbing ability for the CCS process. The dimensions of the column are optimised though sensitivity analysis in Aspen Plus V11, ensuring a large enough diameter size is chosen - too small of a diameter results in hydraulic infeasibility.

3.2.5 Rich/Lean Heat Exchanger

The CO_2-rich solvent stream exiting the absorber is heated to 94°C by a shell-and-tube heat exchanger, where this stream is contacted against the recycled hot lean solvent stream.

3.2.6 Stripper

Like the absorber, the stripper is modelled as a RADFRAC block, operating at a pressure of 1.9bar, with the internal packing material Sulzer MELLAPAK^{TM} 250Y to allow for sufficient desorption of CO_2. The pre-heated CO_2-rich solvent stream enters at the top of the column, where the liquid stream flows down the column. The reboiler duty provides the sensible heat, the heat of CO_2 desorption and the vaporisation heat of water (Lin and Rochelle, 2014); CO_2 is thus desorbed from the solvent stream and exits at the top of the stripper. The hot lean MEA stream exiting the bottom of the stripper is recycled back around to the absorber. The reboiler is designed as a kettle-type reboiler, operating at 125°C. At the specified MEA concentration and lean loading values, the minimum energy consumption of the reboiler is determined as 4.03GJton^{-1}(CO_2), in alignment with literature values reported to be 4.0GJton^{-1}(CO_2) (Soltani, Fennell and Mac Dowell, 2017). The minimum reboiler energy consumption at the given capture efficiency of 95% is obtained through computer-aided optimisation in Aspen Plus V11. A non-linear objective function is set to minimise the reboiler energy consumption, given the decision variables and constraints displayed in Table 3.2.1. To solve this objective function, a portion of the rich-solvent stream is diverged before entering the stripper. This can be seen in Figure 3.1.2. Thus, this is also how the optimal MEA concentration and lean loading values are found.

Table 3.2.1: Decision variables with constraints for minimising the reboiler energy consumption

Decision variable	Lower bound	Upper bound
Rich-stream split fraction	0.05	0.25
Stripper diameter	1.62 m	1.77 m
Stripper operating pressure	1.7 atm	1.9 atm
Molar boilup ratio	0.1	0.2

The reboiler heating supply is obtained by bleeding steam from between the HP and LP turbines from the power plant, where steam is cooled from 286°C (at 10 bar) to 180°C. Total condensation is assumed with no further sub-cooling. For 95% capture efficiency, a ratio of 1.81 $\frac{kg(steam)}{kg(CO_2\ captured)}$ was obtained. Idem, Gelowitz and Tontiwachwuthikul, (2009) reported ratios within the range 1.9-2.5 $\frac{kg(steam)}{kg(CO_2\ captured)}$, yet as the steam in this process is being condensed from such a high temperature this affects the quantity of steam required, thus providing a lower ratio compared to literature.

3.2.7 CO_2 Compression

Before transportation and storage, CO_2 must be compressed to a dense phase, above the critical pressure (73.82 bar). Goto, Yogo and Higashii, (2013) collected and compared several target pressures reported in various papers for CO_2 compression, whereby the most common target pressure was 110bar – this value is therefore chosen in this model.

3.2.8 Cooling System

Cooling utility water is required as a means to reject heat from the coolers across the CCS process, hence a cooling system must be implemented. A forced-draft cooling tower (FDCT) is chosen for this system. This type of tower allows for a closed cooling water system, not requiring circulating water from a lake or river and so is

more applicable across various locations. In FDCTs, a fan is mounted on the side of the tower and is used to force air through the tower. FCDTs are filled with structured packing material such as Koch-Glitch's FLEXIPAC@-2Y (Purushothama, 2009).

The cooling duty is estimated through conducting an energy balance around each cooler, which can be summed to find the total cooling duty ($\dot{q}$) required and thus the flowrate of cooling water ($\dot{m}_c$):

$$\dot{q}_i = \dot{m}_{i,f}\, c_{pf}\left(T_{i,f1} - T_{i,f2}\right) = \dot{m}_{i,c}\, c_{pc}\left(T_{i,c1} - T_{i,c2}\right) \quad (3.2.8.1)$$

$$\dot{q} = \sum_{i=1}^{n} \dot{q}_i \quad (3.2.8.2)$$

$$\dot{m}_c = \frac{\dot{q}}{c_{pc}(T_{c2} - T_{c1})} \quad (3.2.8.3)$$

where c and T represent the specific heat capacity and temperature respectively; the subscripts *f, c* denote the process fluid and cooling water streams; *1, 2* denotes the inlet and outlet streams and *i* denotes each cooling block where i = 1,…,n – with n being the total number of coolers. The cooling water temperature range is set at 20°C - 25°C.

3.3 Heat Integration

Heat generated throughout the process can be transferred between streams and integrated within the process, in order to minimise utility requirements, environmental footprint and improve feasibility as a whole. An analysis under the first law of thermodynamics provides an initial estimate of the heat duty required, with a negative value of $Q_{min} = -9.80\text{MW}$ indicating excess heat requiring cooling, of the magnitude of at least 9.80MW. The heat exchanger network designed must also satisfy the second law of thermodynamics, resulting in the introduction of a minimum allowable temperature difference $\Delta T_{min} = 10K$. By constructing temperature intervals and the associated cascade diagram, any pinch points present restricting the transfer of energy can be identified. From the Grand Composite Curve in Figure 3.3.1, no pinch point is present and so heat can be exchanged across all streams involved. This leads to a total integration of the heating duty of 0.67MW, for a final cooling duty requirement of 9.80MW, representing a 12.1% improvement over the non-integrated system. The heat exchanger network designed to achieve this performance is illustrated in Figure 3.3.2.

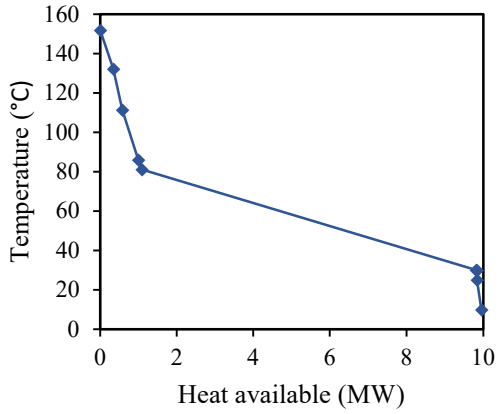


Figure 3.3.1: Grand Composite Curve for the simulated plant's network

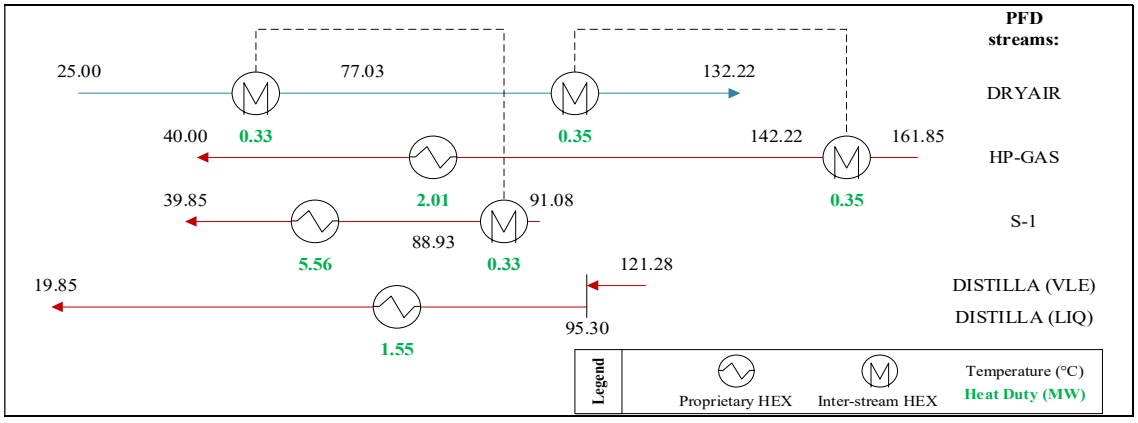


Figure 3.3.2: Heat-integrated heat exchanger network

3.4 Plant Layout

Table 3.4.1: Heat transfer areas for PCC heat exchangers

Heat Exchange Area (m²)						
2COOLWN2	**PRECOOL2**	**HEX**	**1COOLWN2**	**REC-COOL**	**REBOILER**	**DIS-COOL**
1.70	45.65	149.98	20.70	129.59	446.15	29.78

To achieve an accurate and representative plant sizing, sizing of the relevant equipment must be carried out – most of the equipment can be sized accurately in Aspen Plus. Heat exchangers, cooling towers and the compressor are manually sized to achieve specific sizing of these units. Heat exchangers make use of the design equation 3.4.1:

$$Q = U \cdot A \cdot \Delta T_{LM} \qquad (3.4.1)$$

with a minimum approach temperature of 10K selected to ensure feasible heat exchanger operation and areas and an overall heat transfer coefficient of 850Wm^{-2}K^{-1} (Aspen Plus, 2000), thus resulting in acceptable areas for all heat exchangers present. From this, the area A can be calculated from the heat duty Q and logarithmic mean temperature difference ΔT_{LM}. As such, the area (m^2) required for heat transfer is given in Table 3.4.1. For all cases, shell-and-tube heat exchangers are selected, being the industry standard and presenting a range of feasible operating conditions (Shah and Sekulic, 2003).

Hill provides correlations for the calculation of the base area of a cooling tower A_{CT} (Hill, Pring and Osborn, 1990), given by

$$A_{CT} = \frac{\dot{m}_c}{F_{loading}} \cdot F_{CR} \cdot F_{AP} \cdot F_{AT} \qquad (3.4.2)$$

with the liquid loading factor $F_{loading}$, ambient temperature factor F_{AT} , approach temperature factor F_{AP} and cooling range factor F_{CR} provided from the correlation plots A.3.4.1 and A.3.5.1 of Hill, Pring and Osborn, (1990). This results in a cooling tower base area of 108.5m^2.

The CO_2 compressor of magnitude 2MW can be included as a screw gas compressor package sourced from GEA, with dimensions 2.7 x 8.5 x 3.5m (Maggiore, 2016).

In the design of a plant layout, safety is a primary consideration where appropriate inter-unit and interequipment safety distances must be incorporated. These safety distances are given in Tables 3.4.2 (Lees, 2012) and 3.4.3 (IRI, 1991), outlining the minimum horizontal ground separation requirements in metres. Note that '/' implies no spacing requirement, and 'CP' that reference must be made to the relevant codes of practice; see Section C.6 of Mecklenburgh, (1986).

Table 3.4.2: Inter-unit safety distances

Cooling tower	Storage tank[1]	Process equipment[2]
7.5		
30	CP	
30	CP	2

1) High-pressure bullet
2) High-flash point

Table 3.4.3: Inter-equipment safety distances

Compressor	High hazard pump	Column, drum, accumulator	Air cooled HE	Heat Exchanger (HE)
9.144				
15.24	1.524			
15.24	4.572	4.572		
9.144	4.572	4.572	/	
9.144	4.572	3.048	4.572	1.524

4 Results

4.1 Plant Land Footprint

Implementation of the safety distances as detailed in Section 3.4 leads to a plant sizing shown in Figure 4.1.1.

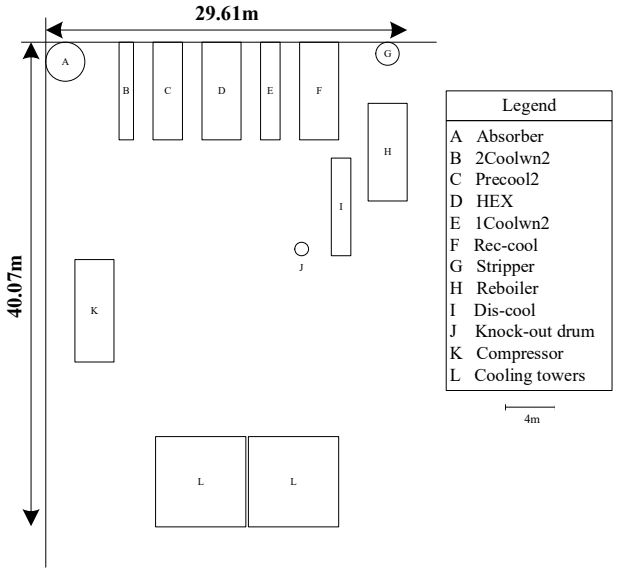


Figure 4.1.1: Diagram illustrating PCC plant layout and sizing

First, the largest diameter column – here being the absorber – is fixed into one corner of the plant layout. Following from this, the remaining process units can be incorporated in a sequence mirroring the process flow diagram as much as possible to achieve minimum spacing and consequent investment due to land requirements; increased piping, cabling, drainage and other support systems; and additional or larger pumps to overcome increased friction losses amongst other factors (Lees, 2012). The cooling system is then appended, defining the lower boundary of the plant layout. A rectangular layout is thus defined and optimised manually, achieving dimensions of 40.07 x 29.61m. This equates to a specific plant area of 232.6m^{2}MW^{-1}. By integrating the heat available to reduce the cooling requirements, energy

requirements are reduced; however, the plant area increases slightly due to the introduction of extra heat exchangers and the consequent separation imposed, thus extending the dimensions slightly.

Finally, the current plant does not include storage tanks – carbon capture processes often transport the captured CO_2 to suitable storage or usage sites through pipeline networks; storage can also typically be carried out via deep geological storage or mineral storage. However, some cases may call for the use of storage tanks, both for the solvent and the CO_2. 2 solvent storage tanks may be included of base dimensions 2.47 x 2.47m to allow for the flexible operation of the plant (Flø, Kvamsdal and Hillestad, 2016), decoupling the power plant and capture plant operation while partially decoupling the absorber and stripper. This can minimize the power plant energy penalty related to carbon capture when undergoing load changes to adapt to variations in energy demand. Additionally, bullet tank cars may be used to store the CO_2 on-site and allow for subsequent transportation. These can load a weight of 60tonnes with base dimensions of 2 x 6m, thus requiring 2 tanks to suit the process (APEC, 2009). Therefore, if these tanks are to be included in the plant, Figure 4.1.2 must be appended 34.63m south of the cooling towers, adhering to safety limitations as per Lees, (2012). This in turn leads to an updated total plant dimension of 109.65 x 29.61m.

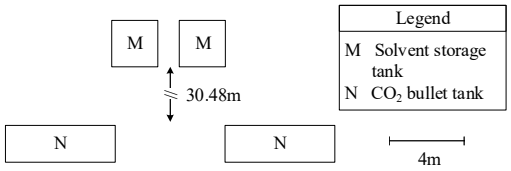


Figure 4.1.2: Diagram illustrating additional storage tank layout

As seen by these dimensions and Figures 4.1.1 and 4.1.2, a large proportion of the plant area is due to the safety limitations – this is particularly true with the inclusion of the storage tank layout. As such, formulating a formal optimization model to achieve the minimal feasible layout computationally may be favourable, and may result in a non-rectangular layout of decreased dimensions.

4.2 Energy Penalty

The impact of retrofitting PCC to a power plant can be evaluated through calculating the energy penalty: this can be measured in various ways. For this work, the energy penalty based on the overall reduction in electricity generated by the power plant, before and after retrofitting PCC, has been calculated. For 95% CO_2 capture rate, the total reduction in power output was found to be 45.7%. Compared to other fuel-fired power plants with 90% capture efficiency, this RDF-to-energy plant incurs an energy penalty of approximately double. This can be seen in Table 4.2.1. As well as this, the power plant efficiency penalty after retrofitting CCS was calculated: this was found to be 13.9% pts.

Table 4.2.1: Energy penalties for various fuel-fired power plants

Fuel Type	Energy Penalty (%)	Decrease in plant efficiency (% pts)
Coal	29.0	10.0
Natural Gas	21.0	7.9
RDF	45.7	13.9

4.3 Variation of RDF properties

RDF 4 (see Table 3.1.1) was chosen as the fuel to simulate in the model; however, the composition of RDF will often vary given the nature of this fuel source. Therefore, other RDF sources with different chemical compositions were inputted into this model, to compare how the varying characteristics influence the performance of the power plant. Figure 4.3.1 displays the results of how the overall plant efficiency varies with LHV of different RDFs – a larger LHV allows for more efficient combustion and thus greater electricity generation for a given amount of fuel input. Variation in the proximate and ultimate analyses of RDFs will also influence the efficiency of the fuel source, suggesting why there is a non-linear trend between LHV and plant efficiency.

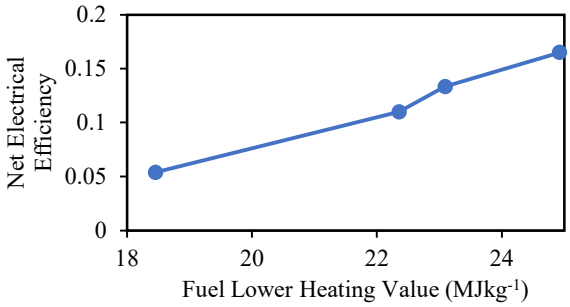


Figure 4.3.1: Lower heating value (LHV) of 4 different RDF sources as a function of net electrical efficiency of an RDF-to-energy plant retrofitted to PCC

5 Discussion

The main objectives to assess the CCR of the RDF-to-energy plant simulated in this work include having sufficient space available on or near the power plant site to accommodate carbon capture equipment, as well as being technically feasible. As displayed in Figure 5.1, retrofitting PCC still allows for a production of over 5MW_e, however a significant energy penalty of 45.7% is imposed. The most energy consuming parts of the process, owing to more than 90% of the total energy consumed in the process, is the regeneration energy of the stripper reboiler and the compression of CO_2 to 100bar, accounting for over 50% and 40% respectively.

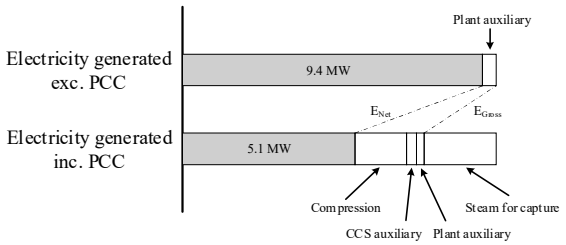


Figure 5.1: Electricity generation exc./inc. PCC

Despite obtaining an optimised reboiler energy consumption of 4.03GJton^{-1}(CO_2), in alignment with literature, there is a significant decrease in the electricity that can be generated by the power plant as a large flow of steam is bled from between the turbines: therefore, efficiency of the overall process is impacted. This is somewhat expected, given the reported high regeneration energy of MEA solvent. However, a larger energy penalty is observed for the RDF-to-energy plant than for coal and natural gas-fired power stations. This is most likely due to having a less efficient fuel source coupled with smaller scale operation, thus the reduction in power generation is more substantial. The overall efficiency of this RDF power plant, before retrofitting PCC, was found to be approximately 30%, whereas coal and natural gas fired power stations report efficiencies of over 35% and 45% respectively. A less efficient fuel means that more CO_2 is emitted per MW_e produced from combustion, requiring higher capture energy. However, RDF does contain a non-negligible biogenic fraction, hence CO_2 emissions relating to this amount can be deducted from the net CO_2 emissions. This implementation can lead to net negative CO_2 emissions, making it a greener technology than CCS retrofitted to fossil fuel-fired power stations.

The variation of RDF properties will further impact the overall plant efficiency, as shown by varying LHVs of different RDFs in Figure 4.3.1. The chemical composition of RDF is not predictable and thus more unfavourable energy penalties are likely to be obtained in practice than what has been found here, where the process has been optimised for one RDF type.

The land requirement associated with the retrofit of this PCC plant is significant, at 232.6m^{2}MW^{-1}, which is a barrier towards implementing PCC as a carbon capture solution. Typical literature values for the footprint of similar plants present an almost tenfold decrease in specific area, with a 1200MW coal-fired power station case study exhibiting a requirement of 35.86m^{2}MW^{-1} for the inclusion of a CCS outfit achieving a similar level of capture as to what is explored in this work (APEC, 2010). Whilst this specific area relates to a coal-fired station, as opposed to a waste-to-energy plant as studied here, this discrepancy is unlikely to account for the entire difference. It is also expected for the specific area to decrease as the power capacity increases, thus becoming more feasible and contributing to the large discrepancies observed when drawing comparisons to large-scale power plant carbon capture outfits. Indeed, based on Fennell's work (Bui *et al.*, 2018), in the range of 300MW up to 1500MW, the optimal specific plot area decreases from 28m^{2}MW^{-1} to 25m^{2}MW^{-1}. The relatively large sizing of the considered PCC plant is first explained by the manual optimisation carried out, as opposed to employing a formal optimisation program. Additionally, flexibility in selecting, sizing and placing cooling towers as required will allow for decreased areas. Moreover, safety distances are the leading factor in the inflated specific area. The influence of safety distances on overall plant size will be minimised as the power output increases, occupying a decreasing proportion of the total plant area and allowing for more efficient plot sizing.

6 Conclusions

In this paper, an RDF-to-energy plant retrofitted to an MEA-based PCC process, processing 4,500kghr^{-1} of RDF, was simulated. An energy penalty of 45.7% was imposed on the process, operating with an MEA solvent concentration and lean loading set at 30wt% and 22% respectively. Such conditions gave a reboiler duty of 4.03GJton^{-1}(CO_2), supplied by a flow of high temperature steam with a ratio of 1.81 $\frac{kg(steam)}{kg(CO_2\ captured)}$. Despite the PCC process operating optimally, retrofitting PCC to a small-scale power plant whilst utilising a fuel of lower efficiency than fossil fuels proved to have a significant effect on the plant performance. Whilst positive net electricity generation is still achievable, the physical land requirement is disproportionately large owing to the inflated specific plot area. This presents complications towards the feasibility of the project when considering the physical implementation. Further considerations into the geographical storage of CO_2, the technical feasibility of transporting CO_2 to storage areas and the economic feasibility of the process are required to assess the overall carbon capture readiness of this RDF-to-energy plant.

References

Aboudheir, A. *et al.* (2003) 'Kinetics of the reactive absorption of carbon dioxide in high CO2-loaded, concentrated aqueous monoethanolamine solutions', *Chemical Engineering Science*, 58(23–24). doi: 10.1016/j.ces.2003.08.014.

Alie, C. *et al.* (2005) 'Simulation of CO2 capture using MEA scrubbing: A flowsheet decomposition method', *Energy Conversion and Management*, 46(3). doi: 10.1016/j.enconman.2004.03.003.

APEC (2009) *BUILDING CAPACITY FOR CO2 CAPTURE AND STORAGE IN THE APEC REGION A training manual for policy makers and practitioners APEC Energy Working Group Asia-Pacific Economic*

Cooperation Building capacity for CO 2 capture and storage in the APEC region iii.
APEC (2010) *Planning and Cost Assessment Guidelines for Making New Coal-Fired Power Generation Plants in Developing APEC Economies CO 2 Capture-ready*. Available at: www.aurecongroup.com.
Aspen Plus (2000) *Aspen Plus ® Aspen Plus User Guide*. Available at: http://www.aspentech.com.
Bisinella, V. *et al.* (2021) 'Environmental assessment of carbon capture and storage (CCS) as a post-treatment technology in waste incineration', *Waste Management*, 128, pp. 99–113. doi: 10.1016/J.WASMAN.2021.04.046.
Bui, M. *et al.* (2018) 'Carbon capture and storage (CCS): The way forward', *Energy and Environmental Science*. doi: 10.1039/c7ee02342a.
Chemtech, S. (2005) 'Structured packings for Distillation, Absorption and Reactive Distillation', *Structured Catalysts and Reactors*.
'COP 27' (2022) in.
Efika, E. C., Onwudili, J. A. and Williams, P. T. (2015) 'Products from the high temperature pyrolysis of RDF at slow and rapid heating rates', *Journal of Analytical and Applied Pyrolysis*. doi: 10.1016/j.jaap.2015.01.004.
Environ Consultants Ltd. (no date) *UK Refuse Derived Fuel (RDF) Energy from Waste*, *Envrion Consultants LTD*.
European Commission (2011) *Energy Roadmap 2050*. Available at: http://www.adimenlehiakorra.eus/documents/29934/31635/EU+-+Energy+Roadmap+2050+-+Impact+Assessment+and+Scenario+Analysis.pdf/01c8f12f-f9dc-4f7f-87bd-a25c0eacbeb3.
Flø, N. E., Kvamsdal, H. M. and Hillestad, M. (2016) 'Dynamic simulation of post-combustion CO2 capture for flexible operation of the Brindisi pilot plant', *International Journal of Greenhouse Gas Control*, 48. doi: 10.1016/j.ijggc.2015.11.006.
García, R. *et al.* (2021) 'Co-pelletization of pine sawdust and refused derived fuel (RDF) to high-quality waste-derived pellets', *Journal of Cleaner Production*, 328. doi: 10.1016/j.jclepro.2021.129635.
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*. doi: 10.1016/j.apenergy.2013.05.020.
Grammelis, P. *et al.* (2009) 'Pyrolysis kinetics and combustion characteristics of waste recovered fuels', *Fuel*, 88(1). doi: 10.1016/j.fuel.2008.02.002.
Hill, G. B., Pring, E. J. and Osborn, P. D. (1990) 'Cooling tower theory and calculations', *Cooling Towers*, pp. 129–161. doi: 10.1016/B978-0-7506-1005-6.50007-1.
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', in *Energy Procedia*. doi: 10.1016/j.egypro.2009.01.202.
IPCC (2018) *Global Warming of 1.5C*. Available at: https://www.ipcc.ch/sr15/.
IRI (1991) 'Plant Layout and Spacing for'.
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.
Kiranoudis, C. T., Maroulis, Z. B. and Marinos-Kouris, D. (1996) 'Drying of solids: Selection of some continuous operation dryer types', *Computers and Chemical Engineering*, 20(SUPPL.1). doi: 10.1016/0098-1354(96)00040-3.
Lees, F. P. (2012) *Lees' Loss Prevention in the Process Industries*, *Lees' Loss Prevention in the Process Industries*. doi: 10.1016/c2009-0-24104-3.
Lin, Y. J. and Rochelle, G. T. (2014) 'Optimization of advanced flash stripper for CO2 capture using piperazine', in *Energy Procedia*. doi: 10.1016/j.egypro.2014.11.160.
Maggiore, C. (2016) *Gas compression solution Power plant and Oil & Gas applications*.
Materazzi, M. *et al.* (2015) 'Fate and behavior of inorganic constituents of RDF in a two stage fluid bed-plasma gasification plant', *Fuel*, 150. doi: 10.1016/j.fuel.2015.02.059.
Mecklenburgh, J. C. (1986) 'Process Plant Layout', *Journal of Pressure Vessel Technology*, 108(2). doi: 10.1115/1.3264778.
Moioli, S. and Pellegrini, L. A. (2013) 'Regeneration section of CO2 capture plant by MEA scrubbing with a rate-based model', in *Chemical Engineering Transactions*. doi: 10.3303/CET1332309.
Mumford, K. A. *et al.* (2015) 'Review of solvent based carbon-dioxide capture technologies', *Frontiers of Chemical Science and Engineering*. doi: 10.1007/s11705-015-1514-6.
Purushothama, B. (2009) *Humidification and ventilation management in textile industry*, *Humidification and Ventilation Management in Textile Industry*. doi: 10.1533/9780857092847.
Puxty, G. *et al.* (2009) 'Carbon dioxide postcombustion capture: A novel screening study of the carbon dioxide absorption performance of 76 amines', *Environmental Science and Technology*, 43(16). doi: 10.1021/es901376a.
Shah, R. K. and Sekulic, D. P. (2003) *Fundamentals of Heat Exchanger Design*, *Fundamentals of Heat Exchanger Design*. doi: 10.1002/9780470172605.
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. doi: 10.1016/j.ijggc.2017.06.001.
UNFCC (2022) 'Climate Plans Remain Insufficient: More Ambitious Action Needed Now', 26 October. Available at: https://unfccc.int/news/climate-plans-remain-insufficient-more-ambitious-action-needed-now.
UNFCCC (2015) *Paris Agreement*. Available at: https://unfccc.int/files/essential_background/convention/application/pdf/english_paris_agreement.pdf.
Xie, K. *et al.* (2019) 'Recent progress on fabrication methods of polymeric thin film gas separation membranes for CO2 capture', *Journal of Membrane Science*. doi: 10.1016/j.memsci.2018.10.049

Feasibility Study on the ZESTY Reactor for Production of Direct Reduced Iron with Hydrogen

Javier Monteliu and Jien Feung Jason Goh
Department of Chemical Engineering, Imperial College London, U.K.

Abstract Calix Ltd. have partnered with the Imperial College London Fennell Group to advance research on their patented Zero Emissions Steel TechnologY (ZESTY) for the production of hydrogen-based direct reduced iron. This study aimed to propose a feasible ZESTY-iron reactor between **10-50 m** in height, with target annual productions of **30,000 tonnes of iron** from hematite ore. Sensitivity analyses were conducted on an integrated kinetic Aspen Plus model to investigate the effects that reaction temperature, solid inlet temperature, pressure and hydrogen excess ratio have on the key performance metrics of residence time and heat duty. Reactor sizing feasibility was then assessed to filter out inadequate configurations that did not comply with Calix's requirements. Process simulations revealed that there is an optimal range of residence times which leads to feasible reactor designs. Operating under minimal hydrogen excess, at elevated pressures and decreased temperatures showed promising results in terms of leading to smaller, more energy efficient reactors. Ultimately, this study provided proof of concept for the ZESTY-iron reactor by proposing a feasible reactor configuration measuring **26.66 m** in height with **4.44 m** in diameter and supplied a foundational simulation model from which Calix Ltd. can expand their Basis of Design for their demonstration plant.

Keywords Calix Ltd., Direct Reduced Iron, Aspen Plus, Decarbonisation, Hematite Ore

1 Introduction

Human-induced climate change is a dire threat to the natural world and societies, with carbon dioxide (CO_2) being the principal greenhouse gas contributing to global warming [1], [2]. Leading research institutes are working to develop new ways to reduce the emissions of CO_2 within the top greenhouse gas-producing industries. One such example is the iron and steel industry, which constitutes 11% of global CO_2 emissions [3]. Key conclusions from COP27 indicate that this sector must experience a 25% reduction from its current CO_2 emissions by 2030 to comply with objectives set in the Paris Agreement [4].

According to the Global Steel Plant Tracker (GSPT), 91.2% of the world's iron is manufactured via blast furnace technology [3]. This process uses coal and coke to synthesize carbon monoxide as a reducing agent and emits vast amounts of CO_2. The next best production alternative is Direct Reduction of Iron (DRI). This method currently uses natural gases to manufacture a mixture of hydrogen and carbon monoxide which is then reacted with high-grade iron ore to produce sponge iron. These differences allow DRI to emit up to 67% less CO_2 when compared to blast furnaces and aid in the road to achieving COP27 goals [5]. Additionally, DRI production is flexible; it is theoretically capable of operating using pure H_2 as a single reducing agent, effectively accomplishing near-zero CO_2 emissions. Unfortunately, one major hurdle is the production of clean H_2. The optimal method would be green H_2, but it will take time before it becomes cost competitive. As such, the general plan to achieving net zero in the iron and steel industry is to shift away from carbon-intensive blast furnaces, towards natural gas-based DRI, before transitioning to green hydrogen-based DRI [6].

The recent conflict in Ukraine has led Russia, the biggest producer of natural gas, to interrupt supply chains and create a natural gas shortage. Uncertainty with supply lines has incited fear that severely disrupts DRI operations, thus fuelling the search for hydrogen-based DRI production [6]. The current leaders of DRI, MIDREX, are pushing forward with a recently announced agreement with H2 Green Steel to produce the world's first commercial 100% hydrogen-based DRI plant [7]. Other companies such as Calix Ltd. are also looking to develop their own methods for hydrogen-based DRI.

Calix Ltd. has partnered with the Imperial College London Fennell Group to advance research for their patented Zero Emissions Steel TechnologY (ZESTY) reactor. It has recently received funds to conduct a design study for a ZESTY-iron demonstration plant capable of producing 30,000 tonnes of iron per annum [8]. This is a major step forward for Calix Ltd. since it will provide proof of concept for their hydrogen-based DRI counter-current reactor. The objective of this study was to establish the design basis for the ZESTY-iron reactor directly applicable to their demonstration plant. Key design parameters affecting the reduction process were investigated through simulations in Aspen Plus, and overall conclusions can be used to take this project forward into the demonstration phase.

2 Background

Previous partnerships between the Fennell Group and Calix Ltd. focused on a thermodynamic feasibility study on Aspen Plus with limited kinetic modelling of the ZESTY-iron reactor on MATLAB. The thermodynamic investigations concluded that a counter-current moving bed reactor operating between 600-800 °C was

feasible [9]. However, preliminary kinetic studies conducted by Tkachenko found that these temperatures would result in unreasonable residence times, exceeding ZESTY's 10-50 m height guidelines. This study concluded that temperatures in the range of 950-1050 °C would be required to achieve a reduction degree of hematite, X_{hem}, of 90% [10]. From a feasibility standpoint, nickel alloys used for construction of Calix's reactors have an operational upper temperature limit of 1000 °C [11]. To follow on from conclusions drawn by Tkachenko, it was decided that a temperature range between 900-1000 °C would be explored for this project. Thus, a novel reactor design for these temperatures must be proposed.

Outside of past work conducted by the Fennell group in collaboration with Calix Ltd., there is an exhaustive list of kinetic studies on DRI of iron oxides, as reviewed by Spreitzer [12]. These past research studies explore a diverse range of reaction conditions and reducing agents. However, for the purposes of this investigation, only the reduction of hematite into iron using H_2 was considered. This section details the reduction reactions and kinetic expressions pertinent to this research paper.

2.1 DRI from Hematite

DRI from hematite (Fe_2O_3) by H_2 is not a single-step reaction. Instead, at temperatures higher than 570 °C, there is a stepwise reduction pathway from hematite to magnetite (Fe_3O_4), followed by a further reduction to wustite ($Fe_{(1-x)}O$), before iron (Fe) is formed. Note that (1-x) in wustite's chemical formula represents atomic vacancies within the lattice [12]. Table 1 showcases the full stepwise pathway with the overall reduction reaction.

Table 1. Reduction pathway of hematite to iron, including the overall reaction (reaction 4) [12], and the corresponding enthalpies of reaction at 800 °C [13].

Number	Reaction Equation	$\Delta_r H_{900°C}$ (kJ/mol)
1	$3Fe_2O_3 + H_2 \rightarrow 2Fe_3O_4 + H_2O$	-6.02
2	$(1-x)Fe_3O_4 + (1-4x)H_2 \rightarrow 3Fe_{(1-x)}O + (1-4x)H_2O$	46.64
3	$Fe_{(1-x)}O + H_2 \rightarrow (1-x)Fe + H_2O$	16.41
4	$Fe_2O_{3(s)} + 3H_{2(g)} \rightarrow 2Fe_{(s)} + 3H_2O_{(g)}$	57.03

2.2 Kinetic Rate Law

It was important to first identify the main rate-controlling process for the reduction of hematite at temperatures between 900-1000 °C. Turkdogan conducted kinetic studies for hematite particle sizes of 800 μm at 0.96 atm under pure H_2 conditions. Temperatures tested ranged from 300 °C to 1100 °C. It was concluded that above 900 °C the reaction is only limited by intrinsic kinetics up to X_{hem} of 90%, provided particle sizes remain smaller than 800 μm [14]. Above this size, diffusional limitations hinder the reduction reaction.

The kinetic rate equation given by Eq.1 retrieved from Chen accounts for the overall reaction under kinetic limiting conditions [15]. This expression has been reformulated in terms of the rate of hematite reduction, r_{hem}, to adapt it for use within Aspen Plus, and was the governing rate law for modelling within this study.

$$r_{hem} = -\frac{dC_{hem}}{dt} = k_o e^{\frac{-E_a}{RT}} \left(p_{H_2} - \frac{p_{H_2O}}{K_{eq}} \right) C_{hem}, \quad (1)$$

where R is the ideal gas constant; T represents temperature in Kelvin; C_{hem} is the molar concentration of hematite in the solid phase; k_o is a pre-exponential term and E_a is the activation energy, both empirically determined to equal 4.41×10^7 s^{-1} atm^{-1} and 214 kJ mol^{-1} respectively; p_{H_2} and p_{H_2O} are the partial pressure of hydrogen and water in atm; and K_{eq} is the equilibrium constant for reaction 3.

K_{eq} is included in the kinetic rate equation since the reduction of wustite with hydrogen has been experimentally found to be heavily influenced by thermodynamics [15]. The temperature dependence relationship for K_{eq} was obtained from [16],

$$K_{eq} = \frac{p_{H_2O}}{p_{H_2}} = exp\left(\frac{-2070}{T} + 1.3 \right). \quad (2)$$

Kinetic experiments conducted by Chen were performed at 0.85 atm, 1150-1350 °C, with mean particle sizes of 21 μm. It is understood that the work throughout this project extrapolates the use of this kinetic rate law at different conditions. However, more adequate reaction kinetic models were not found in literature and specific experimentation exploring the accuracy of this extrapolation must be conducted in future.

3 Methodology

3.1 Aspen Plus Model

Aspen Plus was selected for process simulation given its ability to effectively model solids–gas reactions in a flowsheeting environment with useful built-in unit operations. However, Aspen Plus has two main limitations for modelling the ZESTY reactor: 1. difficulty with modelling counter-current reactors, 2. long simulations run-times for complex models with many reactor blocks and sensitivity tests. These limitations were considered whilst developing the ZESTY reactor model.

The conceptual design of the ZESTY reactor consists of 3 preheating stages and 12 heating stages operating isobarically with the same residence time each.

Aspen Plus' inability to model a counter-current reactor was overcome by utilising Continuous Stirred Tank Reactor (CSTR) blocks to approximate each of the stages. Gas and solid products from each CSTR were then interchangedly fed in a counter-current arrangement. Fresh hematite and gas feeds were input into stage 1 and 15 respectively. To overcome the second Aspen limitation of long run-times, a simplified 7-CSTR model was first used to run sensitivity tests before optimal configurations were simulated on a complete 15-CSTR model.

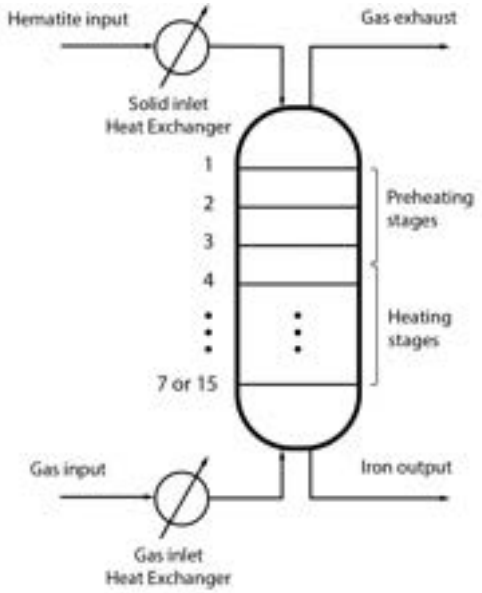


Figure 1. Simplified Aspen Plus ZESTY reactor model.

For the 7-CSTR model, CSTRs 1-3 represent the 3 preheating stages while CSTRs 4-7 each represent 3 stages of the reactor grouped together. The decision for this design stems from the understanding that temperature is an important variable for the reduction reaction. During the preheating stages, heat is exchanged from the hot gases rising up from stage 4 and the cooler solid inlet entering at stage 1. Having an adequate temperature profile across these stages is therefore key for a better representation of the actual system. Meanwhile, the heated stages are set to operate approximately isothermally so temperature differences are less.

To implement the governing kinetic rate law for this project given by Eq.1, the custom rate kinetics feature within Aspen was used. This was the case since the rate expression by Chen did not fit under the standard rate kinetic formats available in Aspen Plus. Another aspect of customisation was required to input certain physical properties of hematite. The component databases in Aspen Plus provided incomplete molar volume and vapour pressure relationships. Therefore, these were manually hardcoded so modelling calculations computed a negligible solid vapour pressure and a constant hematite density of 5100 kg m^{-3} [17].

To assess the accuracy of the modelling conducted, the 7-CSTR model was tested against the experimental data supplied by Chen [15]. The experimental X_{hem} was plotted against the values obtained from Aspen Plus simulations. Overall an R^2 value of 0.995 was observed for the regressed linear relationship, validating the use of this model throughout this study.

3.2 Feasibility Test Approach

Feasibility tests were conducted with the aim of identifying the effect different key process operational paramaters had on the design of the ZESTY-reactor. The main paramaters studied during this investigation were the temperature profile across the heated stages, temperature of the solid inlet, $T_{solid,in}$, pressure, P_T, and H_2 excess ratio. The H_2 excess ratio refers to the number of multiples H_2 input flowrate is from the stoichiometric minimum. The impact of the studied parameters was investigated through two main perspectives: from a kinetic perspective, with aims of minimising reaction residence time, and energy requirements; and from a physical design perspective, sizing reactors that complied with Calix's specifications.

3.2.1 Target Parameters

Target parameters for the feasibility tests were an output production of 30,000 tonnes of iron per operating year (an operating year taken to be 8000 h/yr), 90% overall reduction of input hematite, a reactor size aspect ratio of at least 6 with a reactor height between 10–50 m, and minimal operational energy requirements. These targets were extracted from the ZESTY patent, past research conducted by Tkachenko and well-known reactor engineering guidelines from Douglas [8], [10], [18]. For modelling in Aspen Plus, the overall conversion goal was set throgh design specification blocks, whilst an input hematite feed was calculated to achieve the constant desired throughput.

3.2.2 Identification of Kinetic Relationships

To establish the effect the aforementioned parameters have on residence time and system heat duty, kinetic sensitivity tests were conducted using the 7-CSTR model. Performing sensitivity tests which vary all four operational parameters simultaneously was computationally expensive. Hence, a methodical approach was required to progressively build the understanding of the effect each operational parameter has. For this purpose, each of these were varied independently whilst holding all other constant. As a starting point, preliminary values for operational parameters were extracted from past studies from Ching and Tkachenko [9], [10]. Sensitivity tests were conducted on one operational parameter at a time and kinetically optimal values that minimised residence time and energy were recycled back into the next testing. Overall conclusions for the four operational pa-

rameters were obtained and forwarded onto the sizing of feasible reactors.

3.2.3 Reactor Sizing – Simplifying Assumptions

Reactor sizing was split into two main parts. The first part was focused on a rapid assessment of the reactor system as a whole using the 7-CSTR model. This allowed for a quick comparison between different reactor designs and analysis of the trends each operational parameter has on reactor height and aspect ratio. For this initial assessment a few key assumptions were considered: 1. reactor is isothermal across all stages, 2. physical property parameters are constant at inlet conditions – pure hematite and pure H_2 feeds. This preliminary screening test aided in identifying which reactors were likely to be operationally feasible and would require more comprehensive sizing.

The second part of this reactor sizing exercise utilises the 15-CSTR model and aims to provide a more accurate sizing calculation for reactor configurations of interest. Through modelling of individual stages, Aspen Plus was better able to map temperature profiles and changing physical composition properties at each stage for the ongoing reaction. These values were extracted and used for the sizing of each stage, where the overall reactor height was calculated through the summation of all individual stage heights. Note that the diameter of the reactor, d_r, was held constant throughout. The physical properties of interest include particle density, ρ_p, fluid density, ρ_f, and fluid viscosity, μ_f.

Fundamental concepts and calculations for both reactor sizing sections were identical and are explained through a general approach in Section 3.2.4. Given the inaccuracy of modelling the counter current reactor on Aspen Plus, residence times were obtained through simulations and reactor dimensions were then calculated on a separate Excel document.

3.2.4 Reactor Sizing – Calculations

Key variables that were considered for the sizing of reactors are particle diameter, d_p, residence time, τ, and apparent terminal falling velocity of the solids, u_t. The particle diameter was held constant at a conservative estimate based on explanations regarding particle size limitations in Section 4.2.1. Residence time values were obtained through Aspen Plus simulations, while u_t was derived from Stokes' law,

$$u_t = \frac{1}{18}\frac{d_p^2 g\left(\rho_p - \rho_f\right)}{\mu_f}. \tag{3}$$

Estimating u_t requires several assumptions to be considered. The first was that the flow regime in the reactor was laminar. For falling spherical smooth solids, particle Reynolds number, Re_p , must be less than 0.4 [19]. The second assumption was that constant pressure, temperature and velocity were maintained throughout the reactor/stage. Lastly, it was assumed that most particles would drop along the centreline of the reactor and thus peak gas velocity, $\hat{u}_g$, should be considered in place of mean gas velocity, $\bar{u}_g$ [15].

When calculating peak gas velocity, the ideal gas law was employed. This is reasonably accurate for modelling of high temperature gases at low pressures like in this case. Volumetric gas flowrate, $\dot{V}_g$, is given by Eq.4,

$$\dot{V}_g = \frac{\dot{n}_{T,gas} RT}{P_T}. \tag{4}$$

Following reaction stoichiometry of reaction 4, the molar ratio of gaseous reactants is equal to gaseous products. Hence, total molar gas flowrate, $\dot{n}_{T,gas}$, was constant throughout the reaction and calculated using reactor inlet conditions as given by Eq.5,

$$\dot{n}_{T,gas} = \dot{n}_{H_2O,o} + 3E_{H_2}\dot{n}_{hem,o}X_{hem}, \tag{5}$$

where $\dot{n}_{H_2O,o}$, is the inlet molar flowrate of water, E_{H_2} is the specified excess hydrogen ratio, $\dot{n}_{hem,o}$ is the inlet molar flowrate of hematite, and X_{hem} is the overall reduction degree of hematite. Considering ZESTY is a tubular reactor and using Eq.4 and 5, an expression for the peak gas velocity was derived. Note that flow was taken to be fully developed so $\hat{u}_g$ is twice $\bar{u}_g$.

$$\hat{u}_g = -\frac{8\left(\dot{n}_{H_2O,o} + 3E_{H_2}\dot{n}_{hem,o}X_{hem}\right)}{P_T \pi d_r^2} RT. \tag{6}$$

Since τ measures the total residence time for solid particles, particle falling velocity, u_p, was employed to calculate the reactor height, l, shown in Eq.8,

$$u_p = u_t + \hat{u}_g, \tag{7}$$

$$l = u_p \tau. \tag{8}$$

Overall reactor diameter, d_r, was varied and reactor/stage heights were computed. Note that in Section 3.2.1, one target parameter is to design tubular reactors with an aspect ratio of at least 6. Thus, a second reactor height was calculated by using d_r and an aspect ratio of 6. The Excel Solver function was then utilised to minimise the squared error (SE) between these two heights. For converging scenarios, the reactor configuration was deemed feasible, however, if no convergence was reached, it was deemed infeasible. This analysis guided which reactor configurations should be brought forward to be simulated on the 15-CSTR model for a more comprehensive reactor sizing.

3.3 Energy Profile

From an efficiency and sustainability point of view, it was important to consider limiting the operating energy input into the reactor system. The previous study conducted by Ching investigated the effect that different fuelling profiles would have on reactor performance

using a thermodynamic basis [9]. This study tested the three energy profiles proposed in Ching's report and presents a new profile that considers the kinetic relationships observed over the course of this research. The four energy profiles are presented in Figure 2.

The methodology for this section is similar to Section 3.2.2. All energy profile tests were first performed on the 7-CSTR model to identify the optimal profile for further reactor sizing using the 15-CSTR model. This two-step process was implemented due to Aspen Plus limitations stated in Section 3.1.

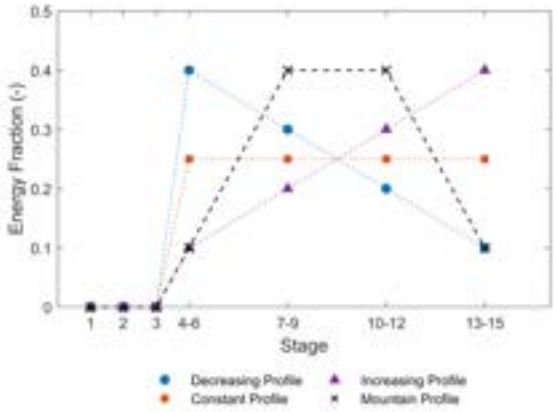


Figure 2. Energy profiles tested in the 7-CSTR model.

4 Results and Discussion

4.1 Kinetic Feasibility Analysis

4.1.1 Temperature Profile

The first step to developing a kinetically optimal reactor was to select a temperature profile for the heated stages. In Section 2.1, it was discussed that this model simplifies the step-wise pathway into a single-step endothermic reaction. Having higher temperatures across the reactor was therefore predicted to result in faster kinetics as well as favouring thermodynamics of the model. To validate this, a sensitivity analysis for various temperature combinations across the different heated stages were run on the 7-CSTR model with initial operational parameters fixed at a $T_{solid,in}$ of 400 °C, pressure of 1 atm, H_2 excess ratio of 3.5 and no water (H_2O) input at the inlet. These parameters were taken from past studies conducted for Calix Ltd. by Ching and Tkachenko [9], [10]. Temperatures of heated CSTRs were varied among 900 °C, 950 °C and 1000 °C.

It was found that residence times decreased when average temperature of the heated stages were increased; confirming the initial hypothesis that higher temperatures led to lower τ. There were slight differences between the residence times of temperature profiles with similar average temperatures but a differing energy distribution. This hints at the importance of testing various energy profiles, studied in Section 4.3. Ultimately, the best performing temperature profile was 1000 °C isothermal operation, which had a τ of 81.1s. Thus, this profile was used for further kinetic feasibility studies.

4.1.2 Solid Inlet Temperature

This section studies the effect that $T_{solid,in}$ has on τ and the heat duty required to heat the solid feed to the specified temperature of stage 4, Q_{solid}. The latter is the sum of two different contributions: 1. the heat duty required to preheat the solid inlet to specified $T_{solid,in}$, 2. additional energy supplied to stage 4 to maintain its set temperature. For this sensitivity analysis, all heated reactors were held isothermally at 1000 °C, at a pressure of 1 atm, H_2 excess ratio of 3.5, and no H_2O input.

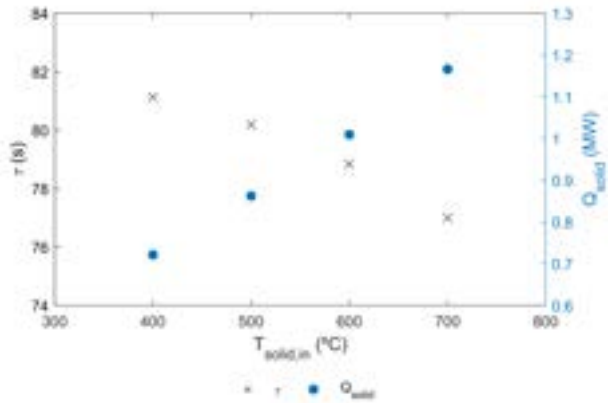


Figure 3. Variation of residence time, τ, and total solid heating duty, Q_{solid}, for different hematite feed temperatures, $T_{solid,in}$, at 1 atm, 3.5 H_2 excess ratio, with no water at the inlet, and 1000 °C isothermal operation of the heated stages.

Figure 3 showcases how τ and Q_{solid} change as $T_{solid,in}$ is varied between 400-700 °C at 100 °C intervals. For higher inlet solid temperatures, residence times were seen to slightly decrease. This follows since higher $T_{solid,in}$ results in higher preheating stage temperatures and thus a mild increase in reduction rates. However, higher inlet solid temperatures led to a significant increase in the total solid heating duty. This is because at high $T_{solid,in}$ there is a small temperature gradient between the exiting gas stream and the incoming solids, leading to minimal heat transfer between phases. Therefore, heat is lost through the gas phase leaving the reactor and more energy has to be supplied into the system. Meanwhile, lower solid inlet temperatures had a larger temperature gradient and heat losses were reduced.

Given the conflicting trends between optimising reactor residence times and solid heating duty, all $T_{solid,in}$ tested were kept under consideration.

4.1.3 Pressure

From simulations, a directly proportional relationship between the residence time and the inverse of pressure

was observed, depicted in Figure 4. This trend follows dependencies shown in Eq.9 extracted from Chen's kinetic rate law, where there exists a directly proportional dependency between r_{hem} and P_T,

$$r_{hem} = -\frac{dC_{hem}}{dt} \propto P_T \left(y_{H_2} - \frac{y_{H_2O}}{K_{eq}} \right). \qquad (9)$$

The rate of reaction is inversely proportional to the residence time for a constant overall reduction degree. Therefore, by inference, τ is inversely proportional to the pressure of the system.

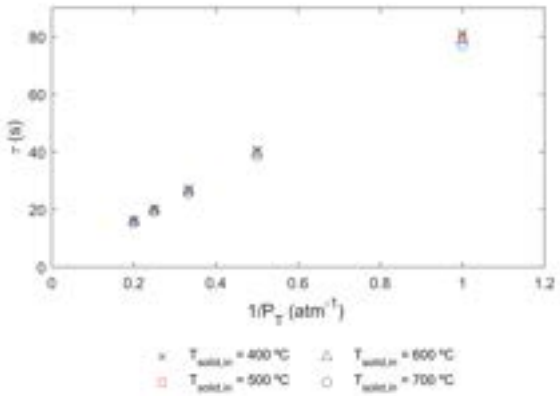


Figure 4. Resulting residence time for different pressures and solid inlet temperatures with a dry H_2 inlet of 3.5 excess ratio and 1000 °C isothermal operation of heated reactor stages.

Figure 4 also highlights the relative impact pressure has on τ compared to $T_{solid,in}$. It is very clear that higher pressures would optimise reaction kinetics and minimise τ much more than by increasing solid inlet temperatures. However, in Section 2.2 it was discussed that Chen's kinetic rate equation is based on experiments conducted at near atmospheric pressures. Further experimentation must be conducted to reaffirm this relationship at higher pressures. Thus, at this point, it is preferrable not to stray too far away from Eq.1 conditions and to limit pressure below 3 atm. Additionally, given the early stages of this reactor design, it would be wise to limit the capital expenditure and operational costs. Gas compressors are CAPEX and OPEX intensive which makes high operating pressures unattractive [18]. The target sources of hydrogen for this reactor are blue or green hydrogen, which can be supplied between 1–3 atm. This further reinforces the idea of keeping pressures below 3 atm for future stages.

4.1.4 H_2 Excess Ratio

The partial pressure of hydrogen has a first order dependency towards the reaction kinetics as shown in Eq.1. Thus, it is important to understand the effect H_2 excess ratio has on τ for constant pressure, as it is varied between 3 and 5 times the stoichiometric minimum.

For the sake of brevity, only 1 atm pressure data points were included in Figure 5. This graph conveys two main trends. The first shows that τ decreases by 22% on average for an increasing H_2 excess ratio between 3.5-5, measured over all $T_{solid,in}$. The second is that as $T_{solid,in}$ values are increased, similar to that seen in Figure 4, there is a small mean decrease of 5.1% in τ. Based on these trends in isolation, it might appear that designing a system with high excess reactant is optimal. However, no consideration has been made towards energy feasibility, where larger gas flowrates require a higher heating duty. Another disadvantage of higher H_2 excess ratios is that equipment, such as reactors or separators when designing a recycle system, are much larger in size [18]. These factors increase the capital and operating expenditure and would heavily affect the economic feasibility of this process. Thus, it is better at this stage to design for low H_2 excess ratios.

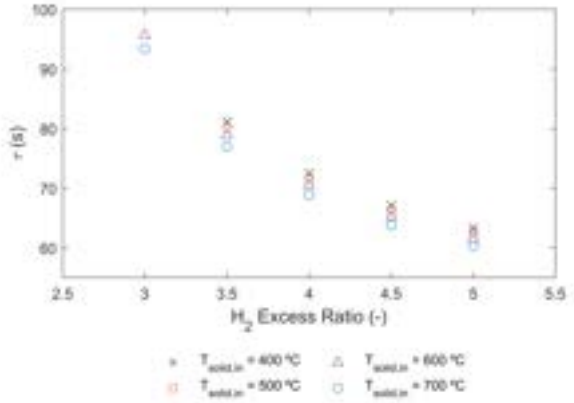


Figure 5. Variation of residence times for different H_2 excess ratios and $T_{solid,in}$ at 1 atm, dry inlet gas feed, and 1000 °C isothermal reactor operation for the heated stages.

In Figure 5, there are two missing result points at H_2 excess ratios of 3. This is because the simulation could not reach the target reduction degree and resulted in errors when run at $T_{solid,in}$ of less than 600 °C. Upon further investigation, this phenomenon was seen to occur because of thermodynamics, where Eq.2 depicts the influence H_2O has on the system. At lower temperatures K_{eq} is less, and therefore there is a smaller equilibrium concentration of water for constant pressure. This is validated by simulation errors that occurred in Stage 1; where temperature was at its lowest and water concentration highest. These limitations aid in selecting feasible H_2 excess ratios for corresponding $T_{solid,in}$. Note that only discrete data points for H_2 excess ratios were tested during this analysis.

4.1.5 Water Inlet Adjustment

The error warnings in Section 4.1.4 indicate that water content limitations will be a huge factor when designing a recycle stream in future iterations. Although a full separation and recycle stream is outside the scope of

this project, changing inlet proportions of H_2O would affect feasibility and must be considered for the purposes of designing a robust reactor.

Using Stage 1 temperature values, T_{S1}, from simulation runs in Section 4.1.4, K_{eq} for various $T_{solid,in}$ and H_2 excess ratios were calculated. Given that a binary gas phase mixture of H_2 and H_2O was considered, the following equation was derived which calculates the thermodynamic maximum water concentration at the outlet of stage 1, $y_{H_2O,eq}$,

$$y_{H_2O,eq} = 1 - \frac{1/K_{eq}}{(1 + 1/K_{eq})}. \tag{10}$$

Since X_{hem} and target iron throughput were both set parameters, the total molar flowrate of reacted H_2 and produced H_2O were constant between simulation runs. Therefore, the maximum inlet molar fraction of water, $\dot{n}_{H_2O,max}$, was computed using Eq.11,

$$\dot{n}_{H_2O,max} = \dot{n}_{T,gas} y_{H_2O,eq} - 3\dot{n}_{hem,o} X_{hem}. \tag{11}$$

Figure 6 plots the relationship between H_2 excess ratios and $y_{H_2O,max}$ at varying $T_{solid,in}$, following Eq.11. For higher H_2 excess ratios and $T_{solid,in}$, the thermodynamic limiting temperature at stage 1 was increased. Overall, having a higher T_{S1} meant that higher $y_{H_2O,eq}$ and higher inlet water flowrates were feasible. Additionally, higher H_2 excess ratios further resulted in an increased hydrogen concentration, for which more water at the inlet was required to achieve the same equilibrium concentration.

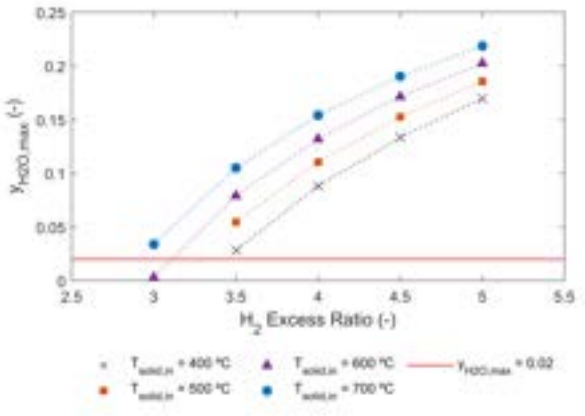


Figure 6. Maximum water inlet molar fraction, $y_{H_2O,max}$, before thermodynamic equilibrium is reached at the gas outlet of stage 1, for varying H_2 excess ratios and $T_{solid,in}$.

The red line in Figure 6 indicates a change in input H_2O molar fraction from 0% in previous simulations to 2%. This is a conservative value and encourages a more robust design as typical separation processes recover up to 99% of H_2 from water mixed gas streams [20]. The resulting effect of this change on overall residence time was negligible. Note that $T_{solid,in}$ of 600 °C at a H_2 excess ratio of 3 fell beneath the inlet water molar fraction and was thus excluded from further kinetic feasibility studies. All other configurations presented in Figure 6 were brought forward for reactor sizing.

4.2 Reactor Sizing

4.2.1 Particle Size Selection

Previous sections optimised reaction operational parameters by minimising τ and energy requirements. To continue with reactor sizing, an additional parameter must be explored – particle size denoted by d_p. Recall from Section 2 that d_p had been mentioned in the context of rate control limitations and that d_p smaller than 800 μm at temperatures above 900 °C are kinetically rate controlled. This size limitation is what has enabled this investigation to progress using the kinetic equation by Chen. However, there is another effect that changing d_p has on reactor sizing, namely the gas flow regime. In Section 3.2.4, creeping flow for terminal falling velocity was assumed. Changes to d_p cause an effect in u_t of the solid particles, which impacts the gas flow regime inside the reactor. Particle sizes that are too large invalidate laminar flow assumptions, whilst sizes that are too small cause entrainment and elutriation up the reactor column. Hence, a conservative d_p of 110 μm was selected.

4.2.2 Initial Reactor Sizing (7-CSTR)

The kinetic feasibility tests identified that pressure and H_2 excess ratios have huge influence on τ. By extension, it was hypothesized that they would have the largest impact on reactor dimensions as well. To test this idea, reactors were sized using the methodology stated in Section 3.2.3 for varying $T_{solid,in}$, pressures and H_2 excess ratios at 1000 °C isothermal operation of the heated stages. Figure 7 showcases if the different designs met ZESTY-iron reactor target parameters of 10-50 m in height and an aspect ratio of 6.

Reactor sizes were seen to become operationally infeasible at high pressures and high H_2 excess ratios. An increase in pressure resulted in two main consequences: 1. large decrease in τ as seen in Section 4.1.3, 2. decrease in the volumetric gas flowrate within the reactor. These effects led to a drastic reduction in reactor volume making it impossible to fit the aspect ratio of a tubular reactor within Calix's height limits. Meanwhile, an increase of H_2 excess ratio led to a decrease in τ and an increase in gaseous volumetric flowrate. The former shortened the height of the reactor whilst the latter increased reactor diameters. The combination of these factors has a counterproductive effect towards the aspect ratio. Therefore, lower pressure conditions and minimal H_2 excess ratios were seen to be favourable. All feasible reactor (indicated in green) where brought forward for comprehensive sizing analysis.

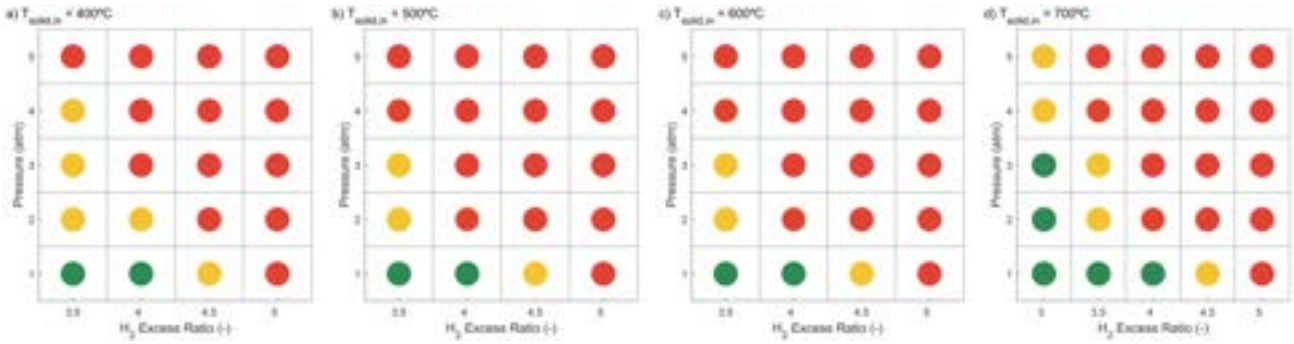


Figure 7. Initial plots representing reactor sizing feasibility at varying pressures and H_2 excess ratios for $T_{solid,in}$ of a) 400 °C, b) 500 °C, c) 600 °C, and d) 700 °C. The gas feed was set with a water molar fraction of 2% and the heated stages throughout the reactor were operated at 1000 °C isothermal. Feasible reactors are shown in green. Closely converging reactors with an SE of less than 50 m^2 (refer to section 3.2.4) are shown in amber. Infeasible reactors are shown in red.

4.2.3 Comprehensive Reactor Sizing (15-CSTR)

All models concluded in Section 4.2.2 were simulated on the 15-CSTR model to identify the best-performing reactor. The methodology used is outlined in Section 3.2.3. Upon inspection, it was found that higher H_2 excess ratios resulted in larger reactor dimensions and higher heat duty requirements. As such, $T_{solid,in}$ reactor configurations not running at their minimum respective H_2 excess ratios were discarded. Secondly, from Section 4.1.5, reactors with $T_{solid,in}$ of 500 °C and 600 °C do not allow for smaller H_2 excess ratios compared to 400 °C. This resulted in reactors that had negligible size differences from the 400 °C $T_{solid,in}$ reactor but with higher heat duties. These reactors were also removed from consideration. Lastly, the 700 °C $T_{solid,in}$ reactors operating above atmospheric pressure proved to be infeasible when run on the 15-CSTR simulation. This is most likely due to a slight decrease in τ with respect to the 7-CSTR model, shifting them from the green to orange zone. Thus, the reactor sizes of the two remaining models are presented in Table 2.

Hypothetically, if τ were to be increased by adjusting other operational parameters like temperature of the heated reactors, then higher pressure configurations would become feasible. If this were the case, these designs could possibly have smaller dimensions than reactors operated at 1 atm. This idea forms the basis of the investigations conducted in Section 4.3.

Table 2. Reactor sizes and total system energy duties for the two kinetically optimal reactor configuration.

$T_{solid,in}$ (°C)	P_T (atm)	H_2 Excess Ratio (-)	$d_r(m)$	$l(m)$	Q_{TOT} (MW)
700	1	3.0	5.99	35.96	3.94
400	1	3.5	7.04	42.24	3.90

It can be concluded that the 700 °C $T_{solid,in}$ reactor design resulted in the smallest reactor dimensions with a minimal increase in heat duty. This forecasts a lower capital expenditure cost with a negligible difference in operation cost. Overall, these two reactors were deemed the most kinetically optimal of all tested reactors based on energy, size, and predicted cost.

Note that the heat duties portrayed in Table 2 have inaccuracies due to the assumption that H_2 inlet is supplied at 25 °C. In future models, once a proper separation and recycle system is set up, this temperature is likely to increase and reduce the total heat duties for both reactors.

4.3 Energy Profile Analysis

4.3.1 Testing Different Energy Profiles

This section builds upon the idea of how performance of a reactor can be enhanced that was first posited in Section 4.2. The idea stems from the understanding that τ can neither be too low as to be infeasible due to the reactor design's inability to meet an aspect ratio of 6, nor too high where reactor height would exceed 50 m. This balance between shifting operational parameters to ensure τ fits into this optimal range allows for manipulation of rector performance (reactor size and operational energy). Throughout the sizing study conducted in Section 4.2, the impact pressure, H_2 excess ratio and $T_{solid,in}$ had on reactor sizing was explored. The one factor that has not been expanded upon is the energy profile across the heated reactors, which was set in Section 4.1.1 to operate isothermally at 1000 °C. This study attempted to reconfigure the energy profile with the aim of increasing energy efficiency whilst minimising reactor size compared to the two most feasible reactors illustrated in Table 2.

As the basis of this investigation is to operate a smaller and more energy efficient reactor, a lower total heat duty must be established. A value of 0.374 MW was obtained by operating a 700 °C $T_{solid,in}$ reactor at 2 atm, with a H_2 excess ratio of 3, isothermally at 970 °C. This overall quota was set and energy was redistributed along the heated reactor stages in accordance with the four heating profiles represented in Figure 2. The results of this analysis are presented in Table 3, where a lower τ indicated better performance.

Table 3. Average temperature of the heated stages, T_{avg}, and residence time, τ, for the different energy profiles at 2 atm, $T_{solid,in}$ of 700 °C and H_2 excess ratio of 3 with a 2% water inlet.

Energy Profile	Heated Stages T_{avg} (°C)	τ (s)
Decreasing	971.0	85.2
Constant	973.9	83.7
Increasing	977.4	79.8
Mountain	983.4	70.5

It can be seen that energy profiles with the highest average temperatures across heated stages resulted in lower τ values. This can be attributed to the bulk of the reaction taking place throughout the heated stages. Hence, it is more efficient to supply heating to the middle-bottom stages of the reactor as this energy will be used to further maximise the rate of reaction where it is highest. The reactor configuration using a mountain energy profile performed the best and was simulated on the 15-CSTR model.

4.3.2 Reactor Performance

The energy distribution profile across the 15-CSTR model replicated the mountain profile from the 7-CSTR simulation in Section 4.3.1. It should be noted that energy fractions were all kept within a ±8% uncertainty, whilst total energy requirement across the stages was decreased by 5% to not exceed the 1000 °C threshold.

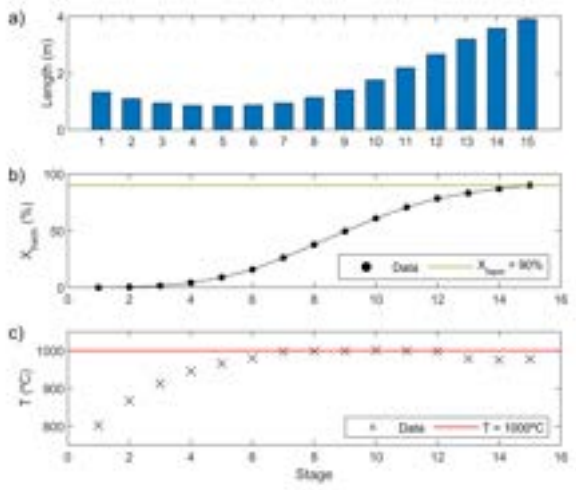


Figure 8. Final proposed reactor's a) height of stages, b) reduction degree and c) temperature profile, for each of the 15 stages. Operating conditions are 2 atm, $T_{solid,in}$ of 700 °C, 2% water inlet, H_2 excess ratio of 3, and a mountain-shaped energy profile.

Figure 8a showcases the height of each stage within the 15-stage reactor. Noting that the viscosity of H_2O is higher than H_2 and that density of iron is larger than hematite, the parabolic-like shape that reactor stage heights take on this figure provides interesting insight into the changing physical properties across the reactor. Figures 8b and c are strongly interlinked towards the sizing of the reactor, as outlined during the height calculations in Section 3.2.4.

At the bottom stages near the gas inlet (stage 15), most of the conversion of hematite has taken place. Therefore, maximum iron and minimal H_2O concentrations are found, leading to longer stages. Travelling up the reactor, the rates of reduction peak at stage 8, where maximum production of H_2O takes place. As more water is generated, the viscosity of the gas progressively increases. This leads to a reduction of u_t and u_p for the same $\hat{u}_g$, following from Eq.3 and 7. By Eq.8, stage heights decrease, reaching a minimum at stage 5. From here upwards, reduction rates of hematite are minimal, thereby leading to negligible changes in H_2O concentration. However, in Figure 8c temperatures can be seen to drop significantly. This decreases gas viscosities resulting in a slight increase in stage heights. Ultimately, the total reactor height was 26.66 m, with a diameter of 4.44 m. This is a smaller reactor than the ones highlighted in Table 2, thus predicting lower capital expenditure.

When considering operational costs, heat duty for the total system was measured at 3.80 MW, which is less than those recorded in Table 2. Although this only constitutes a minor decrease of 3.5%, it serves as a proof of concept for the hypothesis posed at the start of the energy analysis. The main contributor to total heat duty was the inlet gas feed, accounting up to 64% of the total energy requirement. Therefore, future efforts to reduce energy usage should target the appropriate design of a recycle separation system.

5 Conclusion

The integrated kinetic model for the DRI from hematite simulated using Aspen Plus is integral for the advancement of Calix's goals to establish a Basis of Design for their ZESTY-iron demonstration plant. It has the potential to act as the foundation from which future conceptual designs of complementary subsystems can be built. This model was used to successfully establish a design basis for the ZESTY-iron reactor with target outputs of 30,000 tonnes of iron per operational year with a 90% reduction of pure hematite ore. The key operating parameters of the proposed reactor are: 2 atm pressure, 700 °C $T_{solid,in}$, gaseous inputs at a H_2 excess ratio of 3 with 2% molar fraction of water, which uses a mountain energy profile. This reactor configuration was designed for a height of 26.66 m and a diameter of 4.44 m, with a total system heat duty of 3.80 MW.

During this study, a few key findings were made. The first is that pressure and H_2 excess ratio had the biggest influence on reactor sizing. This is due to the effect these two operational parameters had on τ as they were varied. The second interesting finding was that τ has an optimum range. When τ values were too

low, the target aspect ratio of 6 for tubular reactors could not be reached. Contrariwise, when values were too high, tubular reactors would exceed Calix's height guidelines of 10-50 m. This led to the final crucial finding that a smaller and more energy efficient reactor can be designed by using lower temperature conditions at elevated pressures. The resulting reactor has the potential for improved performance and lower capital and operational expenditures. This leaves room for future research into different configurations of this reactor.

An additional tool that Calix Ltd. and the Fennell Group can investigate could be automating the sizing calculations of reactors for each simulation run. This would enable them to conduct rigorous sensitivity tests with smaller step changes across influential operational parameters, and feasibility would be autonomously tested.

References

[1] Fry I. Climate change the greatest threat the world has ever faced, UN expert warns. OHCHR. https://www.ohchr.org/en/press-releases/2022/10/climate-change-greatest-threat-world-has-ever-faced-un-expert-warns [Accessed 9th November 2022].

[2] Reilly JM, Jacoby Henry D., Prinn RG. Multi-gas contributors to global climate change. MIT; 2003. https://www.c2es.org/document/multi-gas-contributors-to-global-climate-change/

[3] Swalec C. Pedal to the Metal. https://globalenergymonitor.org/wp-content/uploads/2022/06/GEM_SteelPlants2022.pdf

[4] Owen-Burge C. Steel cop27. Climate Champions. https://climatechampions.unfccc.int/steel-cop27/ [Accessed 1st December 2022].

[5] Lorraine L. Direct from Midrex 1st Quarter 2020. https://www.midrex.com/wp-content/uploads/Midrex-2020-DFM1QTR-Final.pdf

[6] Green steel and Europe's natural gas challenge — McKinsey. https://www.mckinsey.com/industries/metals-and-mining/our-insights/safeguarding-green-steel-in-europe-facing-the-natural-gas-challenge; 2022 [Accessed 1st December 2022].

[7] Lorraine L. Midrex and Paul Wurth Selected by H2 Green Steel. MIDREX. 2022. https://www.midrex.com/press-release/midrex-and-paul-wurth-selected-by-h2-green-steel/ [Accessed 9th November 2022].

[8] Barucchi A. Calix files a new patent for zero emissions iron and steel. Calix. 2021; https://calix.global/sustainable-processing/new-patent-for-zero-emissions-iron-and-steel/ [Accessed 10th November 2022]

[9] Ching N, Starikova M. Evaluation of equilibrium and kinetics to assess the feasibility of a novel process for the direct reduction of iron with hydrogen. [MEng Final Year Project] Imperial College London; 2022.

[10] Tkachenko A. Exploring the potential of the direct separation reactor (DSR) for flash reduction of iron ore for the steel industry. [Master's Thesis] Imperial College London; 2022.

[11] Langley Alloys. High temperature nickel alloys. Langley Alloys. https://www.langleyalloys.com/2020/05/05/high-temperature-nickel-alloys/ [Accessed 10th October 2022].

[12] Spreitzer D, Schenk J. Reduction of iron oxides with hydrogen – a review. steel research international. 2019;90(10): 1900108. https://doi.org/10.1002/srin.201900108.

[13] Patisson F, Mirgaux O. Hydrogen ironmaking: how it works. Metals. 2020;10(7): 922. https://doi.org/10.3390/met10070922.

[14] Turkdogan ET, Vinters JV. Gaseous reduction of iron oxides: Part III. Reduction-oxidation of porous and dense iron oxides and iron. Metallurgical Transactions. 1972;3(6): 1561–1574. https://doi.org/10.1007/BF02643047.

[15] Mohassab Y, Chen F, Elzohiery M, Abdelghany A, Zhang S, Sohn HY. Reduction kinetics of hematite concentrate particles by co+h2 mixture relevant to a novel flash ironmaking process. In: Hwang JY, Jiang T, Pistorius PC, Alvear F. GRF, Yucel O, Cai L, et al. (eds.) 7th International Symposium on High-Temperature Metallurgical Processing. Cham: Springer International Publishing; 2016. https://doi.org/10.1007/978-3-319-48093-0_28. [Accessed 9th December 2022].

[16] Hara Y, Tsuchiya M, Kondo Shin ichi. Intraparticle Temperature of Iron-Oxide Pellet during the Reduction. 1973 Dec; 1261-1270.

[17] Caenn R, Darley HCH, Gray GR. Chapter 1 - introduction to drilling fluids. In: Caenn R, Darley HCH, Gray GR (eds.) Composition and Properties of Drilling and Completion Fluids (Seventh Edition). Boston: Gulf Professional Publishing; 2017. p. 1–34. https://doi.org/10.1016/B978-0-12-804751-4.00001-8. [Accessed 5th November 2022].

[18] Douglas J. Conceptual design of chemical processes. International. New York: McGraw-Hill Education; 1988.

[19] Brechtelsbauer DrIngC. Reaction Engineering II - Heterogeneous and Multiphase Reactor Design Lecture Script V 6.1. Lecture Script presented at; Imperial College London.

[20] Recovering hydrogen – and profits – from hydrogen-rich offgas. https://www.aiche.org/resources/publications/cep/2018/january/recovering-hydrogen-and-profits-hydrogen-rich-offgas [Accessed 25th October 2022].

Understanding and Modelling Ionic Liquids Using SAFT-γ Mie

Ethan Hoe[1], Alexander Smulian[1]

[1]*Department of Chemical Engineering, Imperial College London, South Kensington Campus, London SW7 2AZ, United Kingdom*

Abstract Ionic liquids (ILs) are salts that are liquid at room temperature with unique and useful properties that have led to much research into their use. In this paper, we develop a model within SAFT-γ Mie for predicting the thermodynamic properties of ILs. Our model describes ILs as fully dissociated ions with explicit accounts for electrostatic interactions. The focus of our model is on 1,3-dialkylimidazolium tetrafluoroborates. We explored two representations of the ILs at varying degrees of coarse-graining: a sphere model and a group contribution model. Using experimental data for density, isobaric heat capacity, and speed of sound, we estimated parameters for imidazolium and tetrafluoroborate functional groups. During the construction of the model, we investigate the importance of accurately modelling the dielectric effect to the validity of the predicted results. We show that electrostatic interactions have significant effects on the calculated properties. We demonstrate the inadequacy of the sphere model and show that the group contribution model provides good predictions for a variety of imidazolium-based ILs. Thus, our work demonstrates the efficacy of using a group contribution method with dissociated ions interacting electrostatically to model ILs in SAFT-γ Mie.

1. Introduction

Ionic liquids (ILs) have gained increasing interest due to their unique properties. ILs are broadly defined as salts that are liquid at low temperatures, often specified as having a melting point below room temperature (25˚C). This is made possible by a typical characteristic shared by most ILs: they contain a bulky asymmetric cation (Shukla and Mikkola, 2020) which inhibits Coulombic interactions, hindering crystallisation.

One of the unique properties of ILs is their extremely low vapour pressure (Aschenbrenner *et al*, 2009). As such, losses due to evaporation are often considered negligible. This has led to numerous investigations of its use as a green solvent as they have minimal evaporative losses during use, hence reducing cost and any associated pollution risks. Separation and regeneration can be easily and sharply done saving on energy usage (Ramdin *et al.,* 2012). ILs are also conductive and electrochemically stable which has attracted interest in their potential use in batteries and electrochemistry (Ray and Saruhan, 2021).

Many ILs are organic compounds which allows for a great deal of customisability and variety while maintaining the desirable properties of an IL. This variety extends to the ability to create highly selective ionic liquids for specific tasks, such as carbon capture (Ramdin *et al.,* 2012). With the sheer amount of variety possible, it is of great interest to scientists to be able to predict these properties and phase behaviours, instead of performing countless tedious lab experiments.

The basis of our research into modelling ILs lies in statistical association fluid theory (SAFT), an equation of state (EOS) rooted in statistical mechanics and perturbation theory. Its main benefit over other classes of EOS is its ability to model complex associating and non-spherical fluids. SAFT comes in many forms, diverging after its first introduction in 1989 (Chapman *et al.* 1989). In this paper, we use SAFT-γ Mie, the state-of-the-art SAFT EOS by Papaioannou *et al.* (2014), a group contribution EOS capable of accurate and simultaneous modelling fluid-phase behaviours and second-order thermodynamic derivative properties.

Previous papers have investigated the modelling of ILs using different versions of SAFT, with different approaches. Ji *et al.* (2012) studied the modelling of imidazolium-based ILs with ePC-SAFT using varying strategies and determined that the best strategy for predicting CO_2 solubility had dissociated ions and factored electrostatic interactions into the EOS. Similarly, Guzmań *et al*. (2015) modelled 1,3-dialkylimidazolium tetrafluoroborates using SAFT-MSA and showed that explicit electrostatic interactions allow for the correct description of trends when increasing alkyl-chain length. More, recent papers by Ashrafmansouri and Raeissi (2021), and Dong *et al*. (2022) modelled ILs as single non-dissociated neutral molecules with the group contribution methods of SAFT-γ; the former used the square well (SW) potential for dispersion interactions and the latter used the Mie potential.

The novelty of our research can be seen in using the electrostatic interactions studied in the earlier papers while also taking advantage of the group contribution methods and Mie potential that make up SAFT-γ Mie. Our research focuses on modelling pure-component systems and aims to test the ability of SAFT-γ Mie to describe and predict the properties of ILs. We compare a simple sphere model with a more complex but flexible group contribution model. Additionally, we investigate the effects of the dielectric constant on the model.

2. Theory

2.1 Molecular Model

At its core, molecules in SAFT-γ Mie are composed of fused-spherical segments with dispersion interactions described by Mie potentials. Association interactions are described using specially defined sticky sites. Developed on the framework of a group contribution approach, chemically distinct functional groups (simply referred to as groups) are represented by one or more identical segments. These groups are put together to describe a molecule's structure. An example of this is illustrated in Figure 3. For ionic/electrolyte systems, charged groups also interact through Coulomb potentials.

2.2 Mie Potential

The Mie potential Φ^{Mie} is an intermolecular pair potential that describes the dispersion forces between two segments given by

$$\Phi^{\text{Mie}}(r) = C\varepsilon\left[\left(\frac{\sigma}{r}\right)^{\lambda^{\text{r}}} - \left(\frac{\sigma}{r}\right)^{\lambda^{\text{a}}}\right] \quad (1)$$

where r is the intersegment distance, ε is the depth of the potential well or dispersion energy, σ is the diameter of the segment (corresponding to the distance where $\Phi^{\text{Mie}} = 0$), λ^{r} and λ^{a} are the repulsive and attractive range respectively, and C is given by

$$C = \frac{\lambda^r}{\lambda^r - \lambda^a}\left(\frac{\lambda^r}{\lambda^a}\right)^{\frac{\lambda^a}{\lambda^r - \lambda^a}} \quad (2)$$

to keep ε equal to the depth of the well regardless of the values of the exponents λ^{r} and λ^{a} used. A graphical representation of the Mie potential is shown in Figure 1.

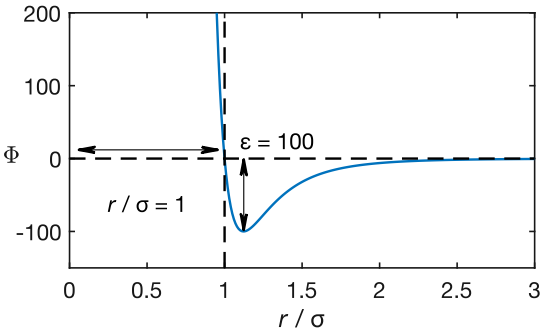


Figure 1. Graphical representation of the Mie potential (12,6); $\varepsilon = 100$, $\lambda^{\text{a}} = 6$, and $\lambda^{\text{r}} = 12$.

2.3 Coulomb Potential

Analogous to the Mie potential, the Coulomb potential u^{Coulomb} is a pair potential used to describe the Coulombic interactions between charged groups. The Coulomb potential between charged groups k and l is given by

$$u^{\text{Coulomb}}(r_{kl}) = \frac{Z_k Z_l e^2}{4\pi\epsilon_0 D r_{kl}} \quad (3)$$

where Z is the charge number, e is the elementary charge, ϵ_0 is the permittivity of free space, D is the dielectric constant, and r_{kl} is the separations of the charges.

2.4 SAFT-γ Mie Equation of State

The SAFT- γ Mie EOS can be thought of as being the sum of different contributions to the total Helmholtz free energy A of a fluid. In this work, the relevant contributions take the form of

$$A = A^{\text{ideal}} + A^{\text{mono}} + A^{\text{chain}} + A^{\text{Born}} + A^{\text{ion}} \quad (4)$$

A^{ideal} is the ideal contribution; it is the Helmholtz free energy if the molecules of the fluid are treated as an ideal gas with only translational, rotational and vibrational kinetic energy contributions. A^{mono} is the monomer contribution; it is the contribution due to the dispersion interactions (attraction and repulsion) between monomer segments described by the Mie potential. A^{chain} is the chain contribution; it is the change in free energy when monomer segments are brought together and are attached to each other to form chains/molecules. Note that association contributions (e.g., hydrogen bonds) are not considered in this work.

The last two contributions correspond to electrostatic interactions unique to electrolyte systems, or in our case ILs. A^{Born} is the Born contribution; it represents the free energy of solvation based on the Born equation. It can be interpreted as the free energy change when a (Born) cavity of diameter $\sigma_{kk}^{\text{Born}}$ is created in a dielectric medium, where then a charge group is inserted. Mathematically, it takes the simple form of

$$A^{\text{Born}} = -\frac{e^2}{4\pi\epsilon_0}\left(1 - \frac{1}{D}\right)\sum_{k=1}^{n_{\text{ion}}}\frac{N_k Z_k^2}{\sigma_{kk}^{\text{Born}}} \quad (5)$$

where n_{ion} denotes the list of charged species, N_i is the number of molecules of charged species i. A^{ion} is the ion contribution, sometimes referred to as A^{MSA}; it describes the contribution due to Coulombic interactions between charged groups through the Mean Spherical Approximation (MSA) method.

For a detailed account of how each contribution is calculated and the theory behind them, refer to the works of Papaioannou *et al.* (2014), Schreckenberg *et al.* (2014), and Haslam *et al.* (2020).

Simply put, the fluid (without association) can be described by nine types of parameters: v^*, S, σ, ε, λ^{r}, λ^{a}, D, Z, σ^{Born}; where v^* is the number of identical segments that make up a group to imitate non-sphericity, and S is the shape factor which describes how much a group contributes to the overall molecule. Although the subscripts are omitted above for generality, it is important to remember that many individual parameters make up each parameter type to describe each and every group and/or the interactions between them.

Parameters in SAFT-γ Mie can be classified into two types – like and unlike. Like parameters characterise the interactions between groups of the same kind (e.g., CH_2 with CH_2) denoted by subscript kk, while unlike parameters characterise the interactions between groups of a different kind (e.g., CH_2 with CH_3) denoted by subscript kl. There are also parameters that describe the group itself denoted by subscript k, which is usually considered to fall under the category of like group parameter.

With the parameters specified, we end up with an expression for the Helmholtz free energy as a function of the state variables temperature T, volume V and composition vector $\mathbf{N}$, $A = A(T, V, \mathbf{N})$

2.5 Thermodynamic Properties

After arriving at an expression for the Helmholtz free energy of the fluid, thermodynamic properties can be calculated using standard thermodynamic relations. In this study, we focus on 3 properties: density, isobaric heat capacity and speed of sound. Density is determined using the following equations:

$$P = -\left(\frac{\partial A}{\partial V}\right)_{T,\mathbf{N}} \quad (6)$$

$$\rho = \frac{M_W}{V} \quad (7)$$

$$M_W = \sum_i \frac{N_i}{N_A} M_{W,i} \quad (8)$$

where P is pressure, N_i is the number of molecules of component i, ρ is mass density, M_W is the total mass of the fluid, $M_{W,i}$ is the molar mass of component i, and N_A is the Avogadro Constant. Isobaric heat capacity is calculated as follows:

$$C_V = -T\left(\frac{\partial^2 A}{\partial T^2}\right)_{V,\mathbf{N}} \tag{9}$$

$$C_P = C_V - T\frac{\left(\frac{\partial P}{\partial T}\right)^2_{V,\mathbf{N}}}{\left(\frac{\partial P}{\partial V}\right)_{T,\mathbf{N}}} \tag{10}$$

where C_V is the isochoric heat capacity, and C_P is the isobaric heat capacity. Lastly, the speed of sound is calculated using:

$$\omega = \sqrt{\frac{V^2}{M_W}\frac{C_P}{C_V}\left(\frac{\partial P}{\partial V}\right)_{T,\mathbf{N}}} \tag{11}$$

where ω is the speed of sound.

3. Method

3.1 Molecular Representation

The ILs are modelled as fully dissociated ions with explicit electrostatic interactions. Previous works (Guzmań et al., 2015; Ji et al., 2012) have indicated that this approach provides the most robustness and representability. Moreover, without having seen any evidence of that, one can still argue that this is the most natural and intuitive way of representing ILs owing to the fact that in reality they do exist as free-moving dissociated ions in the liquid phase (Lee *et al.*, 2015), and by being charged, they are subjected to electrostatic interactions.

In this paper, we consider two representations of the ILs at different degrees of coarse-graining: the simple sphere model and the group contribution model.

Firstly, in the simple sphere model, the cation 1-ethyl-3-methylimidazolium [EMIM] and anion tetrafluoroborate [BF4] are both modelled as individual spheres as illustrated in Figure 2. In other words, they are represented by only one group each.

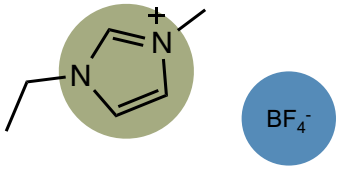


Figure 2. Pictorial representation of the 1-ethyl-3-methylimidazolium [EMIM] cation (green) and tetrafluoroborate [BF4] anion (blue) in the simple sphere model.

In the group contribution model, the cation is broken down into smaller groups as illustrated in Figure 3; [EMIM] is split into two methyl groups (CH_3), one methylene group (CH_2) and one imidazolium group (IM). By singling out IM, the same set of parameters can be applied to other imidazolium-containing cations in a predictive manner. [BF4] is still modelled as a sphere (one BF4 group) since it has a symmetric tetrahedral shape which closely resembles a sphere.

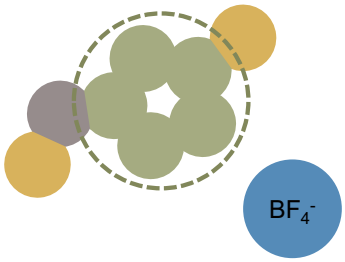


Figure 3. Pictorial representation of the 1-ethyl-3-methylimidazolium [EMIM] cation and tetrafluoroborate [BF4] anion in the group contribution model. The different colours indicate different functional groups: methyl group CH_3 (yellow), methylene group CH_2 (grey), tetrafluoroborate group BF4 (blue), and imidazolium group IM (green). The green dashed circle encompasses the IM group made out of five identical segments.

3.2 Ideal Gas Heat Capacity

The ideal contribution A^{ideal} is specified using a temperature-dependent correlation for the ideal gas isobaric heat capacities. This is done using the Joback method (Joback, 1984) – a group contribution method – with parameters from the work of Ge *et al.* (2008).

3.3 Dielectric Constant and Born Cavity Diameter

The dielectric constant (or relative permittivity) D is used in the calculation of A^{Born} and A^{ion} as a primitive approach to model electrostatic interactions. Experimental data for D for ILs are hard to come by, and when they do there are discrepancies from source to source. For the sake of consistency, we decided to use an average D of 15 for all ILs in this study.

In the SAFT-γ Mie code used (see Section 3.6), D is expressed by an empirical model used by Schreckenberg *et al.* (2014) which takes the following form:

$$D = 1 + \rho_M d_V\left(\frac{d_T}{T} - 1\right) \tag{12}$$

where ρ_M is the molar density of the IL, and d_V and d_T are component specific parameters that represents the density and temperature dependence of D respectively. In our model, D is approximated to 15 by taking d_T as zero and choosing a d_V value that would give a D of 15 at 298.15K and 1atm based on experimental density data at those conditions.

The Born cavity diameter $\sigma_{kk}^{\text{Born}}$ used in the calculation of A^{Born} is approximated as being 7 percent larger than the corresponding segment diameter σ_{kk}.

3.4 Combining Rules

Combining rules are a way to estimate unlike group parameters from their like counterparts. The use of combining rules greatly simplifies the optimization problem by reducing the number of adjustable parameters needed to be estimated. Details of the various combining rules for SAFT-γ Mie can be found in the paper by Haslam *et al.* (2020).

For the new groups introduced in this paper, combining rules are used for all of its unlike parameters with the exception of the unlike dispersion energy, ε_{kl},

between charged groups. This is because ε_{kl} usually deviate considerably from combining rules (Papaioannou *et al.*, 2014), especially between charged groups.

3.5 Initial Estimation

We start off with an initial estimation (i.e., without the use of an optimisation program) of the ε_{kk} and σ_{kk} parameters from previous works for the simple sphere model. The purpose of this step is to see how far one can make predictions by just inferring from existing parameters from a different iteration of SAFT, and to explore the effects each parameter has.

ε_{kk} is obtained by adapting SAFT-γ SW parameters from Ashrafmansouri and Raeissi (2021) using the following relationship:

$$\int_{\sigma}^{\infty} 4\varepsilon^{\mathrm{Mie}}\left[\left(\frac{\sigma}{r}\right)^{12}-\left(\frac{\sigma}{r}\right)^{6}\right]r^2dr=\int_{\sigma}^{\sigma\lambda^{\mathrm{sw}}}\varepsilon^{\mathrm{sw}}r^2dr \quad (13)$$

where $\varepsilon^{\mathrm{Mie}}$ is the equivalent SAFT-γ Mie ε_{kk}; and σ, $\varepsilon^{\mathrm{sw}}$ and λ^{sw} are SAFT-γ SW parameters. An approximation is made here that σ is the same for the SW and Mie EoS. Equation 13 is derived from equating the *Energy Equation* of the Mie and SW potentials.

σ_{kk} is also obtained by adapting SAFT-γ SW parameters from Ashrafmansouri and Raeissi (2021) using the following equation:

$$\sigma^{\mathrm{Mie}}=\left[\sum_{kk}S_{kk}^{\mathrm{sw}}\cdot\sigma_{kk}^{\mathrm{sw}3}\right]^{1/3} \quad (14)$$

where σ^{Mie} is the equivalent SAFT-γ Mie σ_{kk}; S_{kk}^{sw} and $\sigma_{kk}^{\mathrm{sw}}$ are SAFT-γ SW parameters. kk denotes the list of SW groups that make up the simple sphere Mie group. Equation 14 is derived by summing the volumes of individual SW groups and taking the equivalent sphere diameter of the summed volume.

λ^{r} and λ^{a} are taken to be 12 and 6 respectively – the Lennard-Jones potential or Mie (12,6).

SAFT-γ Mie calculations on this step (only) are carried out using gPROMS.

3.6 Parameter Estimation

From here on, SAFT-γ Mie calculations and parameter estimation are carried out using the Julia package Clapeyron.jl (Walker et al., 2022). The parameters are estimated by minimising the least squares objective function F:

$$F=\sum_{i=1}^{N_R}\left(\frac{R_i^{\mathrm{calc}}-R_i^{\mathrm{exp}}}{R_i^{\mathrm{exp}}}\right)^2 \quad (15)$$

where R_i^{exp} is the experimental value, R_i^{calc} is the calculated value, and N_R is the number of data points. This optimisation problem is solved using the Evolutionary Centers Algorithm (Mejía-de-Dios and Mezura-Montes, 2019).

For the simple sphere model, we use density, heat capacity and speed of sound experimental data of [EMIM][BF4] at 1atm to estimate parameters for the groups EMIM and BF4.

For the group contribution model, we use density, heat capacity and speed of sound experimental data of [EMIM][BF4], 1-butyl-3-methylimidazolium tetrafluoroborate and 1-ethyl-3-methylimidazolium acetate [EMIM][Ac] at 1atm to estimate parameters for the groups IM and BF4.

The introduction of [EMIM][Ac] helps overcome the problem of degeneracy/unidentifiability of parameters. An IL with the acetate anion [Ac] is chosen because it can be fully defined with group parameters from previous works. The parameters for CH_3, CH_2 and COO^- groups are taken from the paper by Haslam *et al.* (2020).

During optimisation, λ^{a} is fixed at 6, a common practice as λ^{a} and λ^{r} have a similar qualitative effect on the Mie potential. S of standalone groups such as BF4 and EMIM is set to 1, while S for non-standalone groups such as IM, is estimated as part of the optimisation. ν^* for groups deemed non-spherical such as IM are adjusted manually until reasonable estimated parameters are obtained.

3.7 Deviations

As a way to quantify the quality of our model parameters, we determine the percentage average absolute deviation %AAD given by:

$$\%\mathrm{AAD}=\frac{100}{N_R}\sum_{i=1}^{N_R}\left|\frac{R_i^{\mathrm{exp}}-R_i^{\mathrm{calc}}}{R_i^{\mathrm{exp}}}\right| \quad (16)$$

4. Results and Discussion

4.1 Initial Estimation

In Figure 4, we present the results of the simple sphere model using parameters estimated from previous works (see Section 3.5). These parameters are summarised in Table 1. The parameters, and implicitly the contributions they describe, are introduced in stages starting with (1) only σ and ε (together with other fixed parameters such as λ^{r}, λ^{a}, Z, S, $\nu^*$and Z), followed by the (2) ideal contribution (i.e., the ideal gas isobaric heat capacities), (3) dielectric constant, D, and lastly (4) a non-zero σ^{Born}. The number in the parentheses refers to the lines in Figure 4 and will henceforth be used to refer to their respective stage and changes.

The density predictions for (1) closely match experimental data in terms of both value and slope. This can partly be explained by the fact that the source these parameters were inferred from did not explicitly account of electrostatic interactions; the electrostatic interactions are hence implicitly accounted for in their σ and ε values (and association parameters). However, it is entirely coincidental that the exclusion of their association parameters and the inclusion of Coulombic interactions while using a value of $D=1$ in our model gave such a good first estimation.

(2) has no effect on density. This is because A^{ideal} is only dependent on temperature, so its contribution to the volume derivative of A (Equation 6) is zero.

In (3), density falls by more than 150 units. This is because increasing D implies a stronger dielectric effect that directly reduces the effective strength of Coulombic

Table 1. Summary of model parameters at the 3 main stages of development: initial estimation, simple sphere model, and group contribution model. CR indicates the use of combining rules.

Group	ν_k^*	S_k	σ_{kk}/Å	$\sigma_{kk}^{\mathrm{Born}}$/Å	$\lambda_{kk}^{\mathrm{r}}$	$\lambda_{kk}^{\mathrm{a}}$	$(\varepsilon_{kk}/k_{\mathrm{B}})$/K	Z_k	$(\varepsilon_{kl}/k_{\mathrm{B}})$/K
Initial Estimation									
EMIM	1	1	5.21	$1.07\sigma_{kk}$	12	6	357.60	+1	CR
BF4	1	1	4.64	$1.07\sigma_{kk}$	12	6	167.60	–1	
Simple Sphere Model									
EMIM	1	1	5.16	$1.07\sigma_{kk}$	100.00	6	1312.07	+1	300.00
BF4	1	1	5.00	$1.07\sigma_{kk}$	100.00	6	160.82	–1	
Group Contribution Model									
IM	5	0.76	2.61	$1.07\sigma_{kk}$	13.14	6	484.99	+1	255.83
BF4	1	1	4.18	$1.07\sigma_{kk}$	20.46	6	59.81	–1	

interactions (Equation 3), resulting in a lower density. This relies on the fact that Coulombic interactions of a salt are predominantly attractive. This demonstrates the significance of D and in turn the electrostatic interactions in property prediction.

(4) leads to a slight drop in density. To clarify, setting a non-zero σ^{Born} has the effect of activating A^{Born}, given that the solution when σ^{Born} is zero is undefined (Equation 5). Within Equation 5, there is no explicit volume dependence for A^{Born}– the only place σ^{Born} is used – so it is not expected to affect density. However, D as defined in Equation 12 is dependent on density which is in turn dependent on volume. Hence, A^{Born} has an implicit dependence on volume through D.

Both the heat capacity and speed of sound are significantly underestimated for all stages. This is expected since Ashrafmansouri and Raeissi (2021) did not take into consideration the performance of their parameters with respect to second-order thermodynamic derivative properties, a class of properties that the SW potential often fails to describe accurately as demonstrated by Papaioannou *et al.* (2014).

4.2 Simple Sphere Model

Using the exact same simple sphere model representation as in the initial estimation, the parameters are optimised as detailed in Section 3.6. The optimised parameters are summarised in Table 1, and the results are illustrated in Figure 5.

There are a few caveats about the parameters in Table 1. Firstly, a few of the parameters are at their bounds; these are σ_{kk} for BF4, ε_{kl} and $\lambda_{kk}^{\mathrm{r}}$ for EMIM and BF4. No matter how much the bounds are changed/increased, they still tend towards them. This could be an indication that this representation is flawed, thus the optimiser uses unphysical parameter values to attempt to minimise the difference between experimental data and calculations. Secondly, the parameters are degenerate; they converge to different values each time the optimisation is carried out. The values of ε_{kk} of EMIM and BF4 are also observed to swap places. This is because both groups EMIM and BF4 are estimated using experimental data for a single compound and there is nothing to distinguish the two groups. A possible solution is to introduce another I-

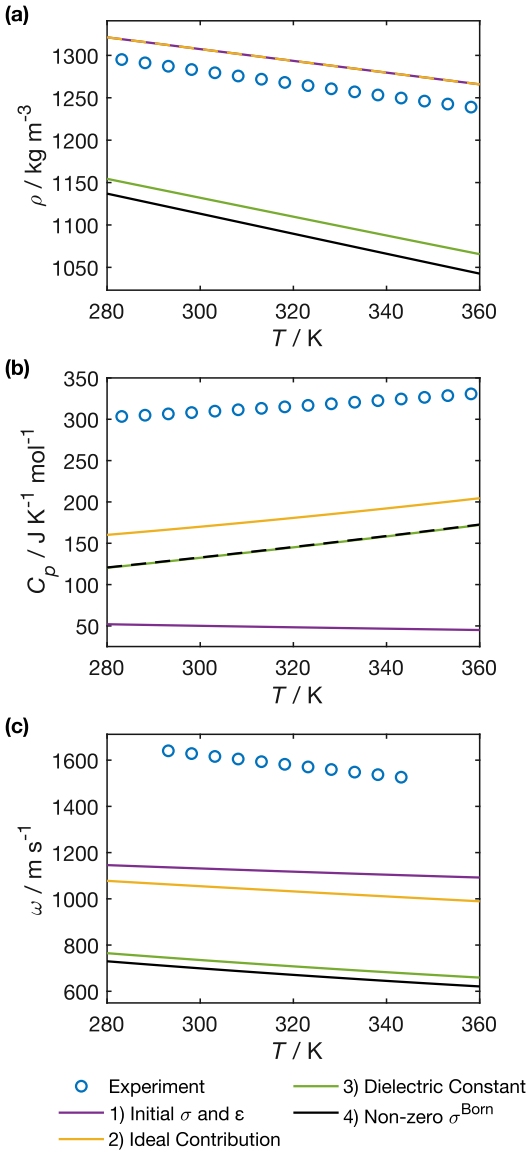


Figure 4. Initial estimation for the simple sphere model: (a) Liquid density, (b) liquid phase isobaric heat capacity, and (c) liquid phase speed of sound at 1atm for [EMIM][BF4]. The various parameters are introduced in a stepped approach, the order of which is indicated by the number in the legend. Solid curves represent SAFT-γ Mie calculations. Symbols represent experimental data. (Neves *et al.*, 2013; Waliszewski *et al.*, 2005; Zarei and Keley, 2017)

L containing only one of the groups; this solution is explored further in the group contribution model.

Looking at Figure 5, although an improvement from the initial estimate can be seen, the heat capacity in particular is still a far cry from experimental data with an %AAD of 43.46%. The results for density and speed of sound are closer to experimental values with an %AAD of 2.52% and 7.88% respectively; however, visually, it can be seen that the slopes are very different suggesting that the deviation will only get worse when the model is calculated at higher/lower temperatures. The %AADs are summarised in Table 2.

From this, it is concluded that the simple sphere model is too crude for a good description of second-order thermodynamic derivative properties of [EMIM][BF4]. It is unable to capture the unique interactions between uncharged groups (such as between methyl and methylene groups) or the interactions between charged and uncharged groups (such as between methyl and BF4 groups), all the while keeping its spherical form; it is unable to describe the density, isobaric heat capacity and speed of sound simultaneously with accuracy.

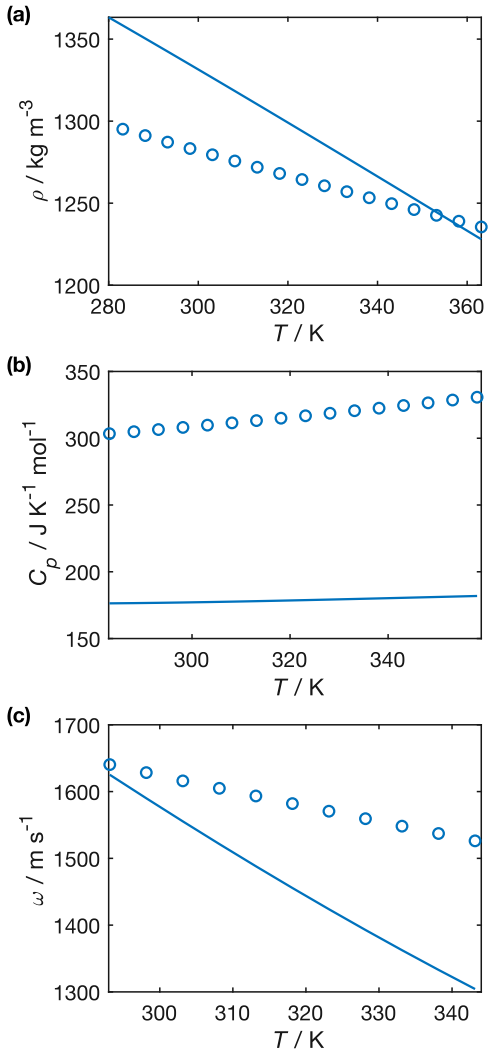


Figure 5. Simple sphere model: (a) Liquid density, (b) liquid phase isobaric heat capacity, and (c) liquid phase speed of sound at 1atm for [EMIM][BF4]. Solid curves represent SAFT-γ Mie calculations. Symbols represent experimental data. (Neves *et al*., 2013; Waliszewski *et al*., 2005; Zarei and Keley, 2017)

Table 2. Percentage average absolute deviation %AAD of liquid phase density, isobaric heat capacity and speed of sound of [EMIM][BF4] using the simple sphere model.

Property	%AAD
ρ	2.52
C_P	43.46
ω	7.88

4.3 Group Contribution Model

In Figure 6, we present the results of the group contribution model. Experimental data for [EMIM][BF4], [BMIM][BF4] and [EMIM][Ac] are used to estimate group parameters for the IM and BF4 groups; the same parameters are then used to make predictions of the same properties for 1-hexyl-3-methylimidazolium tetrafluoroborate [HMIM][BF4], 1-methyl-3-octylimidazolium tetrafluoroborate [OMIM][BF4] and 1-butyl-3-methylimidazolium acetate [BMIM][Ac].

The parameters are summarised in Table 1. At first glance, it can be seen that the parameters are physical/sensible; they are in the range one might expect when comparing with the parameters of previously studied groups. None of them are at the bounds. With the introduction of [EMIM][Ac], the parameters are no longer degenerate; they converge to around the same values each time the parameter estimation is independently carried out.

v^* of 5 for IM is found to be the most optimal. This is based on two criteria: the value of the objective function F (Equation 15) and observing that S no longer tends to 1 (the bound). Physically, this makes sense as the imidazolium ring is made out of 5 atoms.

In Table 3, we summarise the %AAD for the group contribution model. The results are excellent across the board with low %AAD of 0.45–0.66%, 0.19–3.96%, and 0.26–1.58% for density, isobaric heat capacity and speed of sound respectively. Based on these results, it can be concluded that SAFT-γ Mie (with A^{Born}, A^{ion}) is capable of modelling ILs accurately, even with respect to the most stringent test: second-order thermodynamic derivative properties.

However, the model in its current state is not without flaws. Despite having a small %AAD, the gradients of the lines in Figure 6 are all slightly off in the same manner, akin to a systematic error; for example, density is overestimated at low temperatures and underestimated at high temperatures. This observation persists even whe-

Table 3. Percentage average absolute deviation %AAD of liquid phase density, isobaric heat capacity and speed of sound of various ILs containing IM and BF4 group using the group contribution model.

Ionic Liquid	%AAD		
	ρ	C_P	ω
[EMIM][BF4]	0.45	0.66	0.38
[BMIM][BF4]	0.52	0.50	0.44
[EMIM][Ac]	0.45	0.19	0.26
[HMIM][BF4]	0.66	1.16	0.51
[OMIM][BF4]	0.54	2.61	1.58
[BMIM][Ac]	0.57	3.96	1.29

n only experimental data for one IL is used to estimate the parameters. Below, we go through the possible causes of this.

Firstly, in our method, the temperature dependence of D is ignored, and the density dependence is assumed to be linear. As seen in Figure 4, D does have a noticeable effect. If indeed D varies significantly with temperature and density and/or in a way that does not obey Equation 12, it could have undue consequences on the fidelity of the model. Watanabe et al. (2019) showed that D for 1-methylimidazolium acetate decreased by about 5 when the temperature is increased by 60K. As will be discussed in Section 4.4, a small change of 5 in D can cause an increase in %AAD by upwards of 5 folds.

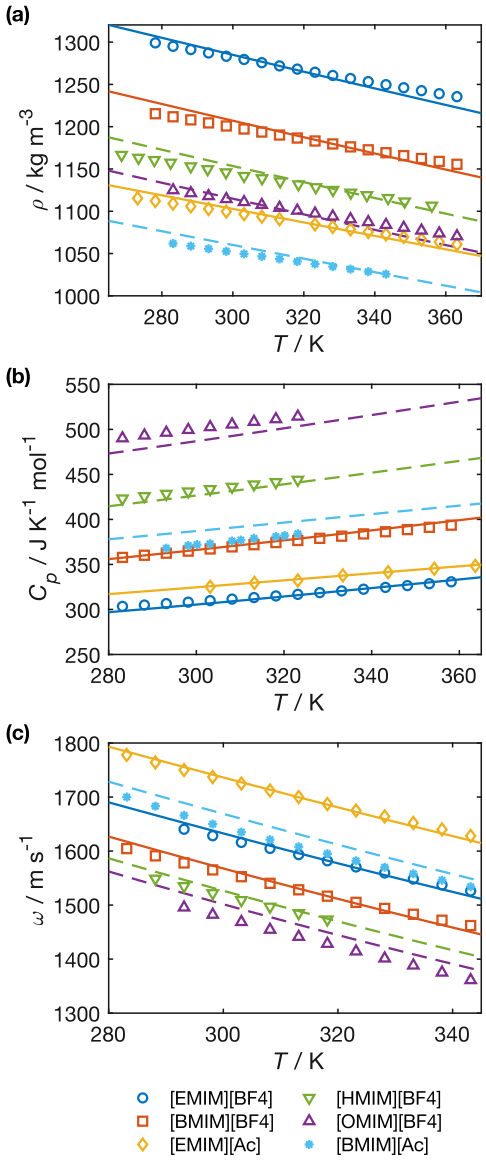


Figure 6. Group contribution model: (a) Liquid density, (b) liquid phase isobaric heat capacity, and (c) liquid phase speed of sound at 1atm for several ILs containing imidazolium IM and/or tetrafluoroborate BF4 groups. Solid curves represent SAFT-γ Mie calculations for ILs used in parameter estimation. Dashed curves represent SAFT-γ Mie predictions for ILs not used in parameter estimation. Symbols represent experimental data. (Neves *et al.*, 2013; Waliszewski *et al.*, 2005; Zarei and Keley, 2017; Vakili-Nezhaad *et al.*, 2012; Kumar, 2008; Froba *et al.*, 2010; Su *et al.*, 2016; Araujo *et al.*, 2013; Mokhtarani *et al.*, 2008; Waliszewski, 2008; Klomfar *et al.*, 2010; Pal and Kumar, 2012; Zorebski *et al.*, 2018)

Another possible reason is the lack of association contributions. Imidazolium and tetrafluoroborate have been shown to form hydrogen bonds (Dong *et al.*, 2006; Zheng *et al.*, 2013). That contribution could be the missing link and could be important when the model is extended to mixtures.

There is an uncertainty in the representation of the ideal isobaric heat capacity obtained from the Joback method. If the ideal heat capacity is inaccurate, this may cause distortion to the parameters as the optimiser attempts to compensate for this error.

Lastly, it could be the simple case that more unlike parameters need to be varied, such as ε_{kl} between IM and CH_3, and λ^r_{kl}.

The model can be extended and validated further by considering the variation with pressure, the extension to mixtures and gas solubilities, and the extrapolation to interfacial and transport properties.

4.4 Effects of the Dielectric Constant

A significant assumption made in our model is the dielectric constant D; how it varies from IL to IL and how it varies with temperature and density. Here, we evaluate the validity of the assumption and its possible consequences by looking at a hypothetical case where D is changed.

In Figure 7, we show how the calculations for the density of [EMIM][BF4] change if we vary the value of D while otherwise using the same parameters as used in Section 4.3. Not only do the values of the predictions change, but the slope of the predictions changes as well.

The resulting %AADs are tabulated in Table 4. At its greatest, decreasing D by 5, increases the %AAD from 0.45% to 2.51%, by a factor of more than 5.

Fitting multiple ILs using the same D is effectively the same as shifting each experimental data set up or down, where the magnitude of the shift depends on how different the true D is from the approximated D.

This is particularly important when working with ILs with vastly different D, such as 1-butyl-3-methylimidazolium iodide with D of 2.87 (Mou *et al.*, 2017), 1,3,-dimethylimidazolium dimethylphosphate with D of 29.6, 2-hydroxyethylammonium lactate with D of 85.6, and 1-(2-hydroxyethyl)-3-methylimidazolium tetrafluoroborate with D of 23.3 (Huang *et al.*, 2011) to name a few.

Therefore, careful attention needs to be given to D and the equations used to calculate D when making predictions and estimating parameters.

Table 4. Percentage average absolute deviation %AAD of liquid phase density at 1atm using different dielectric constants D with parameters estimated at $D = 15$.

D	%AAD ρ
10.0	2.51
12.5	1.02
15.0	0.45
17.5	1.14
20.0	1.92

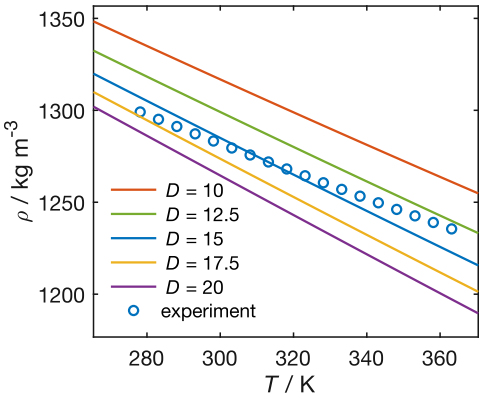


Figure 7. Liquid density of [EMIM][BF4] at different values of dielectric constants. Solid curves represent SAFT-γ Mie calculations using parameters estimated with D = 15. Symbols represent experimental data. (Neves *et al*., 2013)

5. Conclusion

The SAFT-γ Mie EOS with electrostatic contributions is used to model 1,3-dialkylimidazolium tetrafluoroborate ILs as fully dissociated ions. The performance of the model is evaluated against density, isobaric heat capacity and speed of sound experimental data. Two models at different degrees of coarse-graining are assessed: a simple sphere model and a group contribution model. The former is shown to be too crude to accurately represent all three properties simultaneously. The latter, however, provided good agreement with experimental data with %AADs between 0.19% and 0.66 %. The predictive capabilities of the new IM and BF4 group parameters introduced in this paper are evaluated, with %AADs between 0.51% to 3.96% for selected ILs. The electrostatic contributions are demonstrated to have a significant effect on the model calculations, and an inaccurate account of the dielectric effects can affect parameter estimation and predictions. Improvements could be made by accounting for association, pressure variation and an accurate description of the dielectric effect.

Acknowledgement

We would like to thank Prof. George Jackson, Dr Andrew J. Haslam, Pierre J. Walker, and Dr Felipe Antonio Perdomo Hurtado for their guidance and support.

References

Araújo, J. M. M., Pereiro, A. B., Alves, F., Marrucho, I. M., & Rebelo, L. P. N. (2013). Nucleic acid bases in 1-alkyl-3-methylimidazolium acetate ionic liquids: A thermophysical and ionic conductivity analysis. *The Journal of Chemical Thermodynamics*, *57*, 1–8. https://doi.org/https://doi.org/10.1016/j.jct.2012.07.022

Aschenbrenner, O., Supasitmongkol, S., Taylor, M., & Styring, P. (2009). Measurement of vapour pressures of ionic liquids and other low vapour pressure solvents. *Green Chem.*, *11*(8), 1217–1221. https://doi.org/10.1039/B904407H

Ashrafmansouri, S. S., & Raeissi, S. (2021). Extension of SAFT-γ to model the phase behavior of CO2+ionic liquid systems. *Fluid Phase Equilibria*, *538*, 113026. https://doi.org/https://doi.org/10.1016/j.fluid.2021.113026

Chapman, W. G., Gubbins, K. E., Jackson, G., & Radosz, M. (1989). SAFT: Equation-of-state solution model for associating fluids. *Fluid Phase Equilibria*, *52*, 31–38.

Dong, K., Zhang, S., Wang, D., & Yao, X. (2006). Hydrogen Bonds in Imidazolium Ionic Liquids. *The Journal of Physical Chemistry A*, *110*(31), 9775–9782. https://doi.org/10.1021/jp054054c

Dong, Y., Warsahartana, H. G., Hammad, F., & Masters, A. (2022). SAFT-γ Mie model for ionic liquids. *AIChE Journal*, *68*(6), e17697. https://doi.org/https://doi.org/10.1002/aic.17697

Fröba, A. P., Rausch, M. H., Krzeminski, K., Assenbaum, D., Wasserscheid, P., & Leipertz, A. (2010). Thermal Conductivity of Ionic Liquids: Measurement and Prediction. *International Journal of Thermophysics*, *31*(11), 2059–2077. https://doi.org/10.1007/s10765-010-0889-3

Ge, R., Hardacre, C., Jacquemin, J., Nancarrow, P., & Rooney, D. W. (2008). Heat Capacities of Ionic Liquids as a Function of Temperature at 0.1 MPa. Measurement and Prediction. *Journal of Chemical & Engineering Data*, *53*(9), 2148–2153. https://doi.org/10.1021/je800335v

Guzmán, O., Ramos Lara, J. E., & del Río, F. (2015). Liquid–Vapor Equilibria of Ionic Liquids from a SAFT Equation of State with Explicit Electrostatic Free Energy Contributions. *The Journal of Physical Chemistry B*, *119*(18), 5864–5872. https://doi.org/10.1021/jp511571h

Haslam, A. J., González-Pérez, A., Di Lecce, S., Khalit, S. H., Perdomo, F. A., Kournopoulos, S., Kohns, M., Lindeboom, T., Wehbe, M., Febra, S., Jackson, G., Adjiman, C. S., & Galindo, A. (2020). 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. https://doi.org/10.1021/acs.jced.0c00746

Huang, M. M., Jiang, Y., Sasisanker, P., Driver, G. W., & Weingärtner, H. (2011). Static Relative Dielectric Permittivities of Ionic Liquids at 25 °C. *Journal of Chemical & Engineering Data, 56*(4), 1494–1499. https://doi.org/10.1021/je101184s

Ji, X., Held, C., & Sadowski, G. (2012). Modeling imidazolium-based ionic liquids with ePC-SAFT. *Fluid Phase Equilibria*, *335*, 64–73. https://doi.org/https://doi.org/10.1016/j.fluid.2012.05.029

Joback, Kevin G. *A Unified Approach to Physical Property Estimation Using Multivariate Statistical Techniques*, Massachusetts Institute of Technology, 1 Jan. 1984, http://hdl.handle.net/1721.1/15374.

Klomfar, J., Součková, M., & Pátek, J. (2010). Temperature Dependence Measurements of the Density at 0.1 MPa for 1-Alkyl-3-methylimidazolium-Based Ionic Liquids with the Trifluoromethanesulfonate and Tetrafluoroborate Anion. *Journal of Chemical & Engineering Data*, *55*(9), 4054–4057. https://doi.org/10.1021/je100185e

Kumar, A. (2008). Estimates of Internal Pressure and Molar Refraction of Imidazolium Based Ionic Liquids as a Function of Temperature. *Journal of Solution Chemistry*, *37*(2), 203–214. https://doi.org/10.1007/s10953-007-9231-5

Lee, A. A., Vella, D., Perkin, S., & Goriely, A. (2015). Are Room-Temperature Ionic Liquids Dilute Electrolytes? *The Journal of Physical Chemistry Letters*, *6*(1), 159–163. https://doi.org/10.1021/jz502250z

Mejía-de Dios, J. A., & Mezura-Montes, E. (2019). A New Evolutionary Optimization Method Based on Center of Mass. In K. Deep, M. Jain, & S. Salhi (Eds.), *Decision Science in Action: Theory and Applications of Modern Decision Analytic Optimisation* (pp. 65–74). Springer Singapore. https://doi.org/10.1007/978-981-13-0860-4_6

Mokhtarani, B., Mojtahedi, M. M., Mortaheb, H. R., Mafi, M., Yazdani, F., & Sadeghian, F. (2008). Densities, Refractive Indices, and Viscosities of the Ionic Liquids 1-Methyl-3-octylimidazolium Tetrafluoroborate and 1-Methyl-3-butylimidazolium Perchlorate and Their Binary Mixtures with Ethanol at Several Temperatures. *Journal of Chemical & Engineering Data*, *53*(3), 677–682. https://doi.org/10.1021/je700521t

Mou, S., Rubano, A., & Paparo, D. (2017). Complex Permittivity of Ionic Liquid Mixtures Investigated by Terahertz Time-Domain Spectroscopy. *The Journal of Physical Chemistry B, 121*(30), 7351–7358. https://doi.org/10.1021/acs.jpcb.7b04706

Neves, C. M. S. S., Kurnia, K. A., Coutinho, J. A. P., Marrucho, I. M., Lopes, J. N. C., Freire, M. G., & Rebelo, L. P. N. (2013). Systematic Study of the Thermophysical Properties of Imidazolium-Based Ionic Liquids with Cyano-Functionalized Anions. *The Journal of Physical Chemistry B*, *117*(35), 10271–10283. https://doi.org/10.1021/jp405913b

Pal, A., & Kumar, B. (2012). Densities, speeds of sound and 1H NMR spectroscopic studies for binary mixtures of 1-hexyl-3-methylimidazolium based ionic liquids with ethylene glycol monomethyl ether at temperature from T=(288.15–318.15)K. *Fluid Phase Equilibria*, *334*, 157–165. https://doi.org/https://doi.org/10.1016/j.fluid.2012.08.002

Papaioannou, V., Lafitte, T., Avendaño, C., Adjiman, C. S., Jackson, G., Müller, E. A., & Galindo, A. (2014). Group contribution methodology based on the statistical associating fluid theory for heteronuclear molecules formed from Mie segments. *The Journal of Chemical Physics*, *140*(5), 054107. https://doi.org/10.1063/1.4851455

Process Systems Enterprise, gPROMS, www.psenterprise.com/products/gproms, 1997-2022

Ramdin, M., de Loos, T. W., & Vlugt, T. J. H. (2012). State-of-the-Art of CO2 Capture with Ionic Liquids. *Industrial & Engineering Chemistry Research*, *51*(24), 8149–8177. https://doi.org/10.1021/ie3003705

Ray, A., & Saruhan, B. (2021). Application of Ionic Liquids for Batteries and Supercapacitors. *Materials, 14*(11). https://doi.org/10.3390/ma14112942

Schreckenberg, J. M. A., Dufal, S., Haslam, A. J., Adjiman, C. S., Jackson, G., & Galindo, A. (2014). Modelling of the thermodynamic and solvation properties of electrolyte solutions with the statistical associating fluid theory for potentials of variable range. *Molecular Physics*, *112*(17), 2339–2364. https://doi.org/10.1080/00268976.2014.910316

Shukla, S. K., & Mikkola, J. P. (2020). Melting Point of Ionic Liquids. In S. Zhang (Ed.), *Encyclopedia of Ionic Liquids* (pp. 1–9). Springer Singapore. https://doi.org/10.1007/978-981-10-6739-6_109-1

Su, C., Liu, X., Zhu, C., & He, M. (2016). Isobaric molar heat capacities of 1-ethyl-3-methylimidazolium acetate and 1-hexyl-3-methylimidazolium acetate up to 16 MPa. *Fluid Phase Equilibria*, *427*, 187–193. https://doi.org/https://doi.org/10.1016/j.fluid.2016.06.054

Vakili-Nezhaad, G., Vatani, M., Asghari, M., & Ashour, I. (2012). Effect of temperature on the physical properties of 1-butyl-3-methylimidazolium based ionic liquids with thiocyanate and tetrafluoroborate anions, and 1-hexyl-3-methylimidazolium with tetrafluoroborate and hexafluorophosphate anions. *The Journal of Chemical Thermodynamics*, *54*, 148–154. https://doi.org/https://doi.org/10.1016/j.jct.2012.03.024

Waliszewski, D., Stępniak, I., Piekarski, H., & Lewandowski, A. (2005). Heat capacities of ionic liquids and their heats of solution in molecular liquids. *Thermochimica Acta*, *433*(1), 149–152. https://doi.org/https://doi.org/10.1016/j.tca.2005.03.001

Waliszewski, Dariusz. (2008). Heat capacities of the mixtures of ionic liquids with methanol at temperatures from 283.15K to 323.15K. *The Journal of Chemical Thermodynamics*, *40*(2), 203–207. https://doi.org/https://doi.org/10.1016/j.jct.2007.07.001

Walker, P. J., Yew, H. W., & Riedemann, A. (2022). Clapeyron.jl: An Extensible, Open-Source Fluid Thermodynamics Toolkit. *Industrial & Engineering Chemistry Research*, *61*(20), 7130–7153. https://doi.org/10.1021/acs.iecr.2c00326

Watanabe, H., Umecky, T., Arai, N., Nazet, A., Takamuku, T., Harris, K. R., Kameda, Y., Buchner, R., & Umebayashi, Y. (2019). Possible Proton Conduction Mechanism in Pseudo-Protic Ionic Liquids: A Concept of Specific Proton Conduction. *The Journal of Physical Chemistry B*, *123*(29), 6244–6252. https://doi.org/10.1021/acs.jpcb.9b03185

Zarei, H., & Keley, V. (2017). Density and Speed of Sound of Binary Mixtures of Ionic Liquid 1-Ethyl-3-methylimidazolium Tetrafluoroborate, N,N-Dimethylformamide, and N,N-Dimethylacetamide at Temperature Range of 293.15–343.15 K: Measurement and PC-SAFT Modeling. *Journal of Chemical & Engineering Data*, *62*(3), 913–923. https://doi.org/10.1021/acs.jced.6b00496

Zheng, Y. Z., Wang, N. N., Luo, J. J., Zhou, Y., & Yu, Z. W. (2013). Hydrogen-bonding interactions between [BMIM][BF4] and acetonitrile. *Phys. Chem. Chem. Phys.*, *15*(41), 18055–18064. https://doi.org/10.1039/C3CP53356E

Zorębski, E., Musiał, M., Bałuszyńska, K., Zorębski, M., & Dzida, M. (2018). Isobaric and Isochoric Heat Capacities as Well as Isentropic and Isothermal Compressibilities of Di- and Trisubstituted Imidazolium-Based Ionic Liquids as a Function of Temperature. *Industrial & Engineering Chemistry Research*, *57*(14), 5161–5172. https://doi.org/10.1021/acs.iecr.8b00506

Model-Based Design Space for Robust and Flexible CO_2 Capture Systems

Lola Truant and Ka Ying Li
Department of Chemical Engineering, Imperial College London, U.K.

Abstract Carbon capture technologies have been identified as a critical means of tackling climate change. In this work, we designed a vacuum swing adsorption (VSA) cycle for the carbon capture of post-combustion flue gas by assessing the process performance and flexibility of different adsorbents and operating conditions. A pipeline was developed that combines a VSA equilibrium-based model and design space identification to assess the performance of zeolite 13X, Mg-MOF-74 and UTSA-16 in terms of the purity (%) and recovery (%) of CO_2, the CO_2 working capacity of the bed (mol/m^3) and the specific energy usage (kWh/tonne) as well as the flexibility of the system. After screening, the CO_2 purities and recoveries for all three materials were found to be similar to one another, indicating that working capacity, energy usage and flexibility should be used as the key figures of merit for adsorbent screening. The design spaces obtained suggest that there is minimal flexibility in manipulating the evacuation pressure (P_L) as it dictates whether the stiff recovery constraint can be met. To minimise energy usage and overall capture cost, it was found that operating at a higher P_L is preferred, but this would be at the expense of process flexibility. As for adsorbent ranking, UTSA-16 offers the largest feasible operating region, the highest working capacity and lowest energy usage on average and the most flexibility during nominal operation.

1. Introduction

Fossil-based power generation has increased by 70% from 2000 and represents a significant contribution to total CO_2 emissions.[1] Carbon capture and storage (CCS) technologies have been identified as a promising means to mitigate emissions and meet the 2050 net-zero target, but it is critical to optimise the efficiency, cost, and energy impact of these technologies to ensure a sustainable future.[2] Absorption and adsorption are the most technologically advanced CCS techniques used at scale. Adsorption, however, is more energy efficient than absorption and does not require corrosive solvents such as monoethanolamine (MEA).[3]

Adsorption is a separation process where adsorbate molecules in fluid phase adhere to the surface of a solid adsorbent. Different cyclic adsorption configurations exist depending on the methods of adsorbent regeneration. Pressure swing adsorption (PSA) involves pressurizing the feed to above atmospheric pressure for adsorption and reducing it for desorption. In the case of adsorbates with highly non-linear isotherms like CO_2, it is favourable to reduce the desorption pressure further to vacuum level to enhance regeneration i.e., vacuum swing adsorption (VSA).[4] As gas compression to above-atmospheric pressure is energy intensive,[5] feeding the gas at atmospheric conditions is preferred. Cyclic VSA is thus considered in this study.

The choice of adsorbents is a key design variable in adsorption processes. Traditional materials such as zeolites and activated carbon as well as the novel metal organic frameworks (MOFs) are widely used in industry for post-combustion CO_2 capture systems.[6] Zeolite adsorbents are effective for separating out CO_2 in dry flue gas containing non-polar compounds[3] with high uptake at low partial pressures.[5] Zeolite 13X is the current benchmark for CO_2 capture processes[6] due to its low energy consumption.[7] Alternative materials, such as MOFs, have recently been studied extensively due to their excellent CO_2 adsorption capacity and selectivity.[8] In light of the development of more adsorbent materials for CCS,[7,9] selecting the most appropriate adsorbent for optimal process performance based on adsorbent properties, such as CO_2 adsorption capacity, selectivity, regeneration conditions, mechanical/chemical pellet stability, and cost[10] is a complex decision. Adsorbents with high selectivity for CO_2, working capacity and multicyclic stability are desired for an economic adsorption.[3,11,12] The key performance indicators (KPIs) typically used to assess the suitability of adsorbent materials are (i) product purity (ii) product recovery (iii) energy usage and (iv) working capacity.[6,7] Targets of at least 90% recovery and 95% purity are stipulated by the US Department of Energy (DoE) for CO_2 capture systems.[6,7,13] Adsorbents that satisfy the purity-recovery constraints should be further screened with the aim to minimise the overall process costs.[6]

A wide range of approaches have been employed for screening adsorbents and evaluating KPIs, such as experimental adsorption isotherm measurement,[14,15] neural network models[6] and detailed process optimisation.[5,16,17] The modelling and optimisation of cyclic adsorption is significantly computationally expensive, as it involves conflicting objectives, non-linear isotherms and coupled partial differential equations.[4,7,18] In contrast, an equilibrium-based model requires simpler computational algorithms and can yield both analytical and graphical solutions for a much larger range of operating and design

variables. An equilibrium model is implemented in this study to simulate the VSA process.

To evaluate design and operating strategies, multi-objective optimisation is typically used to obtain a single optimal operating point by maximising/minimising process KPIs. However, a suboptimal operating point might provide a larger feasible operating region and thus process flexibility. Therefore, traditional optimisation does not allow for an assessment of the full scale of operating parameters, neglecting the flexibility and robustness of the system. Our work employs model-based design space methodology, a novel framework developed by Sachio et al.[19] that accelerates process design and introduces operational flexibility as a new performance indicator. This gives insight into the extent of feasible operating regions for different adsorbents that satisfy the recovery and purity constraints. Assessing process flexibility allows the consideration of controllability in the early design stages. Controllability is an important practical consideration which is so far unaccounted for in the literature on the design of VSA processes.

This study aims to design an adsorption-based carbon capture process for a typical coal-fired power plant flue gas. A novel approach that combines an equilibrium-based model and design space identification (DSI) framework is used to evaluate process flexibility and alternative operating strategies. We simulate the four-step VSA cycle using the equilibrium-based model and assess the process performance using the four KPIs: CO_2 purity, CO_2 recovery, specific energy usage and working capacity for different adsorbent choices and operating conditions. The results are then used to identify the design spaces, or range of operating conditions under which purity and recovery constraints are met for different adsorbents.

This paper is structured as follows. Section 2 details the workflow in combining the equilibrium-based model and DSI framework. The process performances for different adsorbents are then presented in Section 3. Results are compared and discussed in Section 4.

2. Methodology

2.1 Adsorption Isotherm

The governing process considered in this work is a four-step VSA cycle for post-combustion carbon capture. The mixture to be separated is a dry flue gas of 15% moles CO_2 and 85% moles N_2. The process performance of using the benchmark adsorbent zeolite 13X was assessed along with two representative MOF materials, Mg-MOF-74 and UTSA-16. These are widely studied adsorbents that provide a good distribution of isotherm shapes while meeting the requirements of $PU_{CO_2} \geq 95\%$ and $RE_{CO_2} \geq 90\%$.[12] The extended dual-site Langmuir (DSL) model was used to describe the competitive adsorption equilibrium between CO_2/N_2 on each material, as given by:

$$q_i^* = \frac{q_{sb,i} b_i C_i}{1 + \sum_{j=1}^{n_c} b_j C_j} + \frac{q_{sd,i} d_i C_i}{1 + \sum_{j=1}^{n_c} d_j C_j} \quad (1)$$

where $q_{sb,i}$ and $q_{sd,i}$ are the saturation capacities of species i on site b and site d, in equilibirum with a fluid concentration of C_i. Adsorption equilibrium constants b_i and d_i are given by the van't Hoff equation:

$$b_i = b_{0,i} e^{-\Delta U_{b,i}/RT} \quad (2)$$

$$d_i = d_{0,i} e^{-\Delta U_{d,i}/RT} \quad (3)$$

where b_0 and d_0 are the pre-exponential factors, and $-\Delta U_{b,i}$ and $-\Delta U_{d,i}$ are the molar internal energies. The temperature is denoted as T and the universal gas constant as R. Assuming the mixture is fed at 1 bar and 298.15K, the CO_2 and N_2 adsorption isotherms are illustrated in Figure 1a and 1b. The isotherm parameters for the three adsorbents were obtained from the literature[7] and detailed in Supplementary Material Table A1. It can be observed that the CO_2 isotherms are much more nonlinear than the N_2 isotherms, indicating the selectivity for CO_2 is higher at lower pressures for all three materials. Mg-MOF-74 displays the largest adsorption capacity for both CO_2 and N_2, followed by zeolite 13X and UTSA-16.

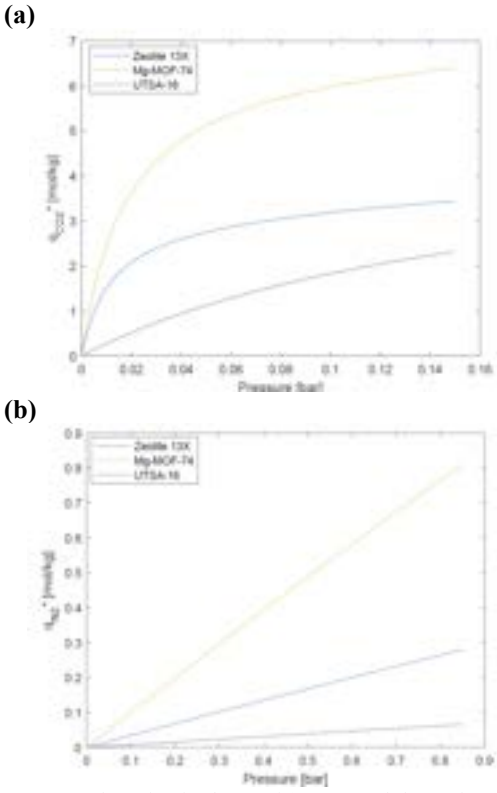


Figure 1. Adsorption isotherm of (a) CO_2 and (b) N_2 for zeolite 13X (blue lines), Mg-MOF-74 (yellow lines) and UTSA-16 (purple lines) at 298.15K.

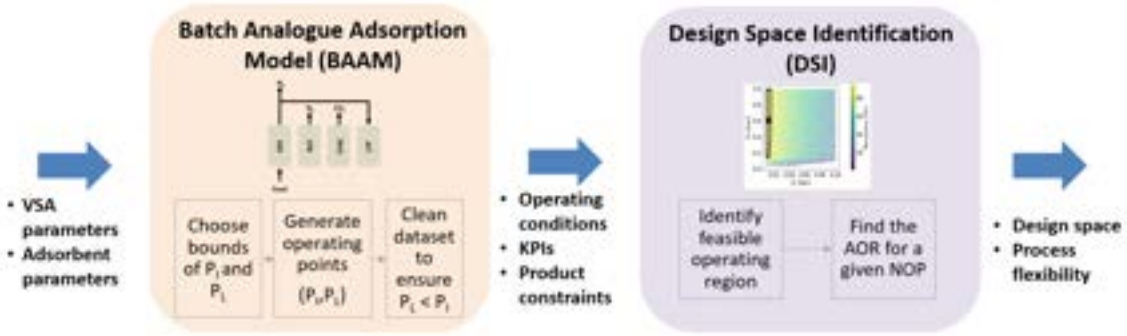


Figure 2. Schematic illustrating the workflow in linking BAAM and DSI models. The inputs and outputs from each model are outlined under the blue arrows. The processing steps within the models are detailed in the grey rectangles.

2.2 Adsorbent Screening Workflow

The model used for simulating the VSA process in this study is adopted from the Batch Adsorber Analogue Model (BAAM) developed by Balashankar et al.[7] and the modelling framework from Maring and Webley.[5] In order to appraise the robustness of different operating strategies, a novel methodology involving model-based design space identification (DSI) is also implemented.[19] The BAAM and DSI models are linked together to screen adsorbents, with the workflow as shown in Figure 2.

The BAAM model in MATLAB simulates the four-step VSA process. The inputs to the BAAM models are the VSA cycle parameters and the adsorbent parameters as detailed in Supplementary Material. Within the BAAM model, the first step is to choose the operating bounds for P_I and P_L. The lower bound for P_L is set at 0.01 bar, as any lower would not be technically achievable in CCS application.[20] The upper bound is taken as 0.1 bar to give a large enough range for process modelling.[21] Since P_H is at 1 bar, close to atmospheric pressure, the upper bound for P_I must be slightly lower, at 0.99 bar. The lower bound for P_I is chosen as 0.04 bar, which is just above the pressure achieved in pilot plant experiments.[7] The quasi-random Sobol sequence is then used to sample 4000 operating points/combinations of P_L and P_I within these bounds. Sobol sequencing is employed as it efficiently samples the operating parameter space with uniform coverage using only a small number of points.[22] To ensure the process is physically feasible, a data cleaning step is added to remove all the points where $P_L > P_I$.

Table 1. The upper and lower bound of the operating parameters in the four-step VSA cycle

Operating parameter	Lower bound	Upper bound
P_L [bar]	0.01	0.1
P_I [bar]	0.04	0.99

The BAAM model is evaluated at each operating point obtained from the Sobol sampling to generate the corresponding KPIs. The operating points and KPIs are then loaded into Python, where the DSI package is used. The operating points are screened using the constraints $PU_{CO_2} \geq 95\%$ and $RE_{CO_2} \geq 90\%$ to identify the design space where the product targets are satisfied. Design space metrics such as the average, maximum and minimum KPI values, the number of samples within the design space as well as its size are obtained to compare the adsorbents quantitively.

For each adsorbent, the nominal operating point (NOP) that maximises the acceptable operating region (AOR) is subsequently found using an iterative approach, along with information on the size and KPI values of this region. This allows us to compare the operation flexibility offered by each adsorbent during nominal operation.

2.3 Batch Analogue Adsorber Model (BAAM)

The four-step adsorption cycle modelled in BAAM consists of: (i) adsorption (ADS), (ii) blowdown (BLO), (ii) evacuation (EVAC) and (iv) light product pressurisation (LPP) as shown in Figure 3. First, the dry flue gas is fed at ambient conditions to saturate the bed at high pressure, $P_H = 1$ bar. Vacuum is then applied to reduce the pressure to an intermediate value, P_I, during blowdown, to remove N_2 from the adsorbent bed. In the evacuation step, pressure is further reduced from P_I to P_L to collect the highly concentrated CO_2 product. Finally, the raffinate stream is added in the LPP step to pressurise the bed from P_L to P_H again. This improves recovery as the CO_2 left in the raffinate stream can be recycled.

The modelling equations used in the equilibrium model are described in detail by Balashankar et al.[7] The equilibrium model allows for rapid simulation and process performance assessment of different adsorbents due to the following assumptions:

- Zero-dimensionality with no spatial gradients in temperature, pressure, and concentration
- Isothermal operation
- Instantaneous gas-solid equilibrium with negligible mass transfer resistance

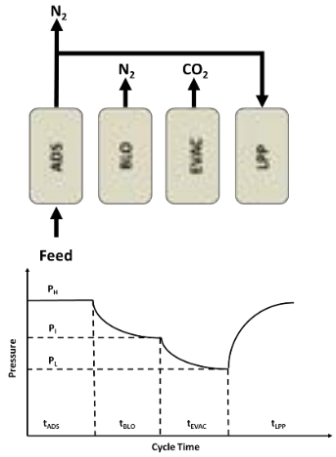


Figure 3. Four-step VSA cycle.

These assumptions reduce the set of coupled partial differential equations in a detailed dynamic model into a set of ordinary differential and algebraic equations. The process performance of the three materials is then evaluated in terms of the purity (%) and recovery (%) of the CO_2 product stream, the CO_2 working capacity of the bed (mol_{CO2}/m^3_{bed}) and the specific energy usage ($kWh/tonne_{CO2}$). The model is implemented in MATLAB and was available in the research group when we started the project.

2.4 Design Space Identification (DSI)

The design space (DS) considered in this study is the feasible operating region where the constraints $PU_{CO_2} \geq 95\%$ and $RE_{CO_2} \geq 90\%$ are met. Design space visualisation thus allows the identification of operating points that satisfy or violate the constraints. This provides insight into the process flexibility prior to carrying out expensive experimental study.[19,23]

A concave hull, alpha shape, is used to construct the design space boundary containing the set of feasible operating points. Alpha, α is an important parameter that dictates how fine/tight the shape is. The shape is convex hull as $\alpha \rightarrow \infty$ and consists of disjointed points for $\alpha = 0$.[24] Bisection method is employed to search for the optimal value for α. Starting with a large α value, the DS shape contains points that violate constraints. The value of α is then reduced to tighten the boundary until there are no violations. This ensures that we can obtain the biggest space possible without any violations with respect to the change in α tolerance. The DSI tool developed in Python allows us to visualise and quantify the DS for an input space with up to three dimensions. The AOR for a given NOP within the DS is identified using the tool. This determines the maximum allowable variations in the input operating parameters (P_I and P_L) under which the CO_2 purity and recovery constraints are still satisfied during nominal operation. The size of the DS and AOR as well as the relevant quantitative metrics can also be extracted. The Python DSI tool was available in the research group when we started the project.

3. Results

3.1 Pareto Fronts and Trade-offs between KPIs

A Pareto front is a non-dominated set of solutions representing the best available trade-off between conflicting objectives.[25] Using the approach described in Section 2, we obtained the CO_2 recovery and purity of each operating point. The trade-off in maximising the purity and recovery resulted in the Pareto fronts as shown in Figure 4a. The circles represent the different operating points of the Pareto fronts, and the green arrows indicate the optimal direction of trade-offs. It can be observed that there is no significant difference between the purity and recovery Pareto fronts of the adsorbents.

Working capacity and energy usage are commonly used as proxies for assessing the economic potential of the adsorption process.[8,12] Once the purity-recovery constraints are imposed, a Pareto front obtained by maximising working capacity while minimising energy usage was plotted for each adsorbent. From Figure 4b, we can see that UTSA-16 provides the best energy usage/working capacity trade-off while meeting the product constraints. Zeolite 13X and Mg-MOF-74 have similar Pareto front behaviours and are suboptimal relative to UTSA-16.

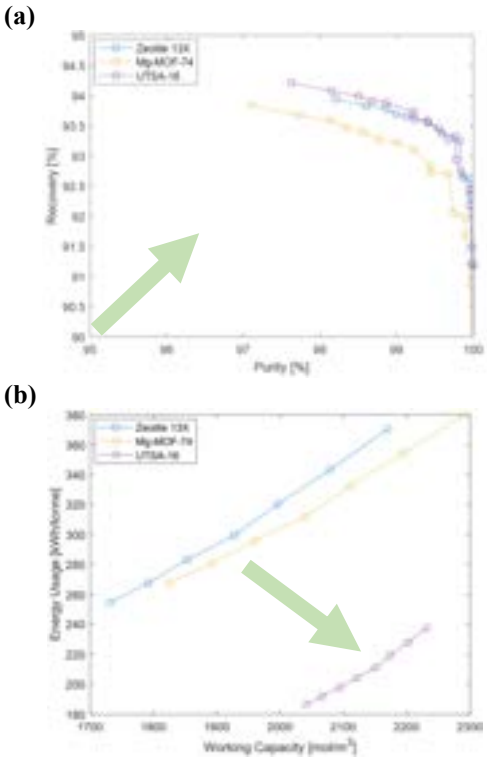


Figure 4. (a) Recovery/purity Pareto fronts and (b) constrained energy usage/working capacity Pareto fronts for zeolite 13X (blue lines), Mg-MOF-74 (yellow lines) and UTSA-16 (purple lines). The circles represent the different operating points of the Pareto front. The green arow indicates the optimal direction of trade-off.

3.2 Design Space and Process Flexibility

The flexibility of the process using different materials was assessed and visualised using the DSI framework described in Section 2.4. The resulting purity and recovery design spaces for the benchmark material, zeolite 13X are shown in Figure 5. The black boundary outlines the DS within which the operating conditions satisfy the constraints $PU_{CO_2} \geq 95\%$ and $RE_{CO_2} \geq 90\%$. The DS shapes and heat maps obtained for Mg-MOF-74 and UTSA-16 were found to exhibit similar behaviour.

The DS shape is rectangular with a very large length to width ratio, i.e., range of P_I to P_L. It is also confined by the lower bound of P_L and upper bound of P_I. It can be observed that the recovery constraint is harder to satisfy than the purity constraint, as most of the operating points that meet the recovery target are only found in the DS. This is not the case for purity, as suggested by the yellow region (representing near 100% purity) outside the DS on Figure 5b.

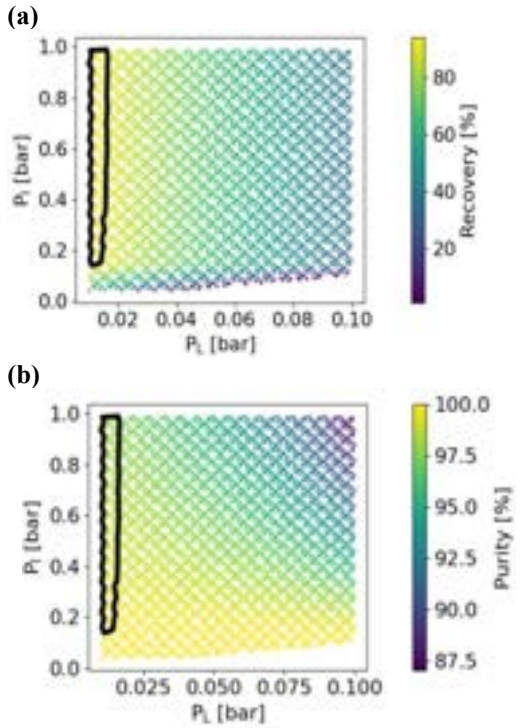


Figure 5. Design space for the benchmark adsorbent, zeolite 13X using (a) recovery as heat map and (b) purity as heat map. The black boundary outlines the design space.

In Figure 6, the DS overlayed with a heat map of working capacity (a-c) or energy usage (d-f) is shown. Using the DSI package (Section 2.4), we found the NOP with the largest AOR for each material. The NOP is indicated in the DS by the cross and the AOR is defined by the dashed rectangle. The maximum and minimum values of the colour bar scale were chosen corresponding to the maximum and minimum KPI values within the DS for each adsorbent. Since UTSA-16 offers an energy usage that is significantly lower than the other materials, a colour gradient would not be visible if the same scale were used for all three materials. The red and black points indicate values that are out of the range shown on the scale. The finding that UTSA-16 has the lowest energy usage (Figure 6f), agrees with the result from the constrained Pareto front (Figure 4b). The colour gradient of the heat map shows the rate of change of KPIs with respect to different operating conditions.

Despite exhibiting similar DS shapes, the DS sizes vary between adsorbents. It is the largest for UTSA-16, followed by zeolite 13X and Mg-MOF-74. The NOPs that yield the largest possible AOR are the same for Mg-MOF-74 and UTSA-16 at around $P_L = 0.013$ bar and $P_I = 0.846$ bar, whereas the NOP for zeolite is at the same P_L but at a lower P_I of 0.611 bar. The heat maps for both working capacity and energy usage in the DS are also comparable for all adsorbents.

4.0 Discussion

4.1 Pareto Fronts

As seen in Figure 4a, the purity/recovery Pareto fronts for all three materials meet the $PU_{CO_2} \geq 95\%$ and $RE_{CO_2} \geq 90\%$ constraints, implying that they are all suitable for postcombustion carbon capture process. Despite having different CO2 adsorption isotherms (Figure 1a), the three materials show similar purity/recovery Pareto fronts, which is in agreement with the literature.[12,26] This is observed because we pre-selected three materials that push the four-step VSA cycle to operate at its best performance. Typically, the purity/recovery Pareto behaviour would differ between materials.[8,20,27] Our selection of adsorbents for screening yielded similar results using the classical approach, highlighting the need for screening based on energy usage, working capacity and process flexibility.

Energy usage and working capacity are key in assessing the process economics, as the former is proportional to the operating cost (OPEX) and the latter is inversely proportional to the capital expenditure (CAPEX). The energy usage/working capacity Pareto fronts (Figure 4b) for zeolite 13X and Mg-MOF-74 are very close to each other even though Mg-MOF-74 has almost double the CO_2 affinity of zeolite 13X (Figure 1a). The similarity between the energy usage/working capacity Pareto fronts for zeolite 13X and Mg-MOF-74 implies that their volumetric-based capacities could be comparable despite having very different mass-based adsorption capacities. It is noted that UTSA-16 provides the best economic potential while meeting the purity-recovery constraints. Moreover, it is observed that the optimality of the Pareto solutions of the materials are correlated with the linearity of their adsorption isotherms. Working capacity is defined as the difference between

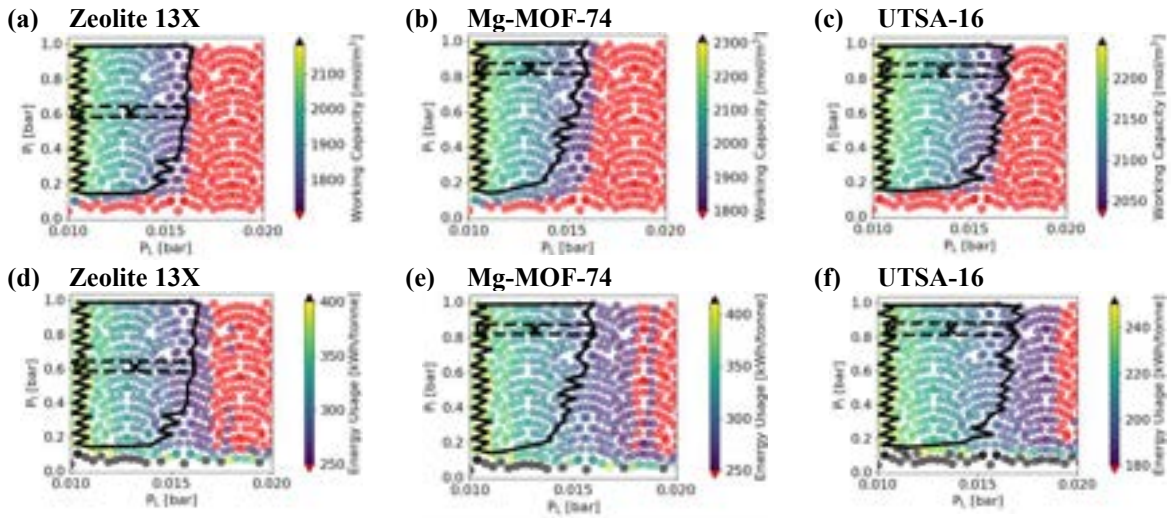


Figure 6. Design space with working capacity as heat map for (a) zeolite 13X, (b) Mg-MOF-74, (c) UTSA-16, and the design space with energy usage as heat map for (d) zeolite 13X, (e) Mg-MOF-74, (f) UTSA-16. The cross represents the NOP and the dashed rectangle represents the AOR.

equilibrium loading at P_I and P_L. In contrast to nonlinear isotherms, a linear isotherm allows reasonable values of P_I and P_L to be chosen to effectively evacuate CO_2 and achieve a high working capacity. The vacuum pumps would consume less energy when operating under less extreme conditions. The fact that UTSA-16 requires the least energy to achieve the product targets suggests that materials with a low N_2 affinity are desirable.

4.2 Purity/Recovery Design Space

Design spaces were identified and studied to evaluate the process robustness and flexibility. As mentioned in Section 3.2, the DS defined based on the purity-recovery constraints for the three materials are similar in shape due to their similar purity-recovery performance. Furthermore, the tall and rectangular shape of the DS suggests there is less flexibility in manipulating P_L than P_I. This is because P_L governs the evacuation step and thus the recovery of CO_2. A close examination of Figure 5a shows that there are fewer operating conditions that meet the recovery target, indicating that recovery is the stiffest constraint. Once P_L is specified to achieve the required working capacity and recovery target, purity can be tuned accordingly by manipulating P_I based on the N_2/CO_2 adsorption isotherms. P_I should be low enough to remove sufficient N_2 for high product purity without compromising the recovery. The purity constraint must be met for downstream geological storage to be feasible, whereas the recovery target is a recommendation of the DoE. The fact that the purity constraint is less stiff implies that even if the system operates slightly outside of DS, the purity target can still be met, keeping the CCS process feasible. Furthermore, the DS is located near the lower bound of P_L in order to meet the recovery target, which would result in a higher OPEX due to the vacuum pumps in the evacuation step.

The heat maps on Figure 5 allow us to assess how the system responds to changes in operating conditions with respect to the KPIs. The colour gradients indicate a higher dependency of purity on P_I, which dictates the amount of N_2 removed during the blowdown step. P_L has a more significant influence on recovery, as it governs the evacuation of CO_2. This is further supported by the results from the Sobol indices (Figure 7). Sobol method is a global sensitivity analysis tool for assessing process design heuristics. We calculated the Sobol indices using the SobolGSA tool[28] based on the 3915 operating conditions (after the data cleaning step) and the corresponding KPI values. First order indices show the most influential manipulated variables (P_L and P_I) towards a certain KPI. Second order indices correspond to the interactions between the manipulated variables. The first and second order indices sum up to one as proof of convergence. We can see that both purity and recovery are predominantly controlled by P_I and P_L respectively, further reinforcing the difficulty in optimising both purity and recovery. The same conclusion can be drawn for working capacity and energy usage.

Despite recovery being the stiffest constraint, as mentioned earlier, it is interesting to note that within the DS, recovery is invariant to the position of the operating point, whereas a higher purity can be achieved at a lower P_I.

4.3 Energy Usage/Working Capacity Design Space

The purity/recovery design spaces provide an initial understanding of the feasibility of the system to meet the CCS targets under different operating strategies, showing that all three materials offer promising carbon capture performances. However, further screening of materials should be based on

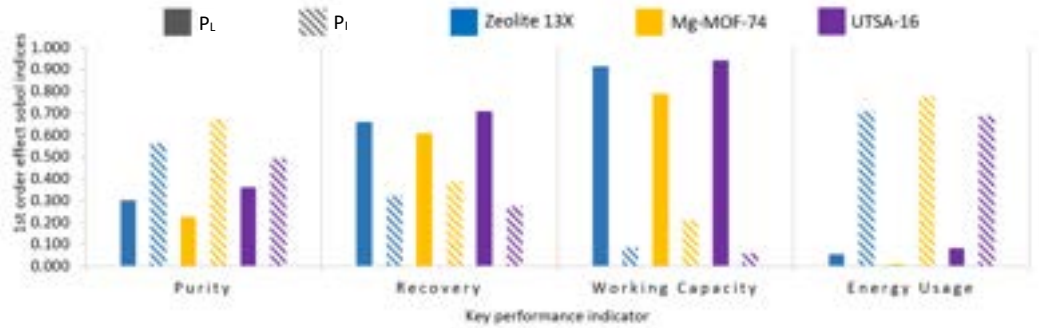


Figure 7. First order effect Sobol indices for purity, recovery, working capacity and energy usage for zeolite 13X (blue), Mg-MOF-74 (yellow) and UTSA-16 (purple). The filled bar corresponds to P_L and the patterned bar corresponds to P_I.

working capacity and energy usage, KPIs that govern process economics. We can easily observe the conflict in maximising working capacity and minimising energy usage and the trade-off that must be made from the colour gradients in Figure 6. As we operate away from the left of the design space, both energy usage and working capacity decrease. During practical operation, around 90% of the overall capture cost is typically attributed to electricity costs.[21] It is therefore desirable to operate at the right-hand side of the design space, i.e., at higher P_L, to minimise energy usage and costs. However, operating close to the design space boundary is done at the expense of process flexibility, indicating there is an inherent trade-off between process controllability and economics.

The colour gradient for working capacity changes in the direction of increasing P_L, but this is not the case for P_I (Figure 6a-c). This reinforces the fact that the working capacity is predominantly controlled by P_L and that the evacuation of CO_2 is primarily dependent on the depth of the vacuum swing. This result is in line with the Sobol indices. The heat maps for energy usage in Figure 6d-f are similar to those for working capacity, but there is also mixing of colours as P_I increases at a fixed P_L. Energy usage is therefore affected by both P_I and P_L, with the latter having the greater impact, agreeing with the fact that OPEX is dictated by the evacuation step. However, Sobol indices show that P_I has a dominant influence on energy usage. This could be attributed to the constraint imposed in BAAM to ensure $P_I > P_L$, such that the values of feasible P_L are governed by P_I. Indeed, the ranges of allowable P_L and P_I overlap and the range of P_I is much larger than the range of P_L (Table 1). This means that having a low P_I can effectively eliminate a high proportion of possible P_L values, but the converse is not true. P_I thus dominates P_L.

4.4 Overall Performance Comparison Between Adsorbents

Table 2 shows the proportion of operating points out of the 3915 samples (after the data cleaning step) located within the design space and the design space size. The metric values for the benchmark material, zeolite 13X are reported in absolute terms and Mg-MOF-74 and UTSA-16 are normalised with respect to zeolite 13X. We can note that UTSA-16 is superior in both DS metrics, offering the largest feasible region of operation.

Table 2. Design space metrics for the three materials. Absolute values for zeolite 13X are shown. Values for Mg-MOF-74 and UTSA-16 are normalised relative to zeolite 13X.

DS metrics	Zeolite 13X	Mg-MOF-74	UTSA-16
	Absolute	*Relative to zeolite*	*Relative to zeolite*
DS size	0.004 bar^2	0.857	1.080
Proportion of samples in DS	6.13%	0.892	1.096

As shown in Table 3, all materials perform similarly in terms of purity and recovery and meeting the product targets. UTSA-16 offers the highest working capacity and the lowest energy usage on average within the DS, indicating that UTSA-16 not only provides the largest feasible operating region but also the best process-scale performance across all KPIs. Mg-MOF-74 consumes more energy than the benchmark but offers a higher working capacity on average. The average energy usage obtained with this pipeline is in the same order of magnitude as the values quoted in the literature for a typical carbon capture plant (~250 – 380 kWh/tonne).[29, 30, 31]

Table 3. Average values of purity, recovery, working capacity and energy usage of the three materials within the design space. Absolute values for zeolite 13X are shown. Values for Mg-MOF-74 and UTSA-16 are normalised relative to zeolite 13X.

Average KPI in DS	Zeolite 13X	Mg-MOF-74	UTSA-16
	Absolute	*Relative to zeolite*	*Relative to zeolite*
Purity	98.861 %	0.992	0.997
Recovery	91.859 %	0.999	1.001
Working Capacity	1930.202 mol/m^3	1.066	1.105
Energy Usage	320.197 kWh/tonne	1.056	0.673

The DS metrics allow us to assess the range of possible operating strategies, whereas the KPI metrics give insight on which adsorbent offers the best process-scale performance. However, once a nominal operating point (NOP) is chosen in the design space, operation flexibility centred around that point can be assessed through the acceptable operating region (AOR). The process using UTSA-16 has the largest resulting AOR width, i.e., the range of possible P_L, at 0.00637 bar, followed by zeolite 13X at 0.00584 bar and Mg-MOF-74 at 0.00553. UTSA-16 therefore offers the most flexible operation. This is, however, a very small acceptable pressure variability, suggesting the VSA process is virtually inflexible.

Overall, UTSA-16 is the best in terms of DS metrics, all four KPI values and flexibility around the NOP. Mg-MOF-74 does not offer any significant process improvement to the benchmark adsorbent zeolite 13X. A similar conclusion is drawn in the literature.[12]

5.0 Conclusion

The aim of this work was to design a vacuum swing adsorption (VSA) cycle for the carbon capture of post-combustion flue gas by assessing the process performance and flexibility for different adsorbents and operating conditions. The current literature focuses on the use of computationally expensive modelling of the VSA process while neglecting flexibility in operation. We developed a novel pipeline that links an equilibrium-based model with a design space identification (DSI) framework. The pipeline completes the simulations and outputs graphical solutions within 15 minutes per material, which is significantly faster than a detailed dynamic model that requires several days of computational time[7]. Using this pipeline, we assessed the performance of zeolite 13X, Mg-MOF-74 and UTSA-16 in terms of the purity (%) and recovery (%) of the CO_2 product stream, the CO_2 working capacity of the bed (mol/m^3) and the specific energy usage (kWh/tonne) as well as the flexibility of the system.

It was found that all three materials meet the purity and recovery constraints given by the US Department of Energy (DoE). However, working capacity, energy usage and flexibility should be used as the key figures of merit for adsorbent screening, as the purity/recovery Pareto fronts and design spaces are similar for all materials. The tall and narrow shape of the design space (DS) indicates that there is less flexibility in manipulating P_L than P_I to achieve the product targets due to the stiffer recovery constraint. The results from the DSI analysis suggest that purity is governed by P_I while recovery, working capacity and energy usage are governed by P_L. The effect of the operating conditions on these KPIs was further assessed using Sobol indices, highlighting the complexity in designing and optimising the VSA process. Minimising the energy usage should be prioritised as it governs the overall carbon capture costs, which could be achieved by operating at the right-hand side of the DS, i.e., at a higher P_L. However, operating nearer to the design space boundary would compromise process flexibility, implying that the process would be more difficult to control in practical operations. Comparing the performances given by the three materials, UTSA-16 is the best in terms of the design space metrics, KPI values and flexibility around a nominal operating point. Mg-MOF-74 does not offer any significant process improvement to the benchmark adsorbent zeolite 13X.

The pipeline developed is a valuable tool for rapid initial adsorbent screening prior to experimental studies. More importantly, it allows us to consider process flexibility in an early design stage, which is an aspect often unaccounted for in the literature for VSA process design. Given that process flexibility is an important element in understanding the controllability and robustness of a process, further work in designing VSA should focus on addressing the trade-off between process flexibility and economics instead of choice of materials. A detailed technoeconomic analysis could be coupled with DSI to visualise how changing the nominal operating point (NOP) and acceptable operating region (AOR) would affect the overall capture cost, thereby achieving the expected economic performance even when disturbances occur. The pipeline we presented in this work is thus an effective tool in designing a robust and flexible adsorption-based carbon capture system.

Acknowledgements

We would like to extend our thanks to Adam Ward and Steven Sachio for their continued support throughout this project.

References

(1) IEA. The role of CCUS in low-carbon power systems https://www.iea.org/reports/the-role-of-ccus-in-low-carbon-power-systems/why-carbon-capture-technologies-are-important (accessed Dec 4, 2022).

(2) IEA. Carbon Capture, Utilisation and Storage https://www.iea.org/fuels-and-technologies/carbon-capture-utilisation-and-storage (accessed Nov 13, 2022).

(3) Wiheeb, A. D.; Helwani, Z.; Kim, J.; Othman, M. . Pressure Swing Adsorption Technologies for Carbon Dioxide Capture. **2016**.

(4) Haghpanah, R.; Majumder, A.; Nilam, R.; Rajendran, A.; Farooq, S.; Karimi, I. A.;

Amanullah, M. Multiobjective Optimization of a Four-Step Adsorption Process for Postcombustion CO2 Capture via Finite Volume Simulation. *Ind. Eng. Chem. Res.* **2013**, *52* (11), 4249–4265. https://doi.org/10.1021/ie302658y.

(5) Maring, B. J.; Webley, P. A. A New Simplified Pressure/Vacuum Swing Adsorption Model for Rapid Adsorbent Screening for CO2 Capture Applications. *Int. J. Greenh. Gas Control* **2013**, *15*, 16–31. https://doi.org/10.1016/j.ijggc.2013.01.009.

(6) Khurana, M.; Farooq, S. Adsorbent Screening for Postcombustion CO2 Capture: A Method Relating Equilibrium Isotherm Characteristics to an Optimum Vacuum Swing Adsorption Process Performance. *Ind. Eng. Chem. Res.* **2016**, *55* (8), 2447–2460. https://doi.org/10.1021/acs.iecr.5b04531.

(7) Subramanian Balashankar, V.; Rajagopalan, A. K.; De Pauw, R.; Avila, A. M.; Rajendran, A. Analysis of a Batch Adsorber Analogue for Rapid Screening of Adsorbents for Postcombustion CO 2 Capture. *Ind. Eng. Chem. Res.* **2019**, *58* (8), 3314–3328. https://doi.org/10.1021/acs.iecr.8b05420.

(8) Yancy-Caballero, D.; Leperi, K. T.; Bucior, B. J.; Richardson, R. K.; Islamoglu, T.; Farha, O. K.; You, F.; Snurr, R. Q. Process-Level Modelling and Optimization to Evaluate Metal-Organic Frameworks for Post-Combustion Capture of CO2. *Mol. Syst. Des. Eng.* **2020**, *5* (7), 1205–1218. https://doi.org/10.1039/d0me00060d.

(9) Hussin, F.; Aroua, M. K. Recent Trends in the Development of Adsorption Technologies for Carbon Dioxide Capture: A Brief Literature and Patent Reviews (2014–2018). *J. Clean. Prod.* **2020**, *253*, 119707. https://doi.org/10.1016/J.JCLEPRO.2019.119707.

(10) Raganati, F.; Miccio, F.; Ammendola, P. Adsorption of Carbon Dioxide for Post-Combustion Capture: A Review. **2021**. https://doi.org/10.1021/acs.energyfuels.1c01618.

(11) Ruthven, D. M. *Principles of Adsorption and Adsorption Processes*; Wiley-Interscience, 1984.

(12) Rajagopalan, A. K.; Avila, A. M.; Rajendran, A. Do Adsorbent Screening Metrics Predict Process Performance? A Process Optimisation Based Study for Post-Combustion Capture of CO2. *Int. J. Greenh. Gas Control* **2016**, *46*, 76–85. https://doi.org/10.1016/j.ijggc.2015.12.033.

(13) US DOE. Carbon Capture Newsletter. 2021.

(14) Harlick, P. J. E.; Tezel, F. H. An Experimental Adsorbent Screening Study for CO2 Removal from N2. *Microporous Mesoporous Mater.* **2004**, *1–3* (76), 71–79. https://doi.org/10.1016/J.MICROMESO.2004.07.035.

(15) Krishna, R.; Van Baten, J. M. A Comparison of the CO2 Capture Characteristics of Zeolites and Metal–Organic Frameworks. *Sep. Purif. Technol.* **2012**, *87*, 120–126. https://doi.org/10.1016/J.SEPPUR.2011.11.031.

(16) Hasan, M. M. F.; First, E. L.; Floudas, C. A. Cost-Effective CO2 Capture Based on in Silico Screening of Zeolites and Process Optimization. *Phys. Chem. Chem. Phys.* **2013**, *15* (40), 17601–17618. https://doi.org/10.1039/C3CP53627K.

(17) Khurana, M.; Farooq, S. Integrated Adsorbent-Process Optimization for Carbon Capture and Concentration Using Vacuum Swing Adsorption Cycles. *AIChE J.* **2017**, *63* (7), 2987–2995. https://doi.org/10.1002/AIC.15602.

(18) Webley, P. A.; He, J. Fast Solution-Adaptive Finite Volume Method for PSA/VSA Cycle Simulation; 1 Single Step Simulation. *Comput. Chem. Eng.* **2000**, *23* (11–12), 1701–1712. https://doi.org/10.1016/S0098-1354(99)00320-8.

(19) Sachio, S.; Kontoravdi, C.; Papathanasiou, M. M. *Model-Based Design Space for Protein A Chromatography Resin Selection*.

(20) Balogun, H. A.; Bahamon, D.; Almenhali, S.; Vega, L. F.; Alhajaj, A. Are We Missing Something When Evaluating Adsorbents for CO2capture at the System Level? *Energy Environ. Sci.* **2021**, *14* (12), 6360–6380. https://doi.org/10.1039/d1ee01677f.

(21) Ward, A.; Pini, R. Efficient Bayesian Optimization of Industrial-Scale Pressure-Vacuum Swing Adsorption Processes for CO2Capture. *Ind. Eng. Chem. Res.* **2022**. https://doi.org/10.1021/acs.iecr.2c02313.

(22) Sobol', I. M.; Shukman, B. V. Random and Quasirandom Sequences: Numerical Estimates of Uniformity of Distribution. *Math. Comput. Model.* **1993**, *18* (8), 39–45. https://doi.org/10.1016/0895-7177(93)90160-Z.

(23) Diab, S.; Gerogiorgis, D. I. Design Space Identification and Visualization for Continuous Pharmaceutical Manufacturing. *Pharmaceutics* **2020**, *12* (3). https://doi.org/10.3390/pharmaceutics12030235.

(24) Edelsbrunner, H.; Miicke, E. P. Three-

Dimensional Alpha Shapes. *Proc. 1992 Work. Vol. Vis. VVS 1992* **1992**, *13* (1), 75–82. https://doi.org/10.1145/147130.147153.

(25) Akbari, M.; Asadi, P.; Besharati Givi, M. K.; Khodabandehlouie, G. Artificial Neural Network and Optimization. In *Advances in Friction-Stir Welding and Processing*; Givi, M. K. B., Asadi, P., Eds.; Woodhead Publishing Series in Welding and Other Joining Technologies; Woodhead Publishing, 2014; pp 543–599. https://doi.org/https://doi.org/10.1533/9780857094551.543.

(26) Nguyen, T. T. T.; Lin, J. Bin; Shimizu, G. K. H.; Rajendran, A. Separation of CO2 and N2 on a Hydrophobic Metal Organic Framework CALF-20. *Chem. Eng. J.* **2022**, *442*, 136263. https://doi.org/10.1016/J.CEJ.2022.136263.

(27) Danaci, D.; Bui, M.; Mac Dowell, N.; Petit, C. Exploring the Limits of Adsorption-Based CO2 Capture Using MOFs with PVSA-from Molecular Design to Process Economics. *Mol. Syst. Des. Eng.* **2020**, *5* (1), 212–231. https://doi.org/10.1039/c9me00102f.

(28) Zaccheus, O.; Kucherenko, S. SobolGSA User Manual. **2021**, No. Version 4.1.1, 0–27.

(29) Lucquiaud, M.; Liang, X.; Errey, O.; Chalmers, H.; Gibbins, J. Addressing Technology Uncertainties in Power Plants with Post-Combustion Capture. *Energy Procedia* **2013**, *37*, 2359–2368. https://doi.org/10.1016/j.egypro.2013.06.117.

(30) He, X.; Britt Hägg, M. Energy Efficient Process for CO2 Capture from Flue Gas with Novel Fixed-Site-Carrier Membranes. **2014**. https://doi.org/10.1016/j.egypro.2014.11.018.

(31) Neveux, T.; Le Moullec, Y.; Corriou, J. P.; Favre, E. Energy Performance of CO2 Capture Processes: Interaction between Process Design and Solvent. *Chem. Eng. Trans.* **2013**, *35*, 337–342. https://doi.org/10.3303/CET1335056.

Supplementary Material

All supplementary materials are available at the end of this paper.

Lead-free Ternary ($Cs_3Bi_2Br_9$) and Double Halide ($Cs_2AgBiBr_6$) Perovskites for Efficient Photocatalytic Reduction of CO_2 to CO

Keliang Xu and Xinyu Zhang

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

Abstract

Lead-free halide perovskites were seen to be promising photocatalysts due to their non-toxicity, reasonable stability and good solar efficiency. In this paper, lead-free ternary perovskites caesium bismuth bromide ($Cs_3Bi_2Br_9$) and double halide perovskite caesium sliver bismuth bromide ($Cs_2AgBiBr_6$) were synthesised under different crystallisation times using the anti-solvent crystallisation method and compared in their performances of photocatalytic reduction of CO_2. One aim of this study was to achieve the correct synthesis as the synthesis of $Cs_2AgBiBr_6$ was known to be challenging. Further characterisations of perovskite crystals were taken using X-ray diffraction (XRD), scanning electron microscopy (SEM) and ultraviolent-visible (UV-VIS) spectrophotometer. Crystals were dispersed on half of the quartz filter and transferred into the reactor, the reaction was taking place under the light intensity that equalled to one sun. The production of CO was measured after one hour using a gas chromatograph (GC), which indicated the photocatalytic performance. It was discovered that the double halide perovskite synthesis was not as ideal due to degradation and the best performance appeared in the $Cs_3Bi_2Br_9$ crystals under one minute crystallisation time which produced 4.65 $\mu mol\ g^{-1}\ h^{-1}$ CO.

Keywords: Lead-free halide perovskites, semiconductor, photocatalysis, CO_2 reduction

1 Introduction and Background

Energy shortages have always been a global concern in recent days. In 2021, wholesale electricity prices in the European Union (EU) soared, and the prices of natural gas and coal rose quite a lot at the same time[1]. However, the demand for fossil fuels is still huge accounts for 82% of primary energy use in 2021[2]. With no doubt, global fossil CO_2 emissions rebounded, it increased by 5.3% in comparison to 2020, reaching 37.9 Gt CO_2[3]. As the Secretary-General stated at the conclusion of COP 27, the world needs to massively invest in renewables and end the addiction to fossil fuels[4].

Solar fuels have therefore received an extremely high level of public attention as the most economically viable, efficient, and environmentally friendly alternative to fossil fuels. One possible approach to producing solar fuels is "artificial photosynthesis". Similar to natural photosynthesis in plants where glucose is produced from water and carbon dioxide under sunlight, energy-poor molecules (H_2O and CO_2) are converted to energy-rich ones[5]. This is also known as the photocatalytic reduction of CO_2. Under the catalysation of semiconductor photocatalyst CO, methanol (CH_3OH), methane (CH_4), formic acid (HCOOH) and etc. can be produced. Metal oxides such as TiO_2, MgO, ZnO and WO_3 are common photocatalysts used in the photocatalytic conversion of CO_2, they have been studied for years and used in the industry[6]. Among all, TiO_2 is the most favoured choice as it is economical, non-toxic, and stable. However, the solar absorption of TiO_2 is actually very poor as it possesses a band gap of ~3.1 eV, absorbing ultraviolet light only[7]. There appears to be a strong desire of developing new semiconductor photocatalysts with better light absorption. Materials with lower band gaps are pursued so that visible light can be absorbed.

A perovskite is a material that has the same cubic crystal structure as the mineral $CaTiO_3$, following the chemical formula ABX_3. 'A' and 'B' represent metallic cations and 'X' is an anion that bonds to both[8]. Researchers first discovered the method to make a stable, thin-film perovskite solar cell with light photon-to-electron conversion efficiencies over 10%, using lead halide perovskites as the light-absorbing layer in 2012, the efficiency grew impressively to 25.2%[8]. Since then, metal halide perovskites gradually became very popular semiconductors to be used in various optoelectronic

fields as they have tuneable band gaps, long carrier diffusion lengths, high carrier mobilities and extraordinary tolerance of defects[9]. These are all characteristics that also make MHPs suitable catalysts for photocatalysis.

$CsPbBr_3$, as one of the typical representatives of MHPs, has been demonstrated as the promising photocatalyst for visible-light-driven photocatalytic CO_2 reduction since the stability of $CsPbBr_3$ was found to be quite high due to excellent PLQY (Photoluminescence Quantum Yield)[10]. However, research showed that the lead from halide perovskite was found to be more dangerous as its ten times more bioavailable compared to other resources of lead contamination that already appeared under the ground[11]. The toxicity of lead is always seen as the most serious concern for LHPs to be used widely. In recent days, more researchers spot on replacing lead in MHPs with metals like Tin (Sn), Bismuth (Bi) and Antimony (Sb)[12]. In the solar cell field, it is estimated that bismuth-based cells could convert light into energy at efficiencies of up to 22%. Bismuth is thus considered to be a suitable non-toxic alternative to lead in halide perovskite[13].

In this research, two Pb-free halide perovskites $Cs_3Bi_2Br_9$ and $Cs_2AgBiBr_6$ (double perovskite) were used as the catalysts in CO_2 photocatalytic reduction. $Cs_3Bi_2Br_9$ is a yellow crystal with a regular perovskite structure with Cs^+ ions in the centre of the cuboctahedron interstices. $Cs_2AgBiBr_6$ is orange and has a slightly different cubic structure that is built of alternating B' and B'' centred octahedrons of $B'X_6$ and $B''X_6$ in a 3D framework known as rock salt ordering[14]. $Cs_2AgBiBr_6$ has a lower band gap of 1.8-2.2 eV[15]. However, the synthesis of double perovskites is known to be more challenging. Hence, one aim of this experiment was to obtain the correct products. The synthesis also underwent four different crystallisation times (1 min, 15 mins, 30 mins, and 60 mins). The CO_2 photocatalytic reduction performances of eight samples were measured and compared. The best result appeared in $Cs_3Bi_2Br_9$ samples with one minute crystallisation time as it displayed the highest average CO production.

2 Experimental Methodology

2.1 Materials

Caesium bromide (CsBr), bismuth bromide ($BiBr_3$), silver bromide (AgBr), dimethyl sulfoxide (DMSO) and 2-propanol (IPA) were purchased from Sigma Aldrich. All the chemicals in this experiment were used without further purification. Precursors which are sensitive to water were kept in glove box and only a small amount was taken off before each synthesis.

2.2 Anti-solvent Crystallisation Synthesis

This experiment tried to produce $Cs_3Bi_2Br_9$ and $Cs_2AgBiBr_6$ by the following two reaction,

$$3CsBr + 2BiBr_3 \rightarrow Cs_3Bi_2Br_9$$

$$2CsBr + BiBr_3 + AgBr \rightarrow Cs_2AgBiBr_6$$

According to stoichiometry, 0.2078 g of CsBr and 0.2922 g of $BiBr_3$ were dissolved in 5 ml of DMSO respectively in two separate small tubes. DMSO, an anhydrous solution, was extracted using shlenk line with nitrogen. These two tubes were placed on a magnetic stirrer for 30 min at a stirring speed of 600 rpm to ensure all bromides were fully dissolved. Afterwards, two tubes were mixed and placed on a magnetic stirrer for another 30 min at 600 rpm. The neck of a 500 ml round flask containing 250 ml of IPA was clamped by an iron stand to make its bottom just touch the surface of a magnetic stirrer which operated at 400 rpm to create a stable vortex. The bromide-DMSO mixture was injected into the already vortexed IPA and crystallised for 1 min. The obtained solution was added into six centrifuge tubes and centrifuged at 10100 rpm for 2 min. After each centrifugation, the clear waste solution was emptied and more previous obtained solution was added and centrifuged until all solution was used up. The crystals left in the tubes were rewashed with anhydrous IPA for three times. Anhydrous IPA was extracted using shlenk line with nitrogen as well. All above procedures were repeated for different crystallisation times (15 min, 30 min and 60 min).

0.2004 g of CsBr, 0.0884 g of AgBr and 0.2112 g of $BiBr_3$ were dissolved in 5 ml, 10 ml and 5 ml of DMSO respectively in three separate small tubes. Followed by same procedures written above.

Eight samples were synthesised in total which are going to be abbreviated as CBB 1 min, CBB 15 min, CBB 30 min, CBB 60 min, CSBB 1 min, CSBB 15 min, CSBB 30 min and CSBB 60 min in this paper.

CBB stands for the ternary perovskite ($Cs_3Bi_2Br_9$) and CSBB stands for the double halide perovskite ($Cs_2AgBiBr_6$).

2.3 Characterisation

Chemical component quantities were determined by a SHIMADZU GC-2030 gas chromatograph (GC). Detailed operations are going to displayed in Section 2.4. Scanning electron microscopy (SEM) images were taken using a ZEISS AURIGA Cross Beam at 5 kV after coating the samples with a layer of chromium (15 nm). The samples were prepared by dispersing the powder on silicon wafers to ensure a homogenous dispersion. X-ray diffraction (XRD) was conducted using an Xpert Pro PANalytical diffractometer operating at 40 kV voltage and 40 mA current using Cu Kα (l ¼ 0.15418 nm) radiation in the 2θ range. Reflectance and absorbance spectra were collected with a SHIMADZU UV-2600 ultraviolent-visible (UV-VIS) spectrophotometer. 0.25 mg of each sample was diluted with 600 mg of barium sulfate, grinded using a mortar and pestle to make them homogeneous until only one colour was seen.

2.4 Photocatalytic Tests

A small amount of sample powders was dissolved in 2 ml of anhydrous IPA and then drop cast onto a quartz filter as even as possible to avoid sample agglomeration. The quartz filter was heated on a hot plate for 15 min to get rid of residual IPA. The sample mass was controlled around 0.8 mg to 1.2 mg. Then, the quartz filter was transferred to the reactor with 40 μl of water drops added away from the sample. The reactor used could contain 20 ml gas and quartz glass window was used to allow the full solar spectrum to reach the surface of the sample. The glass was cleaned with Kimtech Science Precision Wipes and DI water every time before and after the reaction.

Keeping the valve to GC closed, the reactor was evacuated with the vacuum pump to get rid of any gas in the pipes. CO_2 was pumped into the reactor slowly and the flowrate was gradually increased to maximum (50 ml/min). When the pressure went back to zero, the valve to GC was opened and the flowrate was turned to 10 ml/min for 15 min. After 15 min, a before-reaction GC measurement was taken, the gas flow was stopped and the valves before and after the reactor were closed.

After another 15 min, the light was turned on for one hour. AM 1.5G light was used with an intensity of 100 mW/cm^2 which was equal to one sun. After one hour, the valves before and after the reactor were opened and carbon dioxide was pumped in at 5 ml/min for 2.5 min. An after-reaction GC measurement was taken immediately after 2.5 min.

All above procedures were repeated three times for each sample.

3 Results and Discussion

To check if the desired perovskites were synthesised correctly, to have a picture of what the morphology of crystals was, to figure out the differences in band gaps and light absorption abilities, and to test the performances in photocatalytic of CO_2 reduction to CO, all eight samples were analysed by XRD, SEM, UV-Vis and GC and the results will be discussed in this section.

3.1 XRD

The XRD patterns of $Cs_3Bi_2Br_9$ synthesised in different crystallisation times are shown in Figure 1(a). All of the peaks in the XRD patterns of these four crystallisation times were identical to a standard PDF card (PDF #44-0714 from JADE 6.0) The values of FOM of $Cs_3Bi_2Br_9$ on reports produced by JADE 6.0 were all smaller than 2. This revealed a very successful synthesis of $Cs_3Bi_2Br_9$ crystals. There was also a strong interest in the effect of varying crystallisation time on crystallite size. The crystallite size 'd' of a material can be defined by the Scherrer equation for the line broadening of the peak,

$$d = \frac{k\lambda}{\beta \cos\theta}$$

where λ is the wavelength of the X-ray; β, FWHM (the full width at half maximum) of the diffraction peak; θ, diffraction angle; and k, constant[16]. The crystallite size is inversely proportional to the FWHM, hence, a smaller FWHM indicates a more crystalline structure.

Taking the (0,2,2) plane at 31.681° as an example, the FWHM values of each crystallisation time are

listed below. The FWHM values were all provided by Peak Search Report from JADE 6.0.

Table 1: FWHM values summary

Sample	FWMH (Degree)
CBB 1 min	0.202
CBB 15 min	0.177
CBB 30 min	0.179
CBB 60 min	0.197

Decreasing FWHM values from CBB 1 min to CBB 15 min implied that there was a growth of crystallite size in the (0,2,2) plane as time went on during the synthesis. Similar FWHM values were gained for CBB 15 min and CBB 30 min, yet grew up again to 0.197 for CBB 60 min sample. It was suspected that sometime after 15 min, there might be degradation of perovskite happening as the synthesis was taking place in the air. Degradation of perovskite could be affected by many different environmental factors like moisture, heat, UV and etc[17]. It caused chemical instability and some research showed that there was a clear decrease in the XRD peaks intensity and crystal size after exposing the perovskite sample to air[17]. As the IPA used in the synthesis was not anhydrous and the reaction was taking place in the air, there was a strong possibility of degradation. CBB 60 min sample was indeed suspected to be degrading as its (0,2,2) peak was obviously shorter than that of CBB 30 min, with the value of FWHM increasing at the same time.

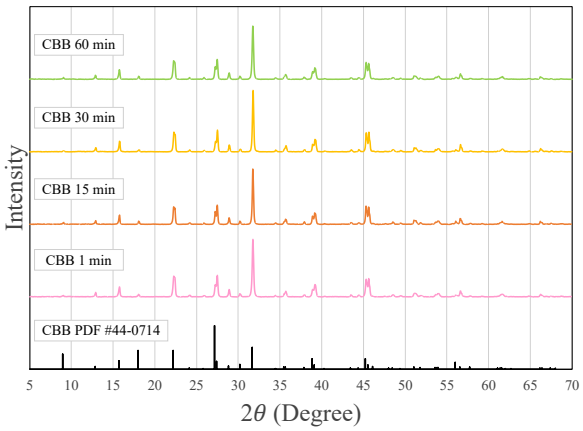


Figure 1(a): XRD pattern of $Cs_3Bi_2Br_9$

The XRD patterns of $Cs_2AgBiBr_6$ synthesised in different crystallisation times are shown in Figure 1(b). As compared to Figure 1(c), a $Cs_2AgBiBr_6$ XRD graph in another research, the XRD in this work was not very identical so further actions should be taken to find out the impurities. It was also discovered that after 1 min, three new peaks appeared at 30.944°, 44.328°and 55.039°, indicating the (2,0,0), (2,2,0) and (2,2,2) planes respectively. These were all peaks appeared in XRD patterns of AgBr. Hence, it was suspected that after 1 min, CSBB started to degrade and therefore AgBr appeared.

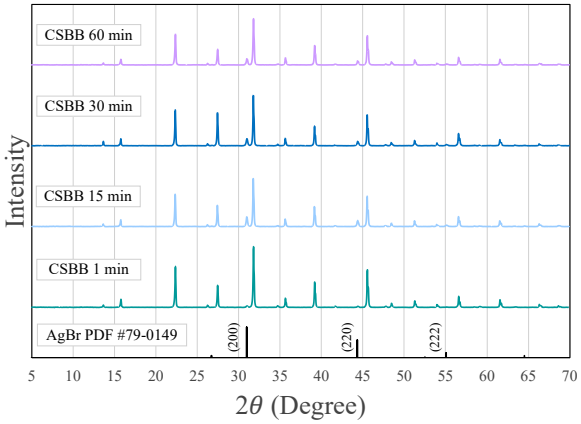


Figure 1(b): XRD pattern of $Cs_2AgBiBr_6$

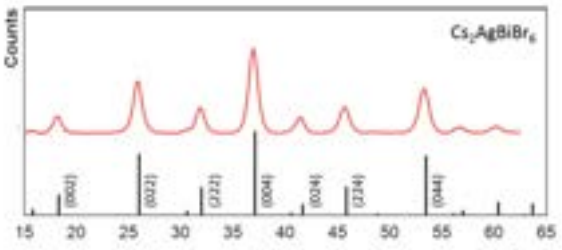


Figure 1(c): Reference and experimental XRD pattern of $Cs_2AgBiBr_6$[18]

3.2 SEM

The representative SEM of images of all eight samples were shown in Figure 2(a) for CBB and Figure 2(b) for CSBB. For $Cs_3Bi_2Br_9$ crystals, the largest size that a single crystal could grow increased slightly as the crystallisation time increased. However, the crystal size of four $Cs_3Bi_2Br_9$ samples was relatively similar but rough in general, with CBB 1 min sample displaying the finest crystal size among all. An increasing number of small crystals were seen to stick to those large ones as time passed. A reasonable guess of the mechanism could be made that small crystals will appear first, and then gather to grow into larger

crystals. However, the crystal size of the CBB 60 min sample became smaller, and that furtherly proved the guess of degradation taking place when crystallisation time increased. As discussed in the XRD characterisation part, the degradation of CBB was suspected to take place at a time after 15 minutes and for sure the CBB 60 min sample had degraded.

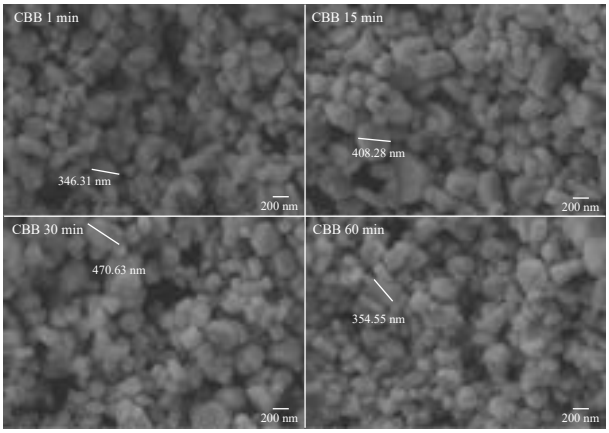


Figure 2(a): CBB SEM images (Mag = 30.00 K X)

$Cs_2AgBiBr_6$ crystals were more than twice as big as $Cs_3Bi_2Br_9$ crystals. The largest size appeared in CSBB 60 min sample, reaching 1.56 μm. In contrast to $Cs_3Bi_2Br_9$, the crystal structures of $Cs_2AgBiBr_6$ were obviously well-defined and a clear growing trend could be identified as the crystallisation time increased. This change can be proved furtherly by Figure 3 that $Cs_2AgBiBr_6$ powders became darker under longer crystallisation time.

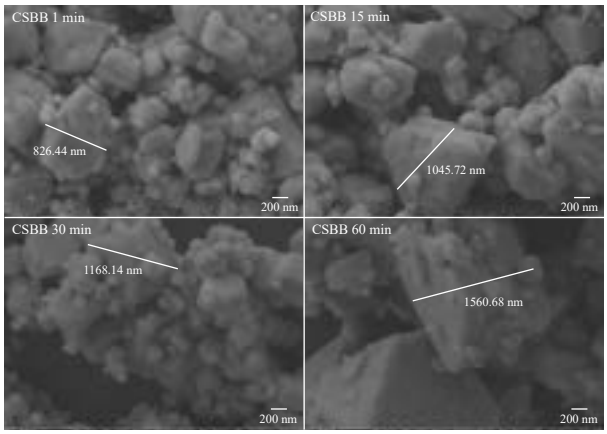


Figure 2(b): CSBB SEM images (Mag = 30.00 K X)

Figure 3: Photo of four CSBB samples where colour differences were clearly seen

3.3 UV-Vis

Since UV-Vis spectroscopy probes electronic transitions between valance band and conduction band, it is a convenient way for analysing the band gaps for semi-conductors[19].

3.3.1 Band Gaps Calculations Based on Reflectance Spectra

The minimum energy difference between the top of the valence band and the bottom of the conduction band is understood as the band gap. For a direct band gap semiconductor, the maximum energy of valence band and the minimum energy of the conduction band occur at the same value of momentum; for an indirect band gap semiconductor, they do not occur at the same value of momentum. Figure 4 simply explains why the value of an indirect band gap is always smaller than the value of a direct one.

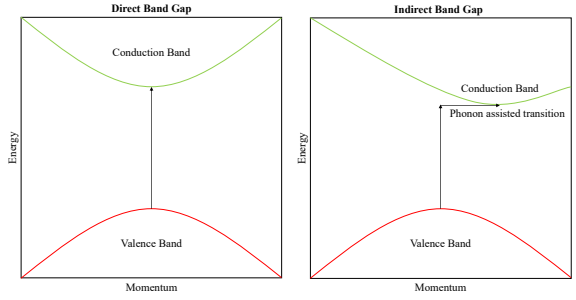


Figure 4: Direct and indirect band gaps[20]

The energy-dependent absorption coefficient, α, can be expressed by the following equation,

$$(\alpha \cdot h\nu)^{1/\gamma} = B\left(h\nu - E_g\right)$$

where h is the Planck constant, ν is the photon's frequency, B is a constant and E_g is the band gap energy. γ is equal to 1/2 for direct transition band gap and 2 for indirect transition band gap.

The Tauc plot of the Kubelka-Munk function was applied to estimate both direct and indirect band gap transitions of CBB and CSBB from their reflectance spectra (Figure 5(a) and Figure 6(a)). $F(R)$ is known as Kubelka-Munk function which is equal to,

$$F(R) = \frac{K}{S} = \frac{(1-R)^2}{2R}$$

where K is the molar absorption coefficient, S is the scattering coefficient and R is the diffused reflectance of material, $R = \frac{\%R}{100}$. Barium sulfate was used as a reference for the reflectance spectra so the R value mentioned in the above equation was actually taken to be the ratio of R_{sample}/R_{BaSO_4}.

By replacing α with $F(R)$ and substituting back to the first equation, the following equation was obtained,

$$(F(R) \cdot h\nu)^{1/\gamma} = B(h\nu - E_g)$$

Graphs with $h\nu$ as the x-axis and $(F(R) \cdot h\nu)^{1/\gamma}$ as the y-axis were plotted for each sample. Taken $\gamma = 1/2$, the x-intercept of the extension of the linear region and the x-axis was the direct band gap (Figure 5(b) and Figure 6(b)). Taken $\gamma = 2$, it was hard to identify the linear region (Figure 5(c) and Figure 6(c)). Therefore, two points lie on the steepest region were chosen, the x-intercept of the line pass through those two points and the x-axis was calculated to be the indirect band gap. The results are summarised in Table 2(a) and Table 2(b).

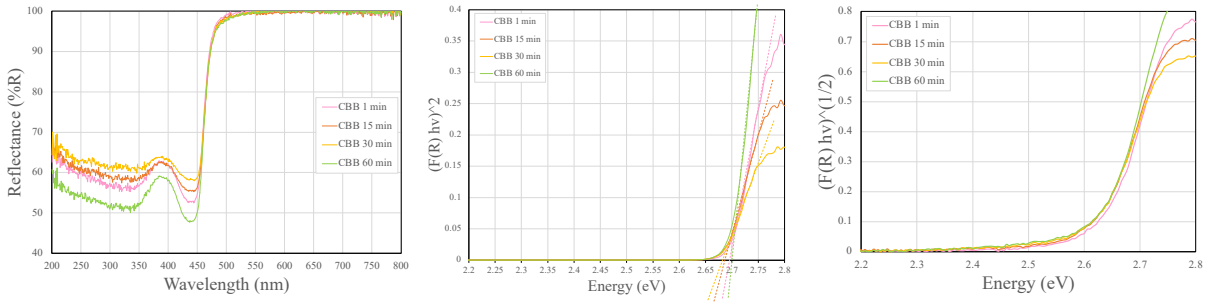


Figure 5(a): CBB reflectance spectra; Figure 5(b): CBB direct band gap; Figure 5(c): CBB indirect band gap

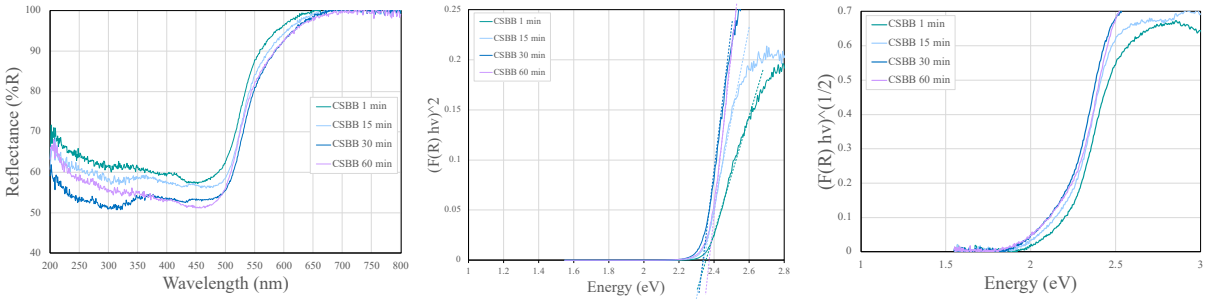


Figure 6(a): CSBB reflectance spectra; Figure 6(b): CSBB direct band gap; Figure 6(c): CSBB indirect band gap

Table 2(a): CBB band gaps summary

Sample	Direct Band Gap (eV)	Indirect Band Gap (eV)
1 min	2.695	2.6334
15 min	2.685	2.6174
30 min	2.680	2.6104
60 min	2.700	2.6186

Table 2(b): CSBB band gaps summary

Sample	Direct Band Gap (eV)	Indirect Band Gap (eV)
1 min	2.350	2.1782
15 min	2.340	2.1687
30 min	2.340	2.1607
60 min	2.360	2.1735

3.3.2 Band Gaps Analysis Based on Absorbance Spectra

As mentioned in Section 3.2, a darker colour was seen for CSBB samples with a longer crystallisation time. It was actually same for CBB but the colour differences were less obvious than what could be distinguished by eyes compared to CSBB (Figure 7). Based on these observations, it was imagined that a darker-coloured sample with a longer crystallisation time would absorb more light and have a lower band gap. However, there was not a tremendous decrease in the band gaps, even the direct band gaps' of CBB 60 min and CSBB 60 min were slightly higher than those of CBB 1 min and CSBB 1 min. Meanwhile, it was hard to conclude a trend in band gaps' change as the direct band gaps' difference of CBB 1 min and CBB 60 min was only 0.005 eV and that of CSBB 1 min and CSBB 60 min was only 0.01 eV; similarly, the indirect band gaps' differences were only 0.0148 eV and 0.0047 eV respectively.

Figure 7: Photo of samples where colour differences were seen

As this experiment aiming at exploring perovskite semiconductors better at absorbing light in the visible range, 380 nm to 700 nm range in the absorbance spectra (Figure 8) will be focused more in this section. CSBB samples absorbed a wider range of visible light than CBB samples. This was also verified by calculations that both direct and indirect band gaps of CSBB were lower than those of CBB.

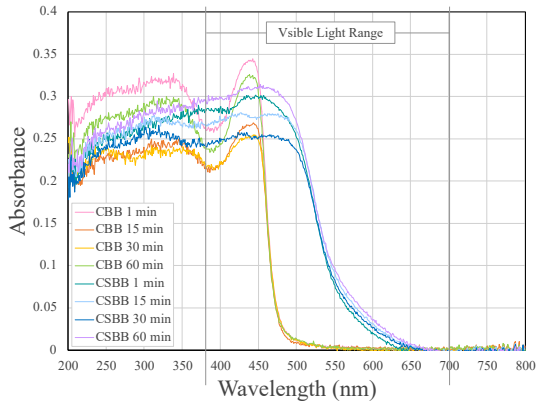


Figure 8: Absorbance spectra of CBB and CSBB

Nevertheless, XRD detected the existence of AgBr in CSBB samples which might contribute to the absorption of visible light. The indirect band gap of AgBr is 2.89 eV[21] which is narrow enough to make it absorb visible light easily[22]. This might be one of the reasons why CSBB absorbed a wider range of visible light than CBB.

3.4 GC

The amount of carbon monoxide production was monitored and measured as it implied the photocatalytic performances of each sample for CO_2 reduction.

3.4.1 General Trend

The sample to be considered as the best photocatalyst in this experiment was obviously CBB with 1 min of crystallisation time which helped to produced 4.65 $\mu mol\ g^{-1}\ h^{-1}$ CO. In general, for both CBB and CSBB, the CO production got less and less as the crystallisation time increased (Figure 9). This trend can be explained by the grow in crystal size which was also mentioned in Section 3.2 shown by SEM images. As the crystallisation time prolonged, the crystals became larger with a smaller surface area available for reactions to take place. As a result, less CO was produced for samples with longer crystallisation time. For CBB samples, the perovskite degradation also accounted for the reduction in CO production.

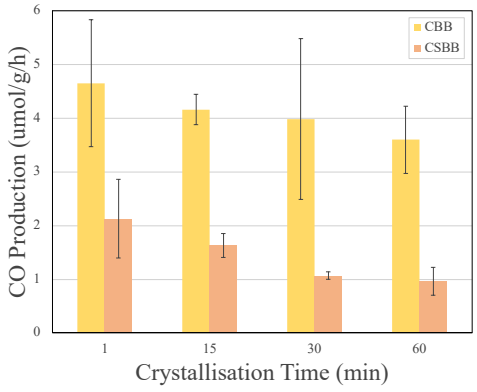


Figure 9: CO production for CBB and CSBB

3.4.2 Compare CBB with CSBB

According to the UV-Vis results discussed in Section 3.3.2, CSBB, owned a lower band gap and absorbed

more light, was hoped to offer a better photocatalytic ability. The GC results were opposite where CSBB only produced less than a half amount of CO compared with CBB under same crystallisation time. This can be explained by the following reasons. Firstly, CSBB crystals seen under SEM were much larger than CBB crystals; for instance, CBB 1 min crystals were approximately 364.31 nm whereas CSBB 1 min crystals were 826.44 nm which was 2.39 times larger. A larger size led to a smaller surface area and a lower production. Secondly, due to the existence of unwanted AgBr in CSBB samples, the sample mass weighed contained not only pure CSBB but also AgBr which indicated the actual CSBB sample mass participated in photocatalysis was less than the mass used in calculations. Namely, the CSBB sample mass might be overestimated and the production might be underestimated.

3.4.3 Control Tests

Two controlled tests were conducted for the samples with best photocatalysis performance for both CBB and CSBB, which were the ones with crystallisation time for 1 min. One of the controlled tests used helium only instead of CO_2 and H_2O, the other also replaced CO_2 with helium but kept H_2O as a reactant. It was predicted that the latter one would produce more CO but the result was opposite for CBB 1 min sample (Figure 10). An unexpected high CO production, 1.24 $\mu mol\ g^{-1}\ h^{-1}$, might be caused by residual IPA which was not fully removed in the quartz filter heating step. IPA might undergo a reduction reaction which contributed to the CO production. Some dark spots were observed in the quartz filter after reaction which proved the presence of remaining IPA in the sample (Figure 11).

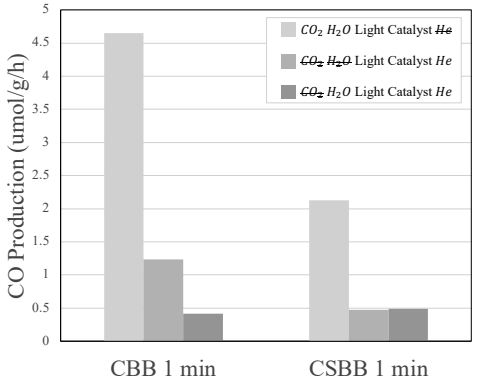


Figure 10: CO production in control tests

Figure 11: Photo taken after reaction proved the presence of remaining IPA

3.4.4 Improved Methodology

In the very start of this experiment, a whole quartz filter and there was no control in the sample mass, in other words, an over needed amount of samples were dropped on the filter and crystals overlapped with each other. Only the crystals on the top surface were active for catalysing the reaction while crystals hidden on the bottom were not effectively playing a role. Table 3 assembles some single experiment data which used different CSBB sample mass. One of the extreme cases used 14.3 mg of CSBB 15 min sample which only produced 0.0565 $\mu mol\ g^{-1}\ h^{-1}$ of CO. Another extreme case used 0.2 mg of CBB 60 min sample, 9.5162 $\mu mol\ g^{-1}\ h^{-1}$ of CO was produced which was even higher than the amount of CO produced using CBB 1 min sample. In the pursuit of a higher but also realistic production, the sample mass was decided to be controlled around 0.8 mg to 1.2 mg. Besides, the reaction time, water content and other parameters were optimised for half a filter, these were needed to be adjusted for using a whole one, which was the reason that only a half quartz filter was used as mentioned in Section 2.4.

Table 3: The effect of sample mass in CO production

Crystallisation Time (min)	Whole or Half Quartz Filter Used	CSBB Sample Mass (mg)	CO Production ($\mu mol\ g^{-1}\ h^{-1}$)
1	Whole	3.4	0.9648
	Half	0.8	1.7517
15	Whole	14.3	0.0565
	Half	0.8	1.5505
30	Whole	7.9	0.0598
	Half	0.1	0.9606
60	Whole	1.7	0.3315
	Half	1.1	0.5094

4 Conclusion and Outlook

To sum up, CBB 1 min sample was considered as the best photocatalyst in efficient reduction of CO_2 to CO which took part in reactions and produced 4.65 $\mu mol\ g^{-1}\ h^{-1}$ of CO. Related crystallisation time to the photocatalytic performances of perovskites, there was no doubt that a longer crystallisation time created a larger crystal size with a smaller surface area and a poorer performance. Based on a lower band gap and a stronger ability in absorbing a wider range of visible light, $Cs_2AgBiBr_6$ was predicted to be a better photocatalyst than $Cs_3Bi_2Br_9$. Nevertheless, CSBB samples helped to produce much less CO than CBB samples. This was due to not only a larger CSBB crystal size resulted in a smaller surface area available for reactions to take place, but also the appearance of AgBr after 15 min when CSBB started to degrade.

To further improve the study of these two perovskites, more trails of GC operations can be carried out. A larger number of sample volume definitely provides a fairer result and a smaller error bar as it was quite large, especially for CBB 30 min sample. In addition, the crystallisation time interval can be narrowed to see if there is a better crystallisation time for synthesising a better photocatalyst. For instance, is there any probability that a crystallisation time between 1 and 15 min does exist which offers a chance to produce the most amount of CO can be a future research topic. Furthermore, the presence of AgBr strongly adversely affected the performances of CSBB samples. A different synthesis methodology such as mechanochemical synthesis can be studied to check if it effectively produces the desired perovskites. Last but not the least, the BET[23] (Brunauer, Emmett and Teller) analysis can be conducted to examine how exactly the surface area of crystals changes and influence their photocatalytic performances.

References

1 Statista. (n.d.). EU: monthly electricity prices by country 2022. [online] Available at: https://www.statista.com/statistics/1267500/eu-monthly-wholesale-electricity-price-country/.

2 Bousso, R. (2022). Global energy consumption topped pre-pandemic levels in 2021, says BP. Reuters. [online] 28 Jun. Available at: https://www.reuters.com/business/energy/global-energy-consumption-topped-pre-pandemic-levels-2021-says-bp-2022-06-28/.

3 joint-research-centre.ec.europa.eu. (n.d.). Global CO2 emissions rebound in 2021 after temporary reduction during COVID19 lockdown. [online] Available at: https://joint-research-centre.ec.europa.eu/jrc-news/global-co2-emissions-rebound-2021-after-temporary-reduction-during-covid19-lockdown-2022-10-14_en.

4 Dropbox. (n.d.). SG COP OUTCOME POSITIVE.mp4. [online] Available at: https://www.dropbox.com/s/kv61ifi4iab66eq/SG%20COP%20OUTCOME%20POSITIVE.mp4?dl=0 [Accessed 6 Dec. 2022].

5 Dempsey, J.L., Winkler, J.R. and Gray, H.B. (2013). 8.15 - Solar Fuels: Approaches to Catalytic Hydrogen Evolution. [online] ScienceDirect. Available at: https://www.sciencedirect.com/science/article/pii/B9780080977744008068 [Accessed 14 Dec. 2022].

6 Fu, Z., Yang, Q., Liu, Z., Chen, F., Yao, F., Xie, T., Zhong, Y., Wang, D., Li, J., Li, X. and Zeng, G. (2019). Photocatalytic conversion of carbon dioxide: From products to design the catalysts. Journal of CO2 Utilization, 34, pp.63–73. doi:10.1016/j.jcou.2019.05.032.

7 Pitre, S.P., Yoon, T.P. and Scaiano, J.C. (2017). Titanium dioxide visible light photocatalysis: surface association enables photocatalysis with visible light irradiation. Chemical Communications, [online] 53(31), pp.4335–4338. doi:10.1039/C7CC01952A.

8 Washington.edu. (2012). Perovskite Solar Cell | Clean Energy Institute. [online] Available at: https://www.cei.washington.edu/education/science-of-solar/perovskite-solar-cell/.

9 Yuan, J., Liu, H., Wang, S. and Li, X. (2021). How to apply to photocatalysis: Obstruction and development of metal halide perovskite photocatalyst. Nanoscale. doi:10.1039/d0nr07716j.

10 Soosaimanickam, A., Rodríguez-Cantó, P.J., Martínez-Pastor, J.P. and Abargues, R. (2021). 2 - Preparation and processing of nanocomposites of all-inorganic lead halide perovskite nanocrystals. [online] ScienceDirect. Available at: https://www.sciencedirect.com/science/article/pi

i/B9780128199770000020 [Accessed 14 Dec. 2022].

11 Li, J., Cao, H.-L., Jiao, W.-B., Wang, Q., Wei, M., Cantone, I., Lü, J. and Abate, A. (2020). Biological impact of lead from halide perovskites reveals the risk of introducing a safe threshold. Nature Communications, 11(1). doi:10.1038/s41467-019-13910-y.

12 Yang, F., Wang, A., Yue, S., Du, W., Wang, S., Zhang, X. and Liu, X. (2021). Lead-free perovskites: growth, properties, and applications. Science China Materials, 64(12), pp.2889–2914. doi:10.1007/s40843-021-1755-4.

13 Renewablesnow.com. (n.d.). Study says bismuth could replace lead in perovskite solar cells. [online] Available at: https://renewablesnow.com/news/study-says-bismuth-could-replace-lead-in-perovskite-solar-cells-576478/#:~:text=July%2018%20%28Renewables%20Now%29%20-%20A%20new%20study [Accessed 14 Dec. 2022].

14 Meyer, E., Mutukwa, D., Zingwe, N. and Taziwa, R. (2018). Lead-Free Halide Double Perovskites: A Review of the Structural, Optical, and Stability Properties as Well as Their Viability to Replace Lead Halide Perovskites. Metals, 8(9), p.667. doi:10.3390/met8090667.

15 Kumar, S., Hassan, I., Regue, M., Gonzalez-Carrero, S., Rattner, E., Isaacs, M.A. and Eslava, S. (2021). Mechanochemically synthesized Pb-free halide perovskite-based Cs2AgBiBr6–Cu–RGO nanocomposite for photocatalytic CO2 reduction. Journal of Materials Chemistry A, [online] 9(20), pp.12179–12187. doi:10.1039/D1TA01281A.

16 www.sciencedirect.com. (n.d.). Scherrer Equation - an overview | ScienceDirect Topics. [online] Available at: https://www.sciencedirect.com/topics/materials-science/scherrer-equation [Accessed 14 Dec. 2022].

17 Mamun, A.A., Mohammed, Y., Ava, T.T., Namkoong, G. and Elmustafa, A.A. (2018). Influence of air degradation on morphology, crystal size and mechanical hardness of perovskite film. Materials Letters, 229, pp.167–170. doi:10.1016/j.matlet.2018.06.126.

18 Bekenstein, Y., Dahl, J.C., Huang, J., Osowiecki, W.T., Swabeck, J.K., Chan, E.M., Yang, P. and Alivisatos, A.P. (2018). The Making and Breaking of Lead-Free Double Perovskite Nanocrystals of Cesium Silver–Bismuth Halide Compositions. Nano Letters, 18(6), pp.3502–3508. doi:10.1021/acs.nanolett.8b00560.

19 Chen, Z. and Jaramillo, T. (n.d.). The Use of UV-visible Spectroscopy to Measure the Band Gap of a Semiconductor Section 1: Introduction to UV-Vis spectroscopy. [online] Available at: https://mmrc.caltech.edu/Cary%20UV-Vis%20Int.Sphere/Literature/Spectroscopy%20Jaramillo.pdf.

20 Doitpoms.ac.uk. (2018). DoITPoMS - TLP Library Introduction to Semiconductors - Direct and Indirect Band Gap Semiconductors. [online] Available at: https://www.doitpoms.ac.uk/tlplib/semiconductors/direct.php.

21 Yu, J., Mietek Jaroniec and Jiang, C. (2020). Surface Science of Photocatalysis. London Elsevier, Ap, Academic Press.

22 Wang, D., Zhao, M., Luo, Q., Yin, R., An, J. and Li, X. (2016). An efficient visible-light photocatalyst prepared by modifying AgBr particles with a small amount of activated carbon. Materials Research Bulletin, [online] 76, pp.402–410. doi:10.1016/j.materresbull.2016.01.003.

23 Particle Technology Labs. (2011). BET SPecific Surface Area. [online] Available at: https://www.particletechlabs.com/analytical-testing/gas-adsorption-and-porosimetry/bet-specific-surface-area.

Polymorphic Phase Transitions for Triglycine: The Effect of Additives and Temperature on the Thermodynamic Stability of Triglycine's Polymorphic Forms

Kelvin Choo (01706258) and Yuyang Cen (01487301)
Department of Chemical Engineering, Imperial College London, U.K.

ABSTRACT

Within the pharmaceutical industry, there is a strong motivation to study the effect of temperature and additives on the morphology of peptides during the crystallisation process to improve the safety, tolerability profiles, and long-term efficacy of peptide-based therapies. This work aimed to investigate the effects of temperature, ethanol concentration and salt additives on the phase transition boundaries between the two polymorphic forms of triglycine: triglycine anhydrate (β-sheeted structure) and triglycine dihydrate (polyproline II / pPII structure). Slurry crystallisation experiments were conducted by mixing triglycine anhydrate in (a) temperature-controlled binary water-ethanol solutions of varying compositions or with (b) salt additives, followed by observation of resultant crystals using microscope images and Raman spectroscopy. The former set of experiments revealed that the dihydrate form was favoured by an increase in ethanol molar fraction and temperature. Further analysis of the addition of four different salts (LiCl, NaCl, KCl and $MgCl_2$) showed that anhydrate formation was favoured at high salt concentrations (>1.5M), with prospective evidence implying that a higher cation charge density results in a lower salt concentration required for phase transition.

Keywords: *Crystallisation of triglycine, polymorphic conformation, phase transition boundary, pPII structure, Raman spectroscopy, crystallisation additives*

1. Introduction

Therapeutic peptides constitute a novel class of pharmaceutical agents generally defined as a short chain of 2-50 amino acids. As synthetically accessible bioactive substances, peptides boast many unique qualities that makes them strong potential candidates when formulating best-in-class therapeutics. Much like biologics, such as proteins and antibodies, therapeutic peptides function by binding to cell surface receptors and triggering intracellular effects with higher specificity and potency than small molecules. Unlike biologics, peptides are generally less immunogenic, and their smaller sizes allow for deeper penetration into human tissue. In addition, peptides tend not to accumulate in organs and therefore, typically have low toxicities (McGregor, 2008). Due to these characteristics, peptide-based therapies have the potential to offer superior patient outcomes in terms of improved safety, tolerability profiles, and long-term efficacy (Bruckdorfer, et al., 2004) (Fosgerau and Hoffmann, 2015).

With the advancement of drug discovery and development strategies, peptide synthesis techniques, and molecular pharmacology in recent years, peptide therapies are playing an increasingly large role in addressing unmet medical needs. Today, there are over 80 peptide therapies on the market and 550-750 in clinical or pre-clinical development across a wide range of indications such as oncology, cardiology and endocrinology (Lau et al., 2018) (Muttenthaler et al., 2021). Despite the apparent boons associated with these advancements and the outstanding pharmacological profile of peptides, there remains many challenges associated with peptide production and manufacturing that limit the proliferation of peptide therapies. In particular, downstream purification remains a major bottleneck in peptide manufacturing and relies on chromatography-based techniques. This often necessitates excessive use of reagents and solvents to obtain high quality peptides, resulting in high costs, poor environmental impact, and large amounts of aqueous waste (Isidro-Llobet et al., 2019).

To address these challenges, the pharmaceutical community has proposed peptide crystallisation as a more cost-efficient and practical alternative to chromatography. However, given the complex properties and large sequence space for peptides (n^{20}), selecting appropriate crystallisation conditions is a laborious, iterative process. The current industry standard involves screening using standard sets of conditions and further fine tuning around conditions with promising results. For a single peptide, this could mean several cycles of optimisation before a satisfactory outcome is achieved, with each lasting days or weeks (Rosa et al., 2020). In addition, peptides often have highly flexible conformations due to their relatively short amino acid chain lengths. The manifestation of different morphological conformations further complicates the screening process as this could affect critical physical and chemical properties of a given therapeutic peptide such as solubility, stability, and biological properties (Guo et al., 2021).

This study aimed to investigate how key conditions affect the most stable conformation of peptides in aqueous systems and further efforts to

develop a more robust, systematic approach to identifying ideal crystal growth condition for peptides, using triglycine as a model peptide. Triglycine exhibits two morphologies: triglycine anhydrate and triglycine dihydrate. Primarily, there were two areas of interest that were investigated via slurry crystallisation experiments. Firstly, the phase transition boundaries between these two conformations were analysed at different temperatures in binary water-ethanol solutions of varying concentrations. Insights from these experiments would be particularly useful as anti-solvents, such as ethanol, are commonly used to assist precipitation of high solubility solutes. Secondly, the critical salt concentrations at the phase transition boundaries were investigated for four different salts (LiCl, NaCl, KCl, $MgCl_2$) to understand the potential role of salt addition in achieving the desired morphological conformation during peptide production.

2. Background

There have been many studies conducted on protein crystallisation. Homogenous nucleation in the absence of any foreign particles is particularly well studied (Karthika et al., 2016). While many new crystallisation theories have emerged to understand the process (Lutsko, 2019), the Classical Nucleation Theory (CNT) is the most common model used to understand the crystallisation of biological molecules (Karthika et al., 2016) owing to its ability to give reasonable predictions of nucleation rates despite its simplicity (Sear, 2007).

Under the CNT, the crystallisation process is described by two separate steps: (a) an initial nucleation step, followed by (b) a growth process (Xu et al., 2021). CNT assumes that the crystal nucleus has the same properties and structure as the stable, mature crystal, and that the nucleus is spherical in shape. While the assumptions are not valid in certain cases, they serve as a fair approximation moving forward. Broadly speaking, crystallisation methods fall into one of the following three categories: batch, vapour diffusion or liquid diffusion (Bergfors, 2009, pp.17). The three methods differ in how the solution chemistry (degree of supersaturation) is adjusted. However, in all cases, the level of supersaturation must be high enough to initiate nucleation and support subsequent crystal growth (Ducruix and Giegé, 1992). In this study, batch crystallisation was chosen due to their ease of setting up (Bergfors, 2009, pp. 19).

Polymorphism is commonly observed during the crystallisation in the industry, specifically in the fields of pharmaceuticals, food and fine chemicals. This phenomenon refers to the existence of multiple crystalline structures for a single compound which, during the crystallisation process, is often influenced by the presence of additives and crystallisation conditions like temperature (Kitamura, 2008). The presence of polymorphs is of particular interest as polymorphs demonstrate different physicochemical properties (Cruz-Cabeza and Bernstein, 2013). Complications may arise for biomolecules with many, flexible conformations, such as in peptides and proteins. In the pharmaceutical industry, a polymorphic phase transition could mean, among other factors, a change in solubility and physical stability, impacting the dosage delivered, drug bioavailability and half-life when administered to patients. This also presents significant regulatory risk to pharmaceutical manufacturers since drug approvals are most often granted only for a single polymorph, highlighting the imperative need to control the morphology of the active pharmaceutical ingredient during the crystallisation process (Barker, 2020). At best, the wrong polymorphic form can cause economic losses to manufacturers, as in the case of Abbott and ritonavir, and at worst, it could even lead to losses of lives or permanent disfiguration, such as in the case of the thalidomide scandal.

The presence of polymorphs introduces an additional layer of complexity in this study. During peptide crystallisation, these factors will influence the thermodynamic stability of the secondary structure of the peptide's morphologies and, generally, the most stable morphology would be crystallised (Cruz-Cabeza and Bernstein, 2013). However, this may not always be the case. Rather, the crystals may pass through a path where the free energy barrier is minimum during nucleation. This means that the crystals obtained from experiments may not be an accurate reflection of what the most stable form is under those specific conditions (Karthika et al., 2016). Ostwald's step rule suggests that the polymorph that crystallises first is usually the metastable one, which more closely resembles the state in the solution and are thus advantaged. This is followed by a transformation to more stable forms over time (Black et al., 2018). As such, it is vital for the experiment set-up to be stirred and left overnight for the crystals to transform into their most stable form.

In the pharmaceutical industry, hydrates constitute an extremely common class of solvates and are viewed as an important area of research. Their prevalence can be attributed to the common use of water as a solvent in pharmaceutical processing (Hilfiker, 2006). As with other polymorphs, hydrates have different physical and chemical properties from anhydrates (Tian et al., 2010). As such, it is vital to understand the factors that affects the stability of the hydrates to control the type of peptide being crystallised to ensure higher efficacy, reduce regulatory risks and improve patient outcomes.

Previous studies done have shown that the stability depends on various factors. One factor is the choice of solvent used. For instance, olanzapine exhibits over 25 crystal forms. Seven of these are pharmaceutically relevant, being three anhydrates (Form I to III), three dihydrates (B, D, and E) and a higher hydrate (Reutzel-Edens et al., 2003). Form I is the most stable form in organic solvents, while in water, Form I is the least stable and it will convert to a dihydrate. In a separate study done on the stability of an indomethacin/methanol system, Veith et al. (2020) found that indomethacin-methanol solvate formation is favoured at lower temperatures, while γ-indomethacin is found at higher temperatures. Despite the existing knowledge on how such factors can influence the stability of polymorphic forms, there is still a relatively poor understanding of the underlying mechanisms at play.

In this study, triglycine was selected as the model peptide to further understand these underlying mechanisms as triglycine dihydrate is relatively easy to crystallise as compared to the hydrated form of other short chain peptides, such as glycine and diglycine (Guo et al., 2021). More importantly, triglycine can adopt a polyproline II (pPII) secondary structure in its dihydrate morphology, in addition to its anhydrate morphology. According to Guo et al. (2021), there is a strong motivation to understand the pPII structure due to its abundance in unfolded proteins and peptides and the high stability it confers. This significantly addresses the need to improve morphological stability to increase the pharmaceutical efficacy of end-products. Hence, the study of triglycine is essential to further knowledge surrounding the pPII structure.

The distinct difference in crystal habits between its dimorphic forms makes visual identification of each polymorph easier. In its anhydrate form, triglycine has a fully extended, β-sheet conformation. This results in a highly regular, "plate-like" crystal habit. On the other hand, the left-handed helical structure of the pPII structure results in an observed "needle-like" crystal habit for triglycine dihydrate (Guo et al., 2021). This morphology was conventionally seen in the aqueous state, but rarely in the solid state. Previous work done in this area has shown that water molecules interact with triglycine's negatively charged carboxylic group and positively charged ammonium groups, stabilising the pPII structure (Guo et al., 2021). The use of different anti-solvents and additives can further the understanding of this phenomena by investigating how they can interact with the side chains of triglycine dihydrate to assist or inhibit their formation.

3. Materials & Methodology

3.1 Materials

Triglycine (purity ≥ 99%, white powder, anhydrous form) was procured from Sigma-Aldrich Co. Ltd. as the solute of choice. As for solvents, 18 MΩ deionised water was obtained from the on-site analytical laboratory and ethanol absolute (purity ≥ 99.8%, clear colourless liquid, molecular biology grade) from VWR Chemicals. The following salts were also procured from Sigma-Aldrich Co. Ltd.: LiCl, NaCl, KCl, and $MgCl_2$ (purity≥ 99%, powder, anhydrous form, molecular biology grade).

3.2 Binary solvent experiments

To explore the critical ethanol concentration for the triglycine phase transformation at different temperatures, binary solvent mixtures with select ethanol concentrations ranging from 1 mol% to 14 mol% were tested at temperatures of 5ºC, 10ºC, 15ºC, 20ºC, and 25ºC as the typical screening temperatures for crystallisation ranges between 4ºC and room temperature (Hampton Research, 2021). All experiments were performed at atmospheric pressure. The ethanol mole fractions tested, as shown in Table 1, were selected based on prior work by Jiang (2021), who previously identified the approximate phase transition boundaries. An excess of triglycine was introduced to each binary solvent mixture based on predetermined solubility data to ensure sufficient triglycine solid in the slurry would be available for subsequent analysis.

Table 1. Experimental conditions tested for triglycine crystallisation in binary solvent mixtures.

Ethanol Molar Fraction / mol%	Temperature /ºC 5	10	15	20	25
1	×	×	×	✓	✓
2	×	×	×	✓	✓
3	×	×	×	✓	✓
4	×	×	✓	✓	×
5	×	×	✓	✓	×
6	×	✓	✓	✓	×
7	×	✓	✓	✓	×
8	×	✓	✓	✓	×
9	✓	✓	✓	×	×
10	✓	✓	×	×	×
12	✓	×	×	×	×
14	✓	×	×	×	×

To accurately assess the most stable conformation for each sample, a series of slurry crystallisation experiments were performed at each temperature, followed by the observation of the resultant triglycine crystals. These samples were immersed in a cylindrical double-jacketed glass vessel, coupled the ARE Heating Magnetic Stirrer and the Grant GR150 R1 Recirculating Water Bath with for agitation and temperature-control respectively (Figure 1).

Figure 1. Experimental set-up depicting the glass vessel and magnetic stirrer (left), and the water bath (right).

The samples were continuously stirred overnight for 24 hours at 5ºC to ensure that equilibrium was achieved and the resultant triglycine was in its most stable conformation. The solids in the slurry were then transferred to a glass slide using a dropper and covered with a cover slip to minimise evaporation. They were visually inspected under the GT Vision GXCAM HiChrome-Met microscope with a 20x magnification to observe the crystal habit of triglycine. Thereafter, the slurry mixtures were filtered using a Büchner funnel and flask connected to a Welch 2014B-01 Vacuum Pump, using Whatman filter papers (Grade Number 1, qualitative) as the filter medium (Figure 2).

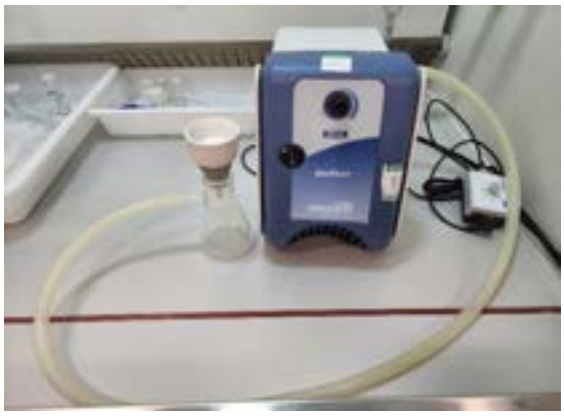


Figure 2. Filtration set-up depicting the Büchner funnel and flask (left), and the vacuum pump (right).

The recovered triglycine residues were transferred to a glass slide and were then analytically characterised via Raman spectroscopy using the Bruker SENTERRA II Raman microscope under 20x magnification. For each sample, 10 to 15 datapoints were taken across different areas of the sample to minimise the incidence of gross errors. This procedure was repeated for 10ºC, 15ºC, 20ºC, and 25ºC.

These results were compared to the initial visual characterisation of the triglycine crystals to identify any deviations from the expected crystal habits, which may uncover insights regarding the crystal growth and interactions between additives and triglycine's functional groups under different conditions.

Table 2. Measurement parameters used for the Raman microscope.

Laser wavelength	532 nm
Power	12.5 mW
Aperture	50 µm
Resolution	4 cm^{-1}
Spectral range	400a ; 50-4260 cm^{-1}

3.3 Salt addition experiments

To understand the effect of the type and concentration of salt on the triglycine phase transformation at 20ºC, a study was conducted using salt concentrations of 1.0M, 1.5M, 2.0M, and 2.5M for four salts: LiCl, NaCl, KCl, and $MgCl_2$. A stock solution of concentration 2.5M was first prepared by dissolving the required mass of salt in 20ml of deionised water. Subsequently, the salt solutions with the concentrations specified above were prepared by diluting an aliquot from the stock solution with deionised water using a pipette.

Table 3. Dilution factors for each desired salt concentration.

Salt conc. (M)	Salt stock solution (ml)	Deionised water (ml)
1.0	2.00	3.00
1.5	3.00	2.00
2.0	4.00	1.00
2.5	5.00	0.00

As with the binary solvent experiments, an excess of triglycine solids was introduced to the salt solutions and the samples were left in the water bath at 20ºC overnight for 24 hours before characterisation of the sample with microscope imaging and Raman spectroscopy the following day.

4. Results

4.1 Binary solvent phase transition trends

Given this study focused on understanding the conditions for phase transition, the different polymorphs were analysed during initial stages of experimentation. The observed microscope images for both the anhydrate and dihydrate forms are displayed in Figure 3(c)-(e), as well as near the phase boundary where a mixture was observed.

The microscope images of the triglycine crystals are shown in Figure 3(a). To verify the exact morphology of the crystal, Raman spectroscopy was conducted on all sample, with focus being placed on three distinct ranges of the spectroscopy, namely being at 1000 cm^{-1}, 1680 cm^{-1}, and 3300 cm^{-1}, as shown in Table 4. The shift in peaks from approximately 1000 cm^{-1} for triglycine anhydrate to 1030 cm^{-1} in triglycine dihydrate corresponds to the C-C bond stretching, which has changed due to the new conformation of dihydrate. The peaks at 1680 cm^{-1} corresponds to the change in structure from a

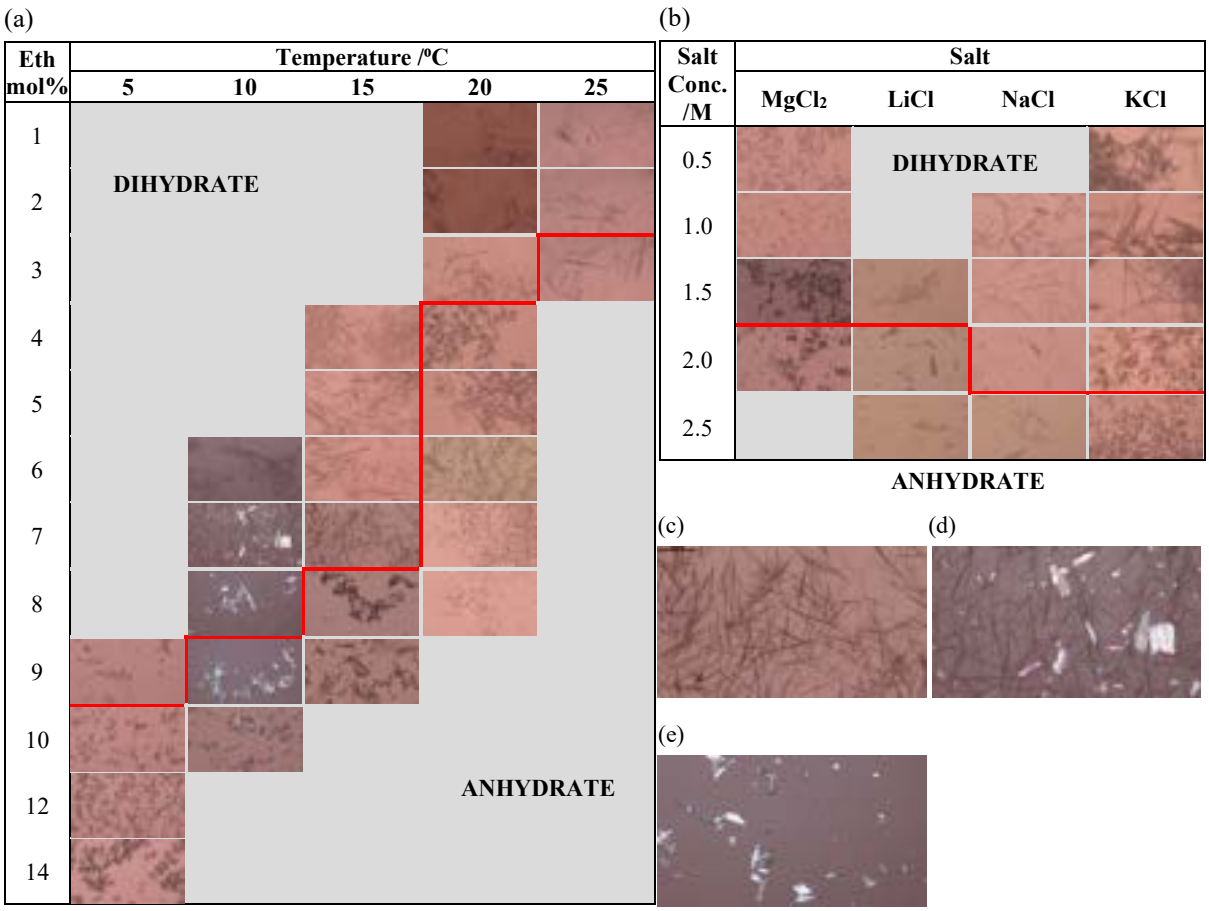


Figure 3. Triglycine phase transition boundaries for (a) different temperatures and ethanol molar fractions, and (b) different concentrations of $MgCl_2$, LiCl, NaCl and KCl. The lines in red represent a transition from triglycine dihydrate to triglycine anhydrate as the dominant morphology. Examples of crystal images under polarised light showing (c) triglycine dihydrate, (d) a mixture and (e) triglycine anhydrate were also included above.

β-sheet in triglycine anhydrate to the pPII conformation present in triglycine dihydrate. Lastly, the peak(s) at 3300 cm^{-1} corresponds to the presence or absence of hydrogen bonds formed between the N-H group of triglycine and water. When water is present, the hydrogen bonding between water and N-H causes the bonded symmetric N-H_2 stretching, which is reflected by the difference in the bands between anhydrates and dihydrates at the 3300 cm^{-1} range. The presence of a single peak at the indicated the presence of a dihydrate, which the presence of a double band indicated the presence of an anhydrate. Based on the above results, a phase transition table can be constructed, which is shown in Figure 3(a).

4.2 Salt addition phase transition trends

As with the above, microscope images of the crystals were taken under different conditions and collated in Figure 3(b).

Table 4. Spectroscopy ranges of interest.

Wavenumber /cm^{-1}	1000	1680	3300
Functional Group	C-C stretching	Peptide bond	N-H_2 stretching
Anhydrate	Single band	Small band, followed by a larger band	Two distinct bands
Dihydrate	Single band above 1000 cm^{-1}	Two distinct bands	Single bands

5. Discussion

5.1 Ethanol interactions with triglycine

From Figure 3(a), anhydrate formation was shown to be favoured by an increase in ethanol content. Prior investigation into protein structures by Singh et al. (2010) suggests that alcohol addition can disrupt nonlocal hydrophobic interactions in the secondary structure of proteins, reducing the stability of the secondary structure. This induces the unfolding of the helical pPII structure in triglycine dihydrate into the fully extended β-sheeted structure characteristic of triglycine anhydrate.

These experiments also confirmed the temperature-dependence of peptide morphology across different ethanol-water mixtures. In general, triglycine anhydrate was observed as the dominant morphology at higher temperatures. Kjaergaard et al. (2010) described a similar observation for select proteins, where a loss of the pPII structure occurred with increasing temperature. The morphology of triglycine proved to be particularly sensitive to temperature between 15 to 20°C, as represented by the large difference between critical ethanol molar fraction at these two temperatures, namely being 8% and 4%, respectively. However, a broader literature review reveals that the underlying mechanism remains poorly understood and represents a significant area for further investigation.

In addition, an increase in ethanol concentration results in the formation of more fragmented, irregular plates, regardless of temperature, as seen in Figure 4. According to El Bazi et al. (2017), adsorption of ethanol at kink sites occurs due to favourable hydrogen bonding interactions with triglycine. This limits the access of triglycine in the solution to the crystal facets and thus, inhibits further growth into a more regular, rod-like shape.

Figure 4. Microscope images showing the crystal habits of anhydrates crystals at a lower ethanol concentration (left) and a higher ethanol concentration (right).

5.2 Salt interactions with triglycine

Although prior findings from literature have established the influence of salt on the morphology for other peptides, this effect has yet to be well investigated for triglycine. This study demonstrates preliminary evidence that the use of salt addition can induce the formation of triglycine anhydrate under conditions where triglycine dihydrate is typically the most stable conformation.

From Figure 3(b), the dihydrate was expectedly observed as the more stable conformation at low salt concentrations, in line with triglycine dihydrate being the dominant conformation in pure water at 20°C. However, it was shown that triglycine anhydrate was the more stable conformation at salt concentrations above 1.5-2M. Han et al. (2021) observed a similar phenomenon with glycine, where salt addition promoted the formation of γ-glycine but could not identify the underlying mechanism. Although finer increments of salt concentrations are necessary to be conclusive, it can be preliminarily concluded that the critical salt concentration at the phase transition boundary decreases with an increase in the cation charge density (Mg^{2+} > Li^{+} > Na^{+} > K^{+}). A theoretical possibility would be the competing effect between the salt ions and water for interactions with triglycine's charged functional groups. Under this theory, the salt ions could stabilise these charged functional groups in place of water and inhibit the formation of the pPII structure.

A separate observation regarding the addition of salts is that the addition of salts seems to increase the solubility of triglycine, suggesting a salting-in effect. This was consistent with what Han et al. (2021) observed in their experiments with glycine crystals. While characterising the solubility of triglycine was not a focal area within this study, it was interesting to note that the solubility of triglycine increased as the concentration of the salt increases. This might be a good area to conduct further research on in the future.

5.3 Triglycine conformation at the phase boundary

While the conformation observed at the phase boundary was most often either the dihydrate or anhydrate, further inspection of the Raman spectra revealed a third form of single crystals that had distinct Raman spectra from both of these expected conformations.

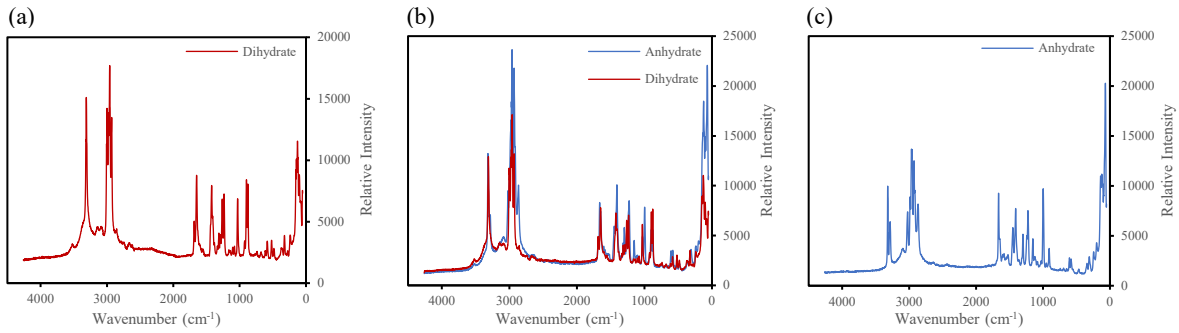


Figure 5. Example Raman spectra of triglycine (a) dihydrate, (b) metastable and (c) anhydrate crystals.

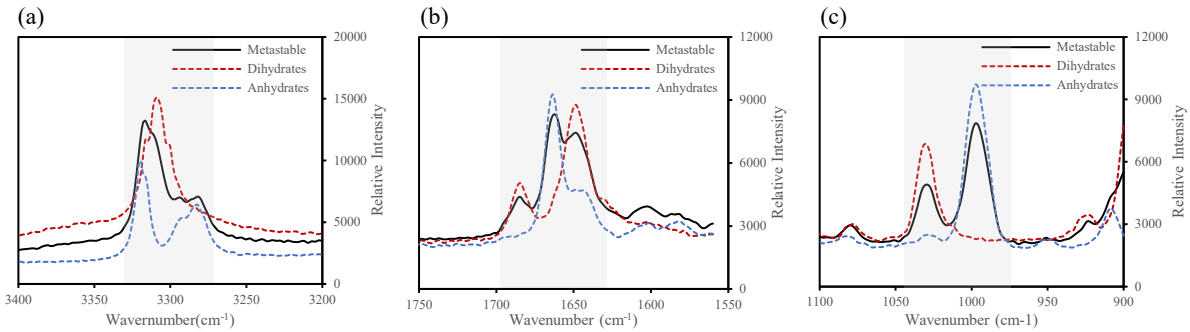


Figure 6. Raman spectroscopy for anhydrate crystals, dihydrate crystals, and metastable crystals at Raman bands of (a) 1000 cm^{-1}, (b) 1680 cm^{-1} and (c) 3000 cm^{-1}, with the main regions of interest highlighted in grey.

Referring to Figure 6, unlike the sharp single-peaks observed for triglycine dihydrate, the Raman bands of these single crystals demonstrated minor peak splitting at 3300 cm^{-1} band. This indicates that the crystalline structure was not fully hydrated, and the sample could not conclusively be determined to be the dihydrate. Yet, the Raman spectra demonstrated remarkable agreement with dihydrate crystals at 1680 cm^{-1}, indicating that these crystals indeed adopted a pPII conformation unique to triglycine dihydrate. Furthermore, these crystals also possessed the characteristic C-C stretching vibrational bands of both the dihydrate and the anhydrate at approximately 1000 cm^{-1}, as shown in Table 5.

Table 5. Comparison between the Raman bands between the metastable state and triglycine's dimorphs.

Conformation	Peak 1	Peak 2
Anhydrate	1000 cm^{-1}	
Dihydrate		1030 cm^{-1}
Metastable	1000 cm^{-1}	1030 cm^{-1}

There are two possible explanations to reconcile these findings. Firstly, this could imply that both the dihydrate and anhydrate co-exist in the crystalline structure in a metastable state, with the resultant Raman spectrum being a superposition of that of each individual morphology. This suggests that single-crystal-to-single-crystal polymorphic phase transition may have occurred, whereby triglycine molecules in a triglycine anhydrate crystal translate to accommodate water in the crystalline structure to form the dihydrate form. This is contrary to the prevailing theory that the polymorphic phase transition is mainly driven by a reconstructive mechanism, whereby triglycine anhydrate crystals dissolve in the solution, followed by the growth of the more stable triglycine dihydrate via nucleation and crystal growth (Krishnan et al., 2015). Secondly, these results could also imply that a third metastable triglycine morphology is possible under certain crystallisation conditions which, to date, has yet to been observed in literature. The indication that the structure is not fully hydrated implies that these crystals could have a monohydrated structure. In this case, the peak splitting at 1000 cm^{-1} could be attributed to the change in length and angle of the C-C bond since the carboxylic group on triglycine molecules are relatively less stabilised compared to triglycine dihydrate (Jorio et al., 2011). However, further analysis using X-ray diffraction and thermogravimetric analysis would be necessary to verify the morphology of these single crystals.

Regardless, this observation of a unique and identifiable Raman spectrum at the phase boundary could potentially be used to speed up the crystallisation screening process by helping to identify crystallisation conditions near the polymorphic phase boundary.

6. Conclusions

This study was concerned with the effects of temperature, water concentration and salt additives on the morphology of triglycine. Conclusions were drawn by varying each of the three variables and subsequent characterisation of the crystals via microscope images and Raman spectroscopy.

For the first variable, it can be concluded that for the same solvent composition, the formation of dihydrates is encouraged at lower temperatures. In addition, increasing the molar fraction of ethanol resulted in triglycine anhydrate being the dominant conformation. Lastly, preliminary findings show that the addition of salts inhibits the formation of dihydrates at high salt concentration, with some indication that the critical phase transition concentration has an inverse relationship with the charge density of the cation. Based on these findings, it is clear that the morphology of triglycine obtained can be manipulated by adjusting any of the three factors. However, while anhydrate formation can be induced with the addition of salts, one should be mindful that the ethanol or salts introduced may be considered as impurities, so further steps downstream in the pipeline may be required to purify the triglycine before the formulation of a given pharmaceutical product.

There are three areas of interest to investigate further going forward. Firstly, investigation into the

kinetics of the phase transition could be conducted. As this work mainly focuses on the endpoint (i.e., the final morphology of the triglycine), little to no work was done on the rate of phase transition. Understanding the kinetic effects of additives could, for example, enable the production of a desired metastable conformation during the crystallisation process. As such, it would be interesting to investigate the effect of addition of salt or solvents on the speed of both the nucleation and growth processes by taking readings at timed intervals to shed more light on the kinetics of this phenomenon. Secondly, the applicability of these findings to other peptides and proteins, particularly those with pPII structures, would be imperative to understand. Especially for non-homopeptides with more complex interactions, these trends may be materially different and would give us insight into the mechanisms driving phase transition boundary trends in more granularity. Lastly, X-ray diffraction and thermogravimetric analysis can be conducted to verify the identity of the metastable form observed. Doing so will improve existing knowledge of the phase transition process from the anhydrate form to the dihydrate pPII conformation.

Acknowledgements

We would like to extend our sincere appreciation to the members of the Heng Research Group for their continual guidance and support. We are especially thankful for the tutelage of Mingxia Guo and are humbled to have had the opportunity to contribute to their seminal work on peptide crystallisation.

References

1. Barker, V. (2020). *Polymorphs at the EPO - Where Are We Now? | European IP Blog*. [online] Finnegan | Leading IP Law Firm. Available at: https://www.finnegan.com/en/insights/blogs/european-ip-blog/polymorphs-at-the-epo-where-are-we-now.html [Accessed 9 Dec. 2022].
2. Bergfors, T.M. (2009). *Protein crystallization*. La Jolla, Calif.: International University Line, pp.17, 19.
3. Black, J.F.B., Cardew, P.T., Cruz-Cabeza, A.J., Davey, R.J., Gilks, S.E. and Sullivan, R.A. (2018). Crystal nucleation and growth in a polymorphic system: Ostwald's rule, p-aminobenzoic acid and nucleation transition states. *CrystEngComm*, 20(6), pp.768–776. doi:10.1039/c7ce01960b.
4. Bruckdorfer, T., Marder, O. and Albericio, F. (2004). From Production of Peptides in Milligram Amounts for Research to Multi-Tons Quantities for Drugs of the Future. *Current Pharmaceutical Biotechnology*, 5(1), pp.29–43. doi:10.2174/1389201043489620.
5. Cruz-Cabeza, A.J. and Bernstein, J. (2013). Conformational Polymorphism. *Chemical Reviews*, 114(4), pp.2170–2191. doi:10.1021/cr400249d.
6. Ducruix, A. and Giegé, R. (1992). *Crystallization of Nucleic Acids and Proteins*. Irl Press.
7. El Bazi, W., Porte, C., Mabille, I. and Havet, J.-L. (2017). Antisolvent crystallization: Effect of ethanol on batch crystallization of α glycine. *Journal of Crystal Growth*, [online] 475, pp.232–238. doi:10.1016/j.jcrysgro.2017.06.021.
8. Fosgerau, K. and Hoffmann, T. (2015). Peptide therapeutics: current status and future directions. *Drug Discovery Today*, [online] 20(1), pp.122–128. doi:10.1016/j.drudis.2014.10.003.
9. Guo, M., Rosbottom, I., Zhou, L., Yong, C.W., Zhou, L., Yin, Q., 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(34), pp.8416–8422. doi:10.1021/acs.jpclett.1c01622.
10. Hampton Research (2021). *Temperature as a Crystallization Variable*. [online] Available at: https://hamptonresearch.com/uploads/cg_pdf/CG101_Temperature_as_a_Crystallization_Variable_2020.pdf.
11. Han, G., Chow, P.S. and Tan, R.B.H. (2021). Understanding the Salt-Dependent Outcome of Glycine Polymorphic Nucleation. *Pharmaceutics*, [online] 13(2), p.262. doi:10.3390/pharmaceutics13020262.
12. Hilfiker, R. (2006). *Polymorphism*. John Wiley & Sons.
13. Isidro-Llobet, A., Kenworthy, M.N., Mukherjee, S., Kopach, M.E., Wegner, K., Gallou, F., Smith, A.G. and Roschangar, F. (2019). Sustainability Challenges in Peptide Synthesis and Purification: From R&D to Production. *The Journal of Organic Chemistry*, 84(8), pp.4615–4628. doi:10.1021/acs.joc.8b03001.
14. Jiang, M. (2021). *The Effect of Ethanol on the Solubility and Crystallisation of Triglycine Peptide The Effect of Ethanol on the Solubility and Crystallisation of Triglycine Peptide*.
15. Jorio, A., Saito, R., Dresselhaus, G. and Dresselhaus, M.S. (2011). *Raman Spectroscopy in Graphene Related Systems*. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA. doi:10.1002/9783527632695.
16. Karthika, S., Radhakrishnan, T.K. and Kalaichelvi, P. (2016). A Review of Classical and Nonclassical Nucleation Theories. *Crystal Growth & Design*, 16(11), pp.6663–6681. doi:10.1021/acs.cgd.6b00794.

17. Kitamura, M. (2008). Control of Polymorphism in Crystallization of Amino Acid. *Developments in Chemical Engineering and Mineral Processing*, 11(5-6), pp.579–602. doi:10.1002/apj.5500110614.
18. Kjaergaard, M., Nørholm, A.-B., Hendus-Altenburger, R., Pedersen, S.F., Poulsen, F.M. and Kragelund, B.B. (2010). Temperature-dependent structural changes in intrinsically disordered proteins: Formation of α-helices or loss of polyproline II? *Protein Science : A Publication of the Protein Society*, [online] 19(8), pp.1555–1564. doi:10.1002/pro.435.
19. Krishnan, B.P. and Sureshan, K.M. (2015). A Spontaneous Single-Crystal-to-Single-Crystal Polymorphic Transition Involving Major Packing Changes. *Journal of the American Chemical Society*, 137(4), pp.1692–1696. doi:10.1021/ja512697g.
20. Lau, J.L. and Dunn, M.K. (2018). Therapeutic peptides: Historical perspectives, current development trends, and future directions. *Bioorganic & Medicinal Chemistry*, 26(10), pp.2700–2707. doi:10.1016/j.bmc.2017.06.052.
21. Lutsko, J.F. (2019). How crystals form: A theory of nucleation pathways. *Science Advances*, 5(4). doi:10.1126/sciadv.aav7399.
22. McGregor, D. (2008). Discovering and improving novel peptide therapeutics. *Current Opinion in Pharmacology*, 8(5), pp.616–619. doi:10.1016/j.coph.2008.06.002.
23. Muttenthaler, M., King, G.F., Adams, D.J. and Alewood, P.F. (2021). Trends in peptide drug discovery. *Nature Reviews Drug Discovery*, [online] pp.1–17. doi:10.1038/s41573-020-00135-8.
24. Reutzel-Edens, S.M., Bush, J.K., Magee, P.A., Stephenson, G.A. and Byrn, S.R. (2003). Anhydrates and Hydrates of Olanzapine: Crystallization, Solid-State Characterization, and Structural Relationships. *Crystal Growth & Design*, 3(6), pp.897–907. doi:10.1021/cg034055z.
25. Rosa, N., Ristic, M., Thorburn, L., Abrahams, G.J., Marshall, B., Watkins, C.J., Kruger, A., Khassapov, A. and Newman, J. (2020). Tools to Ease the Choice and Design of Protein Crystallisation Experiments. *Crystals*, 10(2), p.95. doi:10.3390/cryst10020095.
26. Sear, R.P. (2007). Nucleation: theory and applications to protein solutions and colloidal suspensions. *Journal of Physics: Condensed Matter*, [online] 19(3), p.033101. doi:10.1088/0953-8984/19/3/033101.
27. Shi, L., Holliday, A.E., Shi, H., Zhu, F., Ewing, M.A., Russell, D.H. and Clemmer, D.E. (2014). Characterizing Intermediates Along the Transition from Polyproline I to Polyproline II Using Ion Mobility Spectrometry-Mass Spectrometry. *Journal of the American Chemical Society*, 136(36), pp.12702–12711. doi:10.1021/ja505899g.
28. Singh, S.M., Cabello-Villegas, J., Hutchings, R.L. and Mallela, K.M.G. (2010). Role of partial protein unfolding in alcohol-induced protein aggregation. *Proteins: Structure, Function, and Bioinformatics*, [online] p.n/a-n/a. doi:10.1002/prot.22778.
29. Sjöberg, B., Foley, S., Cardey, B. and Enescu, M. (2014). An experimental and theoretical study of the amino acid side chain Raman bands in proteins. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 128, pp.300–311. doi:10.1016/j.saa.2014.02.080.
30. Soto, L. (2022). *A study on the effect of solvents, temperature, concentration and morphology in the pH of triglycine and the presence of salts in its morphology*.
31. Tian, F., Qu, H., Louhi-Kultanen, M. and Rantanen, J. (2010). Insight into Crystallization Mechanisms of Polymorphic Hydrate Systems. *Chemical Engineering & Technology*, 33(5), pp.833–838. doi:10.1002/ceat.200900572.
32. Veith, H., Luebbert, C. and Sadowski, G. (2020). Correctly Measuring and Predicting Solubilities of Solvates, Hydrates, and Polymorphs. *Crystal Growth & Design*, 20(2), pp.723–735. doi:10.1021/acs.cgd.9b01145.
33. Xu, J., Reiter, G. and Alamo, R.G. (2021). Concepts of Nucleation in Polymer Crystallization. *Crystals*, [online] 11(3), p.304. doi:10.3390/cryst11030304.

Imperial College London

ARTICLE

Techno-economic and Environmental Analysis of Single-Use vs Multi-Use Technologies for ATMP Supply Chain Optimisation

Kye Liew and Isabelle Rider

Department of Chemical Engineering, Imperial College London, London, United Kingdom

Abstract

Stainless steel multi-use technologies (MUTs) are used traditionally in biopharmaceutical manufacturing; however the use of plastic single-use technologies (SUTs) has risen. This paper aims to first assess the economic and environmental viability of SUTs for Advanced Therapy Medicinal Products (ATMPs) production, specifically Adenoviral (AdV) and Lentiviral (LV) vectors. Environmental impact was quantified utilising the ReCiPe 2016 approach to life cycle assessment, whilst economic costs were obtained from SuperPro Designer models. The data was then used as parameters for a multi-integer linear programming (MILP) model to optimise the ATMP supply chains. It was found that SUT-based manufacturing had higher environmental impact and cost in downstream processes, whilst results for SUT manufacturing in upstream and fill and finish processes were product-specific. Optimisation results were also product-specific with AdV having aligned cost and environmental objectives while the objectives for LV were conflicting. This study ultimately aims to provide the framework for future decision support tools to critically assess the impacts of SUT and MUT manufacturing for ATMPs.

Keywords: ATMP, Singe-use Technologies, Multi-use Technologies, Life Cycle Assessment, MILP

1 Introduction

Advanced therapy medicinal products (ATMPs) are a class of therapeutics broadly categorised by gene therapies, cell therapies and tissue-engineered medicines. The majority of ATMPs currently in clinical trials use viral-vectors to deliver genetic material into cells, with 92% specifically utilising Adenoviral vectors for vaccines and Lentiviral vectors for cancer therapy (Capra et al. (2021)), which were the focus of this study. The global ATMP market is currently valued at USD 9.4bn and is projected to grow at a compound annual growth rate of 13.2% reaching USD 22.5bn by 2027 (Grand View Research (2021)). The typical ATMP supply chain is shown in Figure 1.

Figure 1. Typical Viral Vector Supply Chain and Specific Upstream and Downstream Processes

Stainless steel multi-use technologies (MUTs) are traditionally used in biopharmaceutical manufacturing processes, however the adoption of single-use technologies (SUTs) has steadily risen, with adoption growing at a rate of 11% for bioreactors (Langer & Morrow Jr. (2020)). Advantages of SUT systems are the reduced need for cleaning and steaming-in-place (CIP/SIP) processes, increased production flexibility and reduced cross-contamination risk. A key consideration of wider SUT adoption is the potential environmental impact of SUTs through increased plastic production and waste. Environmental impact minimisation must be weighed alongside other possibly conflicting objectives such as costs minimisation and yield maximisation during supply chain planning.

Decision support tools are mathematical models that aid decision-making by exploring trade-offs and the alternatives for decisions (Sarkis et al. (2021)). The objective of this work is to develop a Mixed-Integer Linear Programming (MILP) model to compare the economic cost and environmental impacts of an either MUT or SUT-based upstream (USP), downstream (DSP) or fill and finish (F&F) manufacturing lines for a viral vector-based ATMP. This paper presents results of the life cycle assessment (LCA) and techno-economic calculations used to determine input parameters for the model as well as optimised supply chains for environmental impact and cost minimisation.

2 Background

2.1 Environmental Impacts of Single-Use Technologies

A study by Cytiva Life Sciences (Flanagan (2017)) compared the environmental impacts of SUT and MUT-based monoclonal antibody (MAb) production across multiple regions. The study assessed environmental impact across five categories, climate change, human health, ecosystem quality, resource consumption and water consumption. Flanagan observed that SUTs had lower impact relative to MUTs across all categories, with environmental burden shifting from the use stage of the process to the supply chain. This was attributed to reduced CIP/SIP and an increase in the manufacturing and distribution of single-use consumables. However the study also noted that the results were highly sensitive towards the scale of production as well as the cleanliness of the local electricity grid. Flanagan therefore cautioned against extrapolating these results to other products, underlining the complexity of individual processes as well as the need for location and scale specificity.

2.2 Techno-economic Impact of Single-Use Technologies

Biopharm Services conducted an economic evaluation comparing SUT versus MUT manufacturing for a 2000L scale MAb process (Sinclair & Monge (2002)). It was found that for a new installation, SUTs lowered capital requirements by 20% and the cost of goods by 8%. However, it was noted that the results varied when traditional installations were retrofitted with SUTs and are generally plant-specific.

2.3 Life Cycle Assessment

The LCAs performed in this study follow the ReCiPe 2016 Life Cycle Impact Assessment Method (Huijbregts et al. (2016)). This method values the impact of a process through 18 midpoint indicators and groups them into 3 endpoint areas of protection as shown in Figure 2. ReCiPe 2016 also provides the choice of three cultural perspectives with varying views on issues such as time and expectations, with hierarchist being the consensus perspective. ReCiPe 2016 is an update on its predecessor ReCiPe 2008 and has been observed to have reduced model uncertainty, particularly for water-intensive processes such as pharmaceutical manufacturing (Dekker et al. (2019)).

Figure 2. Overview of the impact categories that are covered in the ReCiPe2016 methodology and their relation to the areas of protection (Huijbregts et al. (2016))

2.4 Flowsheet Model

Sarkis et al. (2022) presented a flowsheet model within SuperPro Designer to quantify the relationship between manufacturing uncertainties of viral vectors and key performance indicators affecting supply chain investment and planning. The model was broadly split into the upstream, downstream and fill and finish manufacturing lines as shown in Figure 1. The vector products modelled were Adenovirus-based vaccines (AdV) and lentivirus-based vectors for ex vivo gene therapy (LV). The model was capable of modelling a range of scales as well as modelling an SUT or MUT-based process. SuperPro Designer relies on built-in algebraic and differential equations to perform its calculations.

2.5 Multi-period Optimisation Model

Outputs from the aforementioned SuperPro Designer model were used as input parameters for the multi-period optimisation model developed by Sarkis. The model was an update of the multi-site snapshot model used in Sarkis et al. (2022), with discretisation being used to enable a demand scenario to be imposed across time periods. The supply chain described in the model consisted of upstream, downstream lines in primary manufacturing, fill and finish lines in secondary manufacturing, intermediate storage nodes (SN) and demand zones (DZ). The model was written in Python using the PyOmo software package.

The initial model aims to maximise revenue through production scheduling and optimised supply chain planning to fulfil the user-specified demand scenario across demand zones and time periods. The model allows for a high degree of user-specified design variables with the key ones listed below:

1. Viral Vector Products (Set i)
   - AdV/LV
2. Scale (Set a)
   - 50L, 200L, 1000L, 2000L
3. Manufacturing Nodes (Set j)
   - USP Section Manufacturing Lines (Set u)
   - DSP Section Manufacturing Lines (Set d)
   - F&F Section Manufacturing Lines (Set g)
4. Time Periods (Set t)
5. Demand for a product at specific time and node ($D_{t,j,i}$)

Economic parameters were split into capital costs of setting up a manufacturing line and the operating costs of running each line. Operating costs were further split into facility, labour and variable costs. The segmentation of the model into separate manufacturing lines allowed for increased flexibility in supply chain planning. The model was adapted to allow it to choose either SUT or MUT manufacturing lines and account for the environmental and economic impact of those choices. This model can be found in Appendix F.

3 Methodology

AdV and LV were chosen as the viral vector products for this study as they accounted for the majority of gene therapies in clinical trials (Capra et al. (2021)) and were well-documented and studied processes. The scale of the process was chosen to be 2000L for cross-comparability of results with results of similar studies conducted on the MAb manufacturing process. Four viral vector manufacturing flowsheets were designed in SuperPro Designer: one SUT and one MUT-based process for both AdV and LV. Data was extracted via SuperPro Designer's reports and used to calculate environmental and economic parameters for the optimisation model. The parameters were also specific to USP, DSP or F&F manufacturing lines to allow the model to consider having individual manufacturing lines be either SUT or MUT-based.

3.1 Environmental Parameters Calculations

It was necessary to identify key contributors of environmental impact with the greatest difference between and SUT & MUT process. Based on research, three specific areas of the process were chosen:

1. Single-use consumables supply chain
2. Stainless steel production
3. CIP/SIP processes

The supply chain of single-use consumables was selected as upon transition from an MUT to an SUT-based process, an increase in environmental impact was observed from the supply chain stage (Flanagan (2017)). It was therefore important to quantify this increase in environmental impact. On the other hand use stage impacts were reduced for the SUT process and water consumption in particularly was observed to fall by 87%, mostly via a reduction in CIP/SIP processes (Sinclair et al. (2008)). Steel use was estimated to be reduced by 62% for SUT processes as estimated in Lopes (2015). All environmental impacts would be converted to a per gram viral vector product basis for use in the optimisation model.

3.1.1 Consumables, Waste and Steel Sizing

The SuperPro Designer model identified the key single-use consumables to be cell bags, storage bags, flasks and test tubes. When the difference in the amount of consumables used was deemed negligible between the MUT and SUT process, the consumable was not accounted for in calculations. It was also assumed that consumables are predominately made of one material. Bags such as bioreactor bags were thus assumed to be entirely made from linear low-density polyethylene (LLDPE), flasks were considered to be polyethylene terephthalate (PET), and test tubes were considered to be polypropylene (PP). In accordance with this, the weight per gram product of each plastic was calculated using Equation 1.

$$\text{Amount of Plastic}(kg/g) = \frac{\sum(\text{Consumable Amount} \times \text{Weight})}{\text{Mass of Viral Vector Produced}} \quad (1)$$

Steel use for the MUT processes were estimated using the volumes of largely steel process equipment, namely bioreactors, blending tanks and centrifuges. Weights of each equipment were then estimated using product data for similarly sized equipment currently available on the market. These weights were then converted to a per gram of product basis by assuming that the equipment had a 20 year lifespan and that production remained the same every year. The corresponding steel use for SUT processes were calculated by reducing the MUT sums by a factor of 0.38 (Lopes (2015)).

Sources of environmental impact from the CIP/SIP processes were chosen to be production of chemicals, water for injection and steam as well as treatment of resultant waste water. Data for each source were taken from SuperPro Designer's environmental impact report (EIR) and were converted to a per gram of product basis.

3.1.2 OpenLCA

LCAs were conducted via the OpenLCA software, using the sizing data collected, in accordance with the ReCiPe 2016 impact assessment method. A hierarchist cultural perspective was selected in accordance with consensus. LCAs for each individual material were conducted using processes from the ecoinvent v3.6 database, which can be found in Appendix A. This yielded values in each midpoint impact category, which were then grouped into their respective endpoint categories. Finally for comparability, the impact to each endpoint category was converted into USD2017 using equivalent monetary values presented in Dong et al. (2018), as shown in Table 1.

Table 1. Monetisation value factors (Dong et al. (2018))

Endpoint Indicator	Environmental Unit	Equivalent Monetary Value (USD2017)
Human Health	DALY	1.46×10^5
Ecosystem Health	species.yr	3.9×10^7
Resource Availability	USD2013	1.1

Due to limitations of processes in the LCA database, only the production of the raw plastic polymer was considered and the additional impact of transforming the polymer into final products were not accounted for. However, this was deemed an acceptable approximation as the environmental impact of the transformation of polymer to finished plastic products is only a quarter of that of raw polymer production (Mori et al. (2013)). For end-of-life impacts, 35% of consumables were sent to sanitary landfills, whilst 65% of consumables were sent off for municipal incineration. This was in line with the regional mix in the United Kingdom. Similar approximations were made for steel use impact and CIP/SIP process impacts when the exact processes were not available within the database.

3.2 Economic Parameters Calculations

Costing data was taken from SuperPro Designer's Itemised Cost Reports (ICR).

3.2.1 Capital Costs

Capital costs (CapEx) was taken from the total capital investment of each process with processes grouped into USP, DSP and F&F manufacturing lines. It was in units of USD2021.

3.2.2 Operating Costs

Operating costs were split into facility, variable, and labour costs with each having its own units. Facility costs, which accounted for plant maintenance and equipment depreciation, in units of USD2021 per month were taken directly from the ICR, whilst variable and labour were calculated with Equation 2 and 3 respectively.

$$C_{VAR}\left(\frac{USD2021}{Gram\ Product}\right) = C_{Materials} + C_{Consumables} + C_{Utilities} + C_{Waste} \quad (2)$$

$$C_{LAB}\left(\frac{USD2021}{Batch\ Product}\right) = C_{Labour} + C_{Lab/QC/QA} \quad (3)$$

3.3 MILP Model Building

A model built by Sarkis was adapted for this study. The updated environmental and economic parameters were used as optimisation input and can be found in Appendix D.2.2 and E.

3.3.1 Set Changes

A new set (Set *s*) was introduced to represent the possible manufacturing line types for USP (*u*), DSP (*d*) and F&F (*g*) lines.

$$Set\ s = \{SU,\ MU\}$$

All cost parameters were indexed by Set *s*, and environmental impact parameters were indexed by scale (Set *a*), product (Set *i*), and Set *s*. Capital cost was changed as well as to be further indexed by product (Set *i*) as significant differences in cost were observed between the SUT and MUT-based processes for both AdV and LV.

The following binary and positive variables were indexed by Set *s* as well to facilitate the updated mdoel:

- Line installation and availability binary variables

$$Z_{t,j,a,u,i,s},\ Z_{t,j,a,d,i,s},\ Z_{t,j,a,g,i,s} \in \{0,1\}$$
$$A_{t,j,a,u,i,s},\ A_{t,j,a,d,i,s},\ A_{t,j,a,g,i,s} \in \{0,1\}$$

- Product allocation binary variables

$$W_{t,j,a,u,i,s},\ W_{t,j,a,d,i,s},\ W_{t,j,a,g,i,s} \in \{0,1\}$$

- Mass balance binary variables

$$N_{t,j,a,u,i,s},\ N_{t,j,a,d,i,s},\ N_{t,j,a,g,i,s} \in \{0,1\}$$

- Batch amount non-negative integer variables

$$B_{t,j,a,u,i,s},\ B_{t,j,a,d,i,s},\ B_{t,j,a,g,i,s} \in \mathbb{Z},\ B \geq 0$$

- Gram amount of product non-negative real variables

$$P_{t,j,a,u,i,s},\ P_{t,j,a,d,i,s},\ P_{t,j,a,g,i,s} \in \mathbb{R},\ P \geq 0$$

- Time constraint non-negative real variables

$$T_{t,j,a,u,i,s},\ T_{t,j,a,d,i,s},\ T_{t,j,a,g,i,s} \in \mathbb{R},\ T \geq 0$$

3.3.2 New Parameters and Variables

Initially, new parameters were defined for each endpoint environmental impact category on a per gram of viral vector product basis for each process section. These parameters were defined using data from the LCA calculations (Section 3.1).

$$E^{HH}_{u/d/g}(\text{DALY/g}),\ E^{EH}_{u/d/g}(\text{species·yr/g}),\ E^{RA}_{u/d/g}(\text{USD2013/g})$$

Additional continuous real variables for the environmental impact for each category from the production of viral vector products across manufacturing lines at specific time period *t*, and specific manufacturing location *j*, is defined as follows:

$$TE^{HH}_{t,j},\ TE^{EH}_{t,j},\ TE^{RA}_{t,j} \in \mathbb{R},\ P \geq 0$$

The totals from each category were summed across all time periods and locations and aggregated utilising the monetisation factors in Table 1. The variable was defined as *TE*.

New binary variables were introduced to account for the choice of SUT or MUT for each process line. These were defined and indexed as follows:

$$S_{t,j,a,u,i,s},\ S_{t,j,a,d,i,s},\ S_{t,j,a,g,i,s} \in \{0,1\}$$

3.3.3 Objective Function

The existing objective function (Equation 4) aimed to maximise profit by maximising revenue from sales and minimising costs, whilst attempting to meet demand.

$$z_0 = (Selling\ Price \times \sum\nolimits_{t,j,i} Sales) - Total\ Costs \quad (4)$$

For this study, two alternative objective functions (z_1, z_2) were formulated, one for costs minimisation and one for total environmental impact minimisation (Equations 5 and 6)

$$z_1 = TC^{cap} + TC^{op} \quad (5)$$

$$z_2 = \sum\nolimits_{t,j} 1.46 \times 10^5 TE^{HH}_{t,j} + 3.9 \times 10^7 TE^{EH}_{t,j} + 1.1 TE^{RA}_{t,j} \quad (6)$$

When either objective function was used, the other was refactored as an equality constraint.

3.3.4 Equality Constraints

The main equality constraint that was added was for calculations of total environmental impact per category. The following equation was used to sum the USP, DSP and F&F production contributions to each impact category for all scales, products and manufacturing line types (Set *a*, *i*, *s*). The constraint for the total human health impact at a specific time and location is shown in Equation 7.

$$TE^{HH}_{t,j} = \sum\nolimits_{u,d,g} \sum\nolimits_{a,i,s} E^{HH} \times P_{a,i,s} \quad (7)$$

3.3.5 Logic Constraints

Logic constraints were used to ensure realistic solutions were found. The sum of the binary variables *S* across manufacturing line types was constrained to be less than or equal to one ensure any manufacturing line was only ever either SUT or MUT-based.

$$\sum\nolimits_s S_{t,j,a,u,i} \leq 1,\ \sum\nolimits_s S_{t,j,a,d,i} \leq 1,\ \sum\nolimits_s S_{t,j,a,g,i} \leq 1 \quad (8)$$

Binary variable *S* was further constrained by the line installation variable, *Z*, such that the model only decides which manufacturing line type to utilise when a line is being installed. An example for USP lines is shown in Equation 9.

$$S_{t,j,a,u,i,s} \leq Z_{t,j,a,u,i,s} \quad (9)$$

To segregate product flows by manufacturing line type, minimum and maximum time constraints were imposed in a similar fashion from the product allocation variables, *W*. An example for USP lines is shown in Equation 10.

$$T_{min} \cdot S_{t,j,a,u,i,s} \leq T_{t,j,a,u,i,s} \leq T_{max} \cdot S_u \quad (10)$$

Lastly the sales and demand constraint was reversed, such that sales met the demand for every time period and location (Equation 11). This was due to the objective no longer being tied to sales maximisation. Due to this change, it was essential for the imposed demand scenario to be within capacity limits for a feasible solution to be found.

$$Sales_{t,j,i} \geq Demand_{t,j,i} \; \forall t, j, i \tag{11}$$

4 Results and Discussion

4.1 General Environmental Analysis

Results were product-specific with SUT and MUT-based process having similar impacts for AdV whilst the LV SUT-based process had significantly higher impact.

Human health impacts were consistently the largest category across products and manufacturing types, this can be attributed to the weighting of the monetisation factors from Dong et al. (2018). Another potential reason for this is the hierarchist perspective chosen for the LCAs, which considers environmental impacts over a period of 100 years. Therefore the midpoint impact of global warming is valued highly, as the time scale of the hierarchist perspective and the time scale of the worst effects of global warming on human health are similar. Another potential cause is the weighting factors formulated by Ponsioen & Goedkoop used in ReCiPe 2016, where climate change was given 44.3% weightage for endpoint impacts (Sala, Cerutti & Pant (2018)). Endpoint impact weighting is ultimately a subjective process and future analysis could be conducted on the impact of different endpoint weighting methods on the model.

4.2 AdV Environmental Parameters

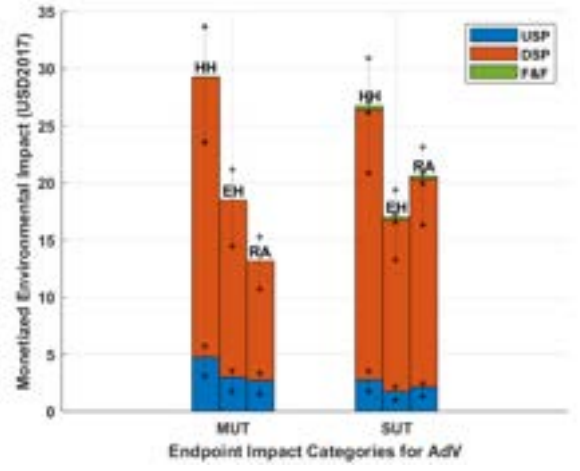


Figure 3. Environmental Impact of AdV Manufacturing per Gram AdV

As seen in Figure 3, the environmental impact of the SUT-based process is smaller than that of the MUT-based process for the human health (HH) and environmental health (EH) categories, whilst being higher for the resource availability (RA) category. This is primarily driven by phosphoric acid production for CIP/SIP accounting for 28% and 25% of impact in HH and EH but only 12% of impact in RA for MUT processes, whilst SUT processes had no impacts from CIP/SIP. RA is more sensitive to impacts from plastic production which is the driving force of environmental impacts for SUT.

4.2.1 Upstream Process Observations (AdV)

Upstream processes account for a small portion of total environmental impact, with the SUT upstream line having smaller impacts across all categories. This is due to upstream processes generally having lower process volumes, SUT processes thus require fewer LLDPE bags. This low environmental impact from plastic production, is further compounded by the absence of impact from CIP/SIP for the SUT process, leading to the SUT upstream line having lower environmental impact.

4.2.2 Downstream Process Observations (AdV)

The downstream process contributes the most to environmental impact for both manufacturing types. This is due to the nature of the downstream processes which involve large volumes of buffers for purification and polishing. There is therefore a larger requirement for either steel blending tanks or LLDPE bags for the SUT process. The environmental impact reduction from no longer needing CIP/SIP processes is matched by the increase in environmental impact of plastic production for the SUT process.

4.2.3 Fill and Finish Process Observations (AdV)

The impact from the F&F section of both manufacturing types is negligible relative to the total impact, as it consists of only one step and no CIP/SIP is required. This being said, the impact from the SUT-based process are larger than those of the MUT-based process. This is due to the nature of AdV vaccines which require dilution into individual doses, thus requiring multiple high-volume LLDPE bags which had greater environmental impact than the steel needed for the blending tanks.

4.3 LV Environmental Parameters

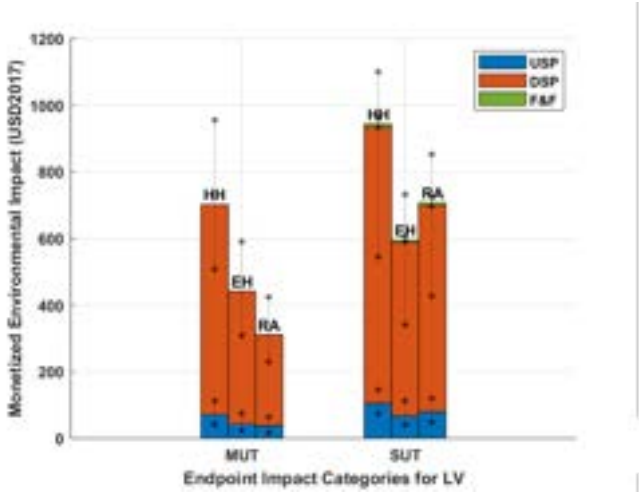


Figure 4. Environmental Impact of LV Manufacturing per Gram LV

As shown in Figure 4, the overall environmental impact from the SUT-based process is higher than that from the MUT-based process across all categories. The per gram impact of LV production is higher than AdV, this is due to LV being a much more highly concentrated product with significantly smaller batch sizes (1.48g to 48.46g batch size for AdV).

4.3.1 Upstream Process Observations (LV)

The upstream process is a relatively small contributor to the overall process impact. However, for LV the SUT process has greater impact than MUT across all categories. This is largely driven by the impact of LLDPE production and waste needed for the SUT process. The resulting environmental impact from this outweighed impacts of CIP/SIP for the MUT process.

4.3.2 Downstream Process Observations (LV)

Downstream process impact is the largest driver for the SUT process having higher environmental impact than the MUT process. This is due to the nature of downstream processes, which involve large volumes of buffers and require several LLDPE bags or steel tanks. Similar to USP, LLDPE production and waste account for the majority of impact.

4.3.3 Fill and Finish Process Observations (LV)

The impact from the F&F section of both processes in negligible compared to the total impact of the process. The SUT process has marginally higher environmental impact, driven by LLDPE production and waste.

4.4 Environmental Parameter Uncertainty Analysis

The main source of uncertainty in this paper is changes in operational ranges due to these processes currently still being under development, leading to uncertainty in batch size, bottleneck steps which might propagate to supply chain decision-making. Therefore, uncertainty analyses were focused on potential changes in the data obtained from the SuperPro Designer flowsheets. Upper and lower bounds of the range of uncertainty were based on values from the global sensitivity analysis of the model conducted by Sarkis, Shah & Papathanasiou (2022). Typical ranges of LV and AdV batch sizes were taken from the report and normalised to batch sizes used in this study to obtain upper and lower bounds, for the monetised environmental impact of each product and impact category. The methodology used for uncertainty analyses and sample calculations can be found in Appendix C.

As can be seen from Figures 3 and 4, the error bars have the potential to yield inverted results; for example, in Figure 4, the lower bound of the range of impact to HH for SUT is lower than the nominal value of impact to HH for MUT, and so it is possible for SUT to ultimately have a smaller environmental impact than MUT for LV manufacturing, affecting the results produced by the optimisation model. In the future, sensitivity analysis of the optimisation model across the operating range of the SuperPro Designer model should be conducted to further understand its impact on optimisation results.

4.5 Comparison of Environmental Results to Literature

Environmental results deviated from results seen in other studies such as those conducted by Flanagan (2017), where SUT-based processes had lower environmental impact. A potential reason for this could be fundamental differences between the MAb and viral vector manufacturing processes. The contrast between these results challenges the comparability of studies for the manufacturing practices of different products. For future work, deeper analysis of the impacts of individual process steps should be conducted to understand the difference in results between the MAb process and the viral vector processes. Another potential reason for the disparity is insufficient coverage of use-stage environmental impacts, which Flanagan (2017) identifies as being the key driver in environmental impact reduction for the SUT process. Future work could expand the range of use stage impacts beyond CIP/SIP and steel use impact to reduce this potential error.

4.6 Economic Parameters

4.6.1 Capital Costs

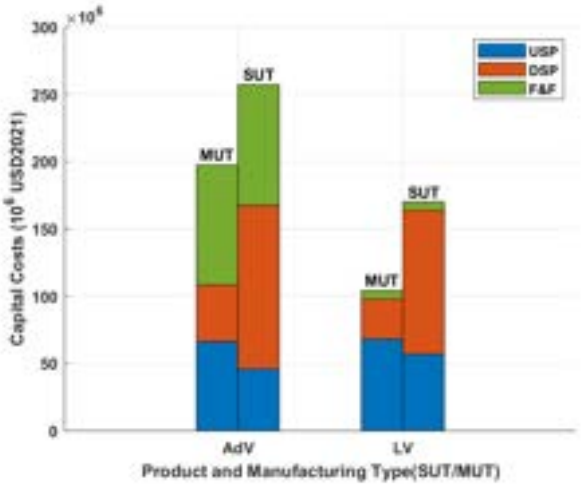


Figure 5. Capital Costs of AdV and LV SUT and MUT processes

For both Adv and LV, having fully SUT-based manufacturing lead to higher capital costs, 30% higher than MUT manufacturing for AdV and 63% higher for LV. This increase is primarily driven by downstream process costs, whilst upstream process was slightly reduced for SUT processes and fill & finish cost remained relatively constant. AdV has higher fill & finish costs due to the nature of the product, with increased complexity for the dilution of the batch into 100,000 individual doses, whilst LV is a highly concentrated product.

The increase in downstream costs is largely driven by higher direct fixed capital (DFC) costs, accounting for 95% of the increase for both AdV and LV. This comprises of direct costs such as plant and equipment costs and indirect costs that include costs of engineering and construction. Equipment costs only account for 7.6% of DFC costs, therefore the increase in DFC costs is due to higher plant and construction costs. Conversely for the upstream and fill & finish processes,

the SUT process has lower capital costs due to lower DFC costs. This suggests that SUT processes are more economically viable for processes with fewer steps and involving lower volumes. In future work, modifying the downstream process in the flowsheet to being only partially SUT-based should be explored. SUT utilisation could be prioritised for processes with high cross-contamination risk and require stringent CIP/SIP, rather than processes involving buffers. This could potentially lower the DFC of the SUT downstream process, lowering total capital costs.

Initial investment costs were reduced by 40% for an SUT-based process relative to a comparable MUT process in a report by Wieland (2022). The discrepancy between the the Wieland (2022) study and the results from this study could be due to inaccuracies in SuperPro Designer's costing inputs or possible selective use of SUT equipment in the Wieland study compared to the fully SUT SuperPro Designer model.

4.6.2 Operating Costs General Observations

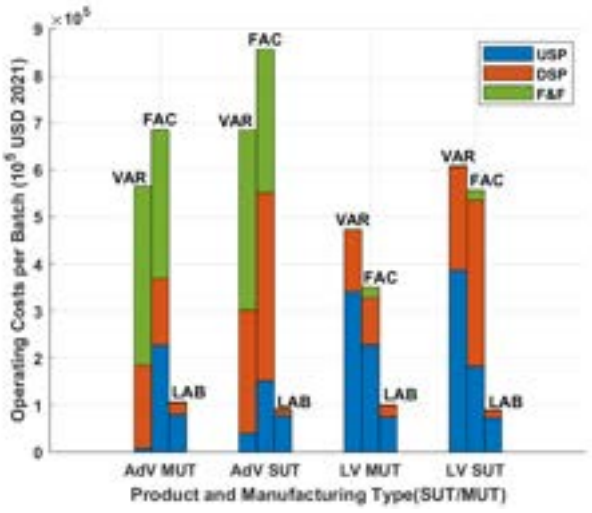


Figure 6. Operating Costs of AdV and LV, SUT and MUT processes

For both viral vectors it is observed that operating costs are higher for SUT processes than for MUT processes, increasing by 36% for LV and 20% for AdV. Increased total downstream costs accounted for the overall increased operational costs for SUT processes, increasing by 130% for LV and 97% for AdV. Total upstream and fill & finish operating costs were lower for both products. The key trends between different manufacturing lines is similar to those observed for capital costs. These observations further supported the segmentation of the model into different manufacturing lines, as upstream lines and fill & finish SUT processes seemed more financially viable. Fill & finish costs were a major contributor to operating costs for AdV, accounting for 42-53%, whilst being relatively negligible for LV. This is due to the dilution of AdV into vaccine doses, requiring complex and expensive fill & finish processes. Across all products and manufacturing types, facility and variable costs were the most significant cost categories for both products with labour cost being relatively negligible.

Lütke-Eversloh & Rogge (2018) observed a reduction in operating costs of around 20% for SUT processes relative to MUT. The study was based on the MAb process, this again highlighting the differences between bioprocesses for different products. To further validate this, an MAb model could be designed within SuperPro Designer to identify if the difference is ultimately due to differences in modelling methods or differences in processes for individual products. Further analysis for the operational costs incurred by individual process steps would also be useful for building a more refined approach to SUT usage for future flowsheets.

4.6.3 Variable Cost Observations

Variable costs are defined as the cost of production and were batch size dependent. Variable cost was a significant cost category, accounting for 42-50% of total operating costs and around 40% of the increase in overall operating cost. For both products, an SUT-based process had higher variable costs. The key components of variable costs were the consumable and material costs, with utilities and waste costs being negligible. The rise in variable costs for an SUT process is wholly driven by significant increases of consumable costs, whilst material costs fell slightly for the SUT process. Similarly to results seen in total operating costs as well as capital costs, the increase was driven by consumable costs in the downstream process, accounting for 67.5% of the increase for LV and 72.9% for AdV The rise in consumable costs is due to significantly higher amounts of single-use consumables needed for the SUT process, whilst the slightly reduction in material costs is due to the elimination of phosphoric acid and sodium hydroxide use for CIP/SIP in the SUT process.

4.6.4 Facility Cost Observations

Facility costs are defined as the cost of operating the plant and are mainly time dependent. Facility costs account for a larger portion of total operating costs for AdV relative to LV, this is due to the aforementioned high dosage nature of the AdV vaccine product. For both AdV and LV, the facility costs of the SUT process were higher than the MUT process, accounting for around 60% of the overall operating costs increase. The facility costs of downstream manufacturing lines increase for the SUT process, whilst decreasing for both the upstream and fill & finish lines, in line with observations of other cost parameters. The biggest increase in downstream costs is in the purification step, this further supports the reduction of SUT usage in purification processes.

4.6.5 Labour Cost Observations

Labour costs accounted for a relatively small portion of total operating costs. For both AdV and LV, labour costs were lower for the SUT process, likely due to the reduced need for CIP/SIP processes. Labour costs were predominantly incurred in the upstream process, this is due to the upstream process steps being more time-consuming therefore requiring more working hours, estimated to be 22 days compared to 5 days for the downstream process.

4.7 MILP Optimisation Results

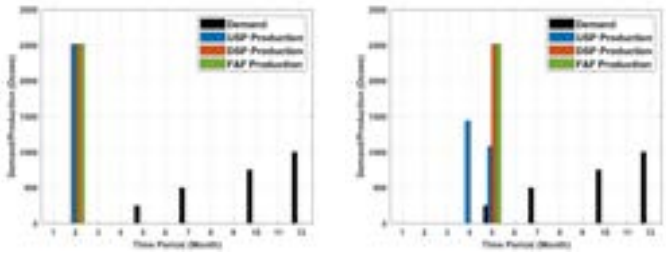
Single-objective optimisation for environmental impact and total costs were conducted and results are shown in Table 2.

Table 2. Optimised solutions for AdV and LV for both objectives

Product, Optimisation Objective	Environmental Impact (USD 2013)	Total Costs (USD 2021)
LV, Environmental Impact Minimisation	$ 15,168.00	$ 353,113,066.80
LV, Cost Minimisation	$ 16,456.32	$ 271,527,502.11
AdV, Environmental Impact Minimisation	$ 5,539.62	$ 361,679,614.80
AdV, Cost Minimisation	$ 5,539.62	$ 361,679,614.80

AdV had higher total costs, while LV had higher environmental impact, consistent with the environmental and techno-economic parameters used. Environmental impact minimisation for LV yielded a slightly lower monetised environmental impact but higher total costs compared to cost minimisation, indicating the objectives were conflicting. In future work, multi-objective optimisation techniques such as the ϵ-constraint method could be used to balance the objectives for LV, with more weight placed on costs as environmental impact is only slightly increased when costs is minimised. For AdV, the model produced a non-unique, degenerate solution whereby the environmental impact and total costs remained the same for both objectives. This is likely due to both objectives being aligned to each other, as environmental impact and techno-economic trends are consistent for AdV.

4.7.1 Production Scheduling



(a) Environmental Impact Minimisation **(b)** Cost Minimisation

Figure 7. Time-based Demand and Production Data for LV

The benefit of multi-period optimisation is observing the model adapt to changing demand. In this study, demand was initially set to zero to provide time for setting up of manufacturing lines, at months 5-6 there was low demand and demand was increased in month 10 and 12. The demand profile for LV is shown in black in Figure 7. For environmental optimisation, the model scaled out production across multiple manufacturing lines to conduct all production in one month (Figure 7a). When cost was minimised, upstream production was split into two campaigns to reduce the number of required upstream manufacturing lines, thus minimising capital cost (Figure 7b). At the moment the model only accounts for setting up of new manufacturing lines, it could be further expanded to account for existing MUT-based manufacturing lines being retrofitted into SUT manufacturing lines.

4.7.2 AdV Environmental Impact Minimisation

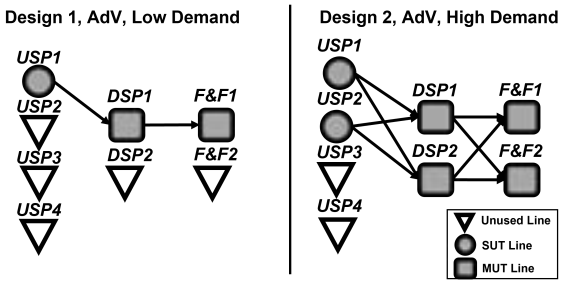


Figure 8. AdV Manufacturing Lines with Minimised Environmental Impact

The manufacturing lines for environmental impact minimisation of AdV are shown in Figure 8, with SUT-based upstream lines and MUT-based downstream and fill & finish lines. Under lower initial demand, the model only utilised one manufacturing line for each process section but as demand increased, manufacturing was scaled out over two manufacturing lines. Upstream lines remained SUT-based and downstream and fill & finish lines remained MUT-based. This is consistent with the environmental parameters in Figure 3, where upstream SUT manufacturing lines had lower environmental impact.

4.7.3 AdV Cost Minimisation

The manufacturing lines for cost minimisation of AdV were the same as the ones for environmental impact minimisation shown in Figure 8. This is consistent with capital costs and operating costs results shown in Figures 5 and 6, where the costs for upstream SUT manufacturing lines were lower.

4.7.4 LV Environmental Impact Minimisation

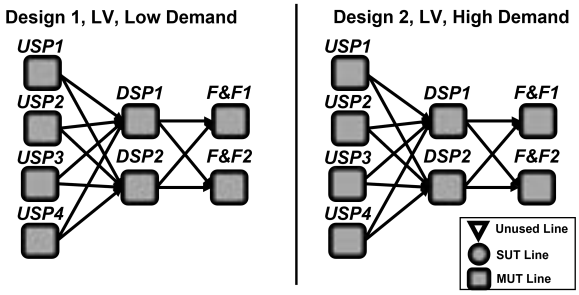


Figure 9. LV Manufacturing Lines with Minimised Environmental Impact

The manufacturing lines for environmental impact minimisation of LV are shown in Figure 9, with fully MUT-based upstream, downstream and fill & finish manufacturing lines. This is consistent with environmental impact results in Figure 4, where MUT manufacturing lines have less impact for all manufacturing lines. There is no change in manufacturing line design for low demand and high demand periods, with the model setting up the maximum number of lines possible to conduct all production within the same month which is shown in Figure 7a. This behaviour is due to there being

no environmental penalty for the setting up of new manufacturing lines since environmental impact is only tied to viral vector production on a per gram basis. This provides scope for future work to expand the model to include the environmental impacts of setting up an SUT and MUT manufacturing line.

4.7.5 LV Cost Minimisation

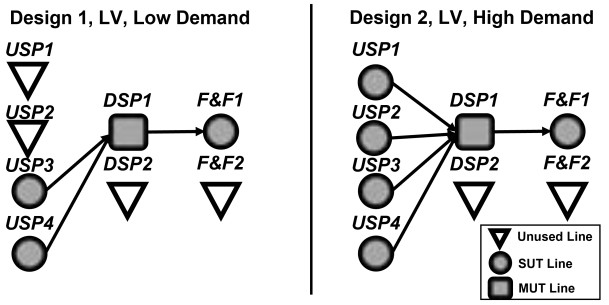


Figure 10. LV Manufacturing Lines with Minimised Costs

The manufacturing lines for cost minimisation of LV are shown in Figure 10, with SUT-based upstream and fill & finish manufacturing lines and MUT-based downstream lines. This is consistent with the capital and operating costs shown in Figures 5 and 6, where both costs are slightly lower for SUT upstream and fill & finish manufacturing lines. The model also adapts to periods of higher demand by setting up additional upstream lines, however it only does so when necessary in order to minimise facility costs. Capital cost is minimised by only operating a single downstream and fill & finish manufacturing line. It is these reduced facility costs that primarily lead to this solution to have lower total costs than the solution in Figure 9.

4.7.6 Validation of Optimal Solution

Table 3. Cost minimised solution for LV with only MUT manufacturing lines

Product, Optimisation Objective	Environmental Impact (USD 2013)	Total Costs (USD 2021)
LV, Cost Minimisation (Fixed MUT)	$ 15,168.00	$ 316,308,181.84

Validation of the optimality of the proposed manufacturing set ups was conducted by fixing the manufacturing type to be either only SUT or MUT. When the model was fixed to SUT only all solutions had higher environmental impact and total cost than the ones shown in Table 2. This confirmed the optimality of the mixed SUT/MUT solutions proposed in Figures 8 and 10. However when the model was fixed to MUT only, cost minimisation for LV produced a solution, shown in Table 3, that had significantly lower total cost but the same environmental impact as the solution in Figure 9. This is due to the model reducing the number of downstream and fill & finish lines, thus reducing the capital costs. This further highlights the potential for multi-objective optimisation of LV production to identify the most optimal solution across both objectives.

5 Conclusions and Outlook

The key conclusions from the environmental life cycle assessment were that the environmental impact of a fully SUT processing chain was higher. However the impact of SUT upstream processes were lower for AdV but not LV. SUT downstream processes consistently had higher environmental impact for both products, suggesting a fully SUT approach may not be suitable for downstream processes due to its number of steps and high volume of buffers. SUT fill and finishing had relatively negligible environmental impact but it was still higher than the the MUT-based process.

On the techno-economic side, SUT cost was lower for upstream processes and significantly higher for downstream processes. For fill and finish, cost were higher for AdV but lower for LV. This was due to the nature of the AdV vaccine product, where the fill and finish section involves the dilution of AdV which require large volumes of buffers and water for injection. This leads to greater cost for an SUT-based fill and finish process due to higher consumable costs.

The optimisation model utilised the environmental and techno-economic parameters to produce optimised manufacturing lines for AdV and LV under low and high demand scenarios. The model produced a non-unique solution for AdV with the same design for environmental impact and cost minimisation, suggesting that the two objectives were aligned for AdV. LV on the other hand had distinct solutions, suggesting the objectives were conflicting. This provided scope for multi-objective optimisation to balance between both objectives and ensure the most optimal solution is found.

In the future, the range of environmental impacts covered could be expanded to improve the model. In particular the environmental impact of setting up of an SUT or MUT manufacturing line and retrofitting of an MUT line into an SUT one could be considered. In order to make the SUT-based downstream process more environmentally and economically viable in the future, a more granular and selective approach to SUT usage in the downstream process could be applied. Equipment with high risk of cross-contamination should be prioritised whilst equipment containing buffers could be excluded.

Acknowledgement

The authors would like to acknowledge the help and support provided by Miriam Sarkis from the Department of Chemical Engineering at Imperial College London. The work she had previously done on optimisation and process scheduling provided the basis for the model used in this study. The support provided by Maria Papathanasiou, Niki Triantafyllou, and Andrea Bernardi was also highly appreciated.

Nomenclature

Abbreviations

AdV Adenoviral vector

CIP Cleaning-in-place

DFC Direct fixed capital

DSP Downstream processing

EER Economic evaluation report from SuperPro Designer

EH Ecosystem health

EIR Environmental impact report from SuperPro Designer

F&F Fill and finish processing

HH Human health

ICR Itemised costing report from SuperPro Designer

LLDPE Linear low-density polyethylene

LV Lentiviral vector

MAb Monoclonal antibodies (drug class)

MILP Mixed-integer linear programming

MUT Multi-use technology

PET Polyethylene terephthalate

PP Polypropylene

RA Resource availability environmental impact

SIP Steaming-in-place

SUT Single-use technology

USP Upstream processing

Sets

a Scales

d DSP manufacturing lines

g F&F manufacturing lines

i Products

j Manufacturing nodes

s Supply chain type (SUT, MUT)

u USP manufacturing lines

Model Variables

TC Total cost *USD*2021

TE Total environmental impact *USD*2017

z Objective for minimisation

Units

DALY Disability-adjusted life year

species · yr Local species loss integrated over time

*USD*2013 USD value in 2013 for Resource Availability

*USD*2017 USD value in 2017 for Total Environmental Impact

*USD*2021 USD value in 2021 for Total Cost

References

Capra, E., Godfrey, A., Loche, A. & Smith, J. (September 2021) Gene-therapy innovation: Unlocking the promise of viral vectors. *McKinsey & Company*. Available from: https://www.mckinsey.com/industries/life-sciences/our-insights/gene-therapy-innovation-unlocking-the-promise

Dekker, E., Zijp, M. C., Kamp, M. E. van de, Temme, E. H. & Zelm, R. van (2019) A taste of the new recipe for Life Cycle Assessment: Consequences of the updated Impact assessment method on food product lcas. *The International Journal of Life Cycle Assessment*. 25 (12), 2315–2324. Available from: DOI:10.1007/s11367-019-01653-3.

Dong, Y., Hauschild, M., Sørup, H., Rousselet, R. & Fantke, P. (October 2018) Evaluating the monetary values of greenhouse gases emissions in Life Cycle Impact Assessment. *Journal of Cleaner Production*. 209. Available from: DOI:10.1016/j.jclepro.2018.10.205.

Flanagan, B. (2017) Single-use technology and sustainability: quantifying the environmental impact in biologic manufacturing. *Cytiva Life Sciences*.

Grand View Research (November 2021) Advanced Therapy Medicinal Products Market Report, 2028. *Grand View Research*. Available from: https://www.grandviewresearch.com/industry-analysis/advanced-therapy-market

Huijbregts, M., Steinmann, Z., Elshout, P., Stam, G., Verones, F., Vieira, M., Zijp, M., Hollander, A. & Zelm, R. (December 2016) ReCiPe2016: a harmonised life cycle impact assessment method at midpoint and endpoint level. *The International Journal of Life Cycle Assessment*. 22. Available from: DOI:10.1007/s11367-016-1246-y.

Langer, E. S. & Morrow Jr., K. J. (February 2020) Rise of Single-Use Bioprocessing Technologies: Dominating Most R&D and Clinical Manufacture. *American Pharmaceutical Review*.

Lopes, A. G. (2015) Single-use in the biopharmaceutical industry: A review of current technology impact, challenges and limitations. *Food and Bioproducts Processing*. 93, 98–114. Available from: DOI:https://doi.org/10.1016/j.fbp.2013.12.002.

Lütke-Eversloh, T. & Rogge, P. (January 2018) Biopharmaceutical manufacturing in single-use bioreactors current status and challenges from a CDMO perspective. *Pharmazeutische Industrie*. 80, 281–284.

Mori, M., Drobnič, B., Gantar, G. & Sekavčnik, M. (August 2013) Life Cycle Assessment of Supermarket Plastic Bags and Opportunities for Bioplastics.

Sala, S., Cerutti, A. K. & Pant, R. (2018) Development of a weighting approach for the Environmental Footprint. *Publications Office of the European Union*. Available from: DOI:10.2760/945290.

Sarkis, M., Bernardi, A., Shah, N. & Papathanasiou, M. M. (May 2021) Decision support tools for next-generation vaccines and advanced therapy medicinal products: Present and future. *Current Opinion in Chemical Engineering*. 32, 100689. Available from: DOI:10.1016/j.coche.2021.100689.

Sarkis, M., Shah, N. & Papathanasiou, M. M. (2022) Characterization of key manufacturing uncertainties in next generation therapeutics and vaccines via global sensitivity analysis. *Journal of Advanced Manufacturing and Processing, (Submitted)*.

Sarkis, M., Tak, K., Chachuat, B., Shah, N. & Papathanasiou, M. M. (2022) Towards Resilience in Next-Generation Vaccines and Therapeutics Supply Chains. *32nd European Symposium on Computer Aided Process Engineering*. Ed. by L. Montastruc & S. Negny. Vol. 51. Computer Aided Chemical Engineering. Elsevier, 931–936. Available from: DOI:https://doi.org/10.1016/B978-0-323-95879-0.50156-9.

Sinclair, A., Leveen, L., Monge, M., Lim, J. & Cox, S. (November 2008) The Environmental Impact of Disposable Technologies Can disposables reduce your facility's environmental footprint? *Biopharm International*, 4–+.

Sinclair, A. & Monge, M. (January 2002) Quantitative economic evaluation of single use disposables in bioprocessing. *Pharmaceutical Engineering*. 22, 20–34.

Wieland, M. (2022) Single-use bioprocessing: Why it pays off to switch to single-use systems now. Available from: https://www.susupport.com/single-use-bioprocessing-switch

Optimisation of Full Oxyfuel Combustion for Cement Production

Darius Ojoko and Olawunmi Olatidoye
Department of Chemical Engineering, Imperial College London, U.K.

Abstract Cement is an essential material for construction buildings. However, the process of production alone is accountable for 5% of global carbon dioxide emissions. The implementation of oxyfuel combustion in the process allows for higher purities of carbon dioxide to be released from the output flue gas, thus allowing the carbon dioxide to be captured more easily using carbon capture systems and potentially being stored or processed further. In this study, a full oxyfuel combustion plant was simulated in ASPEN Plus V11, achieving a CO_2 rich flue gas at full conversion of calcium carbonate. The model was also optimised to meet baseline cement plant specifications so that cement production still met its annual target. The efficacy of the plant operation was also analysed through a set of key performance indicators, which showed general agreement with current cement plants. The plant was also cost analysed and was found to be cost effective when compared with existing cement plants, especially considering that the overall capital expenditure included the cost of CO_2 purification. Full oxyfuel combustion in the cement process was found to be a viable technology for reducing the CO_2 emissions in the production process, however it is unlikely that new cement plants will be built in the near future, thus further study into the construction of retrofitted oxyfuel technologies is recommended.

Keywords – CO_2, clinker, cement, flue gas, oxyfuel,

1 Introduction

Cement is used in the infrastructure of almost all buildings worldwide. The process utilises calcium and silicon which are very abundant, making them suitable for mass production (Perilli, 2019). Traditionally, the process of producing cement requires clinker to be burned with coal. This produces a considerable amount of carbon dioxide, in which the industry contributes to approximately 5% of global carbon dioxide emission (Voldsund, 2021). Decarbonisation of the cement production process will allow for a continued use of a material which the world is highly dependent on.

Oxyfuel combustion in the cement process utilises oxygen and recycled carbon dioxide alongside fuel for combustion in the kiln in the place of natural gas. The clinker burns at a temperature higher than that with ambient air, which allows for a more efficient combustion process (Linde, 2022). One can achieve higher purities of carbon dioxide in the flue gas, aiding the decarbonisation of the process. In addition, nitrogen no longer dilutes the flue gas stream, making it easier to handle and treat. This is beneficial as the carbon dioxide in the flue gas can be extracted more easily by the carbon capture plant, thus a greater amount can be stored instead of being released into the environment.

IKN designed and installed the first successful oxyfuel clinker cooler at the HeidelbergCement plant in Hannover, Germany (CEMCAP, 2018). The aims of the plant were to develop a system that produced clinker with reduced CO_2 emissions and reduced energy consumption. In addition, HeidelbergCement aimed to achieve the same quality of cement produced when cooled in a CO_2 rich atmosphere as with ambient air (HeidelbergCement, 2021). The budget for the project equated to €10M. A pilot oxyfuel cement burner has also been tested by University of Stuttgart in collaboration with Thyssenkrupp. The aim of the downscaled industrial process was to test the impact of high concentrations of CO_2 in the atmosphere and the effect it would have on the calcination process (Carrasco-Maldonaldo, 2021). In October 2022, TotalEnergies and Holcim signed a memorandum of understanding to work towards the decarbonisation of the Obourg Cement plant in Belgium (Global Cement, 2022). The plant currently emits 1.3Mt/yr of CO_2 in which they plan to decrease. One of the technologies included in the decarbonisation will include an 'oxyfuel switchable kiln' as stated by Global Cement (2022). An electrolyser is also in development to produce green hydrogen to produce e-fuels. The switchable kiln would ultimately be fuelled by the oxygen released in the electrolyser (Process Worldwide, 2022).

The main objective of this project was to simulate a cost-effective cement process with full oxyfuel combustion. The modelling of the process was optimised with an aim of achieving a high concentration of carbon dioxide output at 100% conversion of calcium carbonate. The process simulation for the cement plant was carried out using ASPEN Plus V11. A range of key performance indicators (KPIs) was selected to assess the efficacy of the system in areas such as power performance and emission intensity. An economic analysis was carried out to evaluate whether oxyfuel combustion yielded an overall decrease in costs.

2 Background

The cement process consists of raw meal being dried and sent through a series of five preheating cyclones and precalciner. It is then sent through a rotary kiln where it then becomes clinker. The clinker goes through a clinker cooler and clinker mill. The intermediate calcium oxide is formed from the

calcination of calcium carbonate as detailed in Equation 1. This process accounts for 60% of carbon dioxide emissions and takes place at 900ºC (Driver et al., 2022).

$$CaCO_3 \leftrightarrow CaO + CO_2 \quad (1)$$
$$\Delta H_{298k} = +174kJ/mol$$

Carbon dioxide is also produced as a by product of the formation of magnesium oxide from magnesium carbonate in Equation 2 (Driver et al., 2022).

$$MgCO_3 \leftrightarrow MgO + CO_2 \quad (2)$$
$$\Delta H_{298k} = +118kJ/mol$$

Reactions in Equations 3-6 account for the remaining 40% of emissions from this process. These clinkering reactions take place at 1450ºC (Driver et al., 2022).

$$2CaO + SiO_2 \leftrightarrow 2CaO\Delta SiO_2 \quad (3)$$
$$\Delta H_{298k} = -126kJ/mol$$

$$3CaO + SiO_2 \leftrightarrow 3CaO\Delta SiO_2 \quad (4)$$
$$\Delta H_{298k} = -113kJ/mol$$

$$3CaO + Al_2O_3 \rightarrow 2CaO\Delta Al_2O_3 \quad (5)$$
$$\Delta H_{298k} = -7.3kJ/mol$$

$$4CaO + (Fe,Al)_2O_3 \quad (6)$$
$$\leftrightarrow 4CaO\Delta(Fe,Al)_2O_3$$
$$\Delta H_{298k} = -27kJ/kg$$

In 2019, The Global Cement Future Conference suggested that the probability of oxyfuel technology being successfully demonstrated in cement production was between 40-60% (Perilli, 2019). A challenge that can arise from oxyfuel combustion could be the flame length and shape in the rotary kiln. A longer flame length than necessary can reduce the longevity of components and affect clinker reactions (Pneumat Systems Inc., 2021.). Optimising this would require changes to be made to the burner in the kiln to no longer impact the quality of clinker (Bakken & Ditaranto, 2019). From a health and safety perspective, operating oxygen of high purity with fuels can increase the risk to workers when combined with the high temperatures in the kiln (Zeman, 2009). The process will also have to consider an oxygen production facility as well as a CO_2 compression station in terms of capital expenditure, so that the CO_2 can be transported for carbon capture and storage (Zeman, 2009).

Carbon Capture Utilisation and storage (CCUS) technologies play a key role in keeping the global temperature with 2ºC above pre – industrial levels, as per the aims of The Paris Agreement (IEA, 2020). Recently, the UK Government funded 9 companies in 2019 to develop projects that could reduce carbon dioxide emissions to an equivalent of 22,000 cars (GOV.UK, 2019), granting £26 million as part of their aim to achieve Net Zero by 2050. CCUS technologies specifically geared towards the cement industry are set to be commercially available by 2030 (World Economic Forum, 2022).

Alternate technologies such as hydrogen fuel has been considered as an alternative to using fossil fuels in the cement process (Davis, Fennell & Mohammed, 2021). This consists of integrating an ammonia decomposition system to supply hydrogen to the cement plant for fuel. Studies have shown that this can equate to a 44% decrease in CO_2 emissions in comparison to that of the current system (Aziz et al., 2022). Biofuels have also been explored as an alternative to fossil fuels (Hiremath et al., 2022). It was found that they could be substituted without great changes to the overall capital investment. This alternative is only feasible when co-fired up to 20% with coal and is highly dependent on the type of biomass fuel sought after (Cuéllar & Herzog, 2015).

Other technologies integrated in the cement production process include calcium looping, a post-combustion technology that has been applied on a pilot scale to the cement process (Hornberger, Scheffknecht & Spörl, 2017). It has the potential to achieve 95% capture rate of CO_2 whilst maintaining a high energy efficiency, however the process is known for attrition and deactivation of lime (Adjiman et al., 2018). Another post combustion technology considered was biological CO2 capture with algae (Global Cement, 2014). Up to 20% of the oil produced from the process can be used for biodiesel in the cement process, however there are various challenges to be combatted with upscaling the process.

Finally partial oxyfuel combustion has also been studied where only the calciner operates under oxyfuel conditions as supposed to both the calciner and kiln (Carrasco-Maldonaldo et al., 2016). This equates to an estimated 60% concentration of CO_2 in the flue gas and requires a higher level of maintenance in comparison to full oxyfuel operation. Based on the ECRA plant it was estimated that a new installation of an oxyfuel combustion cement plant was €290.7M compared to €103.7M for a retrofit (IEAGHG, 2013). The installation of an air separation unit for the oxyfuel process doubles the amount of energy required per tonne of clinker. The increased energy usage increases the cost of production by 40-50% (Global Cement and Concrete Association, 2022). However, it is assumed that the new installation would incur lower fuel costs due to the higher energy efficiency. Despite this oxyfuel combustion is still considered one of the most economical options for carbon capture in the cement process (Global Cement and Concrete Association, 2022).

3 Methodology

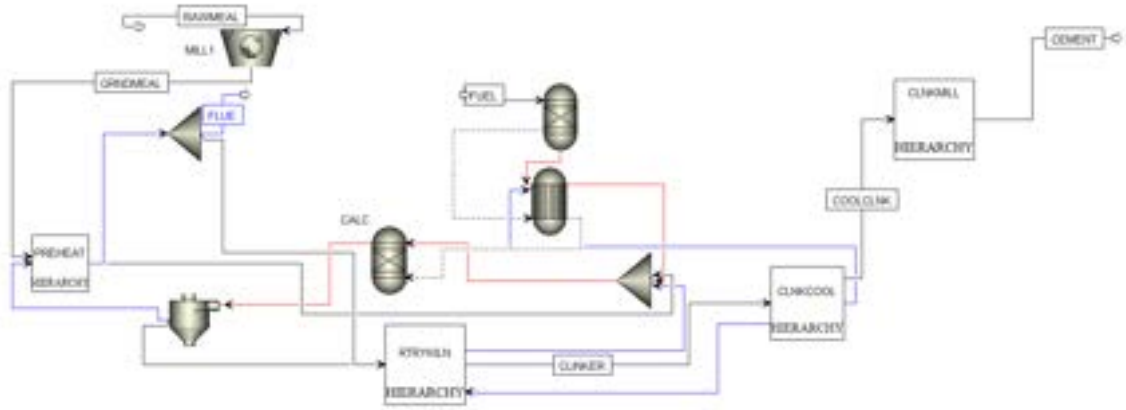


Figure 1. Flowsheet of oxyfuel process

Table 1. Assumed and Simulated Compositions of material streams in the oxyfuel process

Material	Mass Composition									
	$CaCO_3$	SiO_2	Al_2O_3	Fe_2O_3	$MgCO_3$	CaO	MgO	$CaSO_4$	C_3S	Ash
Raw Meal	79.2	13.2	4.4	1.9	1.3	-	-	-	-	
Clinker	-	-	6.8	2.9	-	11.3	1.0	-	77.1	0.9
Gypsum								100.0		
GBFS		33.0	13.0	2.0	7.0	45.0				
Cement	-	10.5	8.3	2.4	2.2	21.3	0.6	6.3	47.9	0.6

Table 2. Assumed fuel composition

HHV ($MJkg^{-1}$)	22.4					
Proximate Analysis (wt%)	**W**	**FC**	**VM**	**Ash**		
	10.0	72.0	21.0	7.0		
Ultimate Analysis (wt%)	**C**	**H**	**N**	**S**	**O**	**Ash**
	68.0	5.0	2.0	1.0	17.0	7.0

3.1 Modelling overview

The cement plant was modelled using the simulation package Aspen Plus V11. The Peng-Robinson equation of state was used to calculate system properties (Driver et al., 2022). This was decided based on literature research as it had been used to model similar processes. The flowsheet for the cement plant is shown above in figure 1. Table 1 and 2 features the compositions of the material and fuel streams for the cement plant.

3.2 Modelling the cement plant

3.2.1 Mills

A closed circuit ball mill was used for milling of the raw meal and the cooled clinker. This was simulated with a fixed energy consumption of 36 kJ/kg (Sprung, 2008). Grindability of the meal was specified at 13.5 kWh/ton (Deniz, 2004). The milled meal was then transferred to the preheating cyclone. In the case of the cooled clinker, the grindability and fixed energy consumption were set to 13.7 kWh/ton and 90 kJ/kg respectively (Deniz, 2004) (Sprung, 2008). Prior to milling, the cooled clinker is mixed with gypsum and GBFS to obtain the desired cement composition.

3.2.2 Preheating Cyclone and Calciner

The preheating cyclone is fed with the ground raw meal at one end and the hot kiln, calciner and fuel gases at the other end in a counter current design. The preheating cyclone which consists of 5 stages is modelled as 5 RGibbs reactors connected in series. The outlet from the topmost stage at 150 ^{0}C is the flue gas stream (predominantly containing CO_2)

which is sent to a CO_2 purification unit. RGibbs reactors were used as they can determine the equilibrium compositions at each stage through the identification of likely products. The calciner was also modelled as an RGibbs reactor for the same reasons. The calciner calcined all the $CaCO_3$. It was fuelled by coal-RDF. Fuelling was modelled with an RYield, RGibbs and SSplit block. The RYield block was used to simulate the decompositions of the coal into its component elements along with 10% water by weight fraction. A calculator block was set up to determine the yields based on the coal composition. The outlet from the RYield block is then passed to an RGibbs block where the combustion reactions are simulated. To simulate the combustion reactions recycled oxygen from the clinker cooler is introduced into the calciner. The mixed phase outlet stream is then fed into the calciner.

3.2.3 Rotary Kiln

The burner within the rotary kiln was modelled the same as that in the calciner, with oxygen being recycled from the hottest end of the clinker cooler to drive the combustion. Firing was assumed to be fixed at a temperature of 1400^0C. To drive the reaction, heat was transferred from the RGibbs block (representing the burner) to the RStoic block. The reactions in the rotary kiln were limited to formation of C_3S as the component bank of the simulation package did not contain the remainder of the cement components. Following reaction, the clinker is sent to the clinker cooler and the gaseous products are recycled into the burner to boost thermal efficiency. The hot gases from the burner were passed through a HeatX block to preheat the kiln feed to improve thermal efficiency of the process. The hot flue gases are then sent to the calciner and subsequently the preheating stages where it is cooled to 150^0C.

3.2.4 Clinker Cooler

The clinker cooler was modelled as three HeatX blocks with the hot clinker as the hot stream and oxygen (25^0C) as the cold stream. To simulate operation of a typical clinker cooler which would be a conveyor belt fitted with blowers, the clinker cooler was modelled in co-current flow. The heated oxygen from the first stage was sent to the rotary kiln. The oxygen from the second stage was sent to the calciner.

3.3 Key Performance Indicators (KPIs)

To assess the cement plant, four indicators were chosen to compare the performance with the baseline cement plant. These are listed in equations 7-10. Final KPI values were obtained at full $CaCO_3$ conversion by varying the fuel flowrate to ensure optimum operation. The specific electrical duty was taken as the combined power consumption by the mills in the system divided by the total mass of cement produced. The specific thermal duty was measured by combining the total fuel flowrate and dividing it by the mass of clinker produced. The higher heating values (HHV_f) for all fuel inlets were 22.4 due to their composition. Finally, the Emission Intensity was measured by comparing the amount of CO_2 in the flue gas with the amount of cement produced.

$$\text{Conversion} = \frac{\dot{m}_{CaCO3,0} - \dot{m}_{CaCO3,n}}{\dot{m}_{CaCO3,0}} \quad (7)$$

$$\text{Specific Electrical Duty (kWh/t}_{cmnt}\text{)} = \frac{\sum P_i}{\dot{m}_{cmnt}} \quad (8)$$

$$\text{Specific Thermal Duty (GJ/t}_{clk}\text{)} = \frac{\sum(\dot{m}_f \times HHV_f)}{\dot{m}_{clk}} \quad (9)$$

$$\text{Emissions Intensity (t}_{CO2}\text{/t}_{cmnt}\text{)} = \frac{\dot{m}_{CO2}}{\dot{m}_{cmnt}} \quad (10)$$

Where $\dot{m}_{CaCO3,0}$ is the inlet mass flowrate of $CaCO_3$, $\dot{m}_{CaCO3,n}$ is the outlet mass flowrate of $CaCO_3$. P_i is the total power, $\dot{m}_{cmnt}$ is the outlet mass flowrate of cement. $\dot{m}_f$ is the inlet mass flowrate of fuel, HHV_f is the higher heating value and $\dot{m}_{clk}$ is the outlet mass flowrate of clinker. $\dot{m}_{CO2}$ is the outlet mass flowrate of CO_2.

4 Results

4.1 Specifications and KPIs

Table 3 illustrates a direct comparison of the full oxyfuel cement plant against a baseline cement plant (IEAGHG, 2013). The oxyfuel plant generally met the specifications of the baseline cement plant with only slight variation in the specific electrical duty. The CO_2 concentration in the flue gas was found to be 87.7% at 16,000 kg/hr fuel station flowrate and 2,500 kg/hr rotary kiln fuel flowrate.

Table 3. Summary of specifications and KPIs for model cement plant against the oxyfuel simulation

Parameter	Scenario		Unit
	Baseline	Oxyfuel	
Clinker production	1.00	1.01	Mt/yr
Cement Production	1.40	1.63	Mt/yr
Specific CO_2	850	878	kg_{CO2}/t_{clk}
Raw meal to Clinker factor	1.60	1.54	-
Conversion	1.00	1.00	-
Specific Electrical Duty	89.00	35.10	kWh/t_{cmnt}

Specific Thermal Duty	3.40	3.18	GJ/t_{clk}
Emission Intensity	0.50	0.54	t_{CO2}/t_{cmnt}

4.2 Fuel sensitivity

The fuel flowrate sensitivities were conducted to assess the effect of varying raw meal flowrate, fuel station flowrate and rotary kiln fuel flowrate on various parameters. The sensitivity test values ranged from 5,000-16,000 kg/hr of fuel station flowrate and the rotary kiln flowrate was varied between 500-6,500 kg/hr, whilst raw meal flowrate was maintained at 200,000 kg/hr.

4.2.1 Conversion

Calcium carbonate conversion was heavily affected by the fuel station flowrates. The inlet $CaCO_3$ flowrate was maintained at 158,400 kg/hr. It was observed that complete conversion did not take place until the flowrate was at least 15500 kg/hr of fuel as shown in Figure 2.

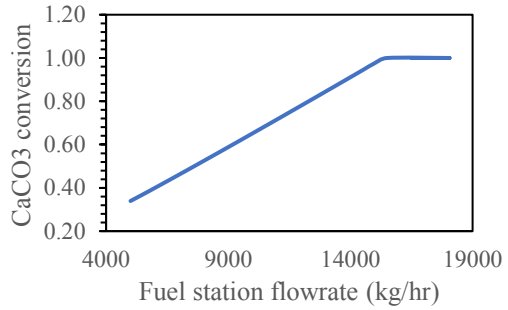


Figure 2. CaCO3 conversion against fuel station flowrate

4.2.2 KPIs

Sensitivity tests were carried out to assess the effect on specific thermal duty and emission intensity. Full conversion of calcium carbonate was maintained during testing. Specific thermal duty values ranged from 3.0-4.4 GJ/t_{clk} which is in line with literature (Driver et al., 2022). Emission intensity also remained between 0.5-0.6 t_{CO2}/t_{cmnt} after full conversion as per Figure 3.

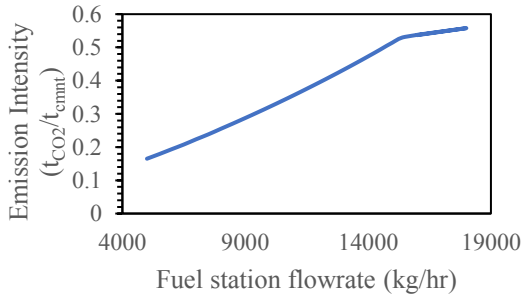


Figure 3. Emission Intensity against fuel station flowrate

The specific electrical duty was found to be 35.1 kWh/t_{cmnt}. When the fuel station flowrate was 12000 kg/hr, the specific electrical duty was at its lowest of 34.5 kWh/t_{cmnt} however this meant that full conversion of $CaCO_3$ did not take place. The specific electrical duty remained between 34.4-35.4 kWh/t_{cmnt} for the duration of testing.

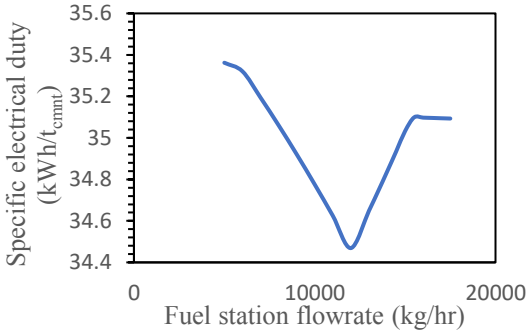


Figure 4. Specific electrical duty varying with fuel station flowrate

4.3 Effect of Recycle

Testing was carried out to observe how recycle of the flue gas stream exiting the topmost stage of the preheating cyclone would affect the process.

An effect was observed in the kiln heat duty which consequently affected the cost of the kiln. Results of these are shown in figures 5 and 6 respectively.

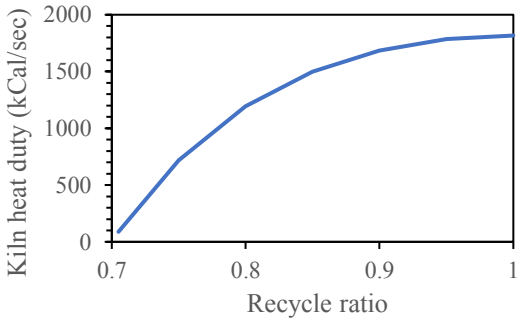


Figure 5. Kiln heat duty varying with recycle ratio

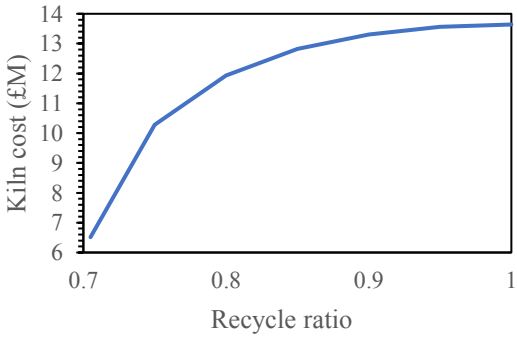


Figure 6. Kiln cost varying with recycle ratio

4.4 Effects of Preheaters

Testing was carried out to observe how a reduction in the number of preheating stages would affect the process.

An effect was observed in the conversion and specific thermal duty, results of which are shown in figures 7 and 8 respectively.

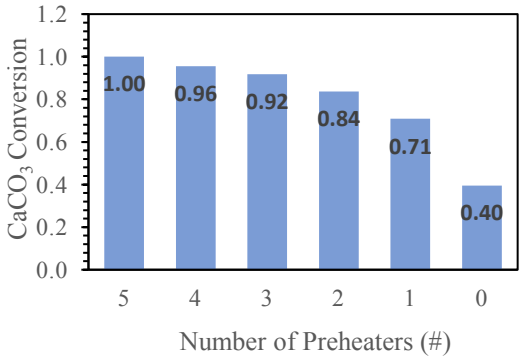


Figure 7. The effect of number of preheaters on $CaCO_3$ conversion

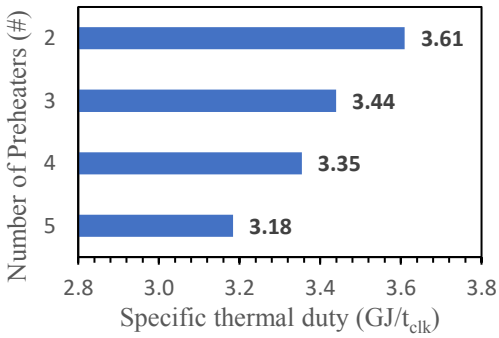


Figure 8. The effect of number of preheaters on specific thermal duty (GJ/t_{clk})

4.5 Economic Analysis

Table 4: Calculated and assumed purchase cost of equipment for oxyfuel plant with varying number of preheaters

Unit Operation	Purchase Cost of Equipment (£M)			
	Oxyfuel plant with five preheating	Oxyfuel plant with four preheating stages	Oxyfuel plant with three preheating stages	Oxyfuel plant with two preheating stages
Pneumatic Conveyor	1.45	1.45	1.45	1.45
Preheater	0.14	0.18	0.20	0.22
Raw Meal Mill	1.19	1.19	1.19	1.19
Calciner	3.24	3.18	3.13	3.05
Rotary Kiln	6.51	6.58	7.84	8.10
Clinker cooler	0.63	0.63	0.63	0.63
Clinker Mill	1.19	1.19	1.19	1.19
Air Separation unit	10.90	10.90	10.90	10.90
CO_2 purification unit	9.50	9.50	9.50	9.50
Total PCE	**34.75**	**34.80**	**36.03**	**36.23**

Table 5: Annual materials and utilities costs for the oxyfuel plant

Input	Unit Price Cost	Unit	Quantity	Annual Cost (£M/y)
Materials (t/hr)				
Raw Meal	5	£/t	200.00	7.92
GBFS	2	£/t	13.28	0.21
Gypsum	6	£/t	66.14	3.14
Coal-RDF	65	£/t	18.50	9.52
Utilities (MW)				
Electricity	58.1	£/MW	47.15	42.49

Table 6. OPEX and associated costs

Costing Parameter	Value (£M)
MC	1.1
LC	4.5
OC	7.7
MSC	23.1
VOC	27.0
DPC	50.1
IPC	5.0
OPEX	55.1

Table 7. CAPEX and associated costs

Costing Parameter	Value (£M)
PCC	34.75
PPC	109.5
FCC	153.3
WCC	43.1
CAPEX	**196.3**

Economic analysis of the process was conducted. This costing was done using methods and factors outlined in Chemical Engineering Economics (Garrett, 1989). Costing of the CO_2 and air purification units was not conducted manually. Values for these were obtained from literature (Gardarsdottir, S. O., 2019). The results of these are summarised in tables 4 to 7

5 Discussion

5.1 Specifications and KPIs

The close agreement of the specifications of the baseline cement plant with the oxyfuel cement plant reinforces the feasibility of operation under full oxyfuel conditions. Table 3 shows that clinker production could be maintained at the same target whilst exceeding annual cement production at full $CaCO_3$ conversion. The CO_2 concentration in the flue gas was slightly lower than anticipated, however with further optimisation high concentrations could be achieved. The specific electrical duty differs the most from the literature value (IEAGHG, 2013), in which the difference could be accounted for by considering the power of the pneumatic conveyor if given the ability to model.

5.2 Fuel sensitivity

5.2.1 Conversion

A fuel station flowrate of 5,000 kg/hr of fuel equated to a conversion of 34% as shown in Figure 2. This dropped the emission intensity fell 0.54 to 0.17 t_{CO2}/t_{cmnt}. The cement mill cost also decreased with the increase in the fuel station flowrate due to the decrease in production as more flue gas was being produced.

5.2.2 KPIs

The specific thermal duty increased with an increase in flowrate. This was as expected as the mass of fuel put into the system increases greater than the increase in clinker produced.

It was observed that even at flowrates as low as 500kg/hr, the specific CO_2 emission were 844 kg/hr due to the raw meal inlet flowrate being 200,000 kg/hr and calcination taking place. This is also due to the process being complete oxyfuel so that only oxygen is supplied to the clinker. As the rotary kiln flowrate was increased incrementally, the specific CO_2 emissions increase to 946 kg_{CO2}/t_{clk}. Specific CO_2 emissions increased more drastically with the increase in fuel flowrate as the conversion of $CaCO_3$ increases, thus more calcination is taking place per unit time, thus showing the fuel station flowrate has a more direct effect on the CO_2 emitted from the process than the rotary kiln.

Specific electrical duty was mainly affected by the fuel station flowrate. The production of cement decreased with the increase in fuel station flowrate but began to decrease faster when the flowrate increased to 12,000 kg/hr as the $CaCO_3$ approached full conversion. As with the power, the mill grinding the raw meal remained at constant power due to production specifications. However, the mill within the clinker's power consumption steadily decreased as the fuel flowrate increased and became constant as full $CaCO_3$ conversion was reached. The effect of both the power consumption and cement production meant the specific electrical duty varied by less than 1% throughout testing, remaining between 34.4-35.4kWh/t_{cmnt}.

5.3 Effects of recycle

Recycling the flue gas from the topmost stage of the preheating cyclone has a dramatic effect on the heat duty of the kiln. This can be seen not just in the actual heat duty value of the kiln but also in the outlet temperature of the kiln. Increasing the fraction being recycled back into the reactor from 0 to 0.295 is able to decrease the outlet temperature of the kiln by 300^0C.

As the recycle ratio is increased, the heat duty of the kiln drops significantly. Increasing the recycle ratio to 29.5% drops the kiln's heat duty by two orders of magnitude. At this ratio the heat duty drops to 88,356.8 cal/s. This happens due to the kiln gas being sent back at significantly lower temperatures than the operating temperature of the kiln (1365^0C). As a result of this it can absorb excess heat from the kiln.

5.4 Effects of preheaters

Testing of the effects of the preheating cyclone on the process was done by reducing the number of stages from five to zero. The most marked effect this had on the process was on the conversion of $CaCO_3$. This went from 100% when all five stages are implemented to zero without preheating.

This is because when full preheating is present, heat being generated in the calciner is used only to drive the calcination reactions. This calcination reactions occur at ~900^0C (Sprung, 2008). However, as preheating stages are reduced the raw meal is introduced into the calciner at increasingly lower temperatures. The heat generated by the fuel now must first heat the raw meal to calcination temperatures. Thus, there is less heat available to drive the calcination reactions.

It is important to note that even at no preheating, the conversion of $CaCO_3$ could possibly still reach a 100% if the fuel flowrate is increased. However, this would increase the both the capital and operating costs of the process. This behaviour is seen in table 4, where a reduction in preheating stages resulted in an increased purchased equipment cost. This is because an increase in fuel flowrate not only directly increases the material costs associated with it, but also the heat duties and fuel blower capacity requirements of the calciner and the kiln thus raising capital costs.

Another feature of the process highlighted by the preheating tests was the relationship between the number of stages and the specific thermal duty. It was noted that the less the number of stages, the greater the specific thermal duty of the process. This is because when the number of stages is reduced, in order to achieve full conversion the mass of fuel inputted into the process is increased.
The reduction in preheating stages also appears to increase the mass of clinker. Whilst this may appear to make a case for a reduced number of stages, upon closer inspection it can be observed that this is simply due to the reduced conversion of $CaCO_3$ which has a high molecular mass.

5.5 Economic Analysis
The viability of this process of course largely depends on its economic feasibility compared to a standard cement plant. Implementation of the oxyfuel process on a large scale requires that it does not dramatically swell the production costs. Tables 4 to 7 in section 3 show the costing results of the oxyfuel process. It is important to note that the value used for electricity in the costing ($225 kWh_e/t_{cmnt}$) was obtained from literature rather than the simulation (Zeman, 2009). This was done because due to the nature of the model, the electrical duty results are not satisfactorily accurate and were primarily used to observe trends.
The simulated plant had a CAPEX of £196.3 M. This is above typical ranges (£158-189 M) for a standard cement plant (Driver, 2022). However, this is below typical values for a new build oxyfuel plant as the units in these must be fitted to prevent gas ingress or egress and so are typically simulated closer to values of £250 M (Global CCS Institute, 2013). Although, as it has been costed on a standard cement plant basis and its value lies within typical ranges for such a model, it can be concluded that the model along with the costing methods used are valid. Therefore, despite uncertainty in the accuracy of the CAPEX value the trends observed are still insightful. Furthermore, as this cost is inclusive of the air separation and CO_2 purification units it is also highly promising despite the uncertainty.
When the number of preheating stages is reduced, the total purchased equipment cost is increased. This is due to the fuel flowrate having to be increased to achieve full conversion. Therefore, as the preheater and fuel blowers in the calciner and kiln are costed based on gas flowrates, there is thus an increase in their associated costs.
From these results it is clear that from a financial perspective, oxyfuel is competitive with current industry standards.

6 Conclusions
This study looked at the application of oxyfuel technology in cement production. This was considered as a means to reduce CO_2 emissions as it would significantly simplify purification of the flue stream as it would no longer be rich in nitrogen. Simulation of an oxyfuel cement plant was carried out in ASPEN Plus V11 to assess performance of the plant under various conditions. Costing of the process was also conducted. $CaCO_3$ conversion and specific thermal duty were found to improve with increasing fuel flowrate and number of preheating stages. An increase in recycle ratio reduced the heat duty and cost of the kiln. Overall costing of the process put CAPEX and OPEX of the plant at £196.3M and £73.4M/yr. Additionally, the process was able to meet production targets. This indicates that this process is feasible and could potentially lead efforts to decarbonise the cement industry.
If this process is to be implemented on a wider scale, there are key research areas that need to be explored. The retrofitting of existing plants is one such key area. As new plants are unlikely to be built, further exploration into how the different gas properties affect individual units in the plant is required. This is paramount as there is a need for the units to be designed in such a way that gas ingress or egress is prevented (Hills, 2015).
Another key way in which this study could be advanced is an exploration into alternative oxygen sources. An exciting alternative is the use of oxygen from electrolytic hydrogen production (Nhuchhen, 2022). Technoeconomic feasibility studies into oxyfuel processes coupled with these would be highly illuminating.
In conclusion, large scale implementation of oxyfuel in the cement industry is years away. Studies such as this indicate that the process has immense potential for decarbonisation. To make it a reality however, further considerations into the challenges affecting its implementation should be researched further.

Acknowledgements
The authors would like to extend their gratitude to Haiyang Jiang and Zhuoyan Wang, (Imperial College London) for their continuous support and guidance throughout this research project.

References
Adjiman, C.S. et al. (2018) Carbon capture and storage (CCS): the way forward. *Energy & Environmental Science*. 11, 1062-1176. https://doi.org/10.1039/C7EE02342A
Aziz, M. et al. (2022) Thermodynamic analysis of hydrogen utilization as alternative fuel in cement production. *South African Journal of Chemical Engineering.* 42, 23-31. https://doi.org/10.1016/j.sajce.2022.07.003
Bakken, J., Ditaranto, M. (2019) Study of a full scale oxy-fuel cement rotary kiln. *International Journal of Greenhouse Gas Control.* 83, 166-175. https://doi.org/10.1016/j.ijggc.2019.02.008
Carrasco-Maldonaldo, F. et al. (2016) Oxy-fuel combustion technology for cement production –

State of the art research and technology development. *International Journal of Greenhouse Gas control.* 45, 189-99. https://doi.org/10.1016/j.ijggc.2015.12.014
Carrasco-Maldonaldo, F. (2021) *Pilot Testing, Simulation, and Scaling of an Oxyfuel Burner for Cement Kilns.* https://www.ifk.uni-stuttgart.de/en/research/publications/thesis-carrasco/ [Accessed 30th November 2022]
CEMCAP. (2018) *CEMCAP in cement world.* https://www.sintef.no/globalassets/project/cemcap/cemcap-in-cement-world.pdf [Accessed 2nd December 2022]
Cuéllar, A.D., Herzog, H.J. (2015) A Path Forward for Low Carbon Power from Biomass. *Energies.* 8, 1701-1715. https://www.researchgate.net/publication/279839407_A_Path_Forward_for_Low_Carbon_Power_from_Biomass
Davis, J.S., Fennell, P.S., Mohammed, A. (2021) Decarbonizing cement production. *Joule.* 5 (6), 1305-11. https://doi.org/10.1016/j.joule.2021.04.011
Deniz, V. (2004) The effect of mill speed on kinetic breakage parameters of clinker and limestone. *Cement and Concrete Research.* 34 (8), 1365-1371. https://doi.org/10.1016/j.cemconres.2003.12.025
Driver, J.G. et al. (2022) Simulation of direct separation technology for carbon capture and storage in the cement industry. *Chemical Engineering Journal*. 449, 137721. https://doi.org/10.1016/j.cej.2022.137721
Gardarsdottir, S. O. et al. (2019) Comparison of Technologies for CO2 Capture from Cement Production—Part 2: Cost Analysis. *Energies.* 12, 542. https://doi:10.3390/en12030542
Garrett, D.E., Appendix 1 - Equipment Cost Estimates, in: Chem. Eng. Econ., 1989. https://www.springer.com/gp/book/9789401165464
Global Cement. (2014) *Pond Biofuels launches bioreactor pilot project at St Marys cement plant.* https://www.globalcement.com/news/item/2680-pond-biofuels-launches-bioreactor-pilot-project-at-st-marys-cement-plant [Accessed 2nd December 2022]
Global Cement. (2022) *Holcim and TotalEnergies to work together on decarbonising upgrade to Obourg cement plant in Belgium.* https://www.globalcement.com/news/item/14721-holcim-and-totalenergies-to-work-together-on-decarbonising-upgrade-to-obourg-cement-plant-in-belgium [Accessed 28th November 2022]
Global Cement and Concrete Association. (2022) *Oxyfuel.* https://gccassociation.org/cement-and-concrete-innovation/carbon-capture-and-utilisation/oxyfuel/#:~:text=Despite%20this%2C%20oxyfuel%20remains%20one%20of%20the%20most,significant%20technology%20barriers%20to%20its%20implementation%20currently%20identified. [Accessed 17th November 2022]
Global CCS Institute. (2013). *DEPLOYMENT OF CCS IN THE CEMENT INDUSTRY*. IEAGHG.
GOV.UK. (2019) *UK's largest carbon capture project to prevent equivalent of 22,000 cars' emissions from polluting the atmosphere from 2021.* https://www.gov.uk/government/news/uks-largest-carbon-capture-project-to-prevent-equivalent-of-22000-cars-emissions-from-polluting-the-atmosphere-from-2021 [Accessed 7th November 2021]
HeidelbergCement. (2021) *Annual Report 2021.* HeidelbergCement
Hills, T. et al (2015) Carbon Capture in the Cement Industry: Technologies, Progress, and Retrofitting. *Environmental, Science and Technology.* 50 (1), 368-377. https://pubs.acs.org/doi/10.1021/acs.est.5b03508
Hiremath, R.B. et al (2022) Sustainable transition towards biomass-based cement industry: A review. *Renewable and Sustainable Energy Reviews*. 163, 112503. https://doi.org/10.1016/j.rser.2022.112503
Hornberger, M. Scheffknecht, G. Spörl, R. (2017) Calcium Looping for CO2 Capture in Cement Plants – Pilot Scale Test. *Energy Procedia.* 114, 6171-6174.- https://doi.org/10.1016/j.egypro.2017.03.1754
IEA. (2020) *Why carbon capture technologies are important.* https://www.iea.org/reports/the-role-of-ccus-in-low-carbon-power-systems/why-carbon-capture-technologies-are-important [Accessed 7th November 2022]
IEAGHG. (2013) *2013-19 Deployment of CCS in the Cement Industry* https://ieaghg.org/publications/technical-reports/reports-list/9-technical-reports/1016-2013-19-deployment-of-ccs-in-the-cement-industry [Accessed 30th November 2022]
Linde. (2022) *Oxyfuel Combustion* https://www.linde-engineering.com/en/process-plants/co2-plants/carbon-capture/oxyfuel/index.html [Accessed 7th November 2022]
Nhuchhen, D.R., Sit, S.P., Layzell, D. B. (2022) Decarbonization of cement production in a hydrogen economy. *Applied Energy.* 119180. https://doi.org/10.1016/j.apenergy.2022.119180
Perilli, D. (2019) *Cement plays the waiting game.* https://www.globalcement.com/news/item/9340-cement-plays-the-waiting-game [Accessed 16th November 2022]
Pneumat Systems Inc. (2021) *Alternative Fuels In Cement Production: Challenges And Solutions.* https://pneumat.com/cement-production-challenges-solutions/ [Accessed 30th November 2021]
Process Worldwide. (2022) *Industry First: Total Energies, Holcim to Decarbonize Cement Plant*. https://www.process-worldwide.com/industry-first-

total-energies-holcim-to-decarbonize-cement-plant-a-0b97eb5e1292433290bc9d2f705ceda8/ [Accessed 30th November 2022]
Sprung, S. (2008). *Ullmann's Encyclopedia of Industrial Chemistry*. Wiley-VCH Verlag GmbH & Co.
Voldsund, M. (2021) *Second-generation oxyfuel technology: Accelerating CCS in the cement industry.* https://blog.sintef.com/sintefenergy/second-generation-oxyfuel-technology-accelerating-ccs-in-the-cement-industry/ [Accessed 10th November 2022]
World Economic Forum. (2022) *The Net-Zero Industry Tracker* https://www.weforum.org/reports/the-net-zero-industry-tracker/in-full/cement-industry [Accessed 7th November 2022]
Zeman, F. (2009) Oxygen combustion in cement production. *Energy Procedia.* 1 (1), 187-194. https://doi.org/10.1016/j.egypro.2009.01.027

Virus-mimicking liposomes for intracellular delivery

Bo Gu, Qianhui Guo
Department of Chemical Engineering, Imperial College London, U.K.

Abstract
Effective drug delivery systems for cancer treatment require abilities to escape endosomes and release drugs with pH change. A novel virus-mimicking liposomal system has been developed to harness the advantage of pH-responsive pseudopeptides. Comb-like polymers were produced by grafting aminododecanoic acid onto a biodegradable metabolite-derived pseudopeptide, poly(L-lysine iso-phthalamide). The long aliphatic side chain successfully anchors the pseudopeptides on the cholesterol-containing liposomes that mimic the viral envelope to create an endosomolytic liposomal system. Favourable particle size control was achieved by extrusion to enhance the performance of liposomes. Optimising the polymer coating concentration allowed the system to exhibit negligible leakage at physiological pH while quickly releasing the payload at pH 6.5. The endosomolytic ability was also demonstrated through haemolysis assay at different pHs. Moreover, cell studies using HeLa cells further confirmed the capability of the liposomal system to effectively deliver both small-molecule and macromolecular drugs. The system has shown great potential for future cancer treatments.

Keywords: Liposomes, pH-responsive, virus-mimicking, intracellular drug delivery

1. Introduction

Therapeutic methods for cancer have been developed for decades. While chemotherapy is the most popular method among all, there are still many concerning issues, such as low specificity and high cytotoxicity.[1] Therefore, drug delivery systems have been a hot topic for researchers to find effective solutions. The viral vector is a successful drug delivery platform that has been extensively studied in recent years for cancer treatment. However, its full potential is yet to be realised due to safety issues.[2] Non-viral delivery methods thus gain increasing attention for their safety and ease of production.[3] Nanotechnology is widely employed for its capability of target-oriented delivery.[4] With advances in nanotechnology, liposomes have been extensively fabricated since the 1990s with the first use of Doxil.[5] Liposomes have high biocompatibility and are capable of loading various types of drugs. However, for liposomes to be successful nanocarriers, they also need to possess the ability to target disease sites, escape from endosomes, and release drugs upon pH changes.[6] Inspired by the feature of anionic peptides from viruses that allow them to disrupt cell membranes to release their content, scientists aim to incorporate the feature to create liposomal systems that can deliver drugs in a targeted, controllable and efficient manner.[3] Liposomal systems with pH-responsive ability are able to control the drug release according to changes in the environment. Under physiological conditions, liposomal systems should aim to reach specific sites to be taken by cells. With pH reduction during endocytosis, liposomal systems would react towards the change to allow drugs to be delivered. Such functions can be developed by coating polymers on the liposomal surface to mimic the viruses.[7] Naturally derived fusogenic peptides could be a choice, but safety problems and production difficulties drive researchers to find other alternatives.[3]

Learning from viral peptides, a series of novel pseudopeptides were synthesised with pH-responsive and membrane-destabilising abilities. Hydrophobic amino acids found in viral peptides were grafted on a biodegradable metabolite-derived pseudopeptide, poly(L-lysine iso-phthalamide) (PLP), first to create a polymer with manipulatable amphiphilicity.[8-10] While the polymer displayed good pH-triggered membrane activity, improvements could still be made to enhance the polymer-membrane interaction. Therefore, with inspiration from natural membrane proteins that contain long fatty acid chains, a second-generation comb-like polymer was synthesised by grafting decylamine (NDA) onto the PLP backbone to enhance the hydrophobicity.[11] To further manipulate the amphiphilicity of pseudopeptides, aminododecanoic acid (ADA) has been grafted onto the carboxylic acid groups on the PLP backbone to produce PLP-ADA polymers.

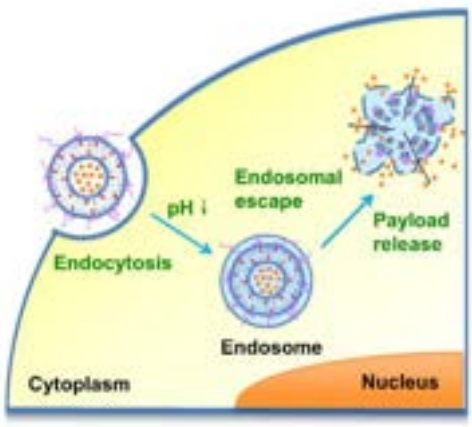


Figure 1. Schematic of the virus-mimicking, endosomolytic liposome and its endosomal escape[3]

This work investigated the effectiveness of using PLP-ADA-coated liposomes for drug delivery. The system was characterised by particle size measurements, polymer coating efficiency and drug encapsulation efficiency. The system was also optimised for its pH-triggered drug release and further tested for intracellular delivery through cell study.

2. Background

In 1965, Bangham et al. published the first description of liposomes. The concept of liposomes was then developed by Gregoriadis.[12] Over the last five decades, liposomal systems have been greatly developed and transferred from laboratory work to commercial use. Many drawbacks were successfully resolved to improve the effectiveness of the system. For example, cholesterol was incorporated into the formulation to produce more rigid lipid bilayers to enhance the stability,[13] thus encapsulating drugs inside liposomes for a longer time. To counter the problem of rapid clearance by cells of the mononuclear phagocyte system, stealth liposomes were developed to increase the circulation half-life. Polyethylene glycol liposomes subsequently demonstrated the ability to treat Kaposi's sarcoma in HIV patients in 1994.[14] The concept of enhanced permeability and retention effect was also introduced by then and became important design guidelines for liposome systems.[15]

Apart from the problems mentioned above, researchers are also concerned about the central topic of effective intracellular drug delivery. Membrane-active liposomes gain huge attention for their ability to escape endosomes through membrane fusion or other types of membrane disruption.[12] Triggered release of drugs is one of the successful solutions that has been extensively used for intracellular drug delivery.

The Chen Group have published some exciting work on virus-mimicking liposomal systems. Pseudopeptides with pH-responsiveness were synthesised and coated on liposomes to produce an effective drug delivery system. The first-generation polymer PP75 directly imitates a hydrophobic amino acid group from influenza viral peptide, and it coats on the liposomal surface through hydrophobic interactions. The second-generation polymer, PLP-NDA18, improves from PP75 by replacing the amino acid group with a long fatty acid side chain to further increase the hydrophobicity. Meanwhile, the side chain allows the polymer to coat on the liposomal surface with membrane insertion, creating more efficient polymer coating through stronger hydrophobic interaction between the aliphatic chain and lipid tails. The two generations of polymers both demonstrated the membrane destabilising ability upon acidification to break endosomes and release small-molecule drug (doxorubicin).[3, 11, 16]

A third-generation polymer, PLP-ADA, was recently produced by the group to further manipulate amphiphilicity. A longer side chain with a carboxylic acid group has been grafted on the backbone of PLP. The same number of carboxylic acid groups is maintained with the grafting, and it is believed to be able to increase the pH responsiveness of the polymer. While some work has been carried out on the function of PLP-ADA polymer within the group, no work has been done on the potential of liposomes coated with this polymer. Therefore, this study is motivated to learn the effectiveness of the virus-mimicking liposomes coated with the newly developed PLP-ADA polymer. We are highly interested in formulating the system to demonstrate pH-responsive drug delivery potential for small-molecule drugs as previously developed systems. Additionally, we want to discover the system's ability to deliver macromolecular drugs such as peptides and proteins.

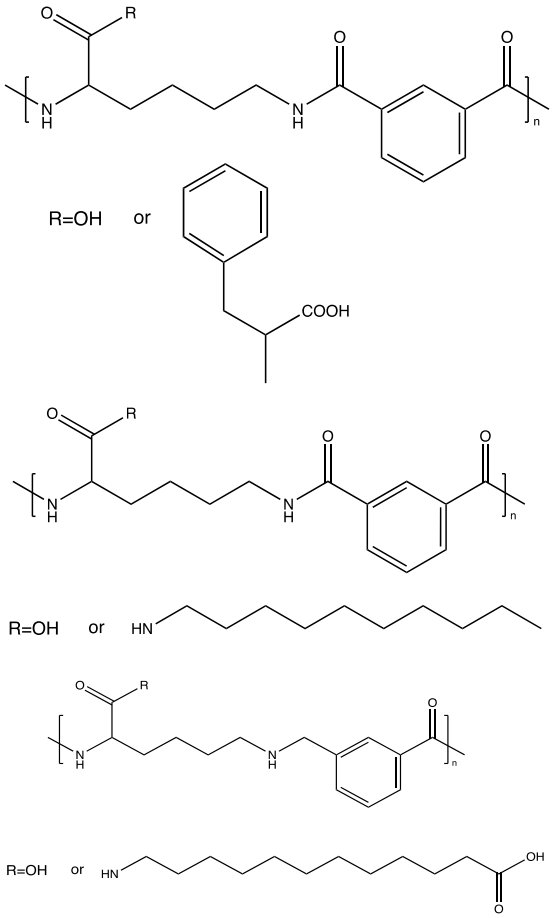


Figure 2. Structures of PP polymer (Top), PLP-NDA polymer (Middle), PLP-ADA polymer (Bottom)

3. Materials and Methods

3.1 Materials

L-α-Phosphatidylcholine from egg yolk (EPC), cholesterol, phosphate-buffered saline (PBS), Dulbecco's phosphate-buffered saline (D-PBS),

Calcein, fluorescein isothiocyanate-dextran (FITC-dextran, average Mw of 40 kDa), penicillin, Dulbecco's modified Eagle's medium (DMEM) and fetal bovine serum (FBS) were purchased from Sigma-Aldrich (Dorset, UK). Chloroform and ethanol absolute were purchased from VWR (Lutterworth, UK). 1,2-dioleoyl-sn-glycero-3-phosphoethanolamine (DOPE) was purchased from Avanti Polar Lipid Inc. (Birmingham, UK). Defibrinated sheep red blood cells (RBCs) were purchased from TCS Biosciences Ltd. (Buckingham, UK). Hoechst 3342 and LysoTracker Red DND-99 were purchased from Fisher Scientific (Loughborough, UK). Triton X-100 was purchased from Alfa Aesar (Heysham, UK). Sodium hydroxide, hydrochloric acid, pre-synthesised pseudopeptidic PLP polymer (poly(l-lysine iso-phthalamide), pre-synthesised pseudopeptidic PLP-ADA30/60 polymer (30%/60% stoichiometric degree of substitution with aminododecanoic acid on the PLP backbone) and pre-synthesised fluorescein pseudopeptidic PLP-ADA30 polymer (Cy5-PLP-ADA30).

3.2 Methods

3.2.1 Preparation of Liposomal Systems

Firstly, the lipid film hydration method was used to prepare bare liposomes.[17] Lipid and cholesterol with molar ratios of 60% and 40% were dissolved in 2 mL chloroform containing 1% (v/v) ethanol. A lipid film was formed by removing the organic solvent through 3 hours of rotary evaporation. Next, 5 mL of PBS buffer was added to the lipid film and hydrated in a 40°C water bath for at least 1 hour. The size of the bare liposomes was controlled to be around 100 nm by extrusion. The hydrated lipid solution was extruded through a 100 nm polycarbonate membrane 32 times to produce stable liposomes. To encapsulate Calcein and FITC-dextran into the liposomes, the model drug solution was added to the sample before extrusion. The virus-mimicking liposomes were then prepared by incubating bare liposomes with pseudopeptide solution at the desired concentration overnight at 4 °C. Excess pseudopeptides and model drugs were removed by dialysis (Float-A-Lyzer, MWCO 300 kDa, Spectrumlabs, USA) in the pH 7.4 PBS buffer for at least 3 hours. To ensure efficient dialysis, the buffer needs to be changed frequently. The samples were stored at 4°C.

3.2.2 Measurement of Particle Size

The hydrodynamic sizes of bare and PLP-ADA30-coated liposomes at pH 7.4 were measured by dynamic light scattering (DLS, Litesizer 500, Anton Paar, UK) at 25 °C. The measurement of each sample was repeated three times for an accurate result. In addition, the samples were diluted 50 times with PBS buffer before measurement to yield a suitable counting rate for the measuring equipment.

3.2.3 Measurement of Coating Efficiency

Unloaded virus-mimicking liposomes were prepared using Cy5-labeled pseudopeptidic PLP-ADA30 polymers according to the methods described in Section 3.2.1. The excess polymers were removed by dialysis in PBS buffer at pH 7.4. Spectrofluorometer (GloMax Explorer Multimode Microplate Reader, Promega, USA) was used to measure the fluorescence intensities of the samples with different initial polymer concentrations (1 mg/ml, 5 mg/ml, 10 mg/ml) before and after dialysis under excitation at 627 nm and emission at 660-720 nm. 100 µL of samples were pipetted into a black 96-well plate, and triplicate measurements are required to ensure accurate results. The Cy5-labeled pseudopeptidic concentrations were calculated through the calibration curve. The polymer coating efficiency was calculated by the equation:

$$\%coating = \frac{C_{pd}}{C_{p0}} \times 100\%$$

where C_{p0} is the pre-dialysed polymer concentration of the sample and C_{pd} post-dialysed polymer concentration.

3.2.4 Measurement of Encapsulation Efficiency

Calibration curves of fluorescence intensity over drug concentration were constructed for Calcein and FITC-dextran, respectively. In order to obtain the Calcein encapsulation efficiency, unloaded Calcein was removed by dialysis. The fluorescence intensities of the sample before and after the dialysis were measured by the spectrofluorometer (GloMax Explorer Multimode Microplate Reader, Promega, USA) under excitation at 490 nm and emission at 510-570 nm. Triton x-100 was used to lyse the liposomes before measurement. 100 µL of samples were pipetted into a black 96-well plate and repeated three times. The Calcein concentrations were calculated through the calibration curve. The encapsulation efficiency of Calcein was calculated using the following equation:

$$\%encapsulating = \frac{C_{dd}}{C_{d0}} \times 100\%$$

where C_{d0} is the pre-dialysed drug concentration of the sample and C_{dd} is the post-dialysed drug concentration.

3.2.5 pH-Dependent Drug Release

The drug release profiles of the virus-mimicking liposomes coating with different polymer coatings at pH 7.4 and pH 6.5 were investigated. The samples and PBS buffer were adjusted to the desired pHs using 2 M HCl and NaOH. 400 µL of the sample was added into a dialysis tube, which was placed in PBS buffer at room temperature. Then, 1 mL of PBS solution was withdrawn and replaced by 1 mL of fresh PBS solution at timepoints 0, 0.5, 1, 2 and 4 h.

The fluorescence intensities of the post-dialysed virus-mimicking liposomes and withdrawn buffers were measured by spectrofluorometer (GloMax Explorer Multimode Microplate Reader, Promega, USA). Triton X-100 was used to lyse the liposomes before measurement. The drug concentrations were calculated through the calibration curves. The percentage of the drug content released was calculated using the following equation:

$$\%release = \frac{C_b V_b}{C_b V_b + C_d V_d} \times 100\%$$

where C_b is the drug concentration of the withdrawn buffer at time t, V_b is the total volume of the dialysis PBS buffer, C_d is the drug concentration of the post-dialysed virus-mimicking liposomes, and V_d is the total volume of the dialysed liposomal solution.

For the Calcein drug release profile, 200 mL of PBS buffer was used to ensure efficient dialysis, and the fluorescence intensities were measured under excitation at 490 nm and emission at 510-570 nm. Since the fluorescent intensity of FITC-dextran is much lower than Calcein, 100 mL of PBS buffer was used for the FITC-dextran release profile, and the fluorescence intensities were measured under excitation at 475 nm and emission at 500-550 nm.

3.2.6 Haemolysis Assay

Haemolysis assay was used to examine the membrane disruptive activity and investigated whether the pseudopeptidic PLP-ADA30 polymer absorbed on the liposomal surface retained its pH-responsive endosomolytic activity.[16] 12 mL of defibrinated sheep red blood cells (RBCs) were added into 2 mL Eppendorf tubes and centrifuged at 3500 rpm for 3 minus (Eppendorf Centrifuge 5424, USA). The supernatant was removed and refilled with pH 7.4 PBS buffer to wash the RBCs. After 4 times washing, 170 µL of RBCs were resuspended into 300 µL of samples. All samples were adjusted to desired pHs using 2 M HCl and NaOH. Two controls were prepared by resuspending RBCs in deionised water for positive control and PBS buffer for negative control. The samples were incubated at 37 °C for 1 hour, then centrifuged at 3500 rpm for 3 minutes. The supernatant of each sample was taken and diluted 10 times with PBS buffer. 100 µL of samples were pipetted into a white 96-well plate and repeated three times. The absorbances were measured by spectrophotometer (GloMax Explorer Multimode Microplate Reader, Promega, USA) at 560 nm. The percentages of relative haemolysis were calculated by the equation:

$$\%haemolysis = \frac{A_e - A_n}{A_p} \times 100\%$$

where A_e is the experimental absorbance, A_n is the negative control absorbance, and A_p is the positive control absorbance.

3.2.7 Cell Culture

HeLa adherent epithelial cells (human cervical cells) were grown in DMEM supplemented with 10% (v/v) FBS and 100 U/mL penicillin unless specified otherwise. Trypsin-EDTA was used to trypsinise the HeLa cells. The cells were maintained in a humidified incubator with 5% CO_2 at 37 °C.[3]

3.2.8 Laser Scanning Confocal Microscopy

In order to access the endosomal escape ability of virus-mimicking liposomes coated with pseudopeptidic PLP-ADA30 polymer, confocal microscopy was performed with tracer molecules, Calcein and FITC-dextran.[3] 2 mL of HeLa cells (1×10^5 cells/mL) were seeded in a glass bottom plate (35 mm, MatTek, USA) and cultured overnight. The cells were then treated with 2 mL of serum-free DMEM containing 0.22 µm filter-sterilised virus-mimicking liposomes, bare liposomes, or free drugs. The Calcein and FITC-dextran concentration of the cells was kept at 40 µM and 4.5 µM, respectively. After incubation at 37 °C for 1 hour, the cells were washed with D-PBS buffer three times. LysoTracker red DND-99 (50 nM) and Hoechst 33342 (1 µg/mL) were added to stain the endosomes/lysosomes and nuclei, respectively. The cells were further incubated for 1 hour and rinsed with D-PBS. The laser scanning confocal microscopy (Zeiss LSM-510 inverted laser scanning confocal microscope, Germany) was used for imaging. Calcein was excited at 490 nm, and the emission at 510-570 nm was collected. FITC-dextran was excited at 470 nm, and the emission at 500-550 nm was collected.

4. Results and Discussions

4.1 Liposome Characterisation

Both EPC and DOPE lipids were used for the preparation of liposomes. However, while the EPC lipid film could easily dissolve in PBS buffer during hydration, the DOPE lipid had very low solubility in PBS buffer. Therefore, it caused difficulties in controlling the quality of liposomes as the amount of DOPE liposomes produced could be different for each batch. Therefore, only EPC lipids were used for all experiments afterwards.

4.1.1 Particle Size Control

The size of liposomes should be cautiously controlled under 200 nm to enhance the residence time in blood and improve the performance of in vivo drug delivery into tumour cells.[18] Figure 3 shows that all liposomes were produced within the desired size range. Bare liposomes had a size of 156.7 ± 2.5 nm, while PLP-ADA30 coated liposomes had a size of 173.9 ± 0.5 nm. The increase in size confirmed that polymers were successfully coated on the liposomes. Furthermore, both bare liposomes and PLP-ADA30 coated liposomes exhibited a decrease in the mean hydrodynamic

diameters when Calcein was encapsulated. The smaller size of the liposomal system can be attributed to the hydrophilicity of Calcein, which led to an interaction between the drug and the hydrophilic head of the lipids. With PLP-ADA30 coated on the liposome surface, the membrane anchoring, comb-like pseudopeptides further strengthened the hydrophilic interaction with the drug due to the presence of the carboxylic acid group on the side chain, presenting a larger decrease in particle size.

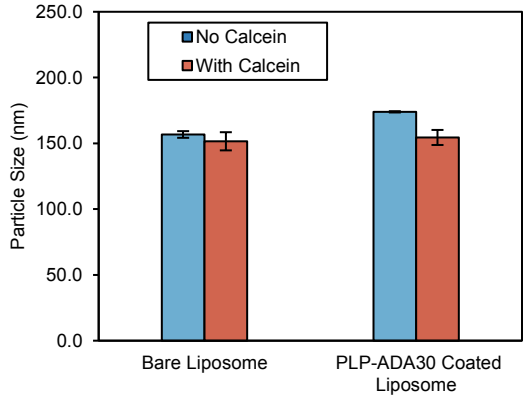


Figure 3. Particle sizes of bare and polymer-coated liposomes with and without Calcein encapsulated

4.1.2 Polymer Coating Efficiency

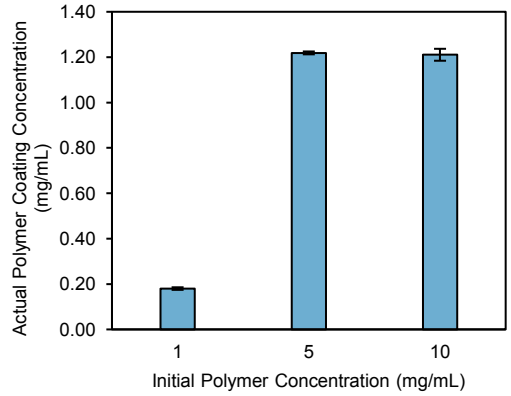


Figure 4. Actual polymer coating concentration of PLP-ADA30 using different initial polymer concentrations

Polymer coating efficiency is essential for the liposomal system as the amount of polymer coated on the liposomes has a critical impact on the system's effectiveness. Therefore, the maximum polymer coating concentration should be quantified to examine the performance of liposomal surface modification. Tests were performed with initial polymer concentrations ranging from 1 mg/mL to 10 mg/mL. This range was selected as both 1 mg/mL and 10 mg/mL could produce satisfactory pH-sensitive liposomes when PLP-NDA18 was used for coating in previous work.[16] As shown in Figure 4, the polymer coating concentration increases as the initial polymer concentration increases. Once the initial polymer concentration reaches 5 mg/mL, the maximum coating concentration is achieved at 1.22 ± 0.01 mg/mL with the highest coating efficiency of 24.4 ± 0.1%. Continuing to increase the initial polymer concentration does not help improve liposomal surface modification, as the actual polymer coating concentration remains at the same value.

4.1.3 Encapsulation Efficiency

Encapsulation efficiency is another essential factor in characterising the liposomal system. Higher encapsulation efficiency is always desired as it will allow more drugs to be kept in the liposomes to be delivered to targeted sites.

5 mM of Calcein stock solution was used to conduct the test, and the liposomes were prepared with an initial Calcein concentration of 0.625 mM. Bare liposomes showed an encapsulation efficiency of 26.3 ± 0.4%, proving the effectiveness of the passive loading method.

Further study was then performed to learn the effect of liposome surface modification on the encapsulation efficiency using different initial polymer coating concentrations. It was found that once PLP-ADA30 was coated on the liposomes, the encapsulation efficiency would largely decrease to around 6%, as shown in Figure 5. As the polymers have a membrane anchoring effect, the insertion of polymer side chains into the lipid bilayers would cause a disturbance in the stability of the liposomes. In addition, as Calcein is a small molecule drug, it can easily escape the entrapment of the liposomes during the coating process, thus resulting in lower encapsulation efficiencies for polymer-coated liposomes. Moreover, Figure 5 also shows that the initial concentration does not have a significant impact on the encapsulation efficiency in the range of 0.5 mg/mL to 5 mg/mL. However, as the initial polymer concentration increases to 10 mg/mL, another decrease in encapsulation efficiency is observed. This is caused by the large amount of free polymers present in the solution, creating membrane instability to leak out the Calcein that is encapsulated before having the surface modification.

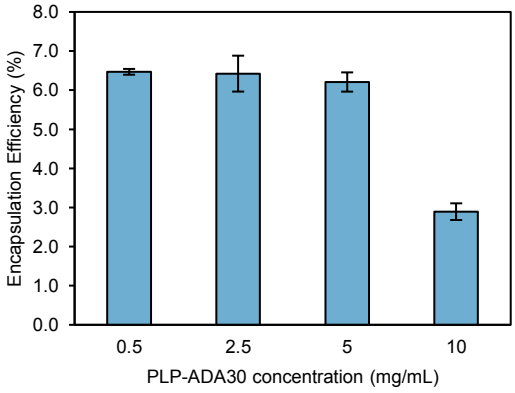


Figure 5. Encapsulation efficiency of Calcein using different initial polymer concentrations

4.2 Drug Release Study

4.2.1 Effect of Different Polymers

For liposomes to be effective for in vivo drug delivery, it is essential to produce liposomal systems that are stable in physiological pH. In this way, they can deliver the desired dosage of the drug into targeted cells instead of losing drugs during blood circulation. Therefore, it is important to first examine the leakage profile of polymer-coated liposomes at pH 7.4 to evaluate the function of the polymer.

PLP, PLP-ADA30 and PLP-ADA60 polymers were used for polymer coating at an initial concentration of 1 mg/mL to test the performance of liposomal systems. From Figure 6, PLP-ADA30 presents the most stable system, giving the lowest leakage percentage throughout 4 hours. Conversely, liposomes coated with PLP polymers were the least stable and had over 40% leakage just after 1 hour. The results were expected because PLP polymers do not have long aliphatic chains grafted on the polymer backbone to strengthen the polymer-membrane interaction. Therefore, both PLP-ADA-coated liposomes exhibited higher stability than PLP-coated liposomes.

Another interesting finding is that as the degree of grafting increases, the stability of the system does not increase. It is postulated that the increasing amount of carboxylic acid group on the side chain brings an unfavourable effect on the hydrophobic interaction between the long aliphatic chain and lipid tails, causing instability of the system and a higher leakage after a shorter period.

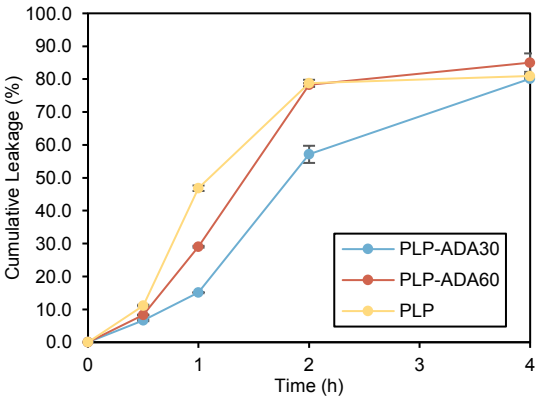


Figure 6. Leakage profile of liposomes coated with different polymers

4.2.2 Effect of Polymer Concentrations

Since PLP-ADA30 showed the best performance in previous experiments, it was used to further investigate the effect of polymer concentration. Four different initial concentrations from 0.5 mg/mL to 5 mg/mL were tested based on the results of polymer coating efficiency tests.

From Figure 7, an obvious disparity of performance can be seen that using an initial concentration of 5 mg/mL gives the most stable liposomal system. There was almost no leakage after 2 hours, and an acceptable leakage of 20.87 ± 0.04% was observed after 4 hours. In the meantime, the other 3 concentrations had leakages over 75% after 4 hours. 5 mg/mL allows the liposomes to reach the maximum polymer coating concentration of 1.22 ± 0.01 mg/mL as mentioned in Section 4.1.2, thus constructing more rigid vesicles to hold drugs inside stably.

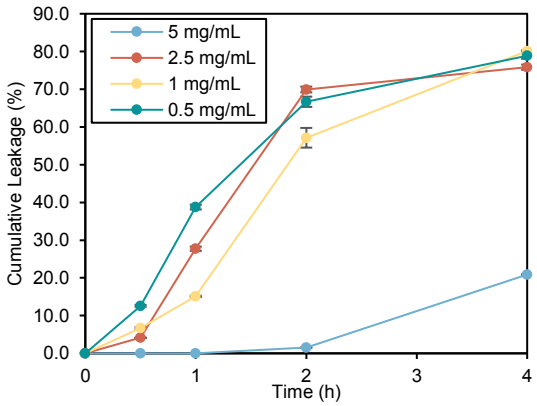


Figure 7. Leakage profiles of liposomes coated with PLP-ADA30 using different initial concentrations

4.2.3 pH-triggered Drug Release

The drug release potential of the modified liposomal system was tested using the same 4 different initial polymer concentrations at pH 6.5 to mimic the early endosome condition. As shown in Figure 8, all modified liposomal systems were able to release the drug content, showing that PLP-ADA30 polymers are effective in creating membrane destabilisation. However, the coating concentration of polymers did not indicate a clear relationship with the degree of drug release due to two contradictory effects. On the one hand, with increasing polymers coated on the liposomes, the pH-triggered polymer activity becomes stronger to release more drugs. However, on the other hand, more polymers inserted in the lipid bilayers also increase the rigidity of the system. Therefore, further research is still required to get a clearer picture of the complex relationship.

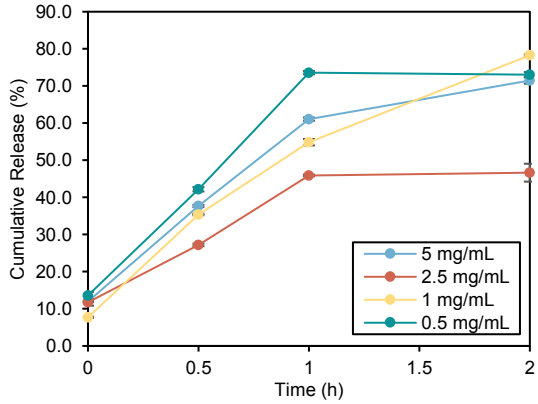


Figure 8. Release profiles of liposomes coated with PLP-ADA30 using different initial concentrations

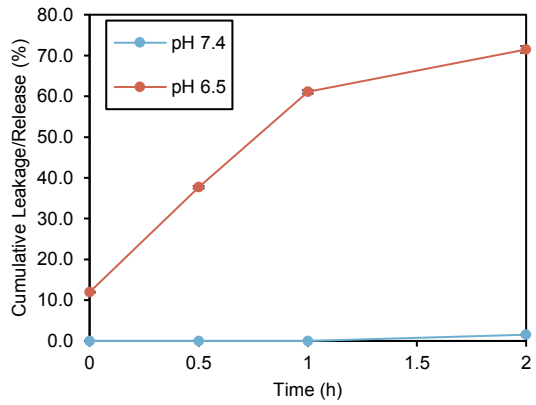


Figure 9. Leakage/Release profile of liposomes coated with maximum coating concentration of PLP-ADA30 and encapsulated with Calcein

Figure 9 shows a clear comparison of the behaviours of the liposomal system treated with maximum coating under different pH conditions. No Calcein was leaked at pH 7.4, while 61.1 ± 0.4% of Calcein was released at pH 6.5 during the first hour. This result is comparable to a previous study published by the research group[11] and successfully demonstrates the effectiveness of PLP-ADA30 polymer. While the long side chain could increase the packing density of the vesicles through stronger hydrophobic interactions with cholesterol and lipid tails at neutral pH,[11] the protonation of the carboxylic acid group during the decrease in pH weakens the interaction and thus increases the permeability of the membrane to release drug outside.

With the successful demonstration of releasing small-molecule drugs using the model payload Calcein, the virus-mimicking liposomal system was further examined for its ability to deliver macromolecular drugs. FITC-dextran, with a molecular weight of 40 kDa, was used as a model payload for proteins. Figure 10 shows that the newly modified liposomal system with the maximum coating of PLP-ADA30 is also effective for pH-triggered drug release. A quick release of 39.2 ± 1.4% was observed at pH 6.5, while the leakage was only 3.2 ± 0.4% after one hour.

It is also noted that the liposomal system encapsulated with FITC-dextran is less effective than the liposomal system encapsulated with Calcein. As FITC-dextran has a much larger size than Calcein, it is harder for the liposomal system to hold the model drug for a longer time. Meanwhile, it is also more difficult for FITC-dextran to pass through the lipid bilayers to be released at pH 6.5.

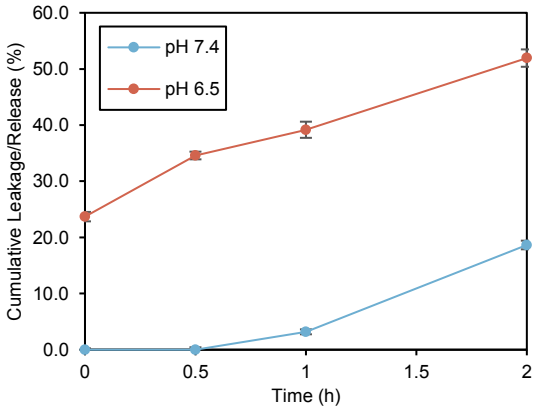


Figure 10. Leakage/Release profile of liposomes coated with maximum coating concentration of PLP-ADA30 and encapsulated with FITC-dextran

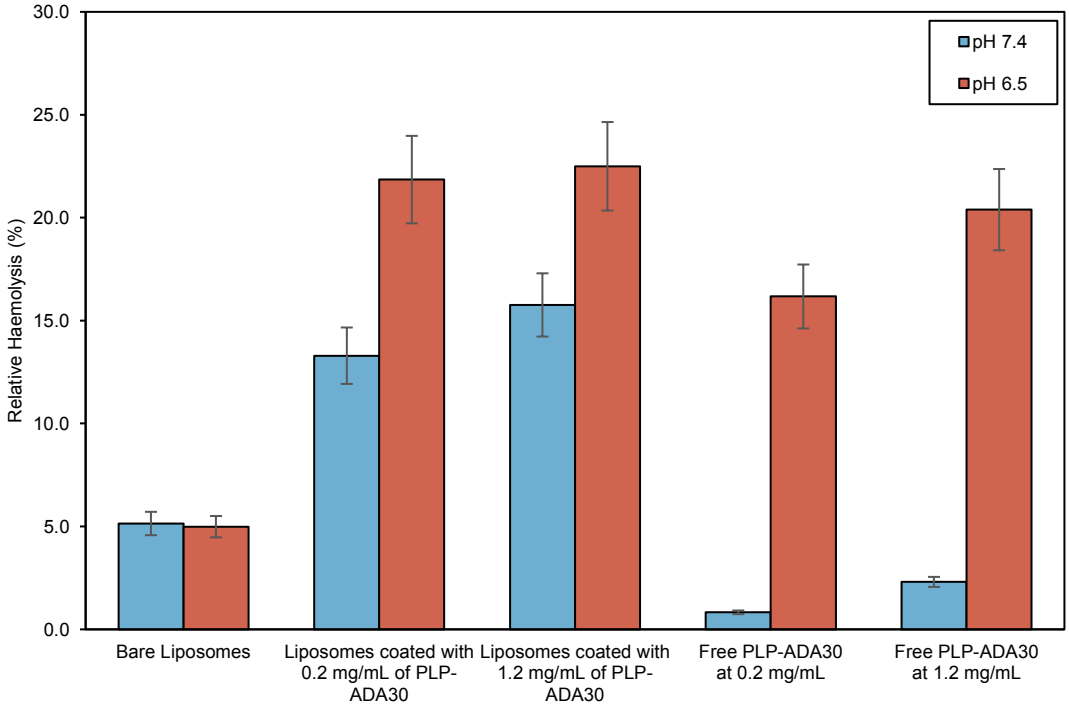


Figure 11. Relative haemolysis assay of liposomes coated with different concentrations of PLP-ADA30 and different concentrations of free PLP-ADA30

4.3 Haemolysis Assay

For liposomes to effectively deliver drugs to the targeted sites, the liposomes must possess the ability to break out endosomes to avoid being degraded.[19] Haemolysis assay was performed using red blood cells as model endosomes to test the membrane activity of the modified liposomal system. As seen in Figure 11, both PLP-ADA30 coated liposomes, and free PLP-ADA30 polymers showed higher membrane disruptive activity upon acidification. Free polymers had almost no membrane activity at physiological pH but a significant increase at pH 6.5, indicating the high effectiveness of PLP-ADA30. As the liposomal system showed similar results at both polymer coating concentrations at pH 6.5, it further proves the capability of the polymer. A low degree of the polymer coating is enough to satisfy the haemolytic requirement. It is also worth noting that the virus-mimicking liposomes exhibited higher endosomolytic activity at both polymer concentrations than free polymer, showing that the liposomal system can be more useful for drug delivery.

4.4 Intracellular Drug Delivery

Upon successfully demonstrating endosome escape ability through haemolysis, the modified liposomal system was further tested for its ability to release drugs into the cell cytoplasm. Calcein and FITC-dextran were used as model payloads separately for the investigation. From Figure 12, PLP-ADA30 coated liposomes showed significantly higher cellular uptake by HeLa cells compared to bare liposomes. Diffused green fluorescence can be seen across the entire cell that was treated with modified liposomes. In contrast, green punctate spots colocalised with red lysosomes for HeLa cells treated with bare liposomes. It can be seen from the merge images that the green and red spot overlap to show many yellow spots, suggesting that bare liposomes are still trapped in the endosomes at much lower concentrations compared to the coated liposomes. HeLa cells treated with free Calcein also showed similar results with low green fluorescence intensity. Free Calcein entered the cells through diffusion, and it can be observed that the cellular uptake was much lower as compared to PLP-ADA30-coated liposomes loaded with Calcein. From Figure 13, it can be seen that the cellular uptake of PLP-ADA30-coated liposomes loaded with 40 kDa FITC-dextran was lower than Calcein, but the green fluorescence still diffused throughout the cell. The images obtained are thus aligned with drug release profiles obtained previously. Since liposomes loaded with Calcein and FITC-dextran exhibited similar behaviours, the pH-responsive virus-mimicking liposomal system again demonstrated its ability to efficiently deliver both small-molecule and macromolecular drugs into cells.

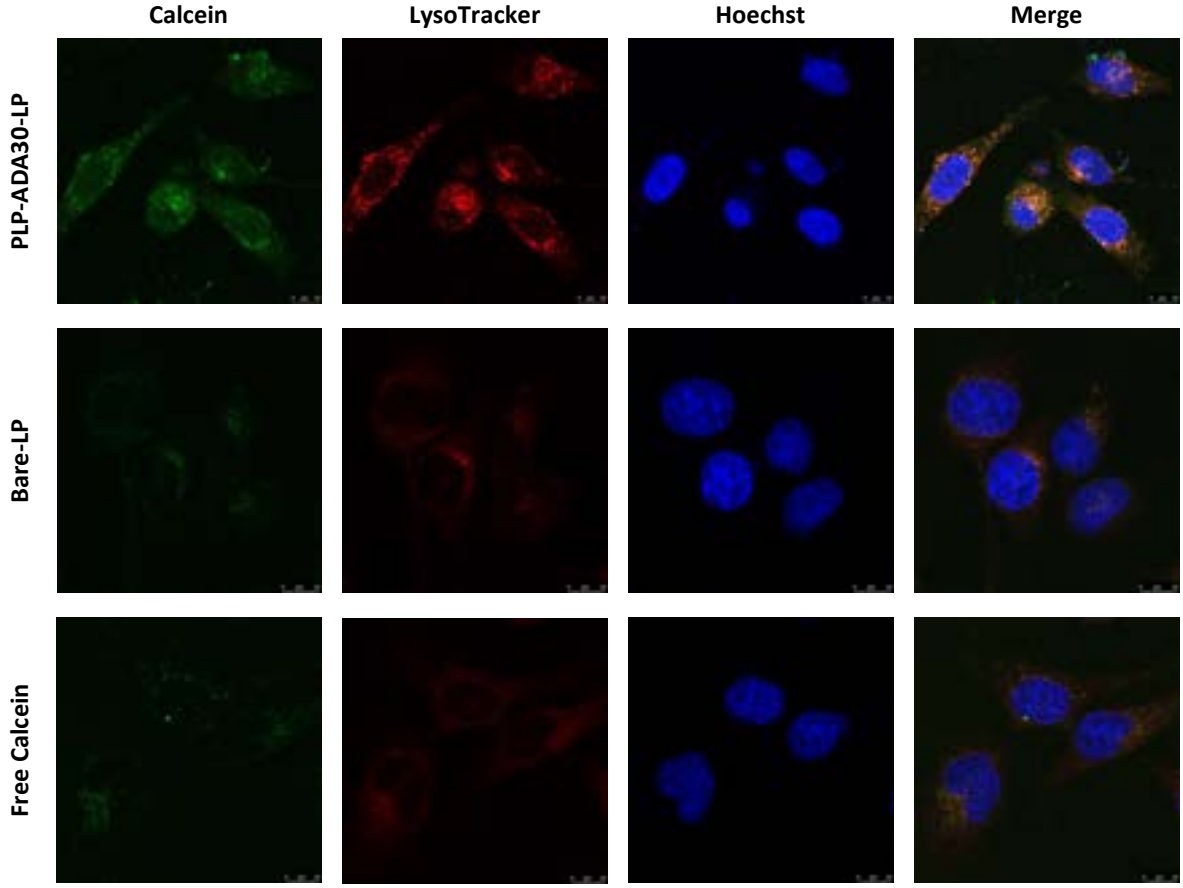


Figure 12. Confocal microscopy images of HeLa cells showing the escape of liposomes encapsulated with Calcein (scale bar 10 μm)

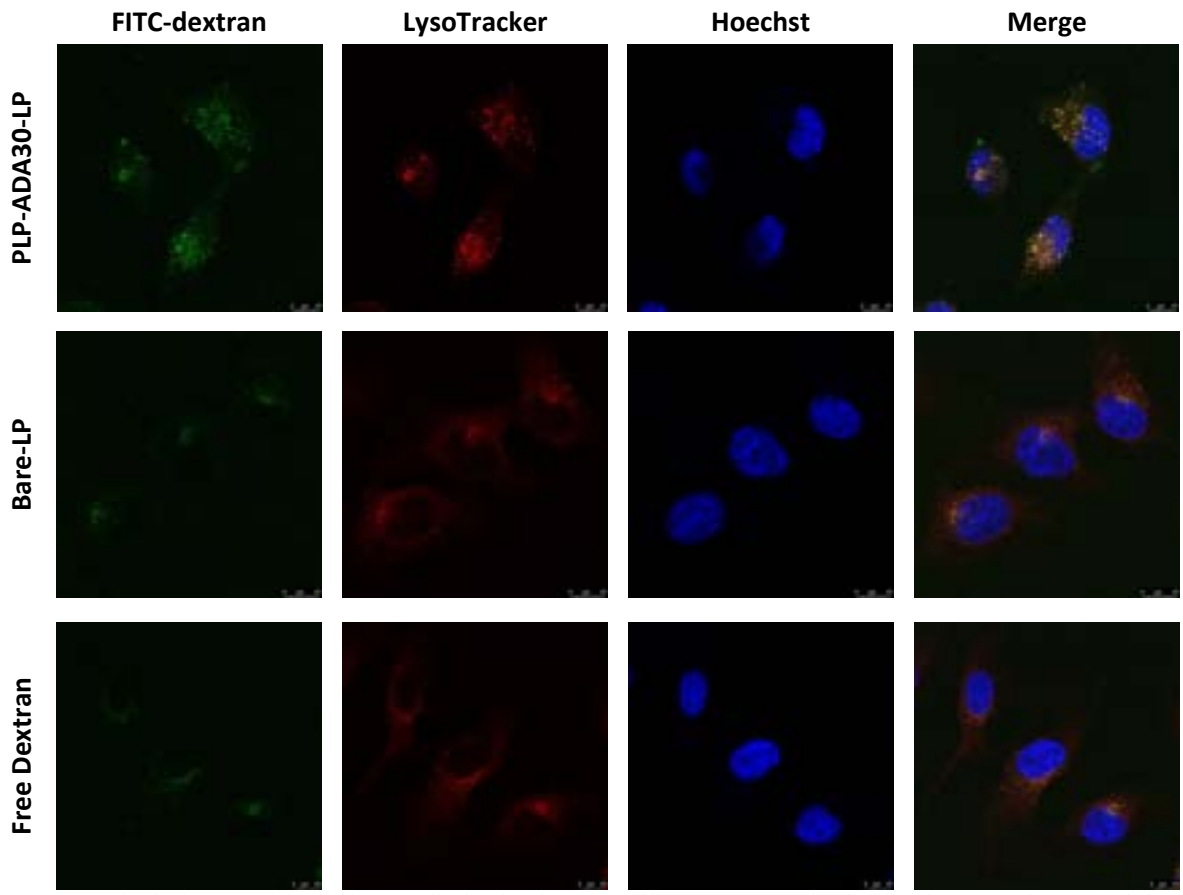


Figure 13. Confocal microscopy images of HeLa cells showing the escape of liposomes encapsulated with 40 kDa FITC-dextran (scale bar 10 μm)

5. Conclusions

A virus-mimicking, pH-responsive liposomal system was successfully developed through this project. The newly developed polymer was able to coat on the surface of liposomes to produce liposomes with desirable particle sizes and enhanced functionalities. The maximum coating concentration was quantified, and the encapsulation efficiency of Calcein was characterised. Using the optimal initial polymer coating concentration of 5 mg/mL, a stable liposomal system was obtained and had negligible leakage after 2 hours. Meanwhile, the effective pH-triggered release was observed for both Calcein and 40 kDa FITC-dextran. The endosomal escape ability was further confirmed by the haemolysis assay as PLP-ADA30 coated liposomes showed higher membrane activity at pH 6.5 than at physiological pH. Confocal microscopy images of HeLa cells demonstrated higher cellular uptake and effective intracellular delivery of the modified liposomal system in a more visible way and reinstated the successful liposome formulation. To examine the full potential of the virus-mimicking liposomal system, future work should be carried out by loading functional payloads into the liposomes. Doxorubicin is a great representative of small-molecule drugs and has been widely used for metastatic and early breast cancer therapy.[20] Saporin, a ribosome-inactivating protein with anticancer effect,[21] can be used as a macromolecular drug to test the liposomal system. Successful drug delivery of the two functional payloads will make the modified system one step closer to the clinical use.

Acknowledgement

We would like to express our deepest appreciation to our research supervisor, Professor Rongjun Chen, for his kind guidance and encouragement throughout the project. Furthermore, we are grateful to our Postdoctoral tutor Dr Apanpreet Kaur, for her essential advice and help. We cannot achieve the promising results without her dedicated support. We also want to thank Miss Jiawei Cui for providing PLP-ADA30 polymers for our research. Finally, we appreciate the countless help from our seniors: Mr Xinyu Lu, Miss Jiawei Cui, Mr Yifan Ding and Miss Yifan Liu.

Reference

1. Shi, J., et al., *Cancer nanomedicine: progress, challenges and opportunities.* Nat Rev Cancer, 2017. **17**(1): p. 20-37.
2. Bulcha, J.T., et al., *Viral vector platforms within the gene therapy landscape.* Signal Transduction and Targeted Therapy, 2021. **6**(1): p. 53.
3. Chen, S. and R. Chen, *A Virus-Mimicking, Endosomolytic Liposomal System for Efficient, pH-Triggered Intracellular Drug Delivery.* ACS Applied Materials & Interfaces, 2016. **8**(34): p. 22457-22467.
4. Patra, J.K., et al., *Nano based drug delivery systems: recent developments and future prospects.* Journal of Nanobiotechnology, 2018. **16**(1): p. 71.
5. Barenholz, Y., *Doxil® — The first FDA-approved nano-drug: Lessons learned.* Journal of Controlled Release, 2012. **160**(2): p. 117-134.
6. Somiya, M. and S.i. Kuroda, *Development of a virus-mimicking nanocarrier for drug delivery systems: The bio-nanocapsule.* Advanced Drug Delivery Reviews, 2015. **95**: p. 77-89.
7. Mi, P., *Stimuli-responsive nanocarriers for drug delivery, tumor imaging, therapy and theranostics.* Theranostics, 2020. **10**(10): p. 4557-4588.
8. Eccleston, M.E., et al., *pH-responsive pseudo-peptides for cell membrane disruption.* J Control Release, 2000. **69**(2): p. 297-307.
9. Chen, R., et al., *The role of hydrophobic amino acid grafts in the enhancement of membrane-disruptive activity of pH-responsive pseudo-peptides.* Biomaterials, 2009. **30**(10): p. 1954-61.
10. Chen, R., et al., *Synthesis and pH-responsive properties of pseudo-peptides containing hydrophobic amino acid grafts.* Journal of Materials Chemistry, 2009. **19**(24): p. 4217-4224.
11. Chen, S., et al., *Membrane-Anchoring, Comb-Like Pseudopeptides for Efficient, pH-Mediated Membrane Destabilization and Intracellular Delivery.* ACS Applied Materials & Interfaces, 2017. **9**(9): p. 8021-8029.
12. Allen, T.M. and P.R. Cullis, *Liposomal drug delivery systems: From concept to clinical applications.* Advanced Drug Delivery Reviews, 2013. **65**(1): p. 36-48.
13. Cullis, P.R., *Lateral diffusion rates of phosphatidylcholine in vesicle membranes: Eeffects of cholesterol and hydrocarbon phase transitions.* FEBS Letters, 1976. **70**(1): p. 223-228.
14. James, N.D., et al., *Liposomal doxorubicin (Doxil): An effective new treatment for Kaposi's sarcoma in AIDS.* Clinical Oncology, 1994. **6**(5): p. 294-296.
15. Wu, J., *The Enhanced Permeability and Retention (EPR) Effect: The Significance of the Concept and Methods to Enhance Its Application.* J Pers Med, 2021. **11**(8).
16. Chen, S., et al., *A pH-responsive, endosomolytic liposome functionalized with membrane-anchoring, comb-like pseudopeptides for enhanced intracellular delivery and cancer treatment.* Biomaterials Science, 2022. **10**(23): p. 6718-6730.
17. Guo, F., et al., *Smart IR780 Theranostic Nanocarrier for Tumor-Specific Therapy: Hyperthermia-Mediated Bubble-Generating and Folate-Targeted Liposomes.* ACS Applied Materials & Interfaces, 2015. **7**(37): p. 20556-20567.
18. Andra, V., et al., *A Comprehensive Review on Novel Liposomal Methodologies, Commercial Formulations, Clinical Trials and Patents.* Bionanoscience, 2022. **12**(1): p. 274-291.
19. Canton, I. and G. Battaglia, *Endocytosis at the nanoscale.* Chem Soc Rev, 2012. **41**(7): p. 2718-39.
20. Lao, J., et al., *Liposomal Doxorubicin in the treatment of breast cancer patients: a review.* J Drug Deliv, 2013. **2013**: p. 456409.
21. Zhang, G.N., et al., *Lipid-Saporin Nanoparticles for the Intracellular Delivery of Cytotoxic Protein to Overcome ABC Transporter-Mediated Multidrug Resistance In Vitro and In Vivo.* Cancers (Basel), 2020. **12**(2).

The Recovery of 5-Hydroxymethylfurfural from Conventional Fructose Dehydration Solvents

Hana Khatib and Bethany Lam
Department of Chemical Engineering, Imperial College London, UK

Abstract

5-Hydroxymethylfurfaral (HMF) is a highly versatile chemical that is produced via the dehydration of fructose. Herein, this study investigates the unexplored isolation step of HMF, by quantifying the recoveries from perturbating solvent systems in batch and continuous processes. The solvent systems explored were low boiling solvents: acetone/water and methanol systems and high boiling solvents: dimethyl sulfoxide (DMSO) and 1,4-dioxane/DMSO systems. For the batch process, rotavaping experiments were conducted for glucose, HMF and acidified HMF solutions. Since crude dehydration reaction effluents are usually acidic, small amounts of sulfuric acid were added to mimic this. For the acidified HMF isolation experiments, the low boiling solvents, acetone/water and methanol, presented high recoveries of 85.76 and 86.66% respectively, as lower operating temperatures were employed, decreasing the tendency of HMF to react and/or degrade. For the high boiling solvents, lower recoveries of 23.31 and 26.87% were obtained for DMSO and 1,4-dioxane/DMSO respectively. This was due to the formation of by-products such as carbonaceous materials called humins, which were observed as dark insoluble particulates. For the continuous process, flash distillation was modelled on Aspen Plus, which recurrently presented higher recoveries for the low boiling solvents. To further understand the behaviour of HMF in solvent systems, a time and heat degradation experiment was devised. For the heat degradation, 1,4-dioxane/DMSO and DMSO demonstrated 40 and 10% degradation of HMF, whilst the low boiling solvents exhibited negligible degradation. This further justifies the use of low boiling solvents for the dehydration of fructose into HMF, providing easier downstream separation and reduced thermal degradation of the valuable product.

Keywords: *5-Hydroxymethylfurfaral (HMF), Recovery, Solvent Removal, Degradation*

1. Introduction

The use of fossil fuels for energy and process feedstock is unique to the chemicals industry where it accounts for 14 and 8% of the total oil and gas demand respectively.[1] Manufacturing, utilisation, and disposal of these fossil-based chemicals all release carbon emissions, estimated annually to be 1.5 gigatonnes of carbon dioxide ($GtCO_2$), contributing to 18% of all industrial CO_2 emissions.[1] Therefore a key part in creating a more sustainable chemicals industry is bio-based chemicals, produced from biomass. Current bio-based chemical and polymer production are estimated to be at 90 million tonnes,[2] with a market of USD 73.16 billion in 2020 and projected to grow to USD 144.63 billion by 2028.[3]

Sugars are a promising feedstock to produce useful chemical building blocks. Food crops such as sugar cane, are classified as first-generation sugars. These have been heavily criticised for their competition with food security, however a comprehensive sustainability assessment from the Nova institute,[4] concluded that they are as valuable as second-generation sugars. These include wood, residual woodchips and cellulosic crops.[5] Both generations provide a strategy to aid in the reduction of carbon emissions,[4] therefore sugars are a pivotal part of research for the bio-chemicals industry to invest in.

In particular, the dehydration of sugars can produce a highly valued, versatile platform chemical known as 5-hydroxymethylfurfural (HMF). The Lignocellulosic Biorefinery Network (LBNet) has identified HMF as one of the UKBioChem10,[6] the top ten green chemicals the UK should develop and commercialise to reduce its dependence on non-renewable feedstocks. The most notable application for HMF is through complete oxidation to produce, 2,5-furandicarboxylic acid (FDCA), which could replace terephthalic acid in the production of polyesters.[6] Further HMF derivatives include hydration into levulinic acid for environmentally friendly herbicides,[6] and hydrodeoxygenation into 2,5-dimethylfuran (DMF) for high-energy biofuel.[7] Utilising HMF appears to be promising in different key markets, therefore,

developing an effective process to produce it is paramount for this extremely functional chemical.

Extensive research on the dehydration of fructose focuses upon the protocol optimisation of reaction conditions such as catalysts, and solvents for maximum HMF yield and selectivity. Whilst research on the downstream processing of the HMF effluents are limited. Advancements in using ionic liquids have been made due to their high dissolution properties and ability to stabilise HMF. Suppressed rehydration to levulinic acid in the presence of water provides higher yields and selectivity towards HMF. However, their extremely low vapour pressures present difficulty in using solvent evaporation.[8]

More conventional solvents including, dimethyl sulfoxide (DMSO) have been extensively used and are reported to give high HMF selectivity due to their ability to minimise by-products by binding to HMF more strongly than water.[9] Initial studies from 1982, showed full conversion of fructose to HMF over 16 hours at 100°C, [10] and more recently, similar results of 100% conversion and 97% HMF selectivity were reported.[11] Despite this, HMF's high affinity in DMSO and its high boiling point (b.p.189°C) suggests HMF separation would be difficult. Also, there are concerns for the environmental toxicity of DMSO for commercial use. [12]

Therefore, the need for successful reaction systems in lower boiling solvents becomes more important when considering the separation process. The high solubility of fructose in 1,4-dioxane (b.p.101°C) led Aellig and Hermans[13] to obtain only 20% HMF yield. Therefore, this resulted in the need to add small amounts of DMSO for a 92% yield, which nullifies using 1,4-dioxane. Furthermore, Van Putten proposed methanol (b.p.64.7°C), obtaining a conversion of 89%, which corresponds to a combined yield for HMF and methoxymethyl furfural (MMF) of 47%.[14] Among the limited research, Dumesic *et al*, proposed an acetone/water system, achieving 97% HMF selectivity and 98% fructose conversion. HMF was then successfully extracted using methyl isobutyl ketone (MIBK) and isolated by vacuum evaporation.[15]

The assurance of these low-boiling systems, producing high HMF yields, now presents the challenge of ensuring the HMF isolation is feasible. HMF is notoriously unstable when exposed to high temperatures with a tendency to react and lead to substantial carbonisation of the product.[16] Therefore, thermal recovery methods such as distillation may potentially cause degradation of this valuable product. This paper aims to address the quantification of recoveries from different solvent systems by investigating the perturbations of separation conditions for effective solvent removal, with the focus on acetone/water (80/20vol%), methanol, and 1,4-dioxane/DMSO (90/10vol%). These have been chosen due to their promising high fructose conversion to HMF and low boiling properties for easier HMF isolation and purification, in addition to using DMSO as a comparative solvent.

2. Methods

2.1 Chemicals

D-Glucose anhydrous (Analytical reagent), methanol (Analytical reagent), sulfuric acid (98%) and acetone (Technical) were purchased from VWR. HMF (≥99%, FG), DMSO (≥99.5%), 1,4-dioxane (99.8%), and D-Sorbitol (≥98%) were all purchased from Sigma-Aldrich. Deionised water (DI) from Veolia Purelab Chorus was used in all experiments.

2.2 Experimental Setup

10 mL solutions of glucose and HMF (0.61, 1.00, 2.04, and 0.44 wt.%) were prepared in acetone/water (80/20 vol.%), methanol, DMSO and 1,4-dioxane/DMSO (90/10 vol.%) respectively as the concentrations were calculated from literature. To more closely mimic reaction effluents which often contain acidic catalysts, 0.192 mL of 1M sulfuric acid solution was used to make up an acid concentration of 16 mM for 10 mL of acidified HMF solutions at the same HMF concentrations. Solvent removal was performed with Buchi Rotavapor R-100 under reduced pressures, (30-500 mbar) using the vacuum pump V-100 in a 250/500 mL round bottomed flask. The substrate was dissolved with 20 mL of deionised (DI) water. Water/oil baths were used to provide a range of operating temperatures, (60-135°C) depending on the solvent system shown in Table 3 (Section 3.2). The solution in the flask was fully immersed in the water/oil bath during evaporation.

Figure 1: *Setup of rotavapor. (A) Water bath used for methanol and acetone/water system with a 500mL rotary flask. (B) Modified setup for silicone oil bath, heated using a hotplate with a magnetic stirrer at 500rpm to reach higher temperatures used for 1,4-dioxane and DMSO systems with an adapter and 250ml rotary flask.*

2.3 Quantification of Recovery

High Performance Liquid Chromatography (HPLC) was used to quantify glucose and HMF in the feed, distillate, and the solutions that contained the redissolved substrate post rotavaping. The instrument was calibrated by preparing and processing different concentrations of glucose (0.11-1.91wt%) and HMF (0.13-1.10wt%). Table 1 lists the details of the respective HPLC instruments used. For the evaporative light scattering detector (ELSD), an external standard of sorbitol was added to adjust for signal variability.

Table 1: *Operating Conditions of HPLC Instruments*

	HMF Detection	Glucose Detection
HPLC instrument	Agilent- 1220 Infinity II- Ultraviolet (UV) detector	Agilent 1260 Infinity II- ELSD
Signal	Variable Wavelength Detector (VWD)	Refractive Index Detector (RID) ELSD
Mobile Phase	0.4mL of 85wt% phosphoric acid (H_3PO_4) in 2L of water	DI Water
Mobile Phase Flowrate	0.45mL/min	0.75mL/min
Column Temperature	65°C	85°C
Volume of Sample Injected	0.5µL	20µL

The recoveries calculation is shown in equation 1.

$$Recovery\ (\%) = \frac{\left(\frac{C_{stock}}{\rho_{solution}}\right) m_{solution}}{\left(\frac{C_{rotavape}}{\rho_{water}}\right) m_{water}} \qquad Eq.\ 1$$

Where C_{stock} and $C_{rotavape}$ are the concentrations (g/L) obtained from HPLC for the sample before and after rotavaping. $m_{solution}$ is the mass (g) of solution rotavaped and m_{water} is the mass of water used to dissolve the substrate. $\rho_{solution}$ and ρ_{water} are the respective densities (g/L) of the solution and water.

2.4 Time Degradation

HMF solutions (10 mL) in the solvent systems were prepared at the concentrations stated in Table 2. Half the sample was kept at room temperature (18-20°C) and the other in the fridge (10-12°C), except for DMSO which was placed in an oil bath at 40°C. Experiments were conducted for up to 26 days.

2.5 Aspen Plus

The property method used was non-random two-liquid (NRTL) as Martcotullio[17] states that for dilute furfurals (<10wt%) at low pressures (<1bar), experimental data matches the closest to NRTL in comparison to other models. The set-up comprised of a mixer to homogenise a HMF and solvent stream which then fed into a flash vessel to simulate the solvent evaporation process. HMF recoveries, purities and solvent removal were quantified.

3. Results & Discussion

3.1 Substrate Concentration in Solvent Systems

Considering that HMF is relatively expensive (~£13/g),[18] a cheaper alternative glucose was used to allow for an initial understanding of solvent behaviour at experimental distillation conditions. The HMF concentrations used for the solvent systems are stated in Table 2 which were calculated from published fructose reaction data. The same concentrations were applied for the glucose experiments, for a direct comparison between substrates.

Table 2: *Composition of Solvent Systems*

Solvent System	wt% of substrate in solution
Acetone/water (80/20vol%)	0.61[15]
Methanol*	1.00, (2.35)[14]
DMSO	2.04[19]
1,4-dioxane/DMSO* (90/10vol%)	0.44, (3.25)[13]

*Glucose was insoluble in the 2nd literature values provided. 1st value stated was used (upper solubility limit)

Glucose was found to be insoluble in the methanol and 1,4-dioxane systems at 2.35 and 3.25wt% respectively. Consequently, the upper solubilities concentrations of 1.00 and 0.44wt% were used. These were obtained by filtering the insoluble glucose to attain a filtrate of the saturated solution. This upper solubility concentration is then attained by processing the filtrate through HPLC.

3.2 Determining Operating Temperatures and Pressures

A range of distillation conditions of temperatures (60-135°C) and pressures (30-500mbar) were evaluated for the distillation of solvent systems. The conditions outlined in Table 3 were found to yield effective solvent removal for glucose.

Table 3: *Optimal temperature and pressure conditions*

Solvent System	Temperature (°C)	Pressure (mbar)
Acetone/water	75	200
Methanol	60	500
DMSO	135	30
1,4-dioxane/DMSO	135	30

As expected, the low boiling solvents required much milder temperatures than the systems containing the high boiling solvent, DMSO. Despite 1,4-dioxane being a lower boiling solvent, due to the presence of DMSO, the same temperature of 135°C was required. It was reported that the small addition of DMSO stabilises the HMF, which in dehydration reactions increases the yield by up to 75%.[13] Increasing concentrations of DMSO in the 1,4-dioxane system also increases the solubility of fructose and reaction rate. This correlation however plateaus, thus a 90/10% volume of 1,4-dioxane/DMSO composition was recommended by Aellig and Hermans.[13] To achieve this high temperature, the experimental set up was modified with a hot plate and and oil bath as shown in Figure 1(B), in substitute to the nominal water bath from the Buchi system. These separation conditions were then used accordingly for the HMF and acidified HMF experiments.

3.3 Rotavaping Experiments

3.3.1 Glucose Recoveries

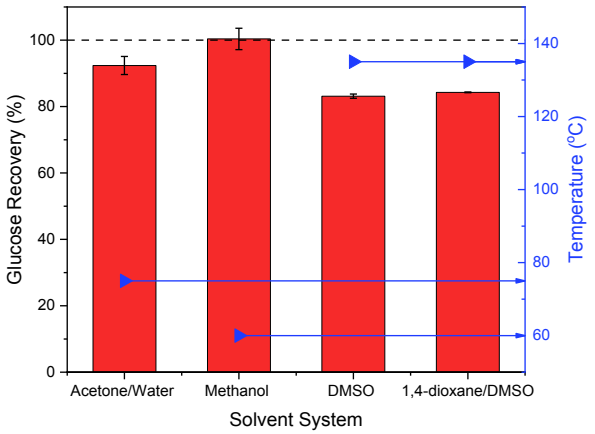


Figure 2: *Recoveries of glucose from solvent systems at different operating temperatures (60-135 °C)*

Encouraging glucose recoveries were obtained for the low boiling solvents (acetone/water and methanol), over 90%, as shown in Figure 2. The discrepancy in recoveries between acetone/water and methanol could be due to the nominal presence of water forming hydrogen bonds with glucose.[20] Thus, small amounts of HMF are vaporised with the water, reducing the recoveries for acetone/water system. Water is used as a co-solvent for acetone, in lieu of a pure acetone system. Dumesic *et al* demonstrated that whilst increased acetone concentrations increase the dehydration rate constant, the solubility of fructose decreases. Hence small amounts of water were added. It was reported that 80/20vol% of acetone/water is capable of dissolving fructose whilst maintaining high reaction rates. [15]

A reduction in recoveries were observed for the higher boiling solvent systems (DMSO and 1,4-dioxane/DMSO) that required temperatures as high as 135 °C, at which Woo *et al* reported glucose degrades to furfaral, HMF, formic acid and lactic acid.[21] This is evident on the chromatogram in Figure 3, when the glucose/DMSO solution was rotavaped for a prolonged period (30 minutes) the recovery obtained was 23%. Shortening the time to 15 minutes, a higher recovery of 83% was obtained, and no by-products were formed as shown in Figure 3. The duration of 30 minutes was initially conducted as prior to this, there was small drop of solution that would not vaporise . However, it was found that at 15 minutes the size of the drop was approximately the same and there was no evidence of humin formation. The correlation between rotavaping time and recovery is plotted on Figure 4. It must be noted that in between the low and high boiling solvent experiments, the signal of the HPLC changed from ELSD to RID. The ELSD detects less than the RID. Thus, the degradation products of glucose for the acetone/water system were perhaps not seen.

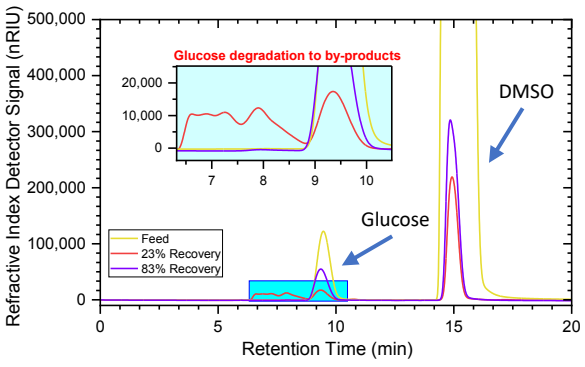


Figure 3: *Comparison of products formed when glucose DMSO solution is rotavaped for 15(83% recovery) and 30(23% recovery) minutes*

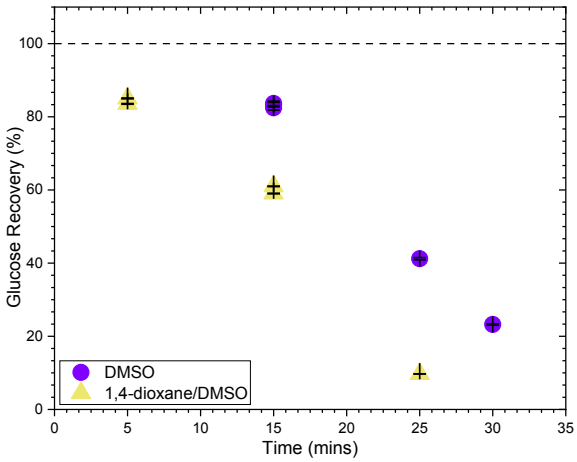


Figure 4: *Recoveries of glucose in respective DMSO and 1,4-dioxane/DMSO decreasing with increased rotavaping*

3.3.2 HMF Recoveries

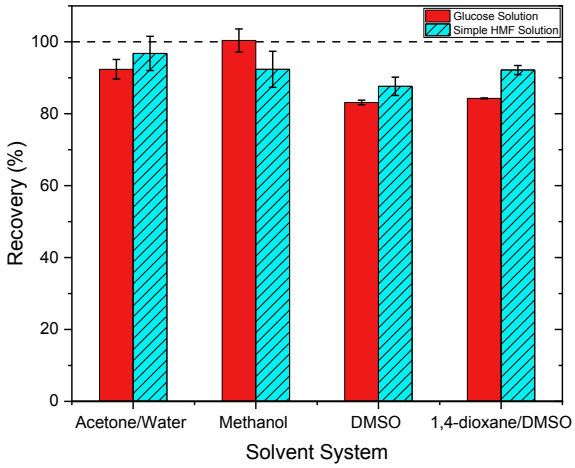


Figure 5: *Recoveries of HMF from solvent systems at different operating temperatures (60-135 °C)*

Shifting to the HMF substrate, as deduced in Figure 5, more consistent recoveries were obtained. The general trend shows that recoveries are very similar between the HMF and glucose experiments. For DMSO and 1,4-dioxane/DMSO system, recoveries obtained are higher than glucose, suggesting HMF is more thermally stable than glucose at higher temperatures. The overall solvent removal from different substrate systems is not dissimilar as the recoveries obtained were similar.

3.3.3 Acidified HMF Experiments

As of now, the simple HMF and solvent systems are an oversimplification of actual crude reaction effluents, as they contain unreacted substrates, by-products and homogeneous acidic catalysts that would affect the separation process.[22] In literature, acidic effluents were found to be 15[15]-17mM[14] therefore adding a layer of complexity, sulfuric acid was added to form 16mM of acidified HMF solutions. All recoveries were found to drop as depicted in Figure 6.

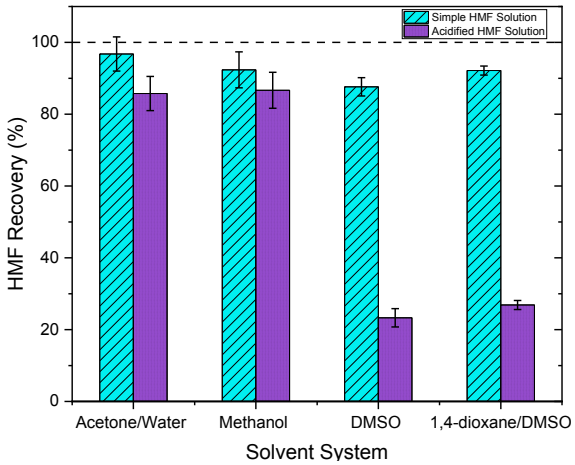


Figure 6: *Impact of recoveries on HMF for 16mM of acidified solutions using sulfuric acid.*

For the low boiling solvents, recoveries dropped by 6-12%. For the methanol system, a portion of the original mass of HMF in the solvent was methylated under the sulfuric acid catalyst, forming methoxymethylfurfural (MMF)[14] which is an HMF-ether. This formation is illustrated on the chromatogram in Figure 7. HMF-ethers are of interest, with valuable uses, such as cetane boosters in diesel blends.[23] In addition, they have found to be more stable than HMF.[23] Onwards, the complete etherification of HMF simultaneous to the isolation stage should be further explored.

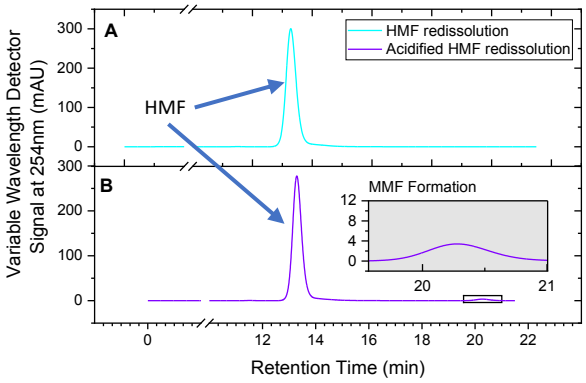


Figure 7: *Methanol solvent (A)Chromatogram of simple HMF recovery (B) Chromatogram of acidified methanol recovery signifying MMF production at ~ 20 minutes*

Conversely, for high boiling solvents, recoveries decreased by up to 70% due to the carbonisation of HMF at high temperatures forming humins.[16] This was clearly visible as brown solid particulates, shown in Figure 8. Humins are a class of carbonaceous materials with a largely unknown molecular structure.[24] They reduce the yield in the dehydration reaction and due to their heterogenous nature, can cause reactor fouling.[14] The dissolved solutions were filtered before being processed in the HPLC. To further identify the by-products formed, gas chromatography (GC) could be employed, as it

is typically used to detect and measure organic compounds.[25] To prevent the production of humins, a heterogenous acid catalyst could be used so that the solution is not acidified. Alternatively, neutralisation of the reaction effluent prior to separation would prevent the catalysed reaction of the humins and perhaps recoveries would be similar to the simple systems.

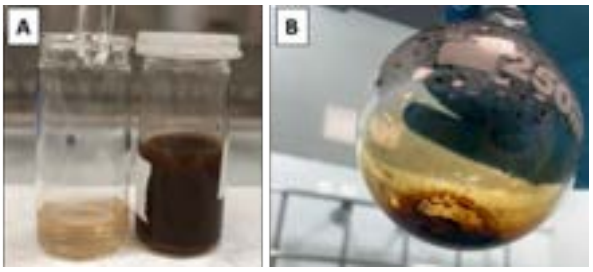


Figure 8: *(A) Left, filtered solution. Right, Solute remained after rotavaping dissolved in distilled water (20mL). (B) Insoluble products formed during rotavaping at 135°C for DMSO solution*

Considering the dehydration reaction of fructose and the ease of HMF isolation, both low boiling solvents prove to be suitable as they have high carbon balances. For the acetone/water system, the minimal formation of humins is reflected in Dumesic *et al*, as the yield (92%) and conversion (96%) are high.[15] Methanol's low boiling nature, high recoveries and production of valuable HMF-ethers makes it suitable. From an economic standpoint, the cost of methanol (~£14.70/L)[26] is higher than acetone/water (~£8.40/L)[27], which should be taken into consideration when selecting the most appropriate solvent.

3.4 Time Degradation Study of HMF

3.4.1 HMF Degradation at Heated Conditions

The versatility of HMF is compromised with its unstable nature, especially enhanced in the presence of an acid,[10] as summarised in the declining recoveries in section 3.3.3. The stability of HMF in the solvent systems is of interest when selecting the most suitable solvent. Therefore, degradation of HMF was investigated at the process temperatures used during the rotavaping experiments as stated in Table 3 over a 24-hour period.

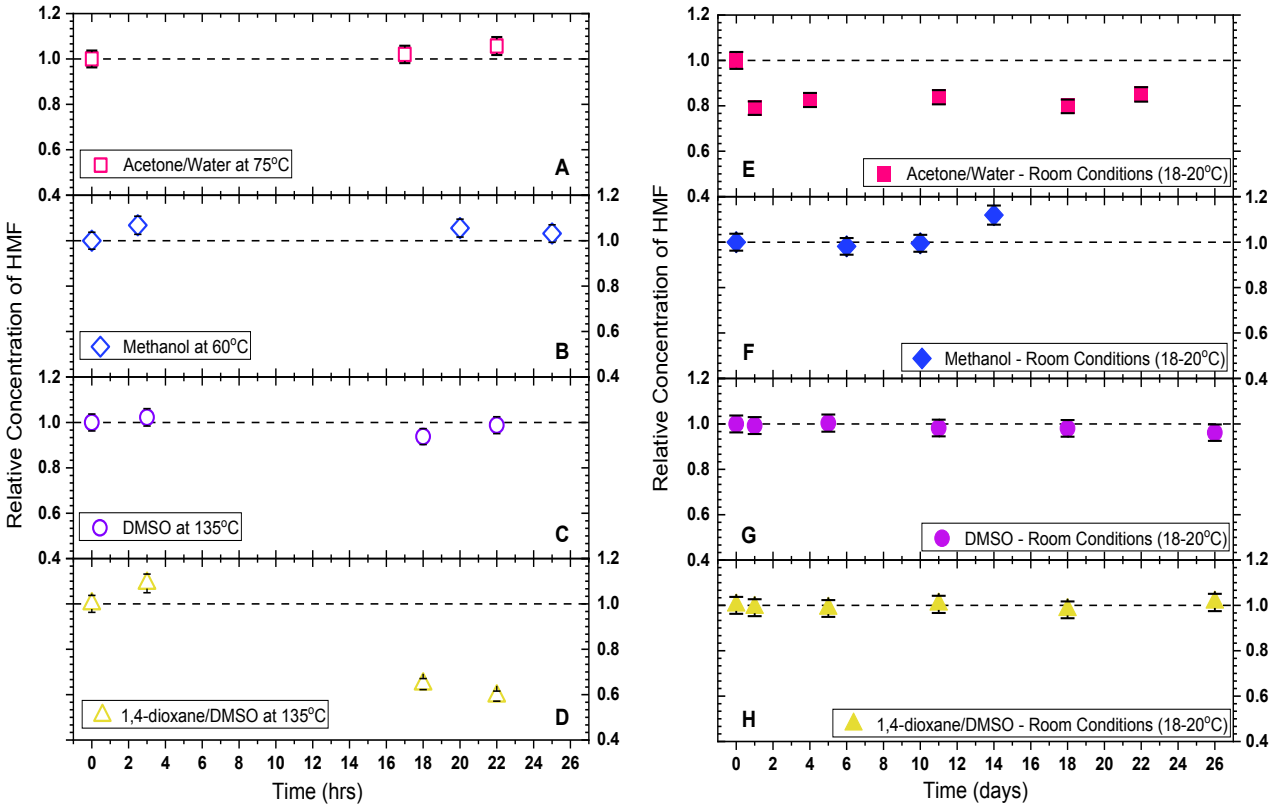


Figure 9: *Degradation of HMF shown by the relative concentrations to the original samples. (A-D): Heated Samples (E-H): Room Temperature Samples*

The fluctuations in the relative concentrations of HMF in the low-boiling solvents shows no degradation occurring, explaining the higher recoveries obtained in the experiments as displayed in Figure 9(A/B). Over a 24-hour period, it is shown in Figure 9(C) that less than 10% of the HMF in DMSO degraded despite being at 135°C. This owes to DMSO possessing a high affinity to bind to HMF, [9] making HMF stable in this solvent.

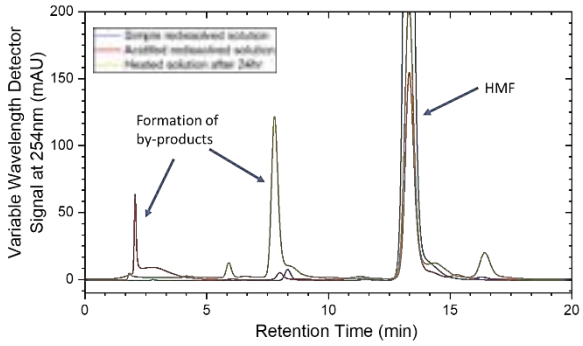


Figure 10: *Comparison of products formed for 1,4-dioxane/DMSO for simple HMF (blue), acidified HMF (red) and heated (yellow)*

On the other hand, for 1,4-dioxane/DMSO (135°C), there was a 40% reduction in HMF concentration, in lieu of products forming as shown in Figure 10. This demonstrates the poor stability of HMF, along with the low recoveries obtained in section 3.3.3. It can be observed that whether in an acidic environment or simply heated at the process condition for a prolonged period, HMF in 1,4-dioxane suffers degradation into different products, as shown in Figure 10. From a safety perspective, 1,4-dioxane is highly flammable, toxic and may cause cancer,[28] which further deters the use of 1,4-dioxane.

3.4.2 HMF Degradation at Storage Conditions

Devising a time degradation study on the effects of HMF under room temperatures of 18-20°C was of value to suggest a reliable solvent for HMF storage. Results are shown in Figure 9(E-H), in which up to a 26-day period, for all solvents investigated, there was minimal HMF degradation. Overall, with the exception of the acetone/water system, variations in HMF concentrations were no greater than 5% of the initial value. Fridge temperature conditions (10-12°C) were also investigated, and similar variations were obtained (Figure S1). For acetone/water, a very steep initial drop-off in concentration was observed after one day.

3.5 Simulation of Flash Distillation of HMF

3.5.1 Comparing Experimental and Simulated Recoveries

Aspen Plus was utilised to compare the experimental batch results to an industrial continuous process by flash distillation. Modelling the process using the same temperatures, pressure, and HMF feed concentrations (At a constant basis flowrate of 100kg/hr) for each solvent system. The simulated recoveries were found to be lower for all systems except for methanol as seen in Table 4. In the case for DMSO and 1,4-dioxane/DMSO, the solution completely vaporises and therefore no product remains in the bottoms, resulting in 0% recovery. HMF is modelled as a liquid, so it vaporises with the solvent into the distillate, hence simulated recoveries are lower than the experimental values. This is in dispute with the HMF being collected as a solid experimentally.

Table 4: *Comparison of experimental and simulated recoveries at same process conditions*

Solvent System	Experimental Recovery (%)	Simulation Recovery (%)
Acetone/water	96.78	78.65
Methanol	92.35	99.16
DMSO	87.64	0
1,4-dioxane/DMSO	92.17	0

3.5.2 Varying Flash Operating Temperatures

To further investigate the difference between the experimental and simulated results, the operating temperature for each solvent system was varied, at their respective set pressures. Solvent removal is the amount of solvent vaporised and this increases with temperature, which results in higher purities of HMF obtained in the product bottoms stream as illustrated in Figure 11. Recoveries decline with temperature, as increasing amounts of HMF are lost to the distillate as it vaporises. To obtain the same experimental recoveries, the methanol system requires a higher temperature of 80°C. On the contrary, lower temperatures are required, for acetone/water (58°C), DMSO (92°C), and 1,4-dioxane/DMSO (58°C) systems.

The experimental temperatures are based on what is measured in the water/oil outside the rotary flask, hence there is uncertainty in actual internal conditions of the solutions. Additionally, there is a temperature gradient between the heating medium, rotary flask, and solution, therefore the point at which the solution vaporises is likely to be lower.

This could possibly explain why the simulated results require a lower temperature for the same solvent removal. An additional probable cause in discrepancies could be based on the NRTL model estimations. It is favourable for predicting liquid activity coefficients. Nonetheless, this is not the case for vapour phase fugacity coefficients.[17]

3.5.3 Recommended Operating Temperatures

For optimal separation, the temperatures at which the recoveries and purities intersect in Figure 11 are recommended. This provides minimal HMF losses to the distillate, with a relatively high purity and solvent removal, as detailed in Table 5. In addition, the heat duties and CO_2 productions of the flash units are reported, with methanol owing to the highest energy use and CO_2 emission. This correlates to methanol having the highest recovery. This could be due the process costing more energy to achieve these higher recoveries.

The flash distillation process gives promising HMF recoveries and purities of above 90%, for the low boiling solvents, as they are also comparable to the experiments conducted. The recoveries for the DMSO and 1,4-dioxane/DMSO systems are minimal, with purities of 50 and 70% respectively, despite the lower heat duties and CO_2 productions. This discourages the use of these solvents; therefore, alternative separation techniques on Aspen should be explored, for instance the use of a crystalliser. This would favour the solid formation of HMF.

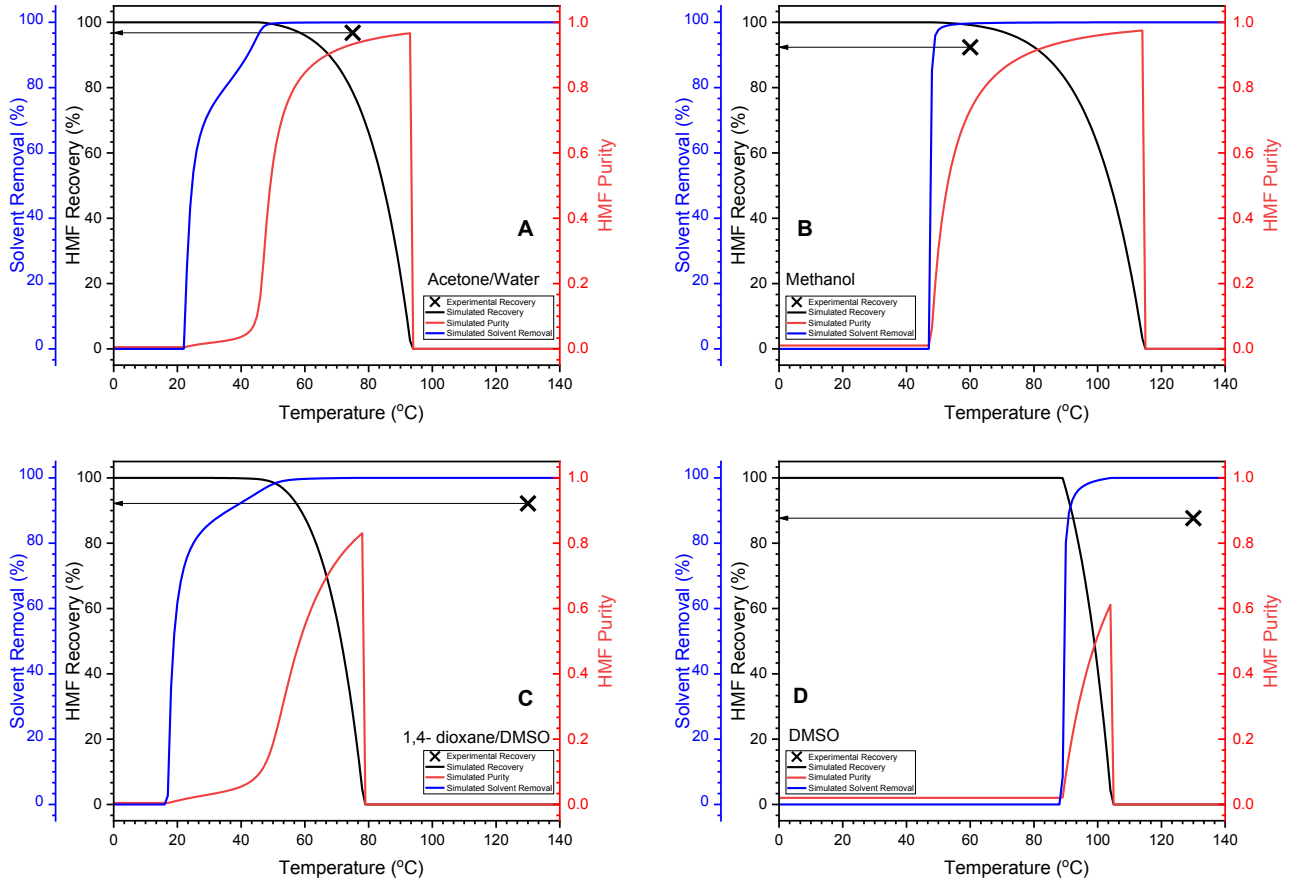


***Figure 11**: Simulated recoveries, purities, and solvent removal from Aspen for (A) Acetone/water(80/20vol%) (B) Methanol (C) 1,4-dioxane/DMSO (D) DMSO. Cross indicates the experimental temperature and recoveries.*

***Table 5:** Operating temperatures that give the optimal recovery, with high purities and solvent removal, with their respective heat duties. Constant feed flowrate of 100kg/hr. Pressures and HMF concentrations same as of experimental.*

Solvent System	Rotavaped Temperature (°C)	Optimal Flash Temperature (°C)	Optimal Recovery Obtained (%)	Purity (-)	Solvent Removal (%)	Heat Duty (kW)	CO_2 Production (kg/hr)
Acetone/water	75	67	90.56	0.90	99.95	31.41	10.90
Methanol	60	81	91.72	0.92	99.91	34.37	11.92
DMSO	135	99	49.17	0.49	98.94	20.93	7.26
1,4-dioxane/DMSO	135	67	69.15	0.70	99.87	14.07	4.88

4. Conclusions & Outlook

This study has demonstrated that the isolation of HMF is feasible for low boiling solvents (acetone/water and methanol) at both batch experimental and continuous industrial scales. HMF displays a high thermal stability in these low boiling solvents, in addition to high recoveries, even in acidic conditions. This is a fundamental consideration, as typical crude reaction effluents are acidic due to the catalysts used. The methanol system seems to be more reactive in acidic conditions than the acetone/water, as HMF-ethers are formed. Nevertheless, HMF was more stable in methanol than acetone/water when left at room temperature for a prolonged period. For a continuous process, the heat duty and CO_2 production for the low boiling solvents are similar. Nonetheless, the production of HMF-ethers can be taken advantage of, due to its increased stability to HMF.[23] To overcome the limited research on HMF-ethers, the isolation of HMF-ethers from reaction effluents could be investigated and compared to the analogous HMF isolations.

High boiling solvents are effective for the reaction process, providing high sugar solubility and HMF yields. However, in acidic conditions, high degradation occurs and humins are formed. Substitution of a heterogenous acidic catalyst would prevent this. Alternatively, the stream prior to separation could be neutralised. Although acidified conditions were employed, the actual investigation of these crude effluents of varying pH, unreacted sugars and by-products could elaborate on which solvent systems prove to be suitable. Among the DMSO and 1,4-dioxane/DMSO systems, DMSO proves to be more suitable for the isolation of HMF, as it is less susceptible to thermal degradation. This is in spite of high fructose solubility in 1,4-dioxane and HMF possessing high affinities to bind to DMSO. A time and heated degradation study of HMF containing sulfuric acid is additionally of interest, as it is proven that more by-products are formed in acidic conditions, thus this would develop the current understanding. For a continuous operation, other separation techniques that favour the solid formation of HMF could be adopted, such as crystallisation or centrifugation.

A limited scope out of a comprehensive list of successful fructose dehydration systems, including ionic liquids and other alcohols have been studied. Biphasic systems involving methyl isobutyl ketone (MIBK) and tetrahydrofuran (THF) have proven to extract HMF from the reaction effluent in situ to suppress side reactions such as rehydration to levunilic acid.[22] Therefore, investigation of the solvent evaporation and verification of the degradation chemistry of HMF in MIKB and THF would provide a further understanding of the importance of solvent extraction prior to thermal recovery.

Overall, the selection of the solvent system would not only depend on the key performance indicators of the fructose dehydration process but also the feasibility of separation, economic and environmental considerations. This provides a comprehensive assessment on which solvents would be suitable for the commercialisation of HMF production.

Acknowledgments

The authors would like to express sincere gratitude to the Hammond research group for warmly welcoming them in the labs. This especially applies to post-grad research supervisor, Laurie Overtoom, for giving guidance and support with the experimental procedure and project in general.

References

(1) International Energy Agency. The Future of Petrochemicals Towards More Sustainable Plastics and Fertilisers; 2018. www.iea.org (accessed 2022-12-04).

(2) de Jong, E.; Stichnothe, H.; Bell, G.; Henning Jørgensen, M.; de Bari, I.; Jacco van Haveren, E.; Lindorfer, J.; an der Johannes Kepler, E. Bio-Based Chemicals A 2020 Update Bio-Based Chemicals With Input from: (Pdf Version) Published by IEA Bioenergy; 2020.

(3) Fortune Business Insights. Bio-based Chemicals Market Size, Share. https://www.fortunebusinessinsights.com/bio-based-chemicals-market-106586 (accessed 2022-11-17).

(4) Dammer, L.; Carus, M.; Piotrowski, S. Sugar as Feedstock for the Chemical Industry What Is the Most Sustainable Option?; 2019.

(5) Fabbri, S.; Owsianiak, M.; Hauschild, M. Z. Evaluation of Sugar Feedstocks for Bio-based Chemicals: A Consequential, Regionalized Life Cycle Assessment. GCB Bioenergy 2022. https://doi.org/10.1111/gcbb.13009.

(6) LBNet. UKBioChem10 - The Ten Green Chemicals Which Can Create Growth, Jobs and Trade for the UK.

(7) Endot, N. A.; Junid, R.; Jamil, M. S. S. Insight into Biomass Upgrade: A Review on Hydro-genation of 5-Hydroxymethylfurfural (HMF) to 2,5-Dimethylfuran (DMF). Molecules 2021, 26 (22). https://doi.org/10.3390/MOLECULES26226848

(8) van Putten, R. J.; van der Waal, J. C.; de Jong, E.; Rasrendra, C. B.; Heeres, H. J.; de Vries, J. G. Hydroxymethylfurfural, a Versatile Platform Chemical Made from Renewable Resources. Chemical Reviews. March 13, 2013, pp 1499–1597. https://doi.org/10.1021/cr300182k.

(9) Portillo Perez, G.; Mukherjee, A.; Dumont, M. J. Insights into HMF Catalysis. Journal of Industrial and Engineering Chemistry. Korean Society of Industrial Engineering Chemistry February 25, 2019, pp 1–34. https://doi.org/10.1016/j.jiec.2018.10.002.

(10) Brown, D. W.; Floyd, A. J.; Kinsman, R. G.; Roshan-Ali, Y. DEHYDRATION REACTIONS OF FRUCTOSE IN NON-AQUEOUS MEDIA. Journal of chemical technology and biotechnol¬ogy 1982, 32 (10), 920–924. https://doi.org/10.1002/jctb.5030320730.

(11) Shimizu, K. ichi; Uozumi, R.; Satsuma, A. En-hanced Production of Hydroxymethylfurfural from Fructose with Solid Acid Catalysts by Sim¬ple Water Removal Methods. Catal Commun 2009, 10 (14), 1849–1853. https://doi.org/10.1016/j.catcom.2009.06.012.

(12) Technologies, L. DMSO 67-68-5 SDS Preview. Safety Data Sheets. https://chemi¬calsafety.com/sds1/sdsviewer.php?id=30161262&name=DMSO.

(13) Aellig, C.; Hermans, I. Continuous D-Fructose Dehydration to 5- Hydroxymethylfurfural Under Mild Conditions. ChemSusChem 2012, 5 (9), 1737–1742. https://doi.org/10.1002/cssc.201200279.

(14) van Putten, R. J.; van der Waal, J. C.; Harmse, M.; van de Bovenkamp, H. H.; de Jong, E.; Heeres, H. J. A Comparative Study on the Reac¬tivity of Various Ketohexoses to Furanics in Methanol. ChemSusChem 2016, 9 (14), 1827–1834. https://doi.org/10.1002/cssc.201600252.

(15) Motagamwala, A. H.; Huang, K.; Maravelias, C. T.; Dumesic, J. A. Solvent System for Effective Near-Term Production of Hydroxymethylfurfural (HMF) with Potential for Long-Term Process Improvement. Energy Environ Sci 2019, 12 (7), 2212–2222. https://doi.org/10.1039/c9ee00447e.

(16) Chheda, J. N.; Dumesic, J. a. Production of Hydroxymethylfurfural from Fructose. Science (1979) 2006, 312 (4), 1933.

(17) Marcotullio, Gianluca. The Chemistry and Technology of Furfural Production in Modern Lignocellulose-Feedstock Biorefineries.; [s.n.], 2011.

(18) Sigma-Aldrich. Sigma-Aldrich Product Specification.https://www.sigmaaldrich.com/GB/en/product/aldrich/w501808?gclid=CjwKCAiAhKycBhAQEiwAgf19ehqkKtJz9Y0qKeCgyfVpS9CGvGKXFmIS1U2717ygDEoYcIW5DBi4BoCQHMQAvD_BwE&gclsrc=aw.ds (accessed 2022-12-03).

(19) Shimizu, K. ichi; Uozumi, R.; Satsuma, A. Enhanced Production of Hydroxymethylfurfural from Fructose with Solid Acid Catalysts by Simple Water Removal Methods. Catal Commun 2009, 10 (14), 1849–1853. https://doi.org/10.1016/j.catcom.2009.06.012.

(20) Chen, C.; Li, W. Z.; Song, Y. C.; Weng, L. D.; Zhang, N. Formation of Water and Glucose Clusters by Hydrogen Bonds in Glucose Aqueous Solutions. Comput Theor Chem 2012, 984, 85–92. https://doi.org/10.1016/j.comptc.2012.01.013.

(21) Woo, K. S.; Kim, H. Y.; Hwang, I. G.; Lee, S. H.; Jeong, H. S. Characteristics of the Thermal Degradation of Glucose and Maltose Solutions. Prev Nutr Food Sci 2015, 20 (2), 102–109. https://doi.org/10.3746/pnf.2015.20.2.102.

(22) van Putten, R. J.; van der Waal, J. C.; de Jong, E.; Rasrendra, C. B.; Heeres, H. J.; de Vries, J. G. Hydroxymethylfurfural, a Versatile Platform Chemical Made from Renewable Resources. Chem Rev 2013, 113 (3), 1499–1597. https://doi.org/10.1021/cr300182k.

(23) Zaccheria, F.; Bossola, F.; Scotti, N.; Evange-listi, C.; Dal Santo, V.; Ravasio, N. On Demand Production of Ethers or Alcohols from Furfural and HMF by Selecting the Composition of a Zr/Si Catalyst. Catal Sci Technol 2020, 10 (22), 7502–7511. https://doi.org/10.1039/d0cy01427c.

(24) van Zandvoort, I.; Wang, Y.; Rasrendra, C. B.; van Eck, E. R. H.; Bruijnincx, P. C. A.; Heeres, H. J.; Weckhuysen, B. M. Formation, Molecular Structure, and Morphology of Humins in Biomass Conversion: Influence of Feedstock and Processing Conditions. ChemSusChem 2013, 6 (9),1745–1758. https://doi.org/10.1002/cssc.201300332.

(25) Bootman, V. Liquid Chromatographs vs Gas Chromatography.LCServices.https://www.lcservicesltd.co.uk/liquid-chroma¬tography-vs-gas-chromatography/#:~:text=GC is typically used to,like polymers%2C nucleotides and tetracyclines.

(26) Methanol ≥99.8%, AnalaR NORMAPUR® ACS, Reag. Ph. Eur. analytical reagen. VWR Chemicals.https://uk.vwr.com/store/product/734013/methanol-99-8-analar-normapur-acs-reag-ph-eur-analytical-reagent.

(27) Acetone ≥99%, TECHNICAL. VWR Chemicals. https://uk.vwr.com/store/product/721128/acetone-99-technical.

(28) Fisher Scientific. 1,4-dioxane 123-91-1. Safety DataSheets.https://chemicalsafety.com/sds1/sdsviewer.php?id=30450043&name=1%2C4-Dioxane.

Effectiveness of catchment tank for a Rainwater Harvesting System (RWHS) in Singapore

Churn Hym Lim and Sean Masci
Department of Chemical Engineering, Imperial College London, U.K.

Abstract
Decentralised rainwater harvesting system are increasingly implemented to reduce stormwater runoff and to fulfil non-potable usage of water onsite. This study investigates the effectiveness of the addition of a catchment tank of varying base areas on the system's ability to treat water onsite and reduce risks of flooding while accounting for economic considerations. Multi-objective optimisation using the utopian point method revealed that the additional optimal catchment tank did not improve the system's rainwater harvesting performance. It also highlighted the need to make full use of downstream systems which was currently significantly underutilised. A narrow but taller tank was more ideal due to the dynamics of the system which affected the flowrate of water passing through the system. The comparison between different rainfall events also pointed out how both total rainfall volume and duration were critical factors when considering the design of the rainwater harvesting system (RWHS). The methodology of assessing the effectiveness of the RWHS can be applied for evaluating other types of catchment tanks to suit different stakeholders and their priorities. A potential area of application for this study would be in designing new RWHS from scratch, enabling the use of smaller downstream tanks.
Keywords: Rainwater Harvesting, Multi-objective optimisation, Utopia Point

1 Introduction
Urban water supply systems are experiencing increasing stress due to population growth, increased urbanisation, and climate change. Rapid population growth and increased urbanisation have led to a surge in demand for water resources and an increase in the construction of buildings, roads, and other civil infrastructures. As a result, the reduction of rainwater absorption capacity from soil in these areas has exposed and made cities more vulnerable to flooding in the presence of extreme rain events. This is further exacerbated by global warming which has increased the frequency, intensity, and duration of rain events. Coupled with economic and political issues related to water, have caused significant increments in volume of wastewater and rainwater within cities, increasing the risk of combined sewer outflow and flooding events (García et al., 2015). In 2021, flooding cost the global economy $82bn, accounting for nearly a third of the total damages from natural disasters (Bevere, 2022). Therefore, sustainable and optimal urban water management has become a goal of strategic planning.

The main goal of these strategies is to maintain the quality and sustainability of water resources and to adapt these strategies for future development. This is particularly important for domestic water usage as it currently accounts for 10% of the total global water demand (Boretti & Rosa, 2019). As such, rainwater is starting to gain attention as an alternative water source due to its relatively low degree of pollution which does not require advanced purification processes (Słyś & Stec, 2020). This, however, would depend on many factors, including air quality, the type of catchment management, the type of roof coverage, etc. Most countries that use rainwater harvesting systems (RWHS) use it mainly as a complementary system to traditional water sources for non-potable use, including for toilet flushing, cleaning work and washing.

There has been much research surrounding the use of RWHS and the different models to mitigate the impact of stormwater events on these systems. Existing models includes the use of decentralised harvesting and detention systems to reduce the rainfall runoff volumes into centralised catchment areas (Soh et al., 2020). This also allows harvested water to be treated onsite to support demands within the building, thereby alleviating the stress on conventional water resources and potential contamination with pollutants carried by urban runoff. This reduces the amount of rainwater reaching the street level and hence reduced runoff volumes and peak flows in areas with RWHS implementation. However, the observed reduction in drainage peak flows attained by RWHS diminishes in long and intense rain events as harvesting tanks are filled in the early phases of the storm and remain filled as rainwater inflows usually exceed the demands from the tanks (Snir, Friedler & Ostfeld, 2022). As such, a full tank loses its drainage flow reduction ability and effectiveness as any subsequent rainwater inflow causes immediate overflow. While RWHS is effective in reducing the risks of stormwater flooding, it is limited to handling the known water problems at its time of design and more must be done to adapt and mitigate this in future with increasing rainfall expected.

The choice of a RWHS and its operations are also subjected to the same economical laws of profit and loss as other investments. Hence, both technical and economic analyses should be considered in the

decision-making process, such as the capital expenditure.

This paper follows up on the works regarding the use of decentralised rainwater detention tanks to reduce rainfall runoff volumes (Soh et al., 2020) and aims to further the development of RWHS by investigating the effect of adding an upstream catchment tank to ensure the system is sufficient in handling the rainfall event. Two key objectives were set: To better understand the dynamics of the RWHS and to develop a methodology that could be used in future designs and models.

Synthetic rainfall patterns were used to model and analyse the system. The model simulates the response of the drainage system during rainfall events and uses the results to derive the required parameters for optimisation to address some of the gaps in the field of an existing passive RWHS network through a real-world case study – a residential estate in Singapore. In this case study, the catchment system will model the DeepRoot Silva Cell, a hybrid system that incorporates both a modular suspended pavement system and soil to promote tree growth to treat water on-site (Figure 1).

Figure 1: Silva Cell schematic (taken from https://innovex.ca/en/products/silvacell/)

The remainder of this paper is structured as follows. The methodology used to perform the optimisation and evaluation of catchment tanks under different rainfall conditions is given in Section 2. An outline and discussion of the case study results are given in Section 3 while the conclusions and future outlook are provided in Section 4.

2 Methods

A suitable rainfall event was chosen to model the system to obtain an ideal catchment tank volume and to assess the effectiveness in meeting the desired key performance indexes (KPIs). However, due to potential trade-offs between the different KPIs, a multi-objective optimisation was performed to obtain a balance between these KPIs. To further evaluate the system's effectiveness, the optimised modelled system with the catchment tank was benchmarked against the system without the catchment tank (base case) and stress tested against rainfall events with larger rain volume.

2.1 Rainwater Harvesting System (RWHS)

2.1.1 Overview

The modelled system involves the addition of 2 tanks from the three-tank water harvesting and detention system (Soh et al., 2020) – a catchment tank and treatment tank. The modelled five-tank water harvesting and detention system will be segmented into 2 main parts – upstream and downstream system. The upstream component comprises of only the catchment tank, while the downstream system comprises of the remaining 4 tanks – separation, detention, harvesting and treatment tank (Figure 2). For the remainder of the paper, the term 'Rainwater Harvesting System (RWHS)' will refer to the system with the additional catchment tank while 'base case' will refer to the system without the catchment tank.

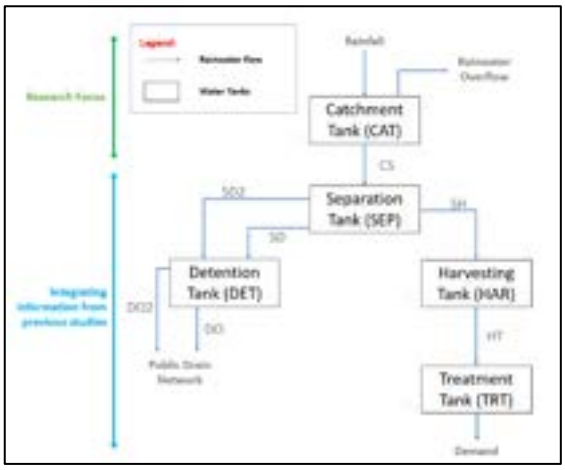


Figure 2: Schematic of the flows and tanks in the RWHS

Rainwater is first captured through a catchment system which will be the focus of the research where the volume of the catchment system will be varied and optimised to analyse the effect on the downstream rainwater harvesting system. The catchment tank also acts as a temporary storage tank for rainwater before it gets directed to the separation tank, the first tank in the downstream system, where it separates water through a separation filter into the detention and harvesting tank. The detention tank will hold discharge the rainwater to the public drainage network (PDN) at a suitable rate. On the other hand, filtered water from the separation tank will flow into the harvesting tank and thereafter be directed into the treatment tank for treatment to satisfy the demand of the building. To prevent excessive and quick overflow of the water in the separation and detention tanks, secondary outlets - weirs are installed in these tanks to allow water to flow out of the tank once it reaches a high tank level.

Assuming the catchment tank fully utilises the effective open area of the roof, the area was fixed at 400m^2 (HDB, 2017)., to ensure that volumetric changes are due to variations in height. To observe the variation in dynamics due to difference in height and to consider the non-maximum utilisation of the entire rooftop space, a tank catchment area of 100m^2 was also simulated. It is important to note that the system was previously optimised and is already robust against overflows in its design (Soh et al., 2020). The dimensions and design parameters of the downstream system are described in Table 1 below:

Table 1: Dimensions of Tanks and orifices

	Components	Area (m^2)	Height (m)
Tanks	Separation	40.00	2.50
	Detention	20.00	2.50
	Harvesting	80.00	3.00
	Treatment	40.00	1.25
Orifice	CS	0.0962	0.00
	SD	0.00785	0.00
	SD2 (weir)	0.900	0.66
	DO	0.0962	0.00
	DO2 (weir)	0.0491	1.80
	SH	0.0962	0.52
	HT	0.0962	0.00

The model was simulated for a 24-hour period where the rainfall event falls within to analyse how the rainfall event affects the dynamics of the system. The water tank is modelled using equations derived from mass balance for the respective tanks. The generalised mass balances for a tank j with area A_j and height H_j are as follows with Q_k representing the flowrate between the tanks through the different outlets. The flowrates were determined using the orifice and weir equations determined by Bernoulli's equation, shown in equations 2 and 3.

Mass Balances:

Accumulation = Inlet Flowrate – Outlet Flowrate

(1) $A_j \frac{dH_j}{dt} = Q_{j,in}(t) - Q_{j,out}(t)$

Flowrate Equations:

(2) $Q_{orifice}(t) = C_d \times a \times \sqrt{2 \times g \times H_{eff}(t)}$

(3) $Q_{weir}(t) = \frac{2}{3} \times C_d \times L \times \sqrt{2 \times g} \times H_{eff}(t)^{1.5}$

C_d: *Discharge Coefficient*
a : *Area of orifice*
g : *Gravitational Constant*
L : *Length of weir*
H_{eff}: *Water level above outlet opening*

2.1.2 Key Performance Indexes (KPIs)

To assess the effectiveness of the RWHS, the system was benchmarked against 6 KPIs, of which the first 3 were main parameters used to optimise the volume of the catchment tank:

1. Rainwater overflow from system
2. Capital expenditure (CapEx) of catchment tank
3. Volume of rainwater harvested
4. Rate of water inflow into separation tank
5. Volume of water discharged into public drain network (PDN)
6. Maximum Tank Utilities

The rainwater overflow from the system was defined as the maximum amount of water overflow from the system in each second. It was critical that significant rainwater overflow to be avoided as floods result in disturbances to people's life, damage to property and even fatality. CapEx is measured by the cost of the catchment tank which is a function of volume. Due to opportunity costs of investments, it is essential to minimise CapEx to generate high returns on investments. The volume of rainwater harvested is the amount of water that enters the harvesting tank over the simulated time period. Having a significant volume of rainwater harvested allows for more water to be treated and used to fulfil demand which results in reduced freshwater usage.

The rate of water inflow into the separation tank provides comparison with the base case to assess the effectiveness of the catchment tank in terms of how efficiently it buffers for intense periods of rainfall on the downstream system. Excess water is discharged at a suitable rate into the public discharge network and excessive discharge may strain and cause the network to fail. Lastly, the maximum occupied capacity of the tank provides an overview on the utility of the tanks which offers insights on how the different tanks may be reduced to reduce costs.

2.1.3 Assumptions

To simulate the modelled RWHS to a real-life scenario, multiple assumptions were made. Firstly, all treated water from the treatment tank will be used to fulfil demand and only for non-potable usage. Excess water from all tanks in the downstream system will result in backflow into the preceding tank and hence will never overflow (e.g. An overflowing separation tank will result in water flowing back into the catchment tank). Only installation costs of catchment tanks were considered in CapEx while other fixed costs such as those associated with digging or piping were not accounted for as it varies from systems, locations and type of surfaces installed on. Although water may leave the system through evapotranspiration, it will be difficult to quantify and model. Coupled with the relatively short simulation time, water leaving through evapotranspiration will be deemed negligible and not be included in the mass balances.

2.2 Rainfall Event Analysis

To simulate real-life rainfall scenarios, a 100-year period of synthetic rainfall data over 5-minute intervals was used as a rainfall signal to the system. To filter out negligible rainfall signal, a single rainfall event was defined as a time period with consecutive rainfall signals that are larger than 0.5m^3. Rainfall signal was then disaggregated to a per second basis to ensure coherence with the mass balance and flow equations.

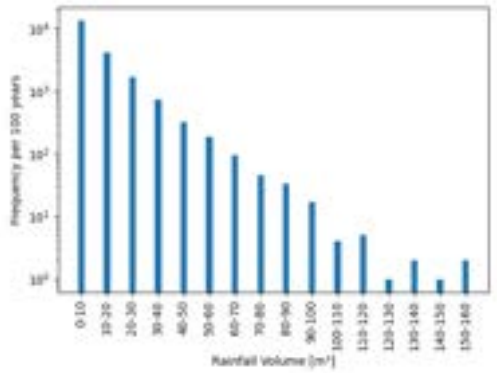


Figure 3: Frequency of rainfall event by total volume

Rainfall events were classified based on the total volume of rain. The RWHS must be able to handle a sufficiently intense rainfall event that occurs relatively frequently to justify the high costs of installations. As such, a rainfall event that occurs 20-30 times per 100-years was identified which translates to volumes in the range 50-60m^3 (Figure 3). Consequently, the rainfall event chosen for simulation in the model, for both the RWHS and base case, will be 60.08 m^3 in volume and 15 minutes long.

To assess how the system reacts to larger rainfall volumes, the system will be stress tested against larger volumes but relatively rare rainfall events. In this case, a significantly higher rainfall volume that occurs approximately 2-3 times per 100-years (Figure 3) was chosen, translating to 153.34m^3 of rainfall and a longer duration of 95 minutes.

2.3 Optimisation of catchment tank volume

Through the simulations on the system while varying tank heights, values were obtained for the main KPIs across a range of volumes. These were used in a 3rd degree polynomial regression analysis to obtain objective functions for harvested water and overflow. These were used to perform a multi-objective optimisation and obtain an optimal catchment tank volume. A regression analysis was not conducted for CapEx as a theoretical relationship was already known. The objective functions for the optimisation are:

$$\begin{aligned} &min && CapEx\ (v) \\ &s.t. && Harvested\ Water\ (v) \geq \varepsilon_{HW} \\ &&& Overflow\ from\ System\ (v) \leq \varepsilon_{OF} \\ &&& CapEx(v) \geq 0 \\ &&& Overflow\ from\ System\ (v) \geq 0 \end{aligned}$$

The ε-constraint method was used where the objective functions of minimising volume of overflow and maximising volume of harvested water were set as constraints by setting inequality around some value ε for each function and constraining it to non-negativity. ε was then varied using parameters that are percentiles of the range of values for harvested water and overflow (Table 2). Excel Solver Non-Linear Generalised Reduced Gradient was then utilised to obtain the optimum volume and the corresponding CapEx, volume of overflow and harvested water.

Table 2: Values for constraints (used as epsilon values)

	100m^2		400m^2	
Percentile	**Volume of Overflow**	**Volume of Harvested Water**	**Volume of Overflow**	**Volume of Harvested Water**
0	0.00	0.00	0.00	0.00
10	6.00	2.62	6.00	2.06
20	12.00	5.24	12.00	4.12
30	18.00	7.86	18.00	6.19
40	24.00	10.48	24.00	8.25
50	30.00	13.10	30.00	10.31
60	36.00	15.72	36.00	12.37
70	42.00	18.34	42.00	14.43
80	48.00	20.96	48.00	16.50
90	54.00	23.58	54.00	18.56
100	60.00	26.20	60.00	20.62

The obtained optimised points were then normalised to allow for equal contribution from each objective function in the optimisation. The normalised optimised data points were then plotted on a 3D graph to obtain the Pareto Front. Since the data points were min-max normalised, the point of origin became the Utopia Point (Szparaga et al., 2019), with maximum harvested water, minimum volume of overflow and CapEx all converging at this point in the normalised data. The optima point for each system is deemed as the point on the Pareto Front which lies closest to the Utopia Point.

2.4 Effectiveness of the catchment tank

The model was thereafter simulated using the optimised catchment tank volumes for the different catchment area and compared against the base case using the same rainfall. All KPIs will be compared on a time basis to analyse the dynamics of the system throughout the simulated 24-hour period. The

optimal volume will then be stress tested against rainfall events with higher rainfall volumes to evaluate the system's ability to handle large rainfall.

3 Results & Discussions

3.1 Performance of catchment tank

The main KPIs were measured across different volumes of the catchment tank for each respective catchment roof area.

Rainwater overflow from the system

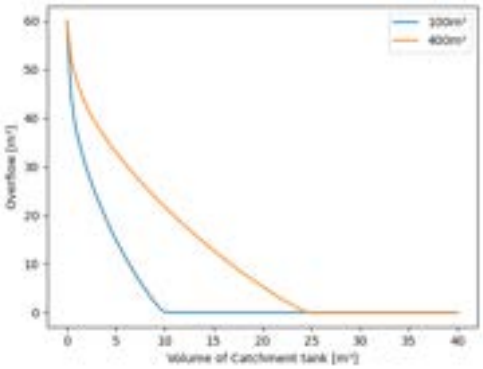


Figure 4: Rainwater Overflow over volumes

At low catchment tank volumes, the RWHS cannot handle the rainfall, leading to overflow. As catchment tank volume increases, the storage capacity of the catchment tank and the time required for rainwater to pass through the catchment tank without overflowing increases as it takes longer for the catchment tank to fill up, reducing the volume of overflow from the system. This was expected as larger volume allows the catchment tank to store more rainwater as it enters the RWHS. However, there comes a point in the system where further increase in catchment tank volume does not result in reduction of rainwater overflow as the system is large enough to handle the total volume of rainfall. As such, to minimise the rainwater overflow of the system, a sufficiently big catchment tank that can handle the total volume of rainfall is required.

The narrower catchment tank was more efficient in handling rainfall as can be seen by the smaller volume of tank required for no overflow from the system (Figure 4). For the same amount of water in the tank, the higher water level in the narrower catchment tank causes the outflow of rainwater leaving the tank to be higher due to a higher effective height. Assuming the same rate of water inflow for both cases, there will be a smaller accumulation and build-up of rainwater in the tank. Hence, for the same catchment tank volume, there was a smaller overflow volume from the narrow tank.

CapEx of catchment tank

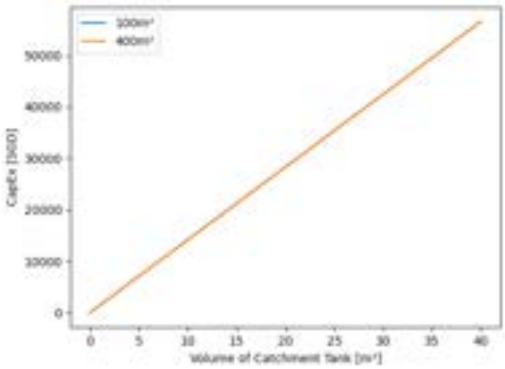


Figure 5: CapEx over volumes

For both cases, the CapEx was a linear function (Figure 5) of the volume of tank (Sherpa, 2021) due to the CapEx only considering installation costs. Further economic analysis can be conducted to account for potential costs savings from reduced freshwater usage to justify a potentially higher CapEx. However, in this study, the main consideration is to keep the CapEx to a minimum.

Volume of rainwater harvested

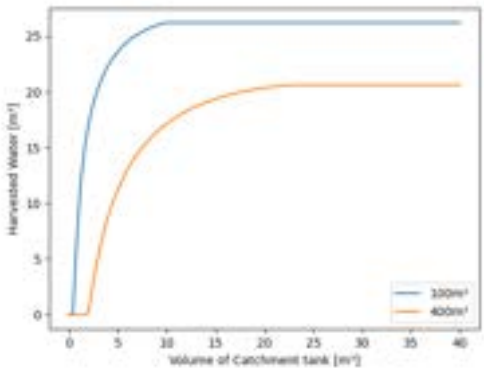


Figure 6: Volume of Harvested Rainwater over volumes

Contrary to rainwater overflow from the system, the volume of harvested water is lowest at the smallest catchment tank volume. It increases with the volume of catchment tank until a point where an increase in catchment tank volume no longer result in further increase in the harvested water due to the fixed volume of rainwater entering the system. Due to larger volume of catchment tank and hence lower volume of overflow, more rainwater can reach the downstream system for harvesting and treatment. However, further adjustments can be made to the downstream system to redirect rainwater into the harvesting tank compared to the detention tank.

For the same volume of catchment tank, the difference in tank area resulted in a higher corresponding water level for the 100m^2 catchment

tank. Since the flowrate of the orifice is a function of the effective height (Equation 2) and the orifice height is the same, the flowrate of rainwater into the separation tank will be higher for this tank. The separation tank will be filled up at a faster rate and the tank level will reach the height of the orifice connected to the harvesting tank (SH) earlier. As such, more water will be directed to the harvesting tank for treatment at a higher rate for the 100m^2 tank.

There was no water harvested for very small catchment tanks for the 400m^2 area as the water level is significantly lower. Combined with a low flowrate into the separation tank, this allows more time for water to be discharged into the detention tank instead. Since the orifice for SD is the lowest in the tank, water will first flow out to the detention tank before the water level can reach the height of the orifice for SH and flow into the harvesting tank.

3.2 Multi-objective optimisation of KPIs

Trade-offs between these KPIs exist where a smaller volume is preferred due to space and cost considerations while a larger volume is preferred to improve the rainwater harvesting performance. Hence, an optimal balance should be obtained in the consideration for building a RWHS.

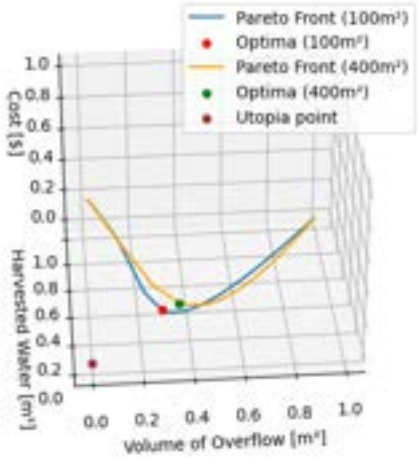


Figure 7: 3D graph showing Pareto Front and respective optima

The calculated optima point for the narrow catchment tank is closer to the Utopia point compared to that of the broader tank (Figure 7) which meant that the narrow tank was more ideal when comparing the 3 main KPIs. This was expected due to the ability for the 100m^2 tank to achieve a lower overflow and higher harvested water at a lower catchment tank volume and CapEx. The tank with smaller base area and taller height would also be preferred in the context of Singapore due to its lack of land area and is one of the most built-up cities in the world.

Table 3: Optimal solutions for volumes and KPIs

Catchment Tank Area (m^2)	Volume (m^3)	CapEx ($)	Overflow from the system (m^3)	Volume of Harvested Water (m^3)
100	4.31	6100	16.55	23.58
400	12.07	17073	17.25	18.56

3.3 Effectiveness of Catchment Tank

The effectiveness of the catchment tank will be evaluated by comparing the different KPIs for the determined optimal volumes against the base case and by stress testing the system against a heavier rainfall pattern.

3.3.1 Base case comparison

The KPIs were measured against a 24-hour simulated timeframe with the start and end of rainfall indicated in the graphs to show how the dynamics change with rain. The KPIs analysed were focused on the timeframe during which the rainfall event occurred.

Overflow from System

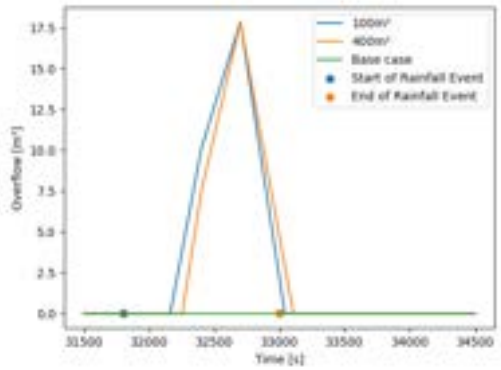


Figure 8: Rainwater overflow over time

No overflow was observed for the base case as the volume of the separation tank was larger than the overall volume of the rain. The volume of optimised catchment tanks of the RWHS for the respective catchment tank areas were similar and significantly smaller, resulting in overflow from the system slightly after the rain started. Due to a smaller optimised volume for the 100m^2 tank (Table 3), the system overflowed earlier than that of the 400m^2 tank. However, the dynamics of the system meant an increased outflow into the separation tank and smaller accumulation in the catchment tank resulting in a slightly smaller overflow despite a smaller volume.

Volume of Harvested Water

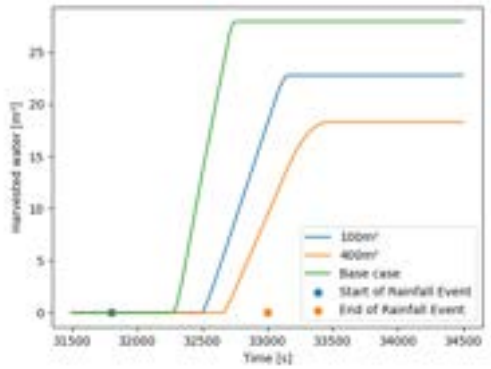


Figure 9: Volume of harvested water over time

The volume of harvested water for the base case was larger than that for the RWHS. As the water flowing into the separation tank was limited by the orifice equation and effective height of water in the catchment tank, the water flow into the separation tank was lower. Hence, more water was harvested earlier as the tank level in the separation tank for the base case reached the height of orifice for SH faster than the cases with catchment tanks.

The relatively constant gradient during the initial harvesting of water meant that the harvesting rate and hence the water level during that period were constant. The gradient decreased at the end of the harvesting process due to the decreasing effective height as the rain has stopped and no more water enters the separation tank. Harvesting stops once the water level falls below the height of the orifice SH.

Separation tank water inflow

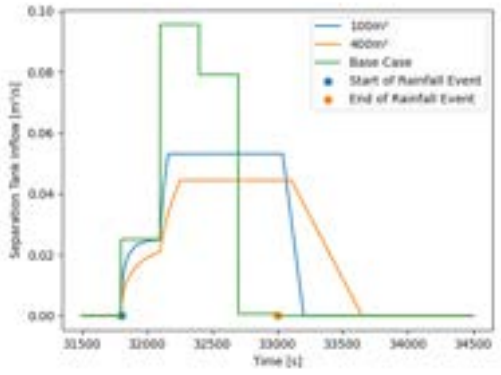


Figure 10: Flowrate of water into separation tank over time

Due to the lack of catchment tank, the separation inflow is simply the rainfall signal for the base case. The presence of the catchment tank helps to smoothen and alleviate the stress on the system at any given time by further spreading out the flow of the rain volume over a larger time period. This makes the system effective in handling rainfall events that are expected to be increasing in intensity in future. The constant peak inflow is partially attributed to the full tank that prevents further increase in flowrate of CS. The broader catchment tank was more effective at cushioning peak rainfall due to smaller effective tank height and flowrate into the separation tank when the tank was full.

Public Drain Network (PDN) flowrate

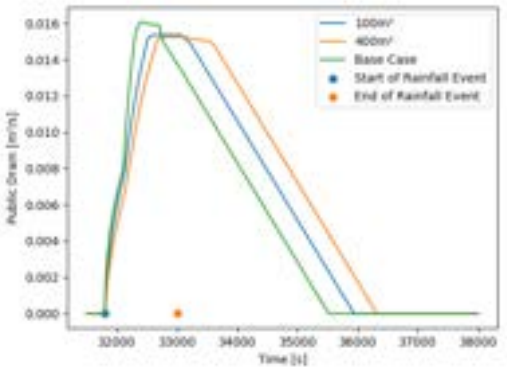


Figure 11: Flowrate to public drains over time

Unlike most of the other flow parameters, the PDN recorded data immediately after the start of the rain due to the orifice of SD being located at the base of the separation tank. This caused the water to flow out from that orifice instantaneously, albeit slowly due to the small orifice area. The base case reached a larger flowrate for PDN as the separation tank can hold the entire volume of rainwater without overflowing, allowing more water to enter the PDN system at a faster rate. As the tank helped to moderate and spread out the water inflow into the separation tank, the rainwater took longer to clear the system through the PDN and hence the discharge rate tailed off later relative to the base case. This should be considered in case studies with frequent rainfall events as water in the system from earlier rainfall events may negate the RWHS effectiveness.

Tank Utilities

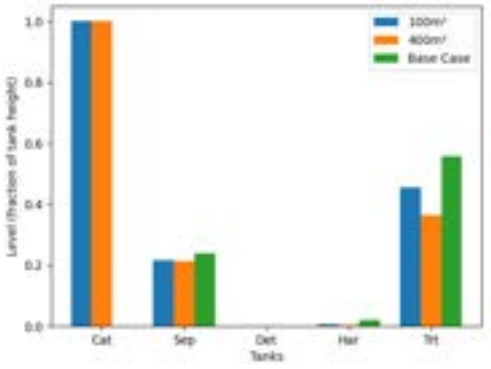


Figure 12: Maximum tank utility

The utility for most tanks except catchment tank (CAT) were generally similar across all cases. Most of the downstream tanks were ineffectively utilised

with less than 50% utility which opens up the possibility of using smaller tanks in the downstream system, especially because Singapore is geographically constrained. The design of the system such as the placement of the orifices can also be adjusted to allow more water to be harvested rather than discharged through the detention tank.

The utility of the separation tank for the RWHS were also relatively similar due to the presence of the SH located at approximately 20% of the tank height, preventing the water level from exceeding that height. The presence of the weir at SD2 also kept the water level from exceeding that level and the installation of a secondary outlet on the catchment tank in the RWHS can be considered. The separation tank utility was slightly more for the base case as the flowrate into the separation tank was higher than that of the RWHS which meant a higher build-up of water. On the other hand, the catchment tank could store some rainwater before it flowed into the separation tank, reducing the utility of the separation tank. Since the catchment tank was significantly smaller than the separation tank, the impact in utility of separation tank was very minimal.

3.3.2 Stress testing

The RWHS was then stress tested against the maximum rainfall volume that was obtained from the synthetic data to evaluate the effectiveness when faced with extreme rainfall volume.

Overflow from system

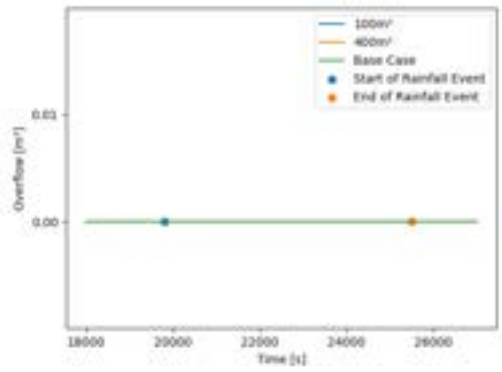


Figure 13: Rainwater overflow over time for stress testing

Despite the higher rainfall volume simulated, there was no overflow from the system for all three cases. The main reason for this was rainfall being spread over a longer duration, which meant the average rainfall per second was much smaller. Since the input flowrate was much smaller, rainwater has more time to pass through the entire system before there was significant accumulation and build-up of the rainwater in the primary tank of the system. This meant that a larger rainfall volume does not necessarily result in overflow from the system. The duration of rainfall event will also affect the rate at which water enters the RWHS.

Volume of Harvested Water

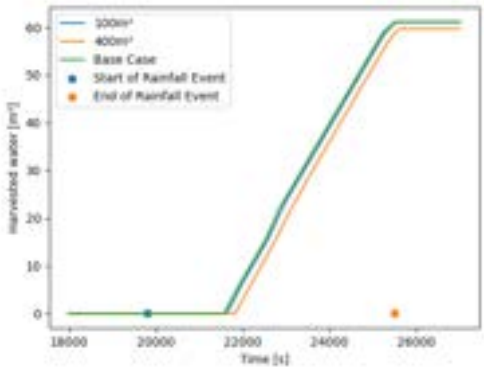


Figure 14: Volume of harvested water over time for stress testing

The volume of harvested water was also very similar in terms of rate and volume harvested across all 3 cases. As the catchment tank was not fully utilised, the flowrate of rainwater into the separation tank was like that of the base case and hence few differences between the 3 cases were observed. The higher volume of harvested water obtained during the stress testing was mainly attributed to the larger volume of rain during the simulated timeframe.

Separation tank water inflow

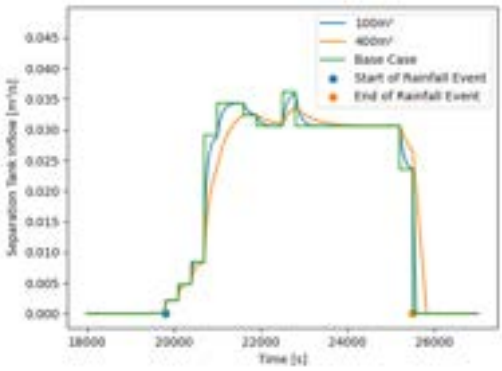


Figure 15: Flowrate of water into separation tank over time for stress testing

The catchment tank was critical in smoothening the rain signals and it was more prominent for the case of 400m^2 catchment tank. However, the peaks were not reduced and less spread out compared to the base case due to the smaller average flowrate. As such, the catchment tank did not store as much rainwater to smoothen the flowrate of rainwater into the separation tank. Relative to the 100m^2 catchment tank, the outflow of the 400m^2 tank is smaller, allowing more water to accumulate in the catchment tank. This spreads out the water inflow into the separation tank across a larger time period.

Public Drain Network (PDN) flowrate

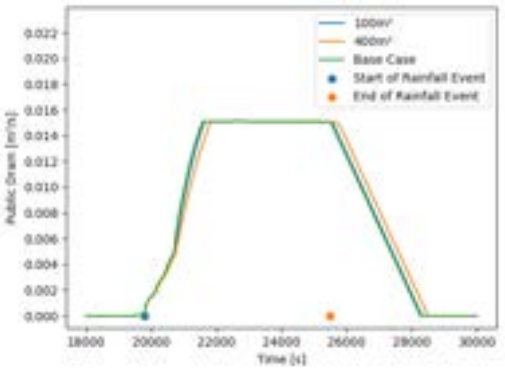


Figure 16: PDN flowrate over time for stress testing

Like the volume of rainwater harvested, the PDN flowrate displayed little differences for all three cases and the behaviour was like that of the base case. This was due to similar separation tank inflow across the 3 cases which resulted in similar effective height in the tank and PDN flowrate.

Tank Utilities

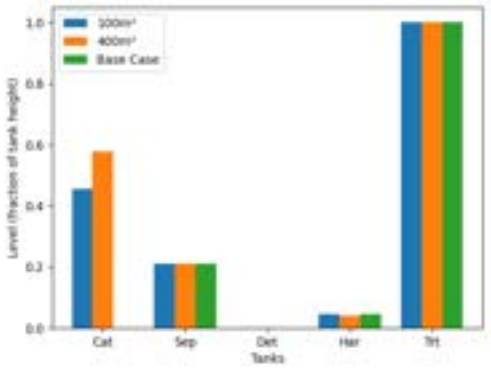


Figure 17: Maximum utility of tanks over 24-hour period for stress testing

The catchment tank's utility (Figure 17) was different due to smaller rain signal per second. As there was no overflow from the system (Figure 13), the catchment tanks were not fully utilised. The lower flowrate for the 100m^2 catchment tank resulted in the build-up of water in the tank, resulting in higher utility. The reduced rainfall per second also meant that the water level in the separation tank only reached the maximum height of SH, resulting in the same utility of separation tank for all 3 cases.

3.4 Qualitative Analysis of Silva Cell

While the simulation and KPIs provide a good quantitative comparison, qualitative analysis should be conducted to understand how the proposed Silva Cell system compares with the base case.

The Silva Cell considered have many merits. The ability to support and sustain biodiversity offers the Silva Cell benefits associated with green roofs such as beautification and air purification. Known as the 'garden city', green roofs have huge room for growth in Singapore due to its lush vegetation, green space and environmental policies, as the government aim to have at least 80% green buildings by 2030 (Up to Us Veolia, n.d.). The system also protects the roof from UV rays and temperature fluctuations, reducing the cost of maintenance and likelihood of re-roofing (Brass, 2022). Moreover, the additional soil layer acts as a fire and heat resistant layer (Eksi et al., 2017) due to high moisture content in plants, improving the thermal efficiency of the building. This is especially significant in Singapore as cooling constitutes almost 25% of an average household's electricity consumption (Muruganathan, 2021). Additionally, Silva Cell can filter unwanted particles, ensuring high water quality which results in less treatment required in the downstream stages. Lastly, a land scarce country like Singapore cannot afford to have inefficient use of land and will require proper planning for all available land area including utilising rooftop for variety of purposes such as carparks. As such, the modular pavement system makes the proposed system very suitable due to its ability to take on a huge load for vehicles.

The RWHS does have some drawbacks. Untreated acidic rain in the upstream system may destroy foliage and corrode pipes in the downstream system. The unpredictable nature of root growth also require frequent monitoring and maintenance to ensure the infrastructure is not damaged by the roots. However, this can be mitigated using technology to reduce damage to existing infrastructure (Ganesan, 2018).

4 Conclusion and Outlook

The modelling revealed that instead of improving the RWHS, the use of the catchment tank reduced the ability to handle rainfall effectively as the larger separation tank in the downstream system could not absorb water from the surface, resulting in overflow and reduced harvested water. It should be noted that the current system is designed to be a passively operated system and robust against overflows. However, this system was found to be inefficiently used due to the large downstream tanks. If the overall system can be redesigned, accounting for the costs of the entire system, the addition or even the replacement of the separation tank with the Silva Cell catchment tank may prove to be useful in buffering rainfall volume and handling high intensity rainfalls. The duration and volume of rain are additional factors to be considered in the RWHS design. The qualitative benefits of the proposed system also far outweigh the potential risks. With proper mitigation measures adopted, these risks can be mitigated, opening up ways where the system can be used for the existing system in the base case.

Since the dynamics affected the effectiveness of the RWHS, further studies can be done to optimise the tank dimensions for the different KPIs. Adjustments to the design of the RWHS such as orifice area or height could help to improve rainwater harvesting performance. KPIs should also be prioritised according to the order of importance based on stakeholder's interest to ensure key targets are met.

Additionally, an active control system can be implemented where actuators such as valves or pumps could be installed in the rainwater harvesting system. These actuators could be activated using a feedback control system to increase the flowrate out or to pump water out of the catchment tank based on the tank's water level, making the system more holistic and adaptable to changing rainfall patterns.

Through the course of this study, it became apparent that retrofitting a catchment system before an existing optimised RWHS is not simple, as it is difficult to make changes to the downstream section. Since a key finding is that the separation tank size can be reduced while still handling significant rainfall, this methodology would perhaps find greater success in application to designing future systems from scratch, which should yield significant cost and material savings if done correctly.

Acknowledgements

The authors of this research paper would like to extend their sincerest gratitude to Professor Nilay Shah, Dr Edward O'Dwyer and Soh Qiao Yan for their continuous support and guidance.

References

1. *Singapore: the world's first garden-city.* https://www.up-to-us.veolia.com/en/singapore-world-first-garden-city [Accessed 25 Nov 2022].

2. Bevere, L. (2022) *Natural catastrophes in 2021: the floodgates are open.* https://www.swissre.com/institute/research/sigma-research/sigma-2022-01.html [Accessed Nov 10 2022].

3. Boretti, A. & Rosa, L. (2019) Reassessing the projections of the world water development report. *Npj Clean Water.* 2 (1), 10.1038/S41545-019-0039-9.

4. Brass, A. (2022) *SUNLIGHT AND UV EFFECTS ON ROOFS: Advancements To Resist Damage.* https://roofingelementsmagazine.com/sunlight-and-uv-effects-on-roofs-advancements-to-resist-damage/ [Accessed Nov 25, 2022].

5. Eksi, M., Rowe, D. B., Wichman, I. S. & Andresen, J. A. (2017) Effect of substrate depth, vegetation type, and season on green roof thermal properties. *Energy and Buildings.* 145 174-187. 10.1016/j.enbuild.2017.04.017.

6. Ganesan, D. (Jul 19, 2018) NParks to step up use of technology in greenery management. *Straits times (Singapore: Daily).* https://tnp.straitstimes.com/news/singapore/nparks-use-technology-improve-tree-management .

7. García, L., Barreiro-Gomez, J., Escobar, E., Téllez, D., Quijano, N. & Ocampo-Martinez, C. (2015) Modelling and real-time control of urban drainage systems: A review. *Advances in Water Resources.* 85 120-132. 10.1016/j.advwatres.2015.08.007.

8. HDB. (Sep 1, 2017) HDB Rolls Out Solar-Ready Roofs for Easier and Faster Installation of Solar Panels. *Singapore Government News.* *https://www.hdb.gov.sg/cs/infoweb/about-us/news-and-publications/press-releases/01092017-hdb-rolls-out-solarready-roofs*

9. Muruganathan, K. (2021) *Commentary: Air-conditioning – the unspoken energy guzzler in Singapore.* https://www.channelnewsasia.com/commentary/air-con-unit-electricity-energy-carbon-emissions-climate-change-1339326 [Accessed Nov 27, 2022].

10. Soh, Q.Y., O'Dwyer, E., Acha, S. & Shah, N. *Optimization and Control of a Rainwater Detention and Harvesting Tank.*

11. Sherpa, G. *Python simulation results.*

12. Słyś, D. & Stec, A. (2020) Centralized or Decentralized Rainwater Harvesting Systems: A Case Study. *Resources.* 9 (1), 5. 10.3390/resources9010005.

13. Snir, O., Friedler, E. & Ostfeld, A. (2022) Optimizing the Control of Decentralized Rainwater Harvesting Systems for Reducing Urban Drainage Flows. *Water (Basel).* 14 (4), 571. 10.3390/w14040571.

14. Szparaga, A., Stachnik, M., Czerwińska, E., Kocira, S., Dymkowska-Malesa, M. & Jakubowski, M. (2019) Multi-objective optimization based on the utopian point method applied to a case study of osmotic dehydration of plums and its storage. *Journal of Food Engineering.* 245 104-111. 10.1016/j.jfoodeng.2018.10.014.

Development of PEG-free lipid nanoparticles in reduced ethanol

Andreas Jirkas and Ourania Papakyriakou
Department of Chemical Engineering, Imperial College London, U.K.

Abstract
Lipid nanoparticles (LNPs) are a key component in modern day vaccines and therapeutic platforms. Amongst the greatest challenges of the development of LNP-based systems include achieving fixed particle size, their lack of stability, cytotoxicity in cells and propensity for physical and chemical aggregation in solution. In this research project, a unique design strategy is deployed using ionic liquids (ILs) with excipients that can enhance the structural stability of proteins and RNA-based lipid nanoparticle vaccines, whilst decreasing ethanol concentration needed and eliminating polyethylene glycol (PEG). The best formulations were optimised by reducing the ethanol concentration. That will positively contribute to emulate effectively commercial production. In this study, a variety of quantitative and qualitative data were utilised to firstly, verify that the lipid film did not remain flat in the buffer, and then to produce an LNP-based system of a pH range 6.0-7.5 that would be tested for transfection, particle size and zeta potential. Finally, F3B8 combination resulted to the lowest ethanol concentration of 20 w/w% while remaining colourless. Even though the LNPs agglomerated, next steps are proposed to avoid agglomeration and attain PEG-free formulations produced using decreased ethanol volume.

Keywords: Lipid nanoparticles, ethanol, polyethylene glycol, excipients, vaccines, ionic liquids, drug allergy

1. Introduction

Vaccine applications frequently constitute the only line of defence against viral infections, as antiviral drugs and treatment success are inadequate. As we have seen over the past 30 months of the COVID-19 pandemic, mRNA nanotechnologies have transformative potential as vaccines and therapeutic platforms. Furthermore, in the pharmaceutical sector, developing lipid nanoparticles (LNPs) has attracted great interest, as these can be used for controlled drug delivery and have been shown to enhance therapeutic effects. However, there are major challenges with the development of LNP-based systems. There is currently a significant bottleneck associated with these technologies: cytotoxicity in cells, attaining fixed particle size, their lack of stability, and propensity for physical and chemical aggregation in solution.

Recently, ionic liquids (ILs) have emerged as a class of solvents used to stabilise and functionalise polymers and drugs. Using ILs with excipients can enhance the structural stability of proteins and RNA-based LNP vaccines. Thus, using ILs with LNPs could enhance the stability of these materials and improve controlled drug release, crucial in pharmaceutical development.

This project aimed to develop polyethylene glycol (PEG)-free lipid formulations and buffers (including ILs and excipients) that could decrease the percentage of ethanol needed in the production of LNPs. Attaining a formulation with the aforementioned characteristics can potentially decrease the cytotoxicity of the formulation and make the formulation accessible to more patients, avoiding incidents with patients allergic to PEG that can suffer immediate serious allergic reactions with symptoms including anaphylaxis upon using PEG vaccines. Quantitative and qualitative techniques have been utilised throughout the project to judge whether the PEG-free LNPs developed in the low ethanol w/w% buffers, had the wanted mechanical and chemical characteristics.

2. Background

During the past years there has been an increased interest in RNA vaccines and therapeutics since they eliminate the risks associated with DNA or live-attenuated vaccines while keeping their advantages [1]. Due to the high instability of RNA once in the body, there have been intense efforts to stabilize it for in vivo applications.

Strategies for RNA delivery include viral vectors, RNA-conjugates, microparticles and nanoparticles with the focus being turned to nonviral vectors since this became possible due to technological and material advancements [1].

Special emphasis has been given to LNPs in the last years. LNPs have also been in the spotlight to the broader public, due to their use in the COVID-19 vaccines. The global LNPs market size was valued at $694 million in 2021 and it is expected to reach $1210 million by 2027 [2].

LNPs are shells made of a monolayer of lipids and their size can be between 20 and 100 nm. Their internal core is made of reverse micelles which encapsulate oligonucleotides [3]. They are recognised as versatile adjuvants that enhance the efficacy of vaccines [4].

Lipids are amphiphilic molecules made of a polar head group linked to a hydrophobic-tail region. Generic LNPs are made by combining cationic or ionizable lipids with other types of lipids.

Cationic lipids are used as their permanently positively charged head, aids the combination of the negatively charged backbone of RNA. Furthermore, they can also bind to the negatively charged cell membrane of mammalian cells to induce the RNA uptake from the cell, in vitro. However in vivo, this exact feature can lead to aggregation of the particles with blood proteins as well as enhanced uptake by the Reticuloendothelial system (RES) [5]. Examples of cationic lipids used for mRNA delivery are 1,2-di-O-

octadecenyl-3-trimethyl-ammonium-propane (DOTMA) and its biodegradable analogue, 1,2-dioleoyl-3-trimethylammonium-propane (DOTAP) [6].

Ionizable lipids remain neutral at bodily pH but are protonated, becoming positively charged when the pH decreases. This feature allows the use of the positive charge and the advantages associated with it, in acidic conditions in vitro, while allowing the LNPs to remain neutral in the blood stream and minimize interaction with negatively charged surfaces in the body. An example of an ionizable lipid is 1,2- dilinoleyloxy- N, N-dimethyl-3-aminopropane (DLin-DMA) [5].

Other types of lipids used in formulations of LNPs include cholesterol and phospholipids as well as PEG. Cholesterol is crucial during cell transfection and helps to stabilise the LNP structure while phospholipids play a crucial role in the formation of the lipid layer and its disruption during endosomal escape [6]. PEG and pegylated lipids act as a barrier between LNPs and other proteins by depositing on the LNP's surface, masking the surface charge of the particle and developing a hydrophilic barrier [7]. PEG-lipids can affect both the size of the particle as well as its zeta potential [5].

Despite the advantages of using PEG and PEG-lipids in LNPs there are certain disadvantages linked to them. Some studies suggested that while PEGylation extends circulation half-life, it does not decrease the binding of proteins to the nanoparticle [8]. Furthermore, there have been studies that suggested the existence of anti-PEG antibodies in healthy individuals that were repetitively exposed to PEG [7]. These antibodies could lead to anaphylactic shocks.

The production of LNPs involves the addition of ethanol to the lipid films, which stabilises the edges of the lipid fragments, resulting to the required shell-structure. The ethanol hydrates the flat lipid films. Upon the addition of ethanol, the mixture must remain colourless, just like vaccines using LNPs are colourless, indicating that the lipids are no longer flat but have rather acquired the wanted shape. Dilution of ethanol with water leads to growth and merge of the fragments to larger bundles which is unwanted [1]. However, injecting large amounts of ethanol to humans is not safe, and LNPs prepared with ethanol need to be washed with buffer to eliminate the ethanol content. This leads to a large buffer waste and can also damage the formed vesicle.

For the above problems associated with PEG and ethanol, this research project aimed to develop PEG-free lipid formulations and buffers that could decrease the percentage of ethanol needed in the production of LNPs.

In general, the structure of lipids is unstable in solution, as lipids can aggregate forming large particles due to the stresses and interactions encountered during the production stage. It was therefore of great importance to develop a buffer solution that would hydrate the lipid films but also prevent lipid aggregation for the LNPs to be effective and efficient.

Numerous organic compounds of low molecular weight used as solvent additives, called "osmolytes", have shown to improve the stability of proteins and aid in the reduction of aggregation. These solvent additives lack affinity for proteins or interact with them in a repulsive way, hence stabilising their structure. Numerous molecules can act as stabilizers including polymers, salts, sugars, surfactants and amino acids [9]. By studying different patents, it was decided that such excipients could also potentially be used to stabilise LNPs [10, 11].

Furthermore, it has been found that ILs help towards the structural and thermal stabilisation of proteins [12]. ILs are organic salts that are in liquid state below 100 °C and are composed of ions in their entirety. They are found to interact with proteins and water mainly through hydrogen bonding [13]. ILs were added to LNPs produced in this research to stabilise them.

To put the above theory into practice, extensive literature review has been carried out to find the lipids needed to produce suitable PEG-free lipid films, the excipients needed to formulate suitable buffer solutions and the ILs that would be combined with the successful formulations. Apart from research, the choice of chemicals was also based on stock availability.

The purpose of the buffer solutions was to replace ethanol's function of stabilising the lipid edges to form the desired lipid shells, while also preventing the aggregation of lipid molecules into larger bodies. The initial goal was to decrease the ethanol concentration used in the production of LNPs to 10 w/w %.

In general, the composition of the buffers contained sugar, surfactant, amino-acid, ethanol and water as the basis [9,10,11,14]. Additional components such as polyols were added in some formulations as they can act as further stabilisers in liquid formulations [15].

Sugars are found to stabilise proteins and are thus an essential part of the buffer formulation, as they appear to be preferentially excluded from the protein vicinity [9]. Possible sugar options include sucrose, trehalose, mannitol, lactose, glucose, maltose, mannose, fructose and others [11]. Sucrose and trehalose were found to be used more often [9] and were therefore the two sugars used. Literature suggested a concentration of sugar between 5 - 75 mg/mL [14], however a concentration of 68 mg/mL was used, based on Kallmeyer, G. et al. (2014) [10].

Adding surfactant to a sugar - amino acid buffer formulation, appears to improve the stability of proteins and leads to lower turbidity values, suggesting less aggregation. The surfactants considered were ones known to be used in the pharmaceutical industry. An acceptable amount of surfactant to stabilise proteins was reported to be 0.05 – 0.5 mg/mL [10]. The surfactants considered were Tween 20, Tween 80 and Pluronic F 68. Tween 20 was chosen as it appeared in most sources and was also in stock. It was used at a concentration of 0.1 mg/mL.

The choice of sugar and surfactant were straight forward with different sources of literature agreeing between them, however the choice of amino acids appeared to be more challenging. Amino acids appear to stabilise proteins due to their interaction with the protein's peptide bond which is unfavourable. Cohesive force mechanism as well as excluded volume effect can be related to the unfavourable interaction observed. The interaction causes stabilizing amino acids to remain in bulk water which is unstable in terms of entropy [9],

hence forcing the protein to take up the least possible volume.

Different papers aiming to stabilise different proteins suggested the use of different amino acids. The discrepancy between sources suggested that the best amino acid depends on the protein to be stabilised and in the case of this study, the lipids to be stabilised.

In general, it was suggested to use basic amino acids such as arginine (Arg), lysine (Lys) and histidine (His) as well as other amino acids [9,10,11,14]. It was also suggested that amino acids that crystalise do not act as stabilisers but rather as bulking agents [9,14]. Different papers report crystallisation of different amino acids for stabilisation of different proteins, thus it was deemed a good idea to try different amino acids on the different lipid formulations, at 10 mg/mL [10].

For the lipid formulations, cholesterol was used, as apart from the reasons mentioned above it is also found to improve the circulation half-lives as it is an exchangeable molecule that can accumulated within a liposome during circulation [16].

The phospholipids used were 1,2-Dioleoyl-sn-glycero-3-phosphocholine (DOPC) and Dioleoyl phosphatidylglycerol (DOPG) which are both unsaturated. The use of unsaturated phospholipids was encouraged by results indicating an improved intracellular delivery and particle uptake when replacing the saturated analogue 1,2-Distearoyl-sn-glycero-3-phosphocho-line (DSPC) [17].

To complete the lipid formulation, an ionizable lipid such as Dlin-DMA was preferred. Due to Dlin-DMA not being into stock, the cationic lipid DOTAP was used instead. According to Sun M. et al., LNPs of DOTAP/ cholesterol composition made it to clinical trials and were classified as “the most efficient gene delivery systems” [17].

Finally, choline chloride ([Cho]Cl) was chosen as the IL to be used for this project. It is a low molecular weight, biocompatible IL with high water solubility of high purity which can be synthesized easily and at low cost [12]. Furthermore, to explore other options, choline dihydrogen phosphate ([Cho][DHP]) and choline bitartrate ([Cho][Bit]) were also used.

3. Materials and Methods

The materials chosen to be used for this paper were determined by existing literature, as outlined in section 2, but also based on stock availability.

The materials used were stored as recommended by the supplier and used without further purification. A list of the materials used is provided in table 3.1 below.

Table 3.1. Main materials used.

Chemical name	Source	Mole fraction purity (%)
Sucrose	Sigma-Aldrich	>99.5
Tween 20	Sigma-Aldrich	>40
L-histidine /HCL	Sigma-Aldrich	>99
Glycerol	Sigma-Aldrich	>99.0
Serine	Sigma-Aldrich	>98
Trehalose	Sigma-Aldrich	>99
Glycine	Sigma-Aldrich	>99.7
Lysine	Sigma-Aldrich	>98
Ethanol	Sigma-Aldrich	>99
Chloroform	Sigma-Aldrich	>99
Cholesterol	Sigma-Aldrich	>99
DOTAP	Sigma-Aldrich	>99
DOPC	Sigma-Aldrich	>99
DOPG	Sigma-Aldrich	>99
[Cho]Cl	Sigma-Aldrich	>99
[Cho][DHP]	Sigma-Aldrich	>99
[Cho][Bit]	Sigma-Aldrich	>99

3.1. Buffer formulations

Aqueous buffers were prepared in deionised water of pH 7.4. The buffer formulations are given in table 3.1.1.

The pH of each buffer was measured using the pH electrode Mettler Toledo InLab Micro (WOLFLABS, UK) and the average results of three measurements were taken. The pH of each formulation was adjusted where needed, by addition of either 0.1 M NaOH or 0.1 M HCl to obtain a pH of approximately 6.5. All formulations were prepared and stored at room temperature until measured or used in further steps.

3.2. Lipid formulations

Lipid films were prepared in a fume cupboard with each sample stored in a glass vial (Thermo Fisher Scientific Inc, USA). The compositions of the lipid films are summarised in table 3.2.1 below.

To prepare the lipid films, the lipids were first dissolved in chloroform. Based on the composition of each film, the lipids were then mixed, and a continuous stream of nitrogen air was utilised to evaporate the chloroform and form a film. Any residual chloroform was removed via desiccation for at least 2 hours. Once completed, the buffer was added to the lipid film.

3.3. Lipid nanoparticles formation

Microfluidic hydrodynamic focusing (MHF) was used to form the LNPs from the hydrated lipids obtained upon mixing the lipid and buffer formulations.

Lipid in buffer solution (LS) was drawn into a 1 mL normject disposable plastic syringe, with PRFE tubing connected (OD:1/16’’). The relevant buffer (B) containing the relevant IL was drawn up into a 5 mL normject disposable plastic syringe, with PRFE tubing connected (OD:1/16’’).

Some of the samples prepared included RNA. The RNA used, was made by VEEV-Fluc pDNA. In the cases used, the RNA was added to B.

LS was injected into the central inlet of an MHF suitable microfluidic device at 100 µL/min via a syringe pump (Harvard). B was injected into the buffer inlet of the same MHF device at 100 µL/min.

The two streams were allowed to equilibrate and form an MHF flow regime where the central lipid stream was flanked by the two sheathing buffer streams at steady flow. The flow rates were then adjusted so that B had a flowrate of 150 µL/min and LS had a flowrate of 50 µL/min. This achieved a flow rate ratio (FRR) of 3.

Collection of samples started after 30 seconds via the outlet tubing present downstream from the crossflow junction.

Table 3.1.1. Buffer components in each formulation. Water was added to each buffer to produce 1 mL of buffer.

Buffer	Sucrose	L-Histidine /HCl	Tween 20	L-Glycerol	Serine	Trehalose Dihydrate	Glycine	Lysine	Ethanol
	mg/mL								
B1	68.0	10.0	0.1	0.0	0.0	0.0	0.0	0.0	100.0
B2	68.0	10.0	0.1	0.0	0.0	0.0	0.0	10.0	100.0
B3	68.0	10.0	0.1	3.0	0.0	0.0	0.0	0.0	100.0
B4	0.0	10.0	0.1	3.0	0.0	68.0	0.0	0.0	0.0
B5	0.0	10.0	0.1	0.0	0.0	68.0	0.0	0.0	0.0
B6	0.0	10.0	0.1	3.0	0.0	68.0	0.0	0.0	100.0
B7	68.0	0.0	0.1	3.0	10.0	0.0	0.0	0.0	0.0
B8	68.0	0.0	0.1	0.0	0.0	0.0	10.0	0.0	100.0
B9	68.0	0.0	0.1	3.0	10.0	0.0	0.0	0.0	0.0
B10	68.0	0.0	0.1	3.0	10.0	0.0	0.0	0.0	100.0

Table 3.2.1. Lipid components (molar composition %) in each lipid formulation

Formulation	DOPC	Cholesterol	DOTAP	DOPG
	mol%			
F1	90	10	0	0
F2	10	50	40	0
F3	50	10	40	0
F4	0	50	40	10
F5	0	10	40	50

3.4. RNA Transfection

The process to introduce nucleic acid to cells by artificial means is called transfection. For transfection, Opti-MEM™ medium, Lipofectamine™ Messenger-MAX™ reagent and RNase-free water were used. HEK 293T cells were prepared in a complete DMEM medium and were allowed to grow to 70-85%.

The transfection was carried out in a 96-well plate, and it was designed to have 100 ng of RNA present in each well (either LNP-encapsulated or not). 6 LNP samples of concentration 1 μg/ml were used in the transfection process, as summarised in table 3.4.1 below:

Table 3.4.1: Samples used in transfection

Name	Composition
20R*	F3 in B8 with 20 w/w% ethanol, 5 w/w% [Cho]Cl and RNA
30R*	F3 in B8 with 30 w/w% ethanol, 5 w/w% [Cho][DHP] and RNA
50R*	F3 in B8 with 50 w/w% ethanol, 5 w/w% [Cho]Cl and RNA
20E*	F3 in B8 with 20 w/w% ethanol, 5 w/w% [Cho]Cl, no RNA
30E*	F3 in B8 with 30 w/w% ethanol, 5 w/w% [Cho][DHP], no RNA
50E*	F3 in B8 with 50 w/w % ethanol, 5 w/w% [Cho][Cl], no RNA

In addition to the samples summarised above, two samples of non-formulated RNA and a cells-only sample, were used as controls. The non-formulated RNA was complexed with lipofectamine. To validate the success of the assay, a sample of Vesicular Stomatitis Virus (VSVG) was also used as a double positive control.

Lipofectamine was diluted in Opti-MEM medium. For every well 0.3 μL of lipofectamine and 5 μL of Opti-MEM were required. Once mixed, the tubes were vortexed and spined down in a microcentrifuge. They were then incubated for 10 minutes at room temperature resulting to Solution A.

Each sample was diluted in Opti-MEM medium to achieve the required 100 ng/μL RNA concentration required. Due to the low volume required, a diluted RNA stock solution was prepared beforehand. For each well, 1 μL of the diluted RNA was mixed with 5 μL of Opti-MEM and was incubated at room temperature for 5 minutes resulting to solution B. Finally, Solution A and Solution B were mixed and incubated for 24 hours at 37 ºC and 5% CO_2.

To assess the success of the transfection, the luminescence of the samples needed to be measured. The RNA in the LNPs included the luciferase gene which encodes a 61-kDa enzyme. In the presence of oxygen, ATP and Mg^{2+}, this enzyme oxidizes D-luciferin which results to a fluorescent product [18].

Once the incubation was complete, 50 μl of the medium were removed, 50 μl of Bright-Glo™ Luciferase Assay Reagent was added to each well and incubated at room temperature for 5 minutes in the dark. The cells were then mixed gently with the substrate, and luminescence was measured using a luminometer. The values were expressed as RLU/mL. RLU is Relative Light Unit.

3.5. Dynamic & Electrophoretic Light Scattering

To investigate the particle size, polydispersity index (PDI) and mean zeta potential of each sample, Dynamic Light Scattering (DLS) and Electrophoretic Light Scattering (ELS) measurements were conducted using Litesizer 500 (Anton Paar Ltd, Germany), at an automatic scattering angle.

500 μL of each sample were diluted with deionised water at pH 7.4 to a lipid concentration of 0.25 mg/mL. The samples were placed in an Ω-shaped capillary Cuvette 225288 (Anton Paar Ltd, Germany) and allowed to equilibrate to 25 ºC for 5 minutes. For each sample measurement, 3 repeats were conducted, with

the Smoluchowski–Kramers approximation utilised, and the average measurement results were reported.

The PDI was used to estimate how uniform the size of the sample is. It is defined as:

$$PDI = \frac{M_w}{M_n} \quad (eq.1)$$

Where M_w is the weight average molar mass and M_n is the number average molar mass.

Table 3.5. 1. Samples tested for size and zeta potential

Name	Composition
20E	F3 in B8 with 20 w/w% ethanol, no RNA
20E*	F3 in B8 with 20 w/w% ethanol, 5 w/w% [Cho]Cl, no RNA
20R*	F3 in B8 with 20 w/w% ethanol, 5 w/w% [Cho]Cl and RNA
30E	F3 in B8 with 30 w/w% ethanol, no RNA
30E*	F3 in B8 with 30 w/w% ethanol, 5 w/w% [Cho][DHP], no RNA
50E	F3 in B8 with 50 w/w % ethanol, no RNA
50E*	F3 in B8 with 50 w/w % ethanol, 5 w/w% [Cho]Cl, no RNA
50R*	F3 in B8 with 50 w/w% ethanol, 5 w/w% [Cho]Cl and RNA

4. Results

4.1. Mixing lipids and buffers

To judge whether a lipid-buffer combination was successful, a visual test was carried out. The mixture should remain colourless suggesting that the required vesicles were formed. Initially, no combination resulted into a colourless solution. However, variations were observed between different lipid-buffer combinations.

To begin with, F1-combinations produced the worst results giving large aggregates that were visible with a naked eye. The effect that different buffers had on F1, was observed by qualitatively observing which combination went cloudy first. F1B4 combination turned cloudy fast, followed by F1B5 formulation, F1B1 and F1B2. F1B3 was visually the best combination but still not completely colourless, as shown in Figure 4.1.1 below. The results were concerning; therefore 1 mL of pure ethanol was added to a freshly made F1. The resulting mixture was better than the buffer mixtures, however it was not completely colourless. For this reason, F1 was no longer investigated.

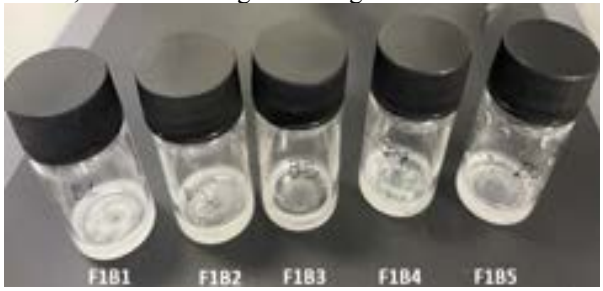


Figure 4.1.1. F1 lipid formulations combined with B1, B2, B3, B4 and B5 from left to right.

F2, F3, F4 and F5 all performed better than F1, but the solutions obtained were once again not colorless. Figure 4.1.2 compares F1, F2 and F3 combinations with B1.

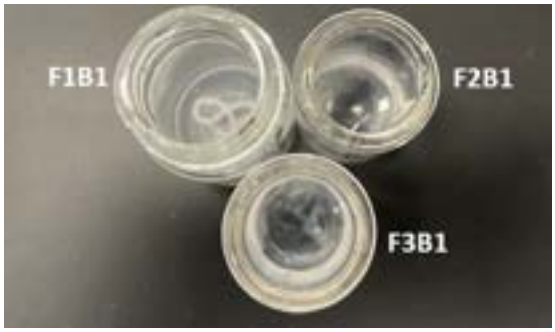


Figure 4.1. 2. Top-view image comparing combinations of B1 with F1 (top left), F2 (top right) and F3 (bottom).

From all the lipid-buffer combinations, it was judged that F3 and F4-containing samples had the best results. F3 and F4 were therefore reproduced, and 1 mL of pure ethanol was added to them. The resulting solutions were colourless, and they were used as a reference point to compare subsequent combinations. This indicated that the PEG-free F3 and F4 lipid formulations could be successfully hydrated.

Furthermore, F3B8, F3B1 and F4B8 were the best candidates, as the mixtures were closer to the colourless result wanted. The formulations were reproduced, but the ethanol concentration in the buffers was increased from 10 w/w% to 50 w/w%. From the new combinations, F4B8(50 w/w% ethanol) turned cloudy but F3B1(50 w/w% ethanol) and F3B8(50 w/w% ethanol) remained colourless. F3B8 formulation remained colourless after reducing the ethanol content in the buffer solution to 30 w/w% and 20 w/w%.

All steps mentioned above, resulted to 4 colourless combinations: F3B1(50 w/w% ethanol), F3B8(50 w/w% ethanol), F3B8(30 w/w% ethanol) and F3B8(20 w/w% ethanol) which were used in subsequent steps.

Figure 4.1. 3. 50 w/w% ethanol B8 added to F4 (left) and F3 (right). F4B8 turned cloudy whereas F3B8 remained colorless.

4.2. Transfection results

Before measuring the luminescence of the samples as described in section 3.4, the samples were placed under a microscope. It was observed that around 90% of the cells present in the LNP-containing samples were dead compared to cells-RNA combination or cells only samples.

The above observation was confirmed by quantitative results. At gain 4000, the luminescence of LNP-containing samples was only 4% compared to the luminescence of the cells-only sample, 0.04% compared to the VSVG sample and 0.004% compared to the RNA-only sample. At gain 3000, no luminescence was detected from the LNP-containing samples.

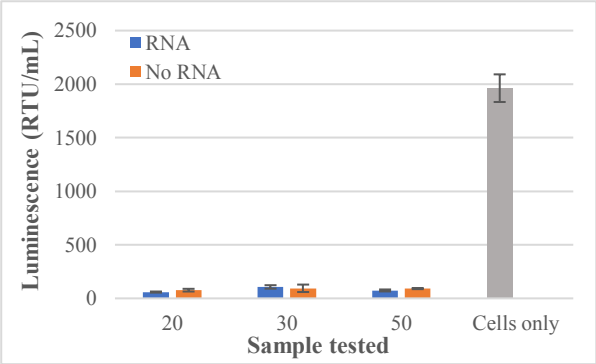


Figure 4.2. 1. Comparison of luminescence (RTU/mL) of LNP-containing samples to cells-only sample measured at gain 4000. Samples from left to right: 20R*, 20E*, 30R*, 30E*, 50R*, 50E*, cells only.

4.3. DLS & ELS results

As it can be observed from figure 4.3.1, samples without IL (20E, 30E, 50E) had more negative mean zeta potential values. The zeta potential values were more positive for samples including ILs. It is worth noting that the mean zeta potential for the 2 RNA-containing samples, were in agreement with the COVID-19 vaccine zeta potential literature data [19]. The mean zeta potential was calculated from the phase analysis light scattering (PALS) [20].

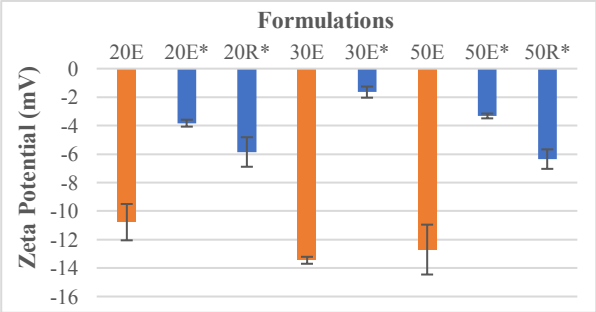


Figure 4.3.1. Zeta potential results for F3 in B8 in varying ethanol concentration, ILs and samples including or not including RNA. Blue columns indicate the samples containing IL.

The Litesizer was utilised to measure the particle size of the samples. The size of particles varied greatly between each run. Measurements were in the order of thousands of nm and resulted to large PDI values for all samples. The standard deviation was significantly larger when compared to literature data [21].

5. Discussion

5.1. Formulations of LNPs

Persistently, for all our formulations PEG was eliminated due to its toxicity to humans and the potential cause of anaphylaxis [22].

F1 failed to remain colourless upon addition of 1 mL of pure ethanol. Twisting the amount of cholesterol and DOPC used could result to formation of a better film. However, it must be kept in mind that this formulation did not include cationic or ionizable lipids and misses out on the potential advantages of using such lipids (as outlined in section 2).

F3 and F4 did remain colourless upon addition of 1 mL of pure ethanol, which was a promising result suggesting that the PEG-free lipid films produced by these formulations could successfully be hydrated forming the required vesicles. F2 and F5 were not tested by the addition of 1 mL of pure ethanol due to finite amount of lipids available for this project, however this procedure could confirm whether the two formulations could be successfully hydrated or not.

Even though F3 and F4 could yield viable LNP-formulations in-vitro, the elimination of PEG could lead to decreased in vivo performance of the LNP. For example, PEGylation protects LNPs from opsonization which reduces the uptake of LNPs from the RES [23], a problem that can occur when using cationic lipids. PEG was eliminated from the lipid formulations however it is not known what the effect of this action will be on opsonization. The in vivo performance of the PEG-free LNPs was not within the scope of this research project, however it is critical to be explored in future steps.

Problems with in vivo application can also rise due to the use of DOTAP, which can be taken up by the spleen and liver and accumulate in the vasculature [23]. While F3 and F4 succeeded at remaining colourless in ethanol, if these formulations were to be further used, the replacement of DOTAP must be considered. As seen in section 2, an ionizable lipid such as Dlin-DMA could replace DOTAP.

5.2. Decreasing ethanol content needed in buffer-lipid mixtures

The initial goal of decreasing the ethanol content required to achieve colourless buffer-lipid mixtures to 10 w/w% was not reached. The best combination was that of F3B8 (20 w/w % ethanol). Even though ethanol needed was not reduced to 10 %w/w, 4 colourless combinations were produced using buffers instead of pure ethanol, which can be considered a success.

As mentioned in section 2, it was hard to predict the exact effect each buffer would have on the lipid formulations. Most of the literature is based on stabilisation of proteins, not lipids, and even for stabilisation of proteins, the choice of excipients used was largely influenced by the protein stabilised. Exact mechanisms of interactions cannot be inferred from the results. However, it is possible to infer what worked and what did not.

The buffer formulations that produced the most colourless mixtures upon their addition to the lipid films were the buffers containing sucrose. It could therefore be extracted that sucrose worked better, under the given circumstances, in comparison to trehalose. Both sucrose and trehalose formed hydrogen-bonds with water and were preferentially excluded from the lipid vicinity. Sucrose and trehalose bind to approximately the same number of water molecules in total [24] and they were expected to have a similar effect. The fact that sucrose was present in the buffers that resulted to colourless results was attributed to the general composition of the buffers rather than the presence of sucrose instead of trehalose.

All the buffer formulations contained 68 mg/mL of sugar. The effect of sugar could be further explored by varying the amount included in the buffer formulations. Studies reported that increasing concentration of sucrose results to better protein stabilisation. The maximum concentration of sucrose used in those studies was 1 M

[25, 26], which corresponds to 342 mg/mL. Even though increasing sucrose concentration could help in the stabilisation of the lipids, making the buffer solution too concentrated may have a detrimental effect to cells during transfection. It is therefore recommended to not exceed a sucrose content of 15 w/w %.

In the 10 different buffer formulations developed, four different amino acids were used. Serine (Ser) which has a polar, uncharged R group, glycine (Gly) which is non-polar, Lys, a basic amino acid with a polar charged R-group and finally L-histidine/HCl (his/HCl), a salt of His, a basic amino acid. Since sugar and surfactant used were the same for most formulations, the choice of amino acid was what determined the results.

His/HCl was most frequently used as it is common to use salts of amino acids in buffers to maintain the pH at constant levels. Out of the six formulations including His/HCl, B1 appeared to work best for F3. All His/HCl formulations were acidic with pH values close to 3 and needed pH adjustment. Gly was only used in B8, the buffer which resulted to the mixture with the lowest ethanol concentration.

Neutral and basic amino acids appeared to result in the most promising buffers, in agreement with literature [8, 13]. Out of the basic amino acids, His is the least basic with Arg being the most. Arg is not classified as a protein-stabilising excipient that lacks affinity for proteins but was proven to be extremely effective in suppressing protein aggregation, by binding weakly to the protein interface [8]. Since lipid aggregation was the problem faced here, Arg could be the key to deriving buffer formulations with 10 w/w % ethanol, as its most basic nature can result to increased number of hydrogen bonds formed. Whether Arg would be effective or not in preventing lipid aggregation, is subject to interactions between Arg and lipid films. The use of Arg does not pose a threat for humans.

The amount of amino acid to be used was not varied in the buffers, as the main focus was to determine which amino acid worked best rather than how much amino acid worked best. The maximum amount of amino acid encountered in literature was 55 mg/mL [13]. In the case of B8 where the use of Gly was successful, the effect of Gly amount could be explored, varying it between 10 – 55 mg/mL, after the optimum amount of sucrose is determined.

Another factor that could contribute to not attaining the 10% ethanol goal could be the use of water. As mentioned in section 2, using water to dilute ethanol leads to merging of fragments to larger bundles which is unwanted. Water was used as its interaction with other excipients used was well known and it was believed that those interactions would help stabilise the lipids. However, water was not excluded from the vicinity of the lipids and its interaction with lipids was not anticipated. Water could be replaced with a neutral buffer such as phosphate buffered saline (PBS) which is commonly used in biological research.

5.3. Size and charge of particles

The size of the LNPs developed for this research project was an order of magnitude larger than it was supposed to be. The suboptimal reproducibility denoted that the large aggregates detected could not be accurately measured as their size was beyond the Litesizer detection limit of 10 μm [19].

Such large sized particles would not be used to deliver the encapsulated oligonucleotides to the target cells as they would be more susceptible to attack by macrophages once opsonized. Desirable chemical, mechanical and electrical properties for drug delivery are possessed by particles with size less than 100 nm [23].

The large PDI and size of the particles could be the result of several factors. The first factor affecting the particle size was the composition of the LNPs. Roces C. B. et al. suggested that increasing the amount of cholesterol and subsequently decreasing the amount of cationic lipid resulted to decreased size and PDI [27]. F3 was made up of only 10 mol % cholesterol. It is possible that the low percentage of cholesterol led to the above unwanted result. A new formulation based on F3 could be derived where cholesterol is increased to 50% with DOPC and DOTAP adjusted accordingly.

Since using PEG affects both the size of the particle as well as its zeta potential, it is possible that the large particles were a consequence of eliminating PEG. PEG would provide strong steric hinderance to the particles and thus stabilise them, preventing agglomeration. Studies have shown that PEG-allergic patients could be vaccinated with a vaccine containing Tween 80 as an excipient. Both aforementioned excipients positively contribute to stabilising the LNPs by acting as emulsifiers [28].

The particle size could also be affected by the microfluidic parameters used in the making of LNPs. The size of aggregates dependent on the diffusion length generated, and the diffusion length dependent on the FRR. High FRR leads to a narrow central steam and thus to the formation of many small aggregations in contrast to low FRR which would lead to a wide central stream and the formation of fewer but larger aggregates [29]. This is why an FRR of 3 instead of 1 was used.

The large PDIs observed confirm that many aggregates were formed, however the size of those aggregates was large. For this reason, it is believed that microfluidic parameters were not the main reason for particle agglomeration.

Zeta potential was measured by ELS. The Smoluchowski–Kramers approximation was selected for the measurements as it is suitable for water-based samples [20]. As a rule of thumb, colloids are considered stable when absolute zeta potential values are over 30 mV [19]. Therefore, the low zeta potential magnitudes observed, indicated the tendency of the samples to aggregate in response to cold chain disruptions. Through the zeta potential data, it has been clearly demonstrated that all the samples consisted of weakly anionic particles. Furthermore, when ILs were added to the samples, the zeta potential was more positive as the ionic strength increased and the pH decreased [13].

5.4. Transfection

The transfection results indicated that the luciferase gene was not expressed by the cells, as no luminescence was detected from the LNP-containing samples. However, around 90% of the cells were reported dead in

the LNP-containing samples, meaning that the samples were toxic to the cells.

The chemicals used in the production of LNPs were all biocompatible and not toxic. It was observed during transfection that the mixture containing the LNPs turned yellow instead of remaining red. The reduced medium Opti-MEM™ used in transfection contained phenol – red, a pH indicator which transitions from yellow for pH < 6.8 to red until pH exceeds 8.2, where it turns fuchsia [30]. As mentioned in section 3.1, the pH of buffers was adjusted to 6.5 which is lower than 6.8. The LNP-containing mixtures were too acidic for the cells and resulted to their death. It is also possible that the size of the LNPs affected the growth of the cells.

Since the cells were killed by the LNP-containing samples, it is not possible to completely assess whether the LNPs would be successful in transferring the RNA to the cell and whether the RNA transferred would be in good enough condition for the luciferase gene to be expressed. Hence, conclusions about the efficiency of RNA encapsulation could not be drawn.

5.5. Effect of Ionic Liquid

As aforementioned, the use of ILs in formulations made the zeta potential more positive due to the higher ionic strength and lower pH. In agreement with literature, acidic solutions have a more positive zeta potential value.

The decrease in absolute zeta potential indicated a greater tendency of particles to agglomerate, as the electrostatic repulsion between two particles would be decreased. The results for the size particle indicated that large agglomerates formed, however due to the suboptimal reproducibility of the results, clear conclusions on whether IL-containing samples increased or decreased the particle size could not be drawn.

The approach to use [Cho]Cl mainly instead of [Cho][DHP] in the formulations has been vindicated by the results. Samples with [Cho]Cl had greater absolute zeta potential values. ILs stabilise macromolecules due to hydrogen bonding as well as electrostatic and hydrophobic interactions. [Cho]Cl was anticipated to be able to form more bonds compared to [Cho][DHP], as the large dihydrogen phosphate chains are expected to prevent close contact of the IL with the macromolecules due to steric hinderance.

Comparing the experimental data with COVID-19 vaccine zeta potential data which was -5.3 mV, it was observed that formulations including [Cho]Cl were almost in agreement with the literature value with the best formulation being 20R* with a zeta potential of -5.8 mV.

Using a higher concentration of ILs in the formulations would have made the zeta potential values more positive therefore using 5 w/w% of ILs was deemed adequate as values obtained were similar to COVID-19 vaccines. Other ILs like choline acetate could be considered, as it is more soluble than [Cho][DHP]. Moreover, the effect on thermal stability of the formulations when varying the ILs should be explored further as certain ILs could significantly improve solubility and thermal stability of proteins in solution [21].

6. Conclusions & Outlook

Even though this research did not manage to produce PEG-free LNPs by using 10 w/w% ethanol, with the desired characteristics, the results were promising. A serious step was taken into exploring the potential of stabilising LNPs with excipients and ILs to decrease the amount of ethanol needed, while also aiming to tackle the problems associated with the presence of PEG.

Especially the fact that F3 and F4 formulations remained colourless upon addition of pure ethanol can be considered a success, as well as the fact that F3B1(50 w/w % ethanol), F3B8(50 w/w % ethanol), F3B8(30 w/w% ethanol), and F3B8(20 w/w % ethanol) combinations remained colourless. These results indicate that the hydration of PEG-free lipid films by using a buffer is possible.

The major problem encountered in the results was the formation of large aggregates by all the LNP samples. However, it appears that a few modifications to the buffers and lipid formulations could be the key for future success.

Increasing the amount of cholesterol to up to 50% in the lipid formulations and decreasing the amount of the cationic lipid should be one of the first modifications to be made. In addition, steric hinderance effect of tween 80 should be explored, by using it as an excipient, to efficiently replace the advantages associated with PEG and size.

Moreover, varying the amount of sucrose in the buffer formulations, testing the effect of Arg and replacing water with PBS could be the changes made to the buffer formulations to ensure that upon addition to the lipid film, the mixture would remain colourless at lower percentages of ethanol.

The use of ILs, especially [Cho]Cl could play an important role in stabilising the LNPs and affecting their thermal stability. Due to the low cost of [Cho]Cl it is worth to be considered as an excipient in buffer formulations.

If following the modifications outlined above results to LNPs with size less than 100 nm, this project can serve as the foundation to a new era in the creation of LNPs, allowing their broader and safer application. For successful candidates, in situ drug delivery must be explored with a series of model drug candidates.

Currently, limited information is available on the effect of excipients on lipids and LNPs as well as what excipients can be used to effectively stabilise LNPs and prevent their aggregation. Further research should be carried out to determine the excipients that work best as well as give more information to the mechanisms used in the interaction between excipients and LNPs.

Production of LNPs without the use of ethanol would mean that no additional buffer will be wasted to wash the LNPs. This would result to cost saving, as the processing steps required would be decreased. To conclude, PEG-free LNPs produced using buffers can lead to the formation of new and improved sustainable vaccines.

7. Acknowledgements
The authors would like to thank Dr Talia Shmool for her guidance and support during the project as well as Colin Pilkington for training and helping with the production of LNPs using microfluidics. We would also like to thank Dr Valarmathy Murugaiah for her help during transfection.

8. References

1. Reichmuth, A.M. et al. (2016) MRNA vaccine delivery using lipid nanoparticles, Therapeutic delivery. U.S. National Library of Medicine. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5439223/ (Accessed: November 16, 2022).
2. Chira, S. et al. (2015) Progresses towards safe and efficient gene therapy vectors, Oncotarget. U.S. National Library of Medicine. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4741561/?report=reader (Accessed: November 16, 2022).
3. Cuthbertson, C.R. and Parsey, M.A.A. (2022) Intro to lipid nanoparticle formulation, caymanchem.com. Available at: https://www.caymanchem.com/news/intro-to-lipid-nanoparticle-formulation#:~:text=LNPs%20are%20dialyzed%20into%20a,7.4%2C%20for%20storage%20and%20use. (Accessed: December 9, 2022).
4. Alameh, M.-G. et al. (2021) Lipid nanoparticles enhance the efficacy of mrna and protein subunit vaccines by inducing robust T follicular helper cell and humoral responses, Immunity. U.S. National Library of Medicine. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8566475/ (Accessed: December 11, 2022).
5. Li, W. and Szoka, F.C. (2007) Lipid-based nanoparticles for Nucleic Acid Delivery - Pharmaceutical Research, SpringerLink. Springer US. Available at: https://link.springer.com/article/10.1007/s11095-006-9180-5 (Accessed: November 18, 2022).
6. Hou, X. et al. (2021) Lipid nanoparticles for mRNA delivery, Nature News. Nature Publishing Group. Available at: https://www.nature.com/articles/s41578-021-00358-0 (Accessed: November 18, 2022).
7. Padín-González, E. et al. (1AD) Understanding the role and impact of poly (ethylene glycol) (PEG) on nanoparticle formulation: Implications for covid-19 vaccines, Frontiers. Frontiers. Available at: https://www.frontiersin.org/articles/10.3389/fbioe.2022.882363/full (Accessed: November 19, 2022).
8. Santos, N.D. et al. (2007) Influence of polyethylene glycol grafting density and polymer length on liposomes: Relating plasma circulation lifetimes to protein binding, Biochimica et Biophysica Acta (BBA) - Biomembranes. Elsevier. Available at: https://www.sciencedirect.com/science/article/pii/S0005273606004901?via%3Dihub (Accessed: November 19, 2022).
9. Ohtake, S., Kita, Y. and Arakawa, T. (2011) Interactions of formulation excipients with proteins in solution and in the dried state, Advanced Drug Delivery Reviews. Elsevier. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0169409X11001852?via%3Dihub (Accessed: November 19, 2022).
10. KALLMEYER, G. et al. (2014) "Stable lyophilized pharmaceutical preparations of monoclonal or polyclonal antibodies."
11. Carpenter, J.F. et al. (1993) "Method for stabilisation of biomaterials."
12. Shmool, T.A. et al. (2022) Ionic Liquid-Based Strategy for Predicting Protein Aggregation Propensity and Thermodynamic Stability, ACS Publications. Available at: https://pubs.acs.org/doi/10.1021/jacsau.2c00356 (Accessed: November 23, 2022).
13. Shmool, T.A. et al. (2021) An experimental approach probing the conformational transitions and energy landscape of antibodies: A glimmer of hope for reviving lost therapeutic candidates using Ionic Liquid, Chemical Science. The Royal Society of Chemistry. Available at: https://pubs.rsc.org/en/content/articlelanding/2021/SC/D1SC02520A (Accessed: November 23, 2022).
14. Massant, J. et al. (2020) Formulating monoclonal antibodies as powders for reconstitution at high concentration using spray-drying: Trehalose/ amino acid combinations as reconstitution time reducing and stability improving formulations, European Journal of Pharmaceutics and Biopharmaceutics. Elsevier. Available at: https://www.sciencedirect.com/science/article/pii/S0939641120302630 (Accessed: December 3, 2022).
15. Kondo, M. and Inoue, K. (1988) "Stabilizing method for monoclonal antibody."
16. Hald Albertsen, C. et al. (2022) The role of lipid components in lipid nanoparticles for vaccines and gene therapy, Advanced drug delivery reviews. U.S. National Library of Medicine. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9250827/ (Accessed: December 4, 2022).
17. Sun, M. et al. (2022) Optimization of DOTAP/chol cationic lipid nanoparticles for mrna, pdna, and oligonucleotide delivery - AAPS pharmscitech, Springer-Link. Springer International Publishing. Available at: https://link.springer.com/article/10.1208/s12249-022-02294-w#:~:text=In%20fact%2C%20LNPs%20are%20currently,tested%20in%20numerous%20clinical%20trials. (Accessed: December 4, 2022).
18. Smale, S.T. (2010) Luciferase assay, Cold Spring Harbor protocols. U.S. National Library of Medicine. Available at: https://pubmed.ncbi.nlm.nih.gov/20439408/ (Accessed: December 9, 2022).
19. Giving it its best shot: Quality Control of antiviral vaccines with the Litesizer (2022). Available at: https://www.theengineer.co.uk/media/na5hv3mv/anton-paar-antiviral-vaccines.pdf (Accessed: December 11, 2022).
20. Santner, J. (March, 2016) Litesizer 500 instruction manual - University of Toledo. Available at: https://www.utoledo.edu/nsm/ic/pdfs/Litesizer%20

500%20Instruction%20Manual%20.pdf (Accessed: December 11, 2022).
21. Shmool, T.A. et al. (2022) Next generation strategy for tuning the thermoresponsive properties of micellar and hydrogel drug delivery vehicles using Ionic liquids, Polymer Chemistry. The Royal Society of Chemistry. Available at: https://pubs.rsc.org/en/content/articlelanding/2022/py/d2py00053a#fn1 (Accessed: December 13, 2022).
22. Sellaturay, P. et al. (2021) Polyethylene Glycol (PEG) is a cause of anaphylaxis to the Pfizer/Biontech mRNA COVID-19 vaccine, Clinical and experimental allergy: journal of the British Society for Allergy and Clinical Immunology. U.S. National Library of Medicine. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8251011/ (Accessed: December 11, 2022).
23. Sharmap, A., Madhunapantula, S.R.V. and Robertson, G.P. (2011) Toxicological considerations when creating nanoparticle-based drugs and Drug Delivery Systems, Expert opinion on drug metabolism & toxicology. U.S. National Library of Medicine. Available at: https://pubmed.ncbi.nlm.nih.gov/22097965/ (Accessed: December 7, 2022).
24. Olsson, C. and Swenson, J. (2020) Structural Comparison between Sucrose and Trehalose in Aqueous Solution, The Journal of Physical Chemistry B. Available at: https://pubs.acs.org/doi/10.1021/acs.jpcb.9b09701 (Accessed: December 8, 2022).
25. Oshima, H. and Kinoshita, M. (2013) Effects of sugars on the thermal stability of a protein, AIP Publishing. American Institute of PhysicsAIP. Available at: https://aip.scitation.org/doi/full/10.1063/1.4811287 (Accessed: December 7, 2022).
26. Lee, J.C. and Timasheff, S.N. (1981) The stabilization of proteins by sucrose, The Journal of biological chemistry. U.S. National Library of Medicine. Available at: https://pubmed.ncbi.nlm.nih.gov/7251592/ (Accessed: December 7, 2022).
27. Roces, C.B. et al. (2020) Manufacturing considerations for the development of lipid nanoparticles using Microfluidics, Pharmaceutics. U.S. National Library of Medicine. Available at: https://pubmed.ncbi.nlm.nih.gov/33203082/ (Accessed: December 9, 2022).
28. Sellaturay, P. (2022) The polysorbate containing AstraZeneca covid-19 vaccine is tolerated by polyethylene glycol (PEG) allergic patients, Clinical and experimental allergy: journal of the British Society for Allergy and Clinical Immunology. U.S. National Library of Medicine. Available at: https://pubmed.ncbi.nlm.nih.gov/34822190/ (Accessed: December 13, 2022).
29. Golden, J.P. et al. (2011) Hydrodynamic focusing--a versatile tool, Analytical and bioanalytical chemistry. U.S. National Library of Medicine. Available at: https://pubmed.ncbi.nlm.nih.gov/21952728/ (Accessed: December 12, 2022).
30. Henn, A. (2021) Temperature, CO2, and pH in cell culture media, BioSpherix. Available at: https://biospherix.com/411-temperature-co2-and-ph-in-cell-culture-media/#:~:text=Contamination%20by%20fast%2Dgrowing%20bacteria,cell%20medium%20also%20turns%20fuchsia. (Accessed: December 9, 2022).

Kinetic modelling of guaiacol hydrogenation with in-situ hydrogen production by glycerol aqueous reforming over $NiSn/Al_2O_3$ catalyst

Zeshu Liu and Yuyao Feng

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

Abstract

Biomass, typically lignin, was expected to be valuable source for renewable biofuels. As lignin have high oxygen content. Upgrading needs to be utilized by hydrotreating. Due to its high concentration in bio-oil and two oxygenated functional groups, guaiacol was selected as a bio-oil model and the hydrogenation of guaiacol was investigated. The conversion of guaiacol and concentration of products were obtained at 230 ◦C, 250 ◦C, and 270 ◦C and the kinetic energy and activation energy for each reaction were calculated. The activation energies of guaiacol to phenol, phenol to cyclohexanone, and phenol to cyclohexanol were found to be 113 kJ/mol, 81.0 kJ/mol, 30.5 kJ/mol respectively.

1. Introduction

The annual increase rate of global consumption of primary energy is about 1.5% from 2007 to 2017 and reaches 2.9% in 2018 according to BP's statistical review (BP, 2019).

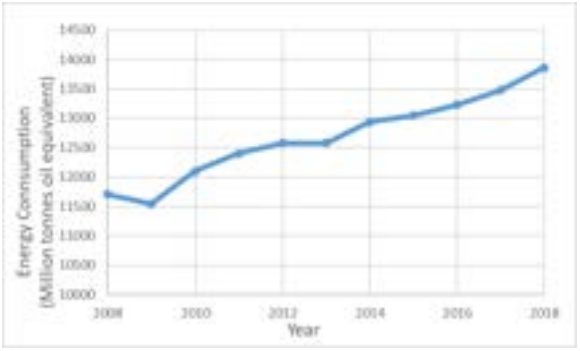


Figure 1. Global consumption of primary energy from 2008 to 2018 adapted from BP's statistics (BP, 2019)

In 2018, the largest source of primary energy is oil taking 33.6%. The second and third largest source is coal and gas taking 27.2% and 23.9% respectively. Clean energies such as nuclear energy, hydroelectricity only takes 15.3%.

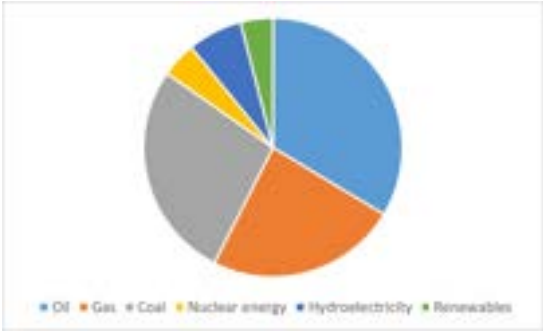


Figure 2. Percentage of global primary energy consumption in 2018 adapted from BP's statistics (BP, 2019)

As the percentage of clean energy is still relatively low, an important topic should be taken up which is the environment issue caused by carbon emission. According to BP's statistics, the carbon emission caused by combusting oil, gas and coal reaches 33890.8 million tonnes in 2018 and is still increasing at a rate of 2% (BP, 2019). As global warming has become an important topic, many governments have proposed carbon emission tax aiming to decrease carbon dioxide emission. In UK, the carbon tax reaches £78/tonne in 2022 (Bridget Beals, 2022).

Considering both environment and economic effects, renewable fuel sources are demanded

to replace traditional fuel sources. As the only renewable organic carbon source in nature, biomass was expected to take important role in the production of renewable biofuels (Chen, 2020).

Biomass composes lignin, cellulose, and hemicellulose. Lignin is the only renewable resource for producing aromatic compound. Nevertheless, paper industries produce a massive amount of lignin every year and 98% of them were directly burned at the same factory (Calvo-Flores and Dobado, 2010). Therefore, it would be worth investigating the utilization of lignin.

2. Background

Two main processes for producing bio-oil from lignin are pyrolysis and hydrothermal liquefaction (Chen, 2020). The pyrolysis of lignin at 450−600 °C and atmospheric pressure was shown to be an economic way to produce bio-oil (Mu et al., 2013).

Bio-oils are organic liquids containing highly oxygenated compounds. Comparing to petroleum derived oil, bio-oils was undesired as fuel source due to its high water content, high viscosity, high acidity, high ash content and, most of all, low heating value (Mortensen et al., 2011). Therefore, upgrading of the bio-oil is required. The bio-oil was upgraded by hydrotreating, catalytic cracking, hydrocracking, supercritical fluids, esterification, emulsion, or extraction (Chen, 2020).

In previous research, guaiacol formed from decomposition of lignin was chosen as the model compound due to its high concentration in bio-oil and the presence of two oxygenated functional groups hydroxy (Csp2OH) and methoxy (Csp2OCH3) (Chen, 2020). To improve the performance of guaiacol as fuel, hydrodeoxygenation was implemented. It was found that comparing to external hydrogen source, hydrogenation of guaiacol with internal hydrogen leads to a higher conversion of guaiacol. Feng et al. found that phenolic compounds can be converted at a yield of 98.22wt% with methanol as liquid hydrogen donor and Raney Ni as catalyst (Feng et al., 2017). Yu et al. shows that depolymerization products of lignin can be converted with nearly 90% selectivity towards cyclohexanol under the same conditions (Yoshikawa et al., 2013). Moreover, some research shows that glycerol, considered as potential renewable hydrogen source, can be used as a hydrogen source through aqueous phase reforming shown below (Putra et al., 2018).

$$C_3H_8O_3 + 3H_2O \rightarrow 7H_2 + 3CO_2 \quad (1)$$

(General equation)

$$C_3H_8O_3 \rightarrow 4H_2 + 3CO \quad (2)$$

(Decomposition of glycerol)

$$CO + H_2O \leftrightarrow H_2 + CO_2 \quad (3)$$

(Water-gas shift)

In the previous work of our lab, hydrogenation of guaiacol with in-situ hydrogen produced by glycerol was implemented with Ni/Al_2O_3 or $Ni - X/Al_2O_3$ (X= Cu, Mo, P) as catalyst (Z. Chen et al., 2020). The glycerol conversion for all catalysts were found to be close to 100%, and the reaction pathway of hydrogenation of guaiacol was proposed as in figure 3.

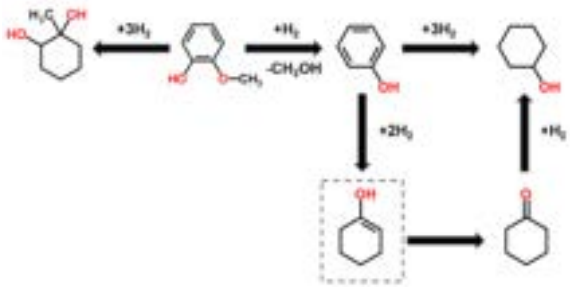


Figure 3. Proposed reaction pathway for hydrogenation of guaiacol with in-situ generated hydrogen from previous research (Z. Chen et al., 2020)

Though in many situations promoted catalyst tend to improve performance of catalyst, the result of previous research shows that all promoted catalysts in fact showed less activity comparing to Ni/Al_2O_3, where the guaiacol conversion decreases from 95% to 50%. Moreover, Ni/Al_2O_3 as catalyst tends to have higher selectivity towards cyclohexanone and cyclohexanol.

Although previous work shows that promoters are likely reducing the activity of Ni/Al_2O_3 in in-situ guaiacol hydrogenation. It was suggested that Sn as promoter may improve performance of Ni/Al_2O_3 in other areas (Reangchim et al., 2019). Therefore, in our

experiment, $Ni - Sn/Al_2O_3$ was decided to be used. As changing catalyst may affect the reaction pathway. The reaction pathway needs to be redetermined.

3. Methods

3.1. Catalyst preparation

As previously mentioned, several promoters were used without any apparent improvement in guaiacol hydro-treating. A new catalyst was used in the experiment ($Ni - Sn/Al_2O_3$). The weight fraction of the components in catalyst was shown in table 1.

Sn	Ni	Al_2O_3
2%	20%	78%

Table 1. weight fraction components in catalyst

To synthesize the catalyst, 10 grams of Al_2O_3, 12.7 grams of $Ni(NO_3)_2 \cdot 6H_2O$ and 0.487 grams of $SnCl_2 \cdot 2H_2O$ were used in the new catalyst preparation. Several steps were taken in the production.

1)Three chemical ingredients were dissolved in an ethanol solution

2)An oil bath (373K-393K) was used to evaporate the ethanol solvent.

3)The solid experienced calcination to remove water hydration, Nitrate, and chloride.

4)Product solids $NiO - SnO/Al_2Cl_3$ were cracked and filtrated to 20 micrometres diameter particles.

5)The particles were reduced in the oven at 823K by hydrogen in four hours for completing reduction and the catalyst $Ni - Sn/Al_2O_3$ was formed

3.2. Experiment

3.2.1 ***Layout of the experimental equipment***

The reactor consists of two parts: a reaction vessel and a lid with several valves, a thermocouple, a magnetic drive, and a pressure gauge. Additionally, cooling water and insulation asbestos protection were added to the magnetic drive and inlet-outlet valves which could remove the excessive heat from the reaction and prevent the vibration.

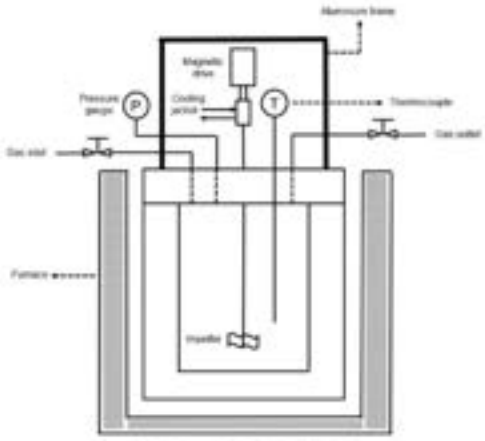


Figure 4. Layout of the same stirred batch reactor from previous research (Chen, 2020)

3.2.2 Reaction conditions & sample extraction

3.2.2.1 Temperature setup

Experiments took part in three different temperatures from 230°C to 270°C to investigate the influence of temperature on the evolution of products and estimate the activation energy of the reaction. Three temperature points were selected: 230°C, 250°C and 270°C. The temperature of the reaction was controlled by adjusting the temperature of the furnace. The temperature of the furnace is higher than the temperature in the reactor and the temperature of the furnace could not be controlled accurately as the result the temperature of the reactor would fluctuate about ± 5°C. 20°C temperature interval would be a suitable difference to study. In the heating up process, after the warming up of the furnace, the initial set-up point of the furnace was 725°C. When the temperature of the vessel reached 160°C, changed the set point to 450°C manually and dropped the reactor part into the furnace.

3.2.2.2 Reactants preparation

As mentioned in the previous section. Glycerol acted as the hydrogen provider for the hydrotreating of guaiacol. To provide sufficient hydrogen in the reaction, an overdose of glycerol was added to the reactor. A glycerol and guaiacol mixed solution was used in the experiment with 1: 2 weight fraction. 1.895ml solution was pumped into the reactor with 0.1611M concentration.

3.2.2.3 Sample extraction

The concentrations of each component at different times during the reaction were detected to investigate the reaction order and kinetic constant of each reaction. 5 samples (0min, 30min, 60min, 90min, 120min) were extracted from the valves of reactor in 230°C and 250°C experiments. Additionally higher reacting temperatures result in more dramatic reactions. For 270°C experiment, extra two samples were collected at 15min and 45min to detect the increasing and decreasing trend precisely. After the sample collection from the reactor, 0.75 grams of sample are extracted by 75ml chloroform to extract all reactants and products, the chloroform solution samples were then used for further analysis.

3.2.3 Method of concentration detection

Gas chromatography-mass spectrometry (GC-MS) analytical method was used in the concentration detection. Different organic compounds would illustrate different heights of peaks at different times (retention time). Firstly GC-MS would test products and reactants by setting concentration (0.01M,0.02M,0.03M,0.04M, 0.08M,0.12M, 0.16M). The area of each peak would be calculated and the calibration line between area and concentration would be plotted respectively.

To confirm the accuracy of the result, flame ionization detection (GC-FID) analytical method was also used to detect the concentration for the experiment at 250 °C. Calibration solutions were prepared at 0.04M, 0.08M, 0.12M, and 0.16M.

4. Results

4.1. Catalyst Recovery

The catalyst recovery is shown in table 2.

Reaction	230 °C	250 °C	270 °C
Mass of Catalyst before reaction (g)	1.5005	1.5007	1.5002
Mass of Catalyst recovered (g)	1.4962	1.4719	1.2136
Percentage recovered	99.7%	98.1%	80.9%

Table 2. Catalyst recovery

The recovery of catalyst for experiment at 230 °C and 250 °C is very high. Nevertheless, the recovery is relatively low for the experiment at 270 °C.

4.2. Concentration calibration

The chromatograms from GC-FID and GC-MS were taken into Openchrome for further inspection. The peaks and peak areas were determined by using built-in functions of Openchrome. Retention time for samples was recorded for determining the identity. The concentration of calibration samples and corresponding peak areas were taken into linear regression.

Figure 5. shows an example of calibration curve for GC-MS.

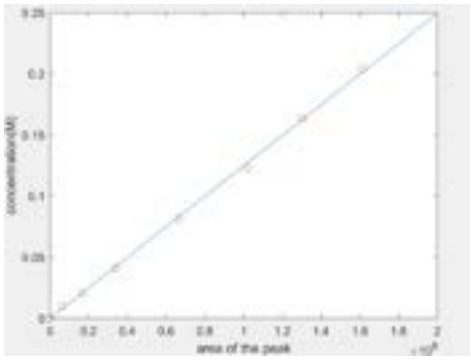


Figure 5. Calibration curve of guaiacol for GC-MS

4.3. Concentration and conversion

The chromatograms of chloroform solution samples were also taken into Openchrome for peak area calculation. However, due to the presence of negative peaks in some of the chromatograms, the auto calculation of Openchrome was inaccurate. Therefore, the chromatograms were exported into excel sheets and Matlab is used to integrated peak areas where the peaks are manually determined. The retention time for each peak is used to determine the identity of compound and the peak area was taken into the calibration curve to calculate concentration. All concentrations were normalized to ensure the law of conservation of mass.

Conversion of reaction (X_i) in batch reactor follow the equation:

$$X_i = 1 - \frac{n_i(t)}{n_i(t=0)} \quad (4)$$

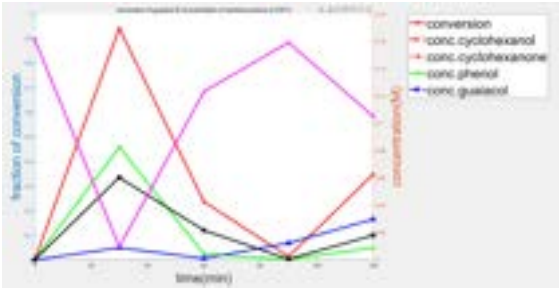


Figure 6. Conversion of guaiacol & concentration of reactant, product at 230°C

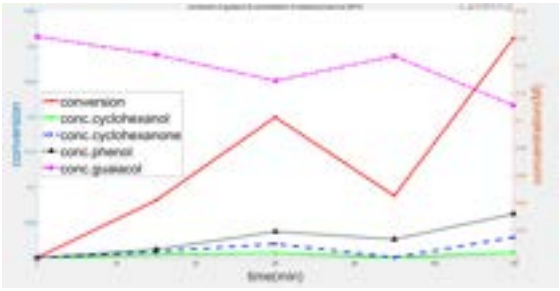


Figure 7. Conversion of guaiacol & concentration of reactant, product at 250°C (GC-MS)

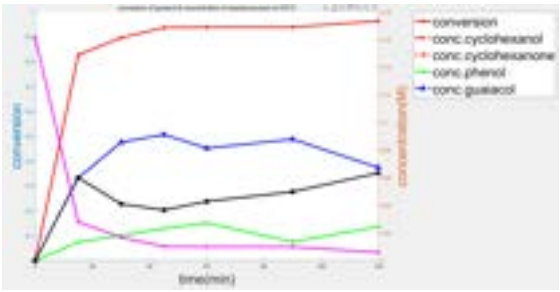


Figure 8. Conversion of guaiacol & concentration of reactant, product at 270°C (GC-MS)

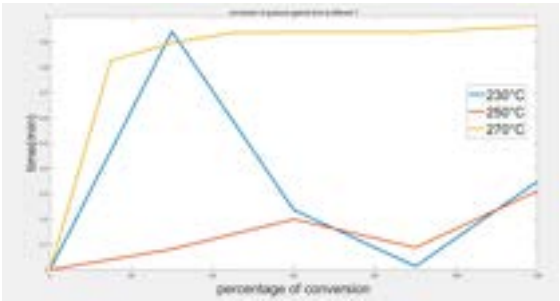


Figure 9. Conversion of guaiacol at different T (GC-MS)

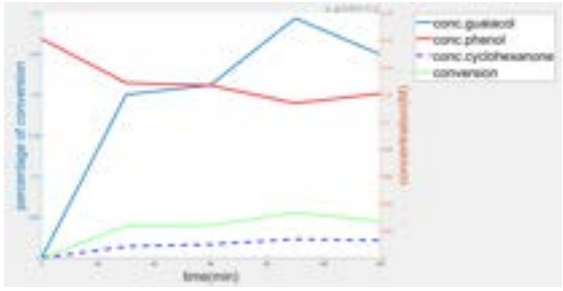


Figure 10. Conversion of guaiacol & concentration of reactant and product at 250°C (GC-FID)

Great number of the results illustrate many fluctuations which does not follow the principles of chemical reactions.

Moreover, results from GC-FID shows that no cyclohexanol was produced. In contrast, results from GC-MS emphasize that both cyclohexanol and cyclohexanone were produced.

5. Discussion

5.1. Catalyst recovery

In previous research using Ni/Al_2O_3 or $Ni-X/Al_2O_3$ (X= Cu, Mo, P) as catalyst, changes in catalyst structures occur after reaction, leading to deactivation of catalyst; Chen et al. illustrates that coke yield is about 7wt% (Z. Chen et al., 2020). The coke formation is mainly caused by poisoning of the active site and/or to pore blockage (Guisnet and Magnoux, 2001). This is undesirable in industrial production. Therefore, it is important to find ways to limit coke formation and regenerate catalyst. Table 2. from the results section shows that instead of gaining mass (forming coke), our catalyst is in fact remaining near to its original weight, especially for catalyst used in 230 °C and 250 °C experiments. The relatively great loss in weight for catalyst used in 270 °C experiment could be caused by measuring errors. However, it is worth to rerun the experiment at 270 °C to investigate reason for the great loss. The result from 230 °C and 250 °C implies that $Ni-Sn/Al_2O_3$ may be less deactivated than Ni/Al_2O_3 and its promoted catalysts. However, due to time limitations, investigations on structure of reformed catalysts were not implemented. Therefore, the conclusion is still unsure and needs further investigation for an accurate result.

5.2. Modelling of reaction system

5.2.1. Reaction Pathway

As previously mentioned, changing catalyst may lead to different reaction pathway. Therefore, the reaction pathway was reinvestigated using the concentrations of reactants and products. Results from GC-FID shows that only three compounds are present in the solution: guaiacol, phenol and cyclohexanone. The reaction pathway was proposed in figure 11.

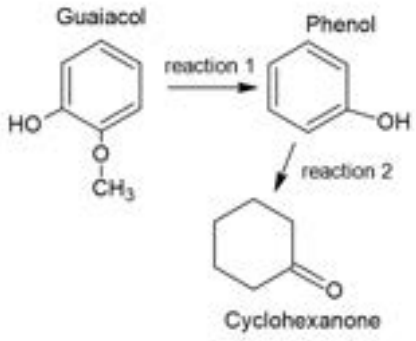


Figure 11. Reaction pathway proposed by GC-FID data analysis

However, data analysis using GC-MS is somehow different, the reaction pathway proposed by using the results from GC-MS is shown in figure 12.

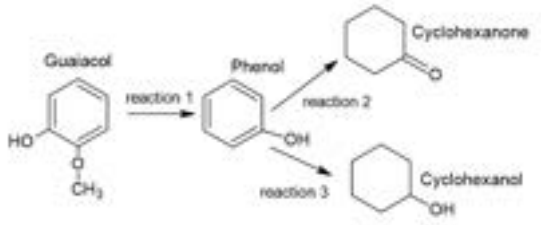


Figure 12. Reaction pathway proposed by GC-MS data analysis

$$Guaiacol \rightarrow Phenol\ (reaction\ 1)\ (5)$$

$$Phenol \rightarrow Cyclohexanone\ (reaction\ 2)\ (6)$$

$$Phenol \rightarrow Cyclohexanol\ (reaction\ 3)\ (7)$$

In the operation of GC-FID, the high purity carrier gas ran out and low purity carrier gas was used instead. This leads to a higher background noise to the chromatogram and may cause worse separation between cyclohexanol's peak and cyclohexanone's peak. Therefore, the reaction pathway in figure 12 was chosen to be the final proposed reaction pathway.

5.2.2. Reaction Kinetic

	0^{TH} order	First order	Second order
Differential rate law for reactant A	$rate = -\frac{\Delta[A]}{\Delta t} = k$	$rate = -\frac{\Delta[A]}{\Delta t} = k[A]$	$rate = -\frac{\Delta[A]}{\Delta t} = k[A]\text{^}2$
Integrated rate law for reactant A	$[A] = [A]_0 - kt$	$[A] = [A]_0 e^{-kt}$ $\ln[A] = ln[A]_0 - kt$	$\frac{1}{[A]} = \frac{1}{[A]_0} + kt$

Table 3. Reaction order equation

Determination of each reaction order in the whole reaction system is necessary for developing the reaction model. Additionally, the kinetic constant (k) could be calculated after determining the reaction order. In the reaction system, the 0^{th},1^{st} and 2^{nd} orders of reaction were mainly focused on. The equation of rate against time were shown above in the table. In terms of concentration vs. time graph, 1^{st} order and 2^{nd} order both exhibit an exponential increase and decrease for product and reactant respectively which was hard to distinguish two orders. As the result, the Integrated rate law was used and three orders would illustrate a straight line in concentration vs. time, ln [concentration] vs. time and 1/[concentration] vs. time separately. Furthermore, the absolute values of the gradient of the line in those three graphs are equal to the kinetic value.

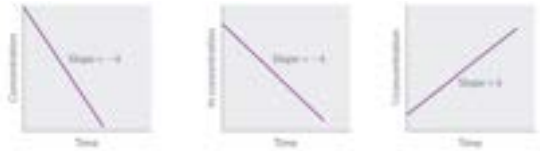


Figure 13. Straight plot to determine rate constant

As the result part mentioned, there are some errors in experimental data that would act as a distraction for determining the reaction order and kinetic constant. Jongerius et al. (Jongerius et al., 2013), Chen et al. (C. Chen et al., 2020) and Zhou (Zhou et al., 2017) illustrated the conversion of guaiacol, concentration of cyclohexanol and concentration of cyclohexanone against time separately. All three graphs exhibited a steady ascent trend. In terms of the experiment at 250°C, all concentrations detected at 90min are inaccurate which makes the R-square value for linearization much smaller than 1.

R-square value	0th order	1st order	2nd order
Reaction 1	0.675	0.666	0.656
Reaction 2	0.385	0.001	0.051
Reaction 3	0.163	0.011	0.050

Table 4. R-square value of linearization for order determination

R-square value (Without data at 90min)	0th order	1st order	2nd order
Reaction 1	0.975	0.985	0.991
Reaction 2	0.954	0.851	0.750
Reaction 3	0.651	0.749	0.713

Table 5. R-square (without data at 90min) value of linearization for order determination

We could determine the order of reactions the first-order formation of cyclohexanol, 0th order formation of cyclohexanone and the second-order reaction of guaiacol. The graphs shown below also exhibited how the data points at 90min distracted the linear regression. The reason for the error data points is the operational error in sample extraction which organic samples were evaporated by the high temperature in the vessel.

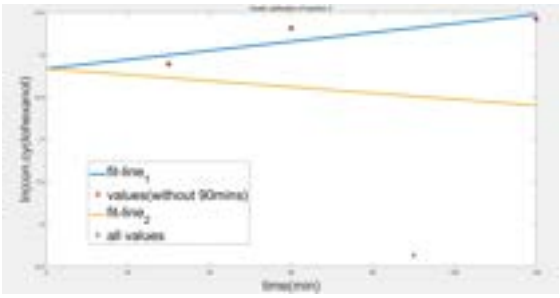


Figure 14. Effect on the best fit line for kinetic analyzation of Reaction 3 by the distractive point

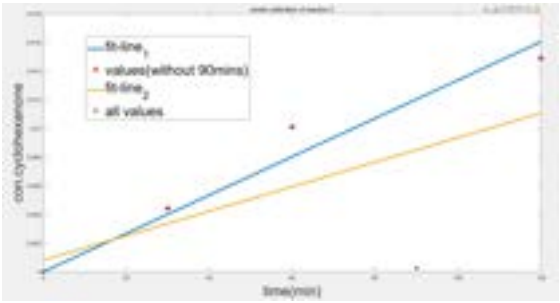


Figure 15. Effect on the best fit line for kinetic analyzation of Reaction 2 by the distractive point

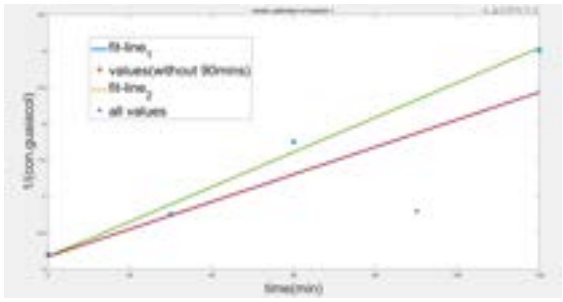


Figure 16. Effect on the best fit line for kinetic analyzation of Reaction 1 by the distractive point

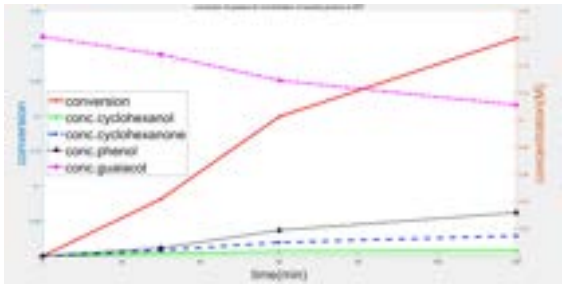


Figure 17. Conversion of guaiacol & concentration of all components vs. time with chosen data sets at 250°C

Following the same method and principle of data analysis, the data used at 270°C were shown in the table below.

	Time point used (min)
Concentration of cyclohexanol	0, 15, 30, 45, 60
Concentration of cyclohexanone	0, 15, 30, 45
Concentration of phenol	0, 15, 30, 45, 60, 120
Concentration of guaiacol	0, 15, 30, 45, 60, 120

Table 6. Selected time of data sets for analyzation at 270°C

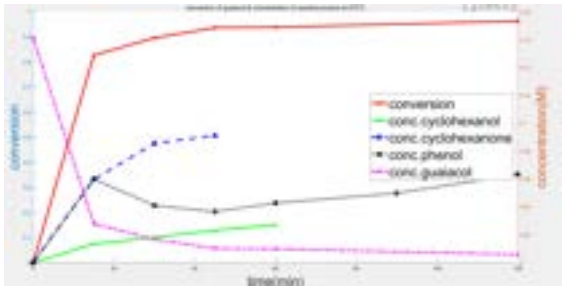


Figure 18. Conversion of guaiacol & concentration of all components vs. time with chosen data sets at 270°C

The error of the data set at 60min is the missing concentration of cyclohexanone which is because of the relatively low boiling point of cyclohexanone(155.6°C) compared with other

organics. Only the cyclohexanone be evaporated at 60min. Furthermore, the cyclohexanone and cyclohexanol were both evaporated at 120min.

	Time point used (min)
Concentration of cyclohexanol	0, 90, 120
Concentration of cyclohexanone	0, 60, 120
Concentration of phenol	0, 60, 90, 120
Concentration of guaiacol	0, 60, 120

Table 7. Selected time of data sets for analyzation at 270°C

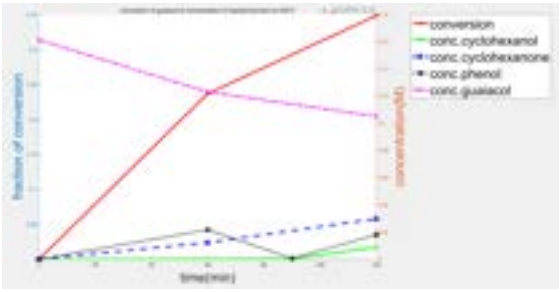


Figure 19. Conversion of guaiacol & concentration of all components vs. time with chosen data sets at 230°C

As shown in figure 6, for results from 230°C, there are great many inaccurate points which caused fluctuations in the conversion of guaiacol and the concentrations of all components.

The data used at 230°C were shown in table 7 above follow by comparing with the results from previous literature (C. Chen et al., 2020; Jongerius et al., 2013; Zhou et al., 2017). Data points were chosen to match with the suggested compound concentration over time to make our analysis more reasonable. However, due to the number of data points we dismissed, the analysis is very inaccurate and needs data from further experiments to improve the accuracy. The kinetic data of 230°C was also found to be fraud later when calculating activation energy.

	Kinetic (M^-1s^-1)		
Experiment	Reaction 1	Reaction 2	Reaction 3
230 °C	4.6×10^{-4}	4.0×10^{-6}	1.2×10^{-3}
250 °C	4.0×10^{-4}	2.2×10^{-6}	8.9×10^{-5}
270 °C	2.2×10^{-2}	4.0×10^{-5}	2.6×10^{-4}

Table 8. Kinetic energy of reactions for all three experiments

5.2.3. Activation Energy

Equation 8 was used to calculate the activation energy

$$lnk = -\frac{E_a}{R}\frac{1}{T} + C \quad (8)$$

Where k is the kinetic of corresponding reaction, E_a is the activation energy, R is the molar gas constant, T is the temperature, and C is constant. The curve of lnk against $\frac{1}{T}$ is plotted.

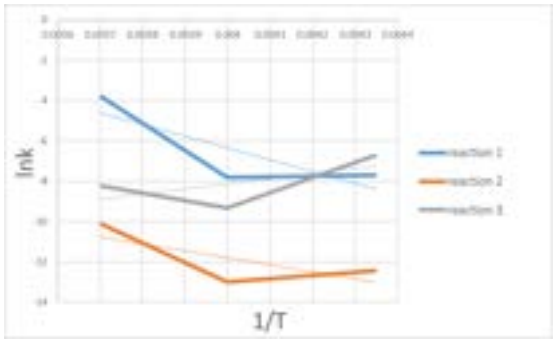


Figure 20. Curve of lnk against $\frac{1}{T}$ for all three reactions

Linear regression curves were used to predict the gradient $-\frac{E_a}{R}$ of each reaction curve, the gradient was then used to calculate activation energy.

$$E_a = -R \times gradient \quad (9)$$

Reaction	1	2	3
Gradient	1.9346	-3432.5	2536.2
Activation Energy (kJ/mol)	-0.016	28.5	-21.1

Table 9. Gradient and activation energy of each reaction

The negative activation energy for reaction 1 and reaction 3 brings concerns as this means rate of reaction decreases with the increase of temperature which does not match with our observation. Gao et al. illustrates that the activation energy of reaction 1 while using platinum as catalyst should be around 99.8 kJ/mol (Gao et al., 2015), which is far from

our result. Moreover, activation energy of phenol to cyclohexanol or cyclohexanone by thermal hydrogenation is around 33kJ/mol (Song et al., 2016). The data points were investigated and founding that data points from 230 °C experiment are taking major role for the unrealistic results. The curve was replotted without using data points from 230 °C.

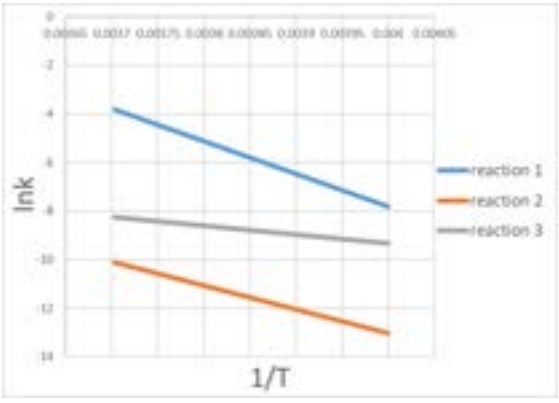


Figure 21. Curve of lnk against $\frac{1}{T}$ for all three reactions without 230C

Reaction	1	2	3
Gradient	-13547	-9746.2	-3673.2
Activation Energy (kJ/mol)	113	81.0	30.5

Table 10. Gradient and activation energy of each reaction without 230C

Without points from 230 °C, the calculated activation energy matches to the one in literature. Due to time restrictions the 230 °C experiment was not reattempted. In future, the experiments need to be redone for better accuracy.

6. Conclusion

A Sn promoted Ni catalyst was applied to the hydrogenation of guaiacol with in-situ generated hydrogen by glycerol aqueous phase reforming. The kinetic energy and activation energy of each reaction was found. By excluding data points from 230°C, the activation energy of guaiacol to phenol, phenol to cyclohexanone, and phenol to cyclohexanol were found to be 113 kJ/mol, 81.0 kJ/mol, 30.5 kJ/mol respectively. Which matches with literature values for similar reactions. However, in future, all experiments need to be redone for several times to get a more accurate and persuasive result. Moreover, the 230°C experiment should be reviewed to clarify the factors accounted for the undesired results. In addition, all our conclusions were based on assuming the in-situ hydrogen provided by glycerol aqueous phase reforming is sufficient all the time. Though this is true in theory, it is worth checking the actual situation. At present, due to restrictions of reactor and safety concerns, the gas samples at each time points are hard to measure. In future research, improvements need to be done enabling sampling of gas during experiment.

7. Acknowledgements

We would like to thank Mr. Jiaxian Luo for assisting us in labs and providing invaluable insights.

8. Reference

BP, 2019. BP Statistical Review of World Energy.

Bridget Beals, 2022. UK sets out future for carbon pricing [WWW Document]. KPMG in the UK. URL https://home.kpmg/uk/en/home/insights/2022/04/uk-sets-out-future-for-carbon-pricing.html (accessed 12.13.22).

Calvo-Flores, F.G., Dobado, J.A., 2010. Lignin as Renewable Raw Material. ChemSusChem 3, 1227–1235. https://doi.org/10.1002/cssc.201000157

Chen, C., Zhou, M., Liu, P., Sharma, B.K., Jiang, J., 2020. Flexible NiCo-based catalyst for direct hydrodeoxygenation of guaiacol to cyclohexanol. New Journal of Chemistry 44, 18906–18916. https://doi.org/10.1039/D0NJ02929G

Chen, Z., 2020. Guaiacol hydrogenation with in-situ generated hydrogen through aqueous phase glycerol reforming.

Chen, Z., Kukushkin, R.G., Yeletsky, P.M., Saraev, A.A., Bulavchenko, O.A., Millan, M., 2020. Coupling

Hydrogenation of Guaiacol with In Situ Hydrogen Production by Glycerol Aqueous Reforming over Ni/Al2O3 and Ni-X/Al2O3 (X = Cu, Mo, P) Catalysts. Nanomaterials 10, 1420. https://doi.org/10.3390/nano10071420

Feng, J., Yang, Z., Hse, C., Su, Q., Wang, K., Jiang, J., Xu, J., 2017. In situ catalytic hydrogenation of model compounds and biomass-derived phenolic compounds for bio-oil upgrading. Renew Energy 105, 140–148. https://doi.org/10.1016/j.renene.2016.12.054

Gao, D., Xiao, Y., Varma, A., 2015. Guaiacol Hydrodeoxygenation over Platinum Catalyst: Reaction Pathways and Kinetics. Ind Eng Chem Res 54, 10638–10644. https://doi.org/10.1021/acs.iecr.5b02940

Guisnet, M., Magnoux, P., 2001. Organic chemistry of coke formation. Appl Catal A Gen 212, 83–96. https://doi.org/10.1016/S0926-860X(00)00845-0

Jongerius, A.L., Gosselink, R.W., Dijkstra, J., Bitter, J.H., Bruijnincx, P.C.A., Weckhuysen, B.M., 2013. Carbon Nanofiber Supported Transition-Metal Carbide Catalysts for the Hydrodeoxygenation of Guaiacol. ChemCatChem 5, 2964–2972. https://doi.org/10.1002/cctc.201300280

Mortensen, P.M., Grunwaldt, J.-D., Jensen, P.A., Knudsen, K.G., Jensen, A.D., 2011. A review of catalytic upgrading of bio-oil to engine fuels. Appl Catal A Gen 407, 1–19. https://doi.org/10.1016/j.apcata.2011.08.046

Mu, W., Ben, H., Ragauskas, A., Deng, Y., 2013. Lignin Pyrolysis Components and Upgrading—Technology Review. Bioenergy Res 6, 1183–1204. https://doi.org/10.1007/s12155-013-9314-7

Putra, R.D.D., Trajano, H.L., Liu, S., Lee, H., Smith, K., Kim, C.S., 2018. In-situ glycerol aqueous phase reforming and phenol hydrogenation over Raney Ni®. Chemical Engineering Journal 350, 181–191. https://doi.org/10.1016/j.cej.2018.05.146

Reangchim, P., Saelee, T., Itthibenchapong, V., Junkaew, A., Chanlek, N., Eiad-ua, A., Kungwan, N., Faungnawakij, K., 2019. Role of Sn promoter in Ni/Al $_2$ O $_3$ catalyst for the deoxygenation of stearic acid and coke formation: experimental and theoretical studies. Catal Sci Technol 9, 3361–3372. https://doi.org/10.1039/C9CY00268E

Song, Y., Gutiérrez, O.Y., Herranz, J., Lercher, J.A., 2016. Aqueous phase electrocatalysis and thermal catalysis for the hydrogenation of phenol at mild conditions. Appl Catal B 182, 236–246. https://doi.org/10.1016/j.apcatb.2015.09.027

Yoshikawa, T., Yagi, T., Shinohara, S., Fukunaga, T., Nakasaka, Y., Tago, T., Masuda, T., 2013. Production of phenols from lignin via depolymerization and catalytic cracking. Fuel Processing Technology 108, 69–75. https://doi.org/10.1016/j.fuproc.2012.05.003

Zhou, H., Han, B., Liu, T., Zhong, X., Zhuang, G., Wang, J., 2017. Selective phenol hydrogenation to cyclohexanone over alkali–metal-promoted Pd/TiO $_2$ in aqueous media. Green Chemistry 19, 3585–3594. https://doi.org/10.1039/C7GC01318C

Abstract

Every year, 4.9 million litres of petroleum are spilled into U.S. waters alone, causing significant damage to ecosystems and the environment. Developing efficient and sustainable methods for remediating oil spills is an active area of research in the scientific community. This report examines the use of waste chicken eggshells as a source for creating a superhydrophobic powder for use in oil-water separation and oil spill remediation. The superhydrophobic powder is prepared by boiling, drying, and blending eggshells, and functionalizing the resulting powder with stearic acid to impart superhydrophobic properties. The maximum measured water contact angle of the powder was 156.7° on a tablet that was created by compressing 5% (w/w) stearic acid-eggshell powder and 32-45 μm particle size with 50kg of force. The effects of varying powder mass, stearic acid concentration, particle size, and emulsion concentration on the effectiveness of oil-water separation are explored. Results showed that a higher mass of powder leads to a higher degree of separation. However, no definitive conclusions could be drawn regarding the effects of varying stearic acid concentration and particle size. The optimized powder had a maximum oil-sorption capacity of 0.64 g_{oil}/g_{powder} and was effective at separating emulsions up to 20% concentration. Further research should be performed on the use of petroleum-based oils and the feasibility of recovering adsorbed oil from the powder for its reusability.

1 Introduction

Approximately 80 million tonnes of eggs are produced globally every year, generating 8.58 million tonnes of eggshell waste per year (Waheed et al., 2020). Chicken eggshells are comprised of 95% calcium carbonate (Butcher and Miles, 2019) and instead of reusing this calcium-rich commodity for a useful application, medium sized egg production companies spend €50,000 to €200,000 per year processing eggshells as waste, where it ends up in landfill sites, creating odours and leading to the growth of harmful bio-organisms (European Commission, 2022). It is therefore imperative to develop methods to turn this abundant waste product into a renewable and valuable commodity.

1.1 Circular Economy

As of November 2022, the world's population has reached 8 billion and is projected to continue increasing until 2100 (Relifweb, 2022) alongside the demand for raw materials. Since the supply of these raw materials is limited, it is crucial for the world to move away from a linear economy and towards the more sustainable concept of a circular economy which entails increasing the life cycle of products, reducing their environmental impact and maximising resource efficiency.

The main principles required for transforming to a circular economy are known as the 3R's principles: Reduction, Reuse and Recycle (Ghisellini et al., 2016). If implemented to real life processes, this will not only benefit the environment, but also boost economies. Circular economy strategies can cut global greenhouse gas emissions by 39% whilst creating 6 million new jobs by 2030, offering an economic opportunity worth $4.5 trillion (McGinty, 2021).

This report investigates reusing waste eggshells to create superhydrophobic powders for use in oil-water separation, perfectly demonstrating a scenario that incorporates the 3R's principles for transforming to a circular economy.

1.2 Knowledge gaps, Aims and Objectives

Despite there being numerous studies researching into the production of superhydrophobic coatings from different types of shell waste (Fang, 2019) and on their antimicrobial (Causby, n.d.) and anti-icing properties (Sanusi, 2021) little literature can be found regarding the production of superhydrophobic powders from waste chicken eggshells and on their effectiveness in oil-water separation.

Oil-water separation has many applications, namely in wastewater treatment and in the petroleum, metal working and food industries (WBDG, 2020).

Every year, 4.9 million litres of petroleum are spilled into U.S. waters alone (Thompson, 2010), destroying habitats, poisoning, and suffocating surface-dwelling animals and contaminating our oceans and their seafood supplies (Edmond, 2021). It is therefore vital to remediate oil-spills as soon as they occur to reduce their devastating effects on the environment. Common methods include using dispersants, booms and skimmers, or in situ burning, but these all come with their flaws. Dispersants tend to be toxic, have a poor biodegradability and in some cases can be more harmful to the environment than the spilt oil itself (Sciencelearn, 2012). Using booms and skimmers requires very specific conditions including calm waters and slow oil speeds and most of the absorbents used have a low efficiency as they tend to also absorb water (Bhushan, 2019). Finally, the in situ burning of oil spills releases a lot of CO_2 and creates thick plumes of black smoke.

It is known that superhydrophobic surfaces have great oil-water separating properties (Latthe et al., 2019), so if the superhydrophobic powder generated from waste eggshells can successfully be used for oil-water separation, it could go a long way towards replacing the harmful and inefficient techniques for oil spill remediation with a much more sustainable, safer and efficient method. Therefore, the 5 main objects of this research project are to:

1. Create a superhydrophobic powder derived from waste chicken eggshells
2. Use the superhydrophobic powder to successfully separate oil-water emulsions
3. Explore the parameters affecting the extent of the oil-water separation and determine their optimised values.
4. Explore the absorption capacity and water content of the wetted powder.
5. Simulate a small scale oil spill and successfully clean up the oil.

2 Background

2.1 Contact Angle and Wettability

The wettability of a surface is defined as the tendency of one fluid to spread on or adhere to a solid surface in the presence of other immiscible fluids and can be expressed more conveniently by its inverse relationship with the measured contact angle, θ^*. For example, a hydrophobic surface with a low wettability would create a high contact angle with water droplets and vice versa.

The contact angle represents the angle that the liquid-vapour interface of a droplet makes with the solid surface and, in its simplest form, can be described by Young's equation in 1805:

$$\gamma_{sv} = \gamma_{sl} + \gamma_{lv}cos\theta_y \quad (1)$$

where γ_{sv}, γ_{sl} and γ_{lv} represent the solid-vapour, solid-liquid and liquid-vapour interfacial energies respectively and θ_y represents the equilibrium contact angle of an ideal surface that is rigid, smooth, insoluble, non-reactive and chemically homogenous.

Since very little surfaces are ideal in practice, Wenzel expanded on Young's equation to account for the roughness of a surface with chemical homogeneity and proposed the equation:

$$cos\theta^* = rcos\theta_y \quad (2)$$

where θ^* is the apparent contact angle and r represents the surface roughness and is defined as the ratio of the actual area to the projected area of the surface. The Wenzel model implies that since $r \geq 1$, the hydrophobicity of an already hydrophobic surface ($\theta_y > 90^\circ$) will increase with the roughness of that surface.

To account for the heterogeneity of a rough surface whereby air is trapped between the droplet and the surface, the Cassie-Baxter equation is used:

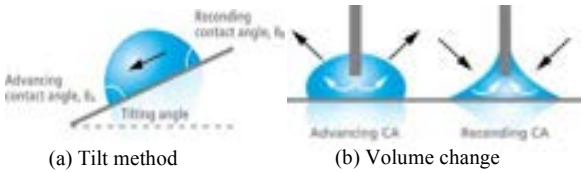
$$cos\theta^* = f_1 cos\theta_y - f_2 \quad (3)$$

where f_1 and f_2 represent the fraction of the surface made up by the solid itself and the air respectively. The Cassie-Baxter model implies that the contact angle and therefore the hydrophobicity of the surface will increase as the air fraction of the surface, f_2 increases.

2.2 Superhydrophobicity and Water Repellency

Since superhydrophobic surfaces tend to be highly water repellent, these terms are commonly used interchangeably even though their definitions are not the same. The hydrophobicity of a surface represents how low its wettability is and can be measured by the static contact angle, whilst the repellency of a surface represents how easily a droplet can roll off the surface and is measured by the contact angle hysteresis. By definition, a superhydrophobic surface forms a water contact angle (WCA) greater than 150°, and a truly water repellent surface has a contact angle hysteresis below 10° (Karapanayiots and Manoudis, 2012).

Contact angle hysteresis arises from the chemical and topographical heterogeneity of the surface (Laurén, 2021) and can be calculated by the difference between the advancing angle, θ_A and the receding angle, θ_R. These angles can be measured by 2 different methods: the tilt method and the volume change method, both of which are illustrated in figure 1.



(a) Tilt method (b) Volume change

Figure 1. Illustration of methods to measure advancing and receding contact angles. (a) Tilt method. (b) Volume change method.

Generally, the volume change method is used more commonly to measure contact angle hysteresis because its results are not affected by the droplet volume size, whilst those from the tilt method are.

2.3 Functionalisation of Waste Chicken Eggshells

Calcium carbonate ($CaCO_3$) is intrinsically hydrophilic. However, it can be functionalised by a reaction with stearic acid (SA) to obtain hydrophobic properties. Stearic acid is a fatty acid consisting of a hydrophilic carboxylic acid head and a hydrophobic aliphatic tail. When $CaCO_3$ and SA react together, the hydrophilic head of the SA chemisorbs onto the calcium cations present on the surface of the $CaCO_3$ particles, forming calcium monostearate. The chemisorbed SA orientates itself in such a way that the hydrophilic head is orientated towards the $CaCO_3$ surface, whilst the hydrophobic, aliphatic tail of the SA is orientated away from it, giving the powder hydrophobic properties. This structure is visualised in figure 2. It is worth noting that the SA may also physisorb to the surface of the solid through weak Van Der Waals forces.

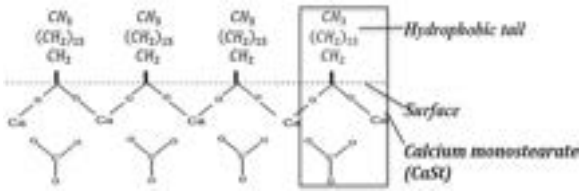


Figure 2. Diagram depicting the chemisorption between stearic acid and calcium carbonate

In theory, the discussed reaction causes the $CaCO_3$ particles to be covered by a monolayer of hydrophobic SA molecules. However, there are many factors that can affect the extent of this monolayer coverage, causing it to have voids which can reduce its hydrophobic potential. Some of these factors include the method of preparation and its conditions, the moisture content of the $CaCO_3$, the $CaCO_3$ particle size, the $CaCO_3$ concentration and the amount of SA required to completely cover the calcite surface with a monolayer (Cao et al., 2016). This report investigates the effects of varying the $CaCO_3$ particle size and the concentration on the hydrophobicity of the powder and its effectiveness in oil-water separation.

3 Methodology

3.1 Statistical Analysis

Analysis of variance (ANOVA) is a statistical method that extends the student's t-test from two data groups to three or more. In this study, groups of data are collected by repeating experiments under the same conditions and an ANOVA test is performed on these groups to determine whether or not there are significant differences between the means of these groups. ANOVA tests are based on the concept of variance, which is a measure of the spread in a set of numbers. It is important to determine whether the differences between the means are statistically significant in order to draw meaningful conclusions from the experimental data. The null hypothesis states that the means of all groups are equal. If rejected, it is implied that the means of at least two groups are different. A significance level of α=0.05 was selected for this report, indicating a 5% probability of committing a type I error and falsely rejecting the null hypothesis. To conduct these tests, a python algorithm was implemented to calculate the F-value and p-value.

3.2 Superhydrophobic Powder Preparation

Eggshells (ES) were first boiled to remove any remaining egg membrane and were then dried in a vacuum oven. The dried ES were then ground into a powder using a blender, and the particle sizes were classified using a vibrating sieve machine. The range of particle sizes investigated in this study were 0-32µm, 32-64µm, 64-75µm, and 70-90µm.

The ES powder was ground together with SA in a mortar and pestle to initiate a reaction between the two. This would functionalise the $CaCO_3$ with SA and establish physisorption and chemisorption between them, giving the powder hydrophobic properties.

3.3 Water Contact Angle Measurement

Tablets of the powder were prepared using the Gamlen tablet press and then blasted with nitrogen gas to clean the surface from loose particles. An automated Ramé-Hart goniometer was used for measuring the WCA using the sessile drop technique with deionised water of known interfacial energy. If necessary, the surface of the tablet was pressed against a sterile surface to remove any unwanted asperities and make it more flat, reducing the inconsistencies between different WCA measurements. The tablet was ensured to be horizontal with a 0° tilt, at which point water droplets of 10µL were electronically dispensed out a pipette. A plastic pipette tip was selected over a metallic one because plastic has a lower surface energy than metal and will thus form less favourable interactions with water molecules, making it more hydrophobic. This allows the water droplet to transfer more easily from the pipette to the tablet.

Since wettability is dependent on both the chemical properties of the solid as well as its surface energy and roughness, the WCA was measured for tablets of different SA concentration, particle diameter and tablet press force whilst keeping all other parameters constant. From these measurements, plots were generated to determine the optimum values of each parameter with the objective of using all the optimum values to create a superhydrophobic surface with WCAs consistently greater than 150 °.

3.4 Contact Angle Hysteresis

Contact angle hysteresis was measured using both the tilt and volume change method as discussed in section 2.2.The tilt method was utilised to measure the WCA of a 15µL droplet on a superhydrophobic tablet. The surface was rotated at a rate of 1 degree per second up to 45 degrees and a measurement of the WCA was taken using the Ramé-Hart advanced goniometer at each interval, totalling 46 readings. In the volume change method, a pipette is placed close to the tablet surface and a droplet of 10µL is dispensed such that the pipette is inside the droplet. An additional 20µL of water is added and removed from the droplet in 1µL increments with contact angles taken at each stage. Figure 3 below shows the droplet before water is added, once it has been added and once it has been removed.

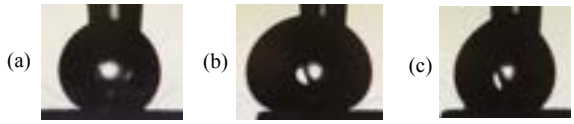


Figure 3. Photographic sequence showing the Wilhelmy method: (a) Before water is added. (b) After adding 20µl. (c) After removing 20µl

3.5 Oil Water Separation

Preliminary tests showed that mixing the superhydrophobic powder with an oil-water emulsion creates floccules of the powder suspended in water, indicating that the powder has an affinity for oil. Four different types of oil were tested to determine the most stable oil-water emulsion, with sunflower oil found to be the most stable. Further testing and quantification of oil-water separation will focus on sunflower oil. Section 3 investigates many different methods for quantifying the extent of separation.

3.5.1 Syringe with Cotton Wool Support

This method aims to separate oil and water whilst retaining powder particles in a cotton wool support. 0.100g of cotton wool was measured using a microbalance. The cotton was then placed at the bottom of a glass syringe and then wetted to reduce its oil-sorption capacity. 0.300g of powder was then dumped on top of the cotton and a 10% oil-water emulsion was prepared by shaking 9mL of water and 1mL of oil together. This emulsion was then poured into the syringe and a plunger was used to push 10mL of liquid through the syringe needle. Most of the powder could be separated from the cotton and the masses of wetted cotton and oil were measured. The samples were then dried in a vacuum oven overnight to just below the bubble point of water at a temperature and pressure of 60°C and 200mbar respectively.

To investigate the capability of only cotton wool in oil-water separation, a control experiment was conducted without the powder and initial results suggested that the cotton absorbed 10-20 times more

oil per unit mass than the powder. Since commercial cotton can be used to separate oil and water (Luo et al., 2022), different methods were therefore explored.

3.5.2 Tea Bag

To quantify separation, a tea bag was filled with 0.500g of powder and then dipped in a 10% oil-water emulsion. It was then dried to measure the mass gain of the tea bag, providing a measure of the oil absorbed. While this method is useful for developing the final product, there were experimental challenges in accurately measuring the separation, as it was difficult to ensure that all of the powder was exposed to the emulsion. Further, the tea bag absorbed a significant amount of oil itself, which could introduce unreliability in the gravimetric analysis.

3.5.3 Filtration

0.500g of powder was mixed in 10mL of a 10% (v/v) oil-water emulsion. It was then attempted to separate the powder from the mixture using gravity filtration, but the oil was unable to penetrate the filter paper and separation could not be achieved, as shown in figure 4.

(a) (b)

Figure 4. Snapshots of the filtration method. **(a)** Whilst powder is mixed with the emulsion. **(b)** During filtration separation

The same outcome was also reached in the control test, where no powder was used. To address this issue, vacuum filtration was attempted on the conically shaped filter paper, but the paper tore when any amount of powder was present. In response to this, the same method was attempted on a flat Büchner funnel. Even with the vacuum pump running at full power, the oil still did not penetrate the filter paper.

Tea bags were also tested as an alternative to filter paper, as they are more porous and allow oil to pass through. However, this increase in porosity also resulted in powder passing through, affecting the accuracy of the gravimetric analysis.

The final vacuum filtration method involved packing a known mass of powder ontop of a mass of cotton wool in a funnel. An emulsion was then filtered through this setup using a vacuum pump to pull excess oil through. However, the filtration approach was ultimately discarded because the cotton still absorbed a significant amount of oil.

3.5.4 Centrifuge

1.000g of powder, 10mL of oil, and 40mL of water were measured and poured into a centrifuge tube, which was shaken to ensure that all of the powder came into contact with the emulsion. The mixture was then centrifuged to create a two-phase liquid and the amount of oil absorbed by the powder was measured. However, it was found that only a very small amount of oil was absorbed, which introduced high uncertainty in the volume measurement, as the equipment was only accurate to ±0.5mL, and the differences observed were smaller than this. Additionally, it was uncertain whether the centrifugation process caused oil to be pushed out of the powder, rendering the method unsuitable.

3.5.5 Syringe with mesh as support

In previous tests, it was found that the powder support absorbed a significant amount of oil. To address this issue, a mesh with an aperture diameter of 32μm was attached to the end of a syringe that was cut open at the base and firmly attached using a cap with a hole for the liquid to exit. Parafilm was used to create a watertight seal and prevent leakage. In these experiments, the oil water separation was investigated on powders with different SA content, particle diameter and quantity of powder. A 10% oil-water emulsion was prepared by mixing 10 mL of oil with 90 mL of water for 20 minutes using a magnetic stirrer. It was important to maintain the stability of the emulsion during transfer and passage through the syringe in order to ensure consistent and controlled experimental conditions. The powder was loaded at the bottom of the syringe and tapped on a hard surface to ensure it was uniformly spread on the mesh. Another syringe was used to withdraw 22.5mL of the emulsion from the stirred beaker and was then poured over the powder. The plunger was used to force 14mL of liquid through the powder and mesh directly into a 15mL centrifuge container. The contents were then centrifuged to create a two-phase liquid, where the water could easily be decanted from the bottom of the centrifuge container through a punctured hole. Some water remained in the container and was dried in a vacuum oven set at 60°C and 200mbar, leaving only the oil. To ensure more reliable results, the experiment was repeated four times and the average was taken for each data group. A gravimetric analysis was undertaken by measuring the masses of the filled centrifuge tube before and after drying as well as the tube's empty mass using a microbalance. From these readings, the masses of water and oil in the filtrate could be calculated, making it possible to determine the concentration of oil in the filtrate. This was compared to control runs without the superhydrophobic powder, revealing that the presence of the powder reduced the amount of oil in the filtrate and that it is effective at separating emulsions.

4 Results

This section presents all data collected from the contact angle measurements and the oil-water separation analysis. It is worth noting that upon analysing data, outliers were omitted if they lied beyond 1.5 times the interquartile range due to deviations from the planned experimental procedure. All error bars displayed in this section represent the standard deviation of a sample excluding its outliers.

4.1 Water Contact Angles

Despite having previously discussed the Wenzel (equation 2) and Cassie-Baxter (equation 3) equations, all contact angles compared in this section will represent the apparent contact angle. Calculating the young's contact angle requires access to an atomic

force microscopy to measure the surface roughness and the surface solid fraction, which was not available under the resource and time constraints of this project.

This report investigates the effect of three parameters on the wettability of the powder: compression force, powder diameter, and SA concentration. These are explored by measuring the WCA while varying one parameter and keeping the others constant. Several measurements were taken for each data point and the average of these readings was plotted against the varied parameter.

4.1.1 Varying Compression Force

The compression force was varied from 50-350kg, whilst keeping the SA concentration and powder diameter constant at 10% (w/w) and 32-45 μm respectively, with its plot shown in figure 5 below.

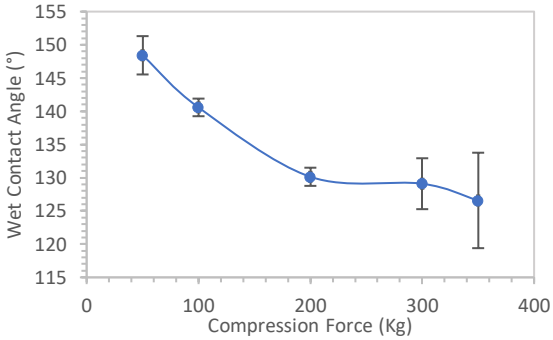


Figure 5. Plot of the mean WCA against compression force

The results show that as the compression force decreases, the WCA increases, with a maximum value of 148.4° ± 2.9° at 50kg. This trend can be explained by both the Wenzel and Cassie-Baxter equations outlined in section 2.1. At lower compression forces, the surface roughness, r and the fraction of the surface occupied by air, f_2 increase. Therefore, the apparent WCA, θ^* will be larger, as observed in figure 5.

Since the WCA improved at lower forces, it was attempted to measure contact angles beyond 50kg and even on the powder itself with no force. However, beyond this point, the tablet became either too fragile to the point that it would crumble, or too rough, resulting in the water droplet picking up loose powder from the surface and adhering to the pipette rather than to the surface. This made it difficult to accurately place the droplet on the surface and perform the experiment with compression forces below 50kg. Instead, if the trend is extrapolated beyond 50kg, it can be hypothesised that the powdered form of the functionalised ES will be even more hydrophobic than its tableted form as it will effectively have no compression force.

4.1.2 Varying Powder SA Concentration

The concentration of SA in the ES powder was varied from 3% to 12% for two different particle diameters (32-45μm and 45-63μm), whilst the compression force was held constant at 50 kg since this value yielded the highest WCA from section 4.1.1. The results of these experiments are summarized in figure 6 below.

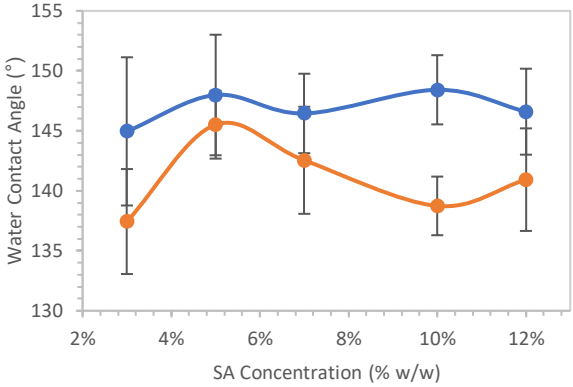


Figure 6. Plot of the mean WCA against SA concentration in the tablet for 2 different powder diameters.

Upon initial observation, it appeared that as the SA concentration increased, a local maximum for both powder sizes was obtained at 5% SA concentration, yielding a WCA of 148.0° ± 5.0° for the 32-45μm powder diameter and 145.5°±2.8° for the 45-63μm powder diameter and that beyond this optimum, there appeared to be no trend. The maximum measured WCA was found to be 156.7° at 5% SA concentration for the 32-45μm particle diameter range, proving that it is possible to create a superhydrophobic tablet in some cases. A photograph of this WCA is shown in figure 7 (a).

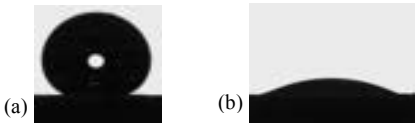


Figure 7. Photographs of contact angles on a superhydrophobic surface. **(a)** Water droplet with a measured WCA of 156.7°. **(b)** Oil droplet with a contact angle of 19.5°.

To investigate if superhydrophobic surfaces possess oleophilic properties, an oil droplet was dispensed onto a tablet under the same previous conditions, and the contact angle was measured and recorded to be 19.5°, as shown in figure 7 (b). This suggests that a superhydrophobic powder is also oleophilic, meaning that it is able to attract and adsorb oil whilst repelling water. To determine if the SA concentration continues to have no significant effect on the WCA at more extreme concentrations, readings were taken at higher concentrations of 50% and 100%, yielding WCAs of 108.7°±2.0° and 91.2°±1.7° respectively. This suggests that as the SA concentration is increased beyond 12%, the powder will eventually become less hydrophobic.

The large error bars of each measurement in figure 6 make it difficult to conclude if there is actually a local optimum at 5%. An ANOVA test was therefore conducted to determine the statistical significance of the observed trend supports that there is insufficient evidence to support the existence of said local optima. The large error bars and lack of correlation between data points in figure 6 are due to the large experimental

error and randomness in the surface roughness of each new measurement, as discussed in section 4.1.4.

Figure 6 also shows that the smaller particle size gives a higher WCA for all SA concentrations. However, the overlap between the error bars of the different powder sizes for 3%, 5%, 7%, 12% make it difficult to conclude if this is truly the case. Further investigation will therefore have to be performed to see how the powder's size affects the WCA. When analysing this, the SA concentration will be kept constant at 5% because, although the data points are not statistically different for the 32-45µm powder, they are for the 45-63µm powder, and 5% SA concentration yields the best WCA for this size.

4.1.3 Varying Particle Size

The WCA was measured for different particle sizes in the range of 0-32µm, 32-45µm, 45-63µm and 63-75µm whilst the compression force and SA concentration were kept constant at 50kg and 5% respectively for the same reasons discussed in section 4.1.2.

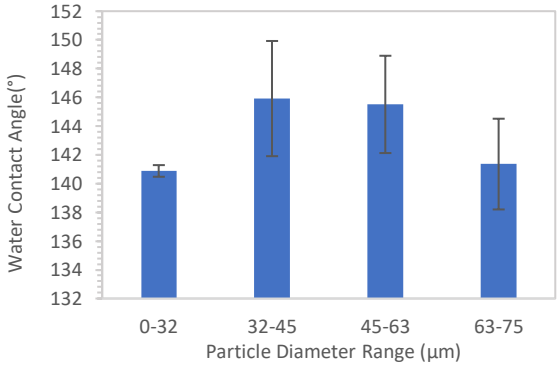


Figure 8. Plot of the mean WCA against particle diameter.

As particle diameter increases beyond this range, the WCA then appears to decrease. Upon initial consideration, this inverse relationship may be explained by the fact that the surface area decreases as the particle diameter increases and that very fine particles tend to agglomerate. Consequently, there is a smaller surface for the SA to react with when it is ground together with calcium carbonate in a mortar and pestle. As a result, more SA would physisorb to the surface rather than chemisorb, potentially decreasing the overall hydrophobicity of the powder. Nonetheless, the results of the ANOVA statistical test at a 95% confidence level indicate that this is not the case and that there is no significant difference between the means of the WCAs in the diameter range of 32-75µm. This implies that the observed differences in the WCAs are likely to be due to random chance and not to any underlying effect of particle diameter on WCA. This suggests that all the SA reacts with the calcium carbonate and chemisorbs onto the surface in the same way, regardless of the particle diameter in the given range.

The minimum at the diameter range of 0-32µm can be explained by the fact that the tableted form of smaller particles was much smoother than the tablets of larger sizes, resulting in a lower apparent contact angle, for the same reasons discussed in section 4.1.1. Furthermore, very fine particle sizes show a strong tendency to agglomerate, encouraging voids to form in the hydrophobic monolayer of the powder (Cao et al., 2016), thus decreasing their hydrophobicity and the WCA.

4.1.4 Contact Angle Measurement Limitations

Although the Ramé-Hart goniometer is capable of measuring contact angles to an accuracy of 1 decimal place, there are many factors that can affect the accuracy of contact angle measurements. This may explain why the standard deviation of the different mean WCA measurements was so high. Some of these factors include the presence of contaminants on the sample surface, surface roughness, inconsistencies between experiments, surface properties, and environmental factors. These factors can introduce error or uncertainty into contact angle measurements and should be carefully controlled or accounted for in order to obtain reliable and accurate results.

After examining the powders closely, it was observed that there were occasionally small metal contaminants in the powder which likely came from the mortar and pestle used in the preparation process. Other common contaminants present on surfaces include dust and dirt which could have been introduced to the sample during its preparation and handling process, or from the air in the surrounding environment. Even after blasting the tablet with compressed Nitrogen gas, these contaminants can still be present on the surface, altering its surface energy and roughness, and this can in turn affect the WCA.

The method of measurement used in this experiment was subject to a high degree of human error, making it difficult to accurately reproduce the results of each experiment. Between experiments, the location where the droplets were placed on the tablet's surface and the timing of when the measurement first began after first contacting the surface varied, introducing a significant amount of random error. To eliminate these inconsistencies, the human factors in the procedure should ideally be replaced with automation. However, this would be challenging, time-consuming, and may not be necessary.

The measurements of the WCA were taken on different days when the temperature, pressure, and humidity may have varied. These environmental factors can also affect the surface tension and interfacial interactions between the liquid and the surface, which can interfere with the WCA measurements, introducing error and uncertainty into the measurements (Diaz, Savage and Cerro, 2017).

4.2 Contact Angle Hysteresis

Both the tilt method and the volume change method were used to measure the contact angle hysteresis on a tablet compressed with 50kg of force consisting of a 5% SA concentration and a particle diameter of 32-45µm. In the former, it was found that the size of the droplet volume greatly affected the roll off angle and therefore the contact angle hysteresis measurement.

More specifically, droplets of 10µl would not roll off the surface, even when it was tilted to an angle of 45°, as shown in figure 9. However, droplets of 15µl would roll off at an average angle of 11.5°±6.4° with an average contact angle hysteresis of 23.1°±2.7°. Since the roll off angle greatly depended on the droplet size, the volume change method was also used to verify the contact angle hysteresis.

Figure 9. Photograph of a 10µl water droplet on a tablet surface tilted at 45°

Figure 10 below shows an example of one measurement taken using the volume change method. As the droplet size is increased, the contact angle converges to around 157° indicating that this is the advancing contact angle. Similarly, as the contact angle is decreased, the contact angle converges to give a receding contact angle of approximately 134°, yielding a contact angle hysteresis of about 23°.

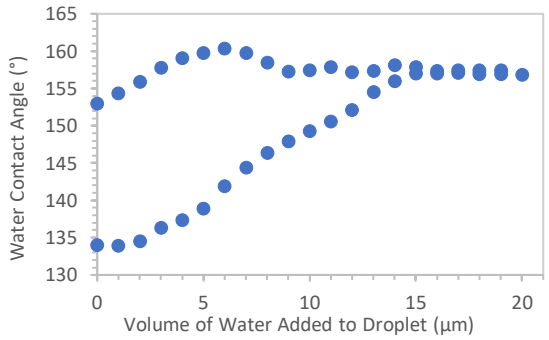


Figure 10. Plot of WCA variation as water is added to the droplet.

Both methods give similar results, but the tilt method is not very repeatable as the droplet only rolled off the surface twice out of the eight times it was performed. The volume change method is therefore more accurate and reliable for measuring the contact angle hysteresis, and is the preferred method.

The obtained contact angle hysteresis value of 23° suggests that the powder is not water repellent, as it is not below 10°. However, further experiments can still be performed to test its effectiveness in oil-water separation, as previous tests indicate that the powder is hydrophobic and potentially superhydrophobic.

4.3 Oil-Water Separation

Many different methods of quantifying oil-water separation were investigated. However, the results presented in this section were obtained using the syringe with mesh support method that has been described in Section 3.5.5. To determine the powder formulation that yields the best performance, the mass of powder, SA concentration, and particle diameter were varied when separating a 10% oil-water emulsion. After identifying the parameters that yielded the best separation, the powder was also tested under varying emulsion concentrations to further evaluate its performance.

4.3.1 Varying Powder Mass

To investigate the relationship between the mass of powder loaded and the concentration of oil in the filtrate, a series of experiments were conducted in which the SA concentration and particle diameter were held constant at 10% and 45-63µm, respectively.

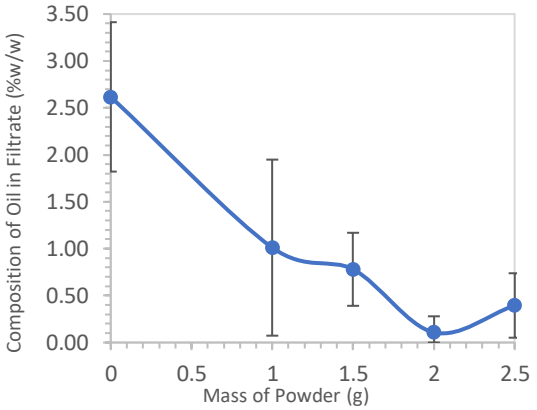


Figure 11. A plot showing how the oil composition of the filtrate varies with mass of powder

Figure 11 shows a negative correlation between the concentration of oil in the filtrate and the mass of powder used. This is likely due to the fact that a larger quantity of powder will adsorb more oil. At a mass of 2g, the filtrate is essentially an oil-free aqueous solution, indicating that this is the minimum amount of powder required for the complete separation of oil and water. The slight uptick in the oil concentration of the filtrate at 2.5g is likely to be an outlier due to experimental inconsistencies. This can be seen by the fact that the 2g datapoint has smaller error bars than the 2.5g datapoint, making it a more reliable indicator. Additionally, it is both intuitively and theoretically reasonable that beyond 2g, there should be a decrease in oil concentration. Since nearly complete separation is achieved at 2g, this mass can be taken to be the saturation point of the powder in a 10% emulsion, and the oil absorption capacity is calculated to be approximately $0.7mL_{oil}/g_{powder}$ in a 10% emulsion.

In the absence of powder, the oil concentration in the filtrate is noticeably higher, but it does not approach the initial concentration of 10%, indicating that the mesh support itself also contributes towards oil-water separation. This observed separation is likely facilitated by the resistance created by the mesh, reducing the velocity of the more viscous oil.

4.3.2 Varying Powder SA Concentration

Functionalised eggshell powder of 45-63µm particle diameter was formulated for varying concentrations of SA. As it was established in the experiment described earlier where the mass of powder was varied, using a 1.5g mass of powder ensures that it will always saturated. This mass will also allow for more oil to pass through to the filtrate, making it easier to measure with a lower uncertainty.

Running the ANOVA test on these groups of data returned a p-value of 0.45, corresponding to a 45% likelihood that the mean values of these data are the

same. This indicates that the variation between SA concentration observed in figure 12 is due to randomness and no conclusion can be drawn from these results.

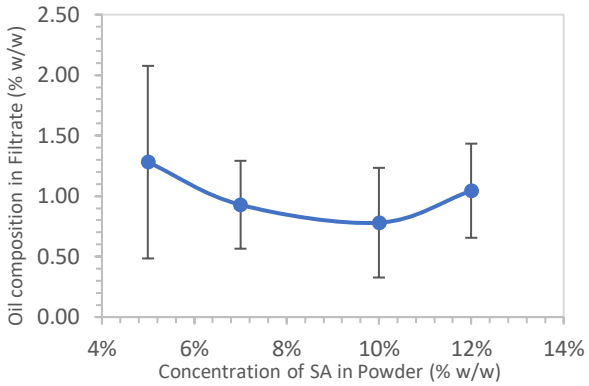


Figure 12. A plot showing how the oil composition of the filtrate varies with SA concentration in the powder.

Tests were also performed on pure ES powder with a 0% SA content, and it was observed that the non-functionalised powder had a very thin, sand-like consistency, which contrasted with the cohesive, cake-like texture of the functionalized powder, suggesting that the non-functionalised powder particles were able to block the mesh apertures and increase the resistance of this filter medium. Consequently, since oil is more viscous than water, the increased resistance to flow will amplify the separation due to viscosity and velocity differences. This hypothesis is supported by the fact that significantly more force was required to push the emulsion through the non-functionalised powder than through the functionalized powder. Additionally, the water content in the non-functionalised powder was 32.8±1.0%, which is much higher than the 13.3±3.3% for the functionalized powder, indicating that the ES powder adsorbed less oil than its functionalized counterpart. It is therefore possible that the separation of the 0% powder was not due to its ability to absorb oil, but rather to its ability to block pores. Because of these differences, it is not appropriate to compare the data for the 0% powder to that of the functionalized powder, and it has been omitted from the above graph.

4.3.3 Varying Powder Diameter

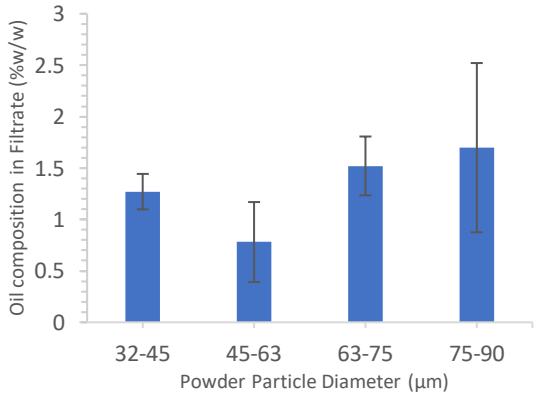


Figure 13. A plot showing how the percentage water in the cake varies as particle diameter increases.

To investigate the effect of varying powder diameter against oil-water separation efficacy, the SA concentration of the powder was kept constant at 10% and 1.5g of powder was used in each experiment.

Statistical analysis of the data suggests that the null hypothesis cannot be rejected, as the ANOVA test returned a p-value of 0.21, which is greater than the accepted maximum threshold of 0.05. Since the variation in the data are not statistically significant, no definitive conclusions can be drawn on the particle size and its effects on oil-water separation.

4.3.4 Varying Emulsion Concentration

The separation of oil and water was studied by varying the emulsion concentration while holding the SA concentration, particle diameter, and mass of powder constant at 10%, 45-63µm, and 1.5g, respectively. The emulsion concentration was varied from 5% to 30%, and the oil composition of the resulting filtrate was determined.

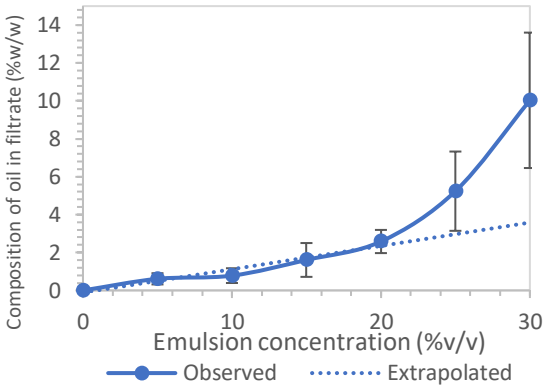


Figure 14. A plot showing how the oil composition of the filtrate varies with emulsion concentration

Figure 14 demonstrates a strong positive linear correlation when the concentration is varied between 0% and 20% with a gradient of 0.1232 and a coefficient of determination of 0.948. The low gradient implies that a small increase in the emulsion concentration would result in a smaller increase in the filtrate concentration, indicating that the powder is effective at separation in this range. Above 20%, the oil composition in the filtrate increases rapidly, suggesting that the powder becomes saturated with oil at and beyond this concentration, allowing the oil to easily pass through the powder.

4.3.5 Powder Cake Analysis

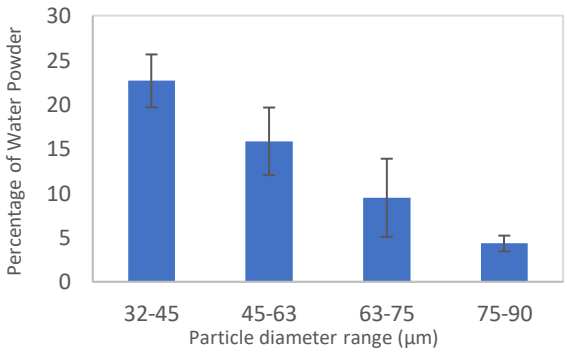


Figure 15. A plot showing percentage of water content against particle diameter

As part of this study, the percentage of water in the powder was determined by recovering the wetted powder cake from the mesh and drying it in a vacuum oven at 60C and 200mbar. The water content in the sample was then determined by comparing the mass of the powder before and after drying.
The data suggests that the water content of the powder decreases linearly as the particle diameter increases. This may be because the overall surface area and therefore the sorption capacity of the powder decreases with increasing particle diameter, resulting in less water being absorbed by the powder.

4.3.6 TGA analysis

A sample of the dry powder obtained after the emulsion was passed through was analysed in a thermogravimetric analysis (TGA).

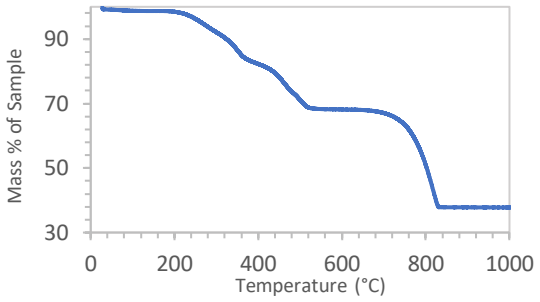


Figure 16. TGA showing the percentage decrease of the intiaial sample weight against increasing temperatures.

TGA tells us that when drying the powder in the previous section, no significant amount of oil was evaporated and that there is potential for oil-powder separation by heating to around 360°C to recover the oil, as a there is a sharp decrease of mass at around 380°C due to the SA evaporating. Heating past 700°C will initiate the decomposition of $CaCO_3$ to CaO
.

4.3.7 Oil-Water Separation Limitations

The mesh method was used in this study over the other methods discussed in section 3.5. However, like all scientific experiments, this method also has its limitations.

In many runs, small amounts of powder were able to pass through the mesh support, particularly when using smaller particle diameters. Since it was assumed that a negligible mass of powder made it through to the filtrate, the calculated filtrate oil concentrations are likely to be underestimates of the true concentration.

Another issue is that as the mass of powder or oil concentration in the emulsion was increased, the plunger required more force to be pushed. This led to inconsistencies between experiments, introducing more error and increasing the deviation between results. Additionally, the need for higher pressures sometimes resulted in leakage through the parafilm, causing some runs to be discarded, reducing the reliability of some datapoints.

It was originally assumed that separation occurred solely due to the powder's ability to adsorb oil. However, it was found that a degree of separation also occurs due to differences in viscosity and velocity. Oil is approximately 100 times more viscous than water, meaning it has a higher resistance to flow. The presence of the powder and mesh support impede the flow of oil much more than water, leading to some separation. Therefore, the sorption capacity of the powder and its effectiveness in oil-water separation are likely to be overestimates in this report.

Regarding the preparation of the functionalized powder, human variability must be considered when using a mortar and pestle. It is practically impossible to ensure that every particle was ground together, and the amount of particles that weren't contacted by the mortar varied between each batch. Also, the distribution of functionalised powder may not be uniform in the batch, leading to inconsistencies in measurements. These inconsistencies can be solved by automating the method using a ball mill instead.

5 Small Scaled Oil-spill Simulation

A small-scale oil spill was simulated by adding 200 mL of seawater to a crystallizing dish. 5 mL of sunflower oil was then added to the surface of the water, and both the water and oil were dyed with blue and red colours, respectively. A mesh containing 10g of functionalised ES powder was then submerged in the dish and mixed with the seawater-oil mixture. This mass of powder used in this experiment was in excess of the estimated oil sorption capacity. As a result of this, the amount of oil in the container decreased, and upon removal of the mesh, almost pure water was observed with the bulk of the oil being separated. On closer observation of the water surface, some residual oil was visible. The used powder was transferred to a petri dish, and it was observed to have a red colour from the dye used.
Overall the small scale test was successful accomplishing objective 5, however it raises questions on large scale implementation of this approach as the used powder has to be collected and cannot be reused. Possible avenues for future research include exploring the feasibility of extracting oil from the used powder and testing the powder with petroleum-based oils.

Figure 17. Images showing the oil before and after and the wetted powder.

6 Conclusion

This study found that lowering the compression force increased the WCA, and the optimum compression force was found to be 50kg. Analysis of variance tests indicated that the concentration of SA and the particle size did not significantly affect the WCA. Although optimum conditions could not be obtained for all the studied parameters, the maximum measured WCA was 156.7°, indicating that it is possible to achieve a superhydrophobic powder from chicken eggshells, and the highest mean WCA obtained was 148.8°±2.9° at a compression force, SA concentration and particle

diameter of 50kg, 10% and 32-45μm respectively. While this does not correspond to a superhydrophobic angle, it is likely that using the powdered form with no compression force will produce WCAs greater than 150°, resulting in a superhydrophobic surface. Based on these 2 observations, it can be concluded that the first main objective of creating a superhydrophobic powder derived from waste chicken eggshells has been achieved.

Several variables for the superhydrophobic powder formulation were investigated to determine the optimal powder for separating oil and water. However, many of the results did not show a statistically significant trend according to the ANOVA test, and the data obtained may be due to experimental randomness. Data suggests that the SA concentration in the powder does not affect oil water separation, therefore the least amount of SA should be considered as it is the most sustainable option. Variation of particle diameter yielded statistically the same mean, however there is a strong indication from the cake analysis that larger particle diameters absorb less water but there is no certainty as to the amount of oil that is absorbed.

The capacity of oil absorbed by the powder was calculated to be 0.7 mL_{oil}/g_{powder} based on the data for varying powder mass. However, this value is an overestimate as the separation is also affected by the viscosity differences between the liquid. Powder cake analysis gave a calculated absorption capacity of 0.22 g_{oil}/g_{powder} ± 0.14 g_{oil}/g_{powder} for particles with a diameter of 75-90 μm. This value however, is subject to experimental error.

The TGA results suggest that eggshell powder and the oil can be potentially separated by heating. This will reduce the need to dispose of the used-up powder and also in an oil spill scenario recover the spilled oil.

Non-functionalized eggshell powder resulted in good oil-water separation, but using this material absorbed almost twice the amount of water than its non-functionalised counterpart, suggesting that separation here occurs due to aperture blockage rather than its ability to absorb oil.

7 Outlook

The use of an atomic force microscopy allows the surface roughness and solid fraction of the surface to be measured. With this extra information, it would be possible to calculate the young's contact angle and normalise the contact angles to a smooth surface, allowing for the comparison of data groups without the variability in surface roughness. This may decrease the deviation across measurements to the point where statistically different results may be obtained and a trends can be observed.

Additionally, the effect of different drop volumes has not been extensively investigated in this report. Initial tests showed that a 15μL droplet will give a larger WCA, suggesting that droplet size could be a meaningful parameter. Further studies should thus be performed on the effect of droplet volume on the hydrophobicity of a surface.

The method used to quantify separation has many limitations. Mainly, separation is influenced not only by the hydrophobicity and oil-sorption on the powder, but also by fundamental filtration principles. It is therefore recommended to explore a method that focuses more on the full immersion of the powder in an emulsion to more accurately quantify the oil sorption capacity of the powder.

The small-scale oil spill remediation test showed promising results, warranting further research into larger scale simulations involving different types of oils, particularly petroleum-based ones. It would also be beneficial to study the feasibility of recovering adsorbed oil from the powder and if it retains its superhydrophobic properties, which would further support the development of a circular economy.

8 References

Bhushan, B. (2019). 'Bioinspired oil–water separation approaches for oil spill clean-up and water purification'. Philosophical Transactions of the Roy

Butcher, G. and Miles, R. (2019). 'Concepts of Eggshell Quality'.

Cao, Z., Daly, M., Clémence, L., Geever, L.M., Major, I., Higginbotham, C.L. and Devine, D.M. (2016). 'Chemical surface modification of calcium

Causby, A.T. (n.d.). 'Investigating the formulation and efficacy of antimicrobial coatings derived from

Chu, Michael (2004-06-10). "Smoke Points of Various Fats - Kitchen Notes". Cooking For Engineers. Retrieved 2013-02-07.

Diaz, M.E., Savage, M.D. and Cerro, R.L. (2017). 'The effect of temperature on contact angles and

Edmond, C. (2021). 'This is how oil spills damage our environment'.

European Commission (2022). 'LIFE supports circularity in the ceramics sector'.

Fang, H. (2019). 'Investigation of superhydrophobic coatings from different types of shell waste'.

Ghisellini, P., Cialani, C. and Ulgiati, S. (2016). 'A Review on Circular Economy: the Expected

Karapanayiots, I. and Manoudis, P. (2012). 'Superhydrophobic surfaces'.

Latthe, S.S., Sutar, R.S., Bhosale, A.K., Sadasivuni, K.K. and Liu, S. (2019). 'Chapter 15 -

Laurén, S. (2021). 'What is contact angle hysteresis?' [online] www.biolinscientific.com. Available at: https://cutt.ly/Q0xZ7yo

Luo, X., He, Z., Gong, H. and He, L. (2022). 'Recent advances in oil-water separation materials

McGinty, D. (2021). 5 'Opportunities of a Circular Economy'.

Relifweb (2022). 'World population to reach 8 billion on 15 November 2022'

Sanusi, J.W. (2021). 'Investigation of Durability and Anti-Icing Properties of Superhydrophobic

Sciencelearn (2012). 'Cleaning up the oil spill'. [online] Science Learning Hub. Available at: https://rb.gy/zxvtak

Thompson, A. (2010). FAQ: 'The Science and History of Oil Spills'.

Fe-Doped TiO_2 Photoanodes Grown by Aerosol-Assisted Chemical Vapour Deposition of a Titanium Oxo/Alkoxy Cluster

Senanur Duman and Aisha Ali
Department of Chemical Engineering, Imperial College London, U.K.

Abstract
Photoelectrochemical (PEC) water splitting is a promising technology for the production of solar hydrogen. Titanium dioxide (TiO_2) is a popular semiconducting transition metal oxide used as a photoanode material in PEC cells. To achieve its full potential, Fe^{3+} doped nanostructured TiO_2 films formed from a $Ti_7O_4(OEt)_{20}$ titanium oxo/alkoxy cluster solution using aerosol-assisted chemical vapour deposition (AACVD) were explored. Fe^{3+} doping at 0.05, 0.1, 0.5, and 1.0 mol% concentrations were investigated. Scanning electron microscopy (SEM) revealed that photoanodes of porous morphology were synthesised. X-ray diffraction (XRD) showed that anatase TiO_2 was successfully formed. PEC measurements indicated that increasing Fe^{3+} dopant concentration increased photocurrent density, however, an improvement upon pure TiO_2 was achieved solely for 1.0 mol%. The beneficial effect of Fe^{3+} doping at 1.0 mol% was also evident in ultraviolet-visible (UV-vis) spectroscopy results, showing a decrease in the band gap of TiO_2 and a red shift in absorption which can improve solar light harvesting.

Introduction

There is ample solar energy potential to provide global energy needs from a renewable, carbon-free power supply.[1] Photoelectrochemical (PEC) water splitting is a promising technology for solar light harvesting which can convert solar energy to storable hydrogen fuel. [2] As a photoanode material in PEC cells, TiO_2 is a popular n-type semiconductor which has been studied extensively due to its high photocatalytic efficiency, low-cost, non-toxicity, chemical inertness, and photostability. [3] However, due to its wide band gap meaning only 5% of the solar spectrum is used, [4] and fast recombination of hole-electron pairs, [5] the large-scale applications of TiO_2 in photocatalysis are limited. [6]

Consequently, over the past few decades, much research has focussed on reducing hole-electron recombination rates and improving light absorption through band gap regulation, morphology control, and the construction of heterogeneous junctions. [7]

Though TiO_2 exists in amorphous form and different crystalline polymorphs including anatase, rutile, and brookite, anatase TiO_2 shows the greatest photocatalytic activity. [8] Therefore, one approach is to synthesise nanostructured anatase TiO_2 photoanodes given the higher surface area that promotes charge transfer across larger solid-liquid interfaces, shortened charge carrier pathways, and induced light scattering which facilitates the generation of multiple hole-electron pairs. [9] There are many possible synthesis methods including sol-gel, micelle and inverse micelle, hydrothermal, and chemical vapour deposition. [10] Of these, chemical vapour deposition (CVD), specifically aerosol-assisted, is a promising fabrication technique for large-scale PEC devices, capable of producing large-area robust films with good reproducibility and relatively low processing cost. [11]

Doping can also enhance the PEC performance of TiO_2 photoanodes by increasing light absorption and supressing recombination, [12] ensuring efficient photocatalysis. [13] Specifically, Fe^{3+} is a favourable doping candidate due to its ease of incorporation into the TiO_2 structure given the similar atomic radii of Ti^{4+} and Fe^{3+}, and Fe^{3+} ions possessing the ability to trap photogenerated electrons and holes, reducing recombination rate. [6]

It is expected that Fe^{3+} doping coupled with aerosol-assisted chemical vapour deposition (AACVD), has the potential to address the limitations of TiO_2 in PEC usage. Hence, to the best of our knowledge, this paper presents the first instance of the formation of Fe^{3+} doped nanostructured anatase TiO_2 photoanodes formed by AACVD.

Background

Water can be split into molecular oxygen and hydrogen using a PEC cell which extracts electrical energy from sunlight. [14,15] A PEC water splitting cell requires one or two semiconducting electrodes which absorb solar photons to produce charge carriers, and a membrane to separate the products of the two half-cell reactions, oxygen and hydrogen. [12,16] When the energy of an incident photon is equal to or greater than the band gap of the semiconductor material, an electron is excited from the valence band to the conduction band, generating a free electron and hole which participate in oxidation (1) and reduction (2) reactions, respectively.

$$2H_2O + 4h^+ \leftrightarrow 4H^+ + O_2 \quad (1)$$

$4H^+ + 4e^- \leftrightarrow 2H_2$ (2)

During this process, a large proportion of hole-electron pairs recombine at the surface and in the bulk material, dissipating energy in the form of light or heat. [17] Fast hole-electron recombination is detrimental to the PEC performance of a semiconductor as it reduces the number of charge carriers that can react in these photocatalytic reactions.

A solar-to-hydrogen (STH) conversion efficiency of 10% is required for commercialisation of PEC water splitting, as well as cheap manufacturing and long-term stability of photoelectrodes. [18] To ensure a high STH conversion efficiency, the conduction and valence band positions should be more negative than the hydrogen evolution potential and more positive than the oxygen evolution potential, respectively, with a band gap of appropriate size to ensure a sufficient portion of the solar spectrum is absorbed. Water splitting requires a 1.23V difference between the redox levels of the oxidation and reduction reactions at 20°C. [12]

The empty d band of Ti^{4+} in TiO_2 means the valence band energy is heavily influenced by the O 2p levels, hence the valence band is low in energy and stable towards oxidation under PEC usage. [12] However, the highly ionic character of TiO_2 produces a band gap that is too large for efficient solar light harvesting as it cannot absorb visible light, which accounts for approximately 40% of the solar spectrum. [17]

Therefore, nanostructured anatase TiO_2 films with high-energy facets exposed such as {0 1 0} or {0 0 1} have sparked great interest, especially the {0 1 0} facet owing to its favourable surface atomic and electronic structure. [19] AACVD has been proven to produce optimal anatase TiO_2 nanostructures with desert rose morphology and high exposure of the most active facet, {0 1 0}. [20]

To perform AACVD, an aerosol is generated by atomising the chemical precursors dissolved in solvent into finely divided sub-micrometre liquid droplets which are distributed throughout a gas medium, to be delivered into a heated zone. [21] The solvent is rapidly evaporated or combusted, and the chemical precursor deposits the desired film. [21]

By adding dopants to the chemical precursor, impurities can be introduced into pure TiO_2 to enhance conductivity. There are two types of doping: n-doping and p-doping.

In n-doping, the dopant atoms act as donors and conductivity is based on free electrons as they are the majority charge carrier. In contrast, for p-doping dopant atoms act as acceptors, and free holes are the majority charge carrier and therefore determine conductivity. The band structure for n-doping involves the Fermi level shifting to just below the conduction band, whereas for p-doping it lies just above the valence band (Figure 1). [22]

TiO_2 has been doped with metal and non-metal dopants to address its limitations. Mahmoud et al [23] adopted the hydrothermal method to dope with Mn, Cd, and Cu. Mn and Cd improved visible light absorption by 35% and 21.9%, respectively, and reduced band gap. Cu, however, only slightly improved light absorption but enhanced charge carrier lifetime. All dopants significantly improved photocatalytic activity towards water splitting. Regue et al [24] found that Mo-doped TiO_2 photoanodes prepared by spray pyrolysis also outperform TiO_2 photoanodes, by a factor of two in terms of photocurrent. Wang et al [25] synthesised N-doped TiO_2 films by magnetron sputtering and improved photocatalytic activity by controlling the preferred orientation of the deposited films.

Iron doping of TiO_2 has also been widely investigated. Romero et al [26] produced Fe^{3+} doped TiO_2 through calcination of a Ti-containing metal organic framework (MIL-125) obtained via hydrothermal synthesis. Increased visible light absorption and reduced recombination rate were achieved, with the optimal photocatalytic performance for water splitting being amongst the highest reported in literature, obtained at 0.5 wt.% Fe^{3+}.

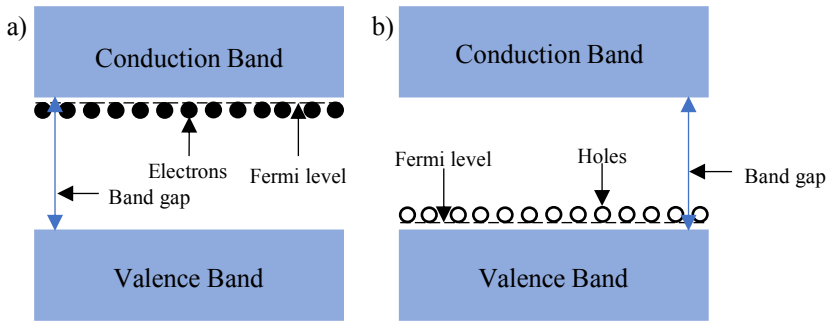


Figure 1: Band structure for n-doping (a) and p-doping (b)

Fe^{3+} doped TiO_2 nanostructures have been formed by methods including sol-gel, [27] magnetron sputtering, [28] and molten salt, [29] with reported optimum Fe^{3+} concentrations as 0.2 at.%, 1.1 at.%, and 0.5 wt.% respectively. A decrease in band gap as well as a shift in the absorption edge towards the visible light region was observed. In all cases, the addition of Fe^{3+} resulted in higher hydrogen production in comparison to pure TiO_2 and other types of metal doped TiO_2 photoanodes.

As indicated by previous studies, different synthesis methods can lead to different outcomes of optimal Fe^{3+} doping. This may be due to different synthesis methods accommodating Fe^{3+} ions in distinctive positions in the TiO_2 structure. [26] The aim of this study is to address a gap in literature by reporting the effect of Fe^{3+} doping on TiO_2 films formed by AACVD for the first time.

Methods

Materials

Ethanol (96 vol%) and toluene (0.02 vol% water) were sourced from VWR Chemicals BDH, and titanium (IV) ethoxide $Ti(OEt)_4$ and iron (III) chloride hexahydrate were sourced from Sigma-Aldrich. Aluminoborosilicate glass (ABS) coated with a fluorine-doped tin oxide (FTO) transparent conductive layer ($7\Omega\ s^{-1}$) was provided by Sigma-Aldrich and cut in 2.5cm-wide and 10cm-long substrates, for their use in the AACVD system. These substrates were cleaned ultrasonically using a RS PRO Ultrasonic Cleaner with Hellmanex III solution, isopropyl alcohol, and deionised water (each for 10mins), and finally dried with compressed air.

Synthesis of $Ti_7O_4(OEt)_{20}$

$Ti_7O_4(OEt)_{20}$ titanium oxo/ethoxy cluster was synthesised via controlled hydrolysis in toluene according to a procedure derived from Eslava et al [30]. A mixture of deionised water (0.68mL) and ethanol (10mL) was added dropwise to a solution of titanium (IV) ethoxide (14mL) in anhydrous toluene (30mL) under a nitrogen atmosphere. The mixture was stirred overnight, and subsequent solvent evaporation yielded a white/yellowish crystalline solid precipitate of $Ti_7O_4(OEt)_{20}$.

Preparation of TiO_2 Photoanode

The $Ti_7O_4(OEt)_{20}$ precipitate was washed in toluene to make a combined solution (185mL). Iron (III) chloride hexahydrate was added for iron doping at 0.05, 0.1, 0.5, and 1.0 mol% concentrations. The mixture was stirred vigorously to ensure the dopant was fully dissolved.

All doped TiO_2 photoanodes were prepared using AACVD. Depositions were carried out onto FTO-ABS substrates, which were placed horizontally inside a tube furnace. Aerosol droplets were produced using a TSI Model 3076 Constant Output Atomiser, with nitrogen used as a carrier gas at a pressure of 0.9 bar. Following deposition at 450°C for 1.5 hours, the substrate was left to cool under nitrogen flow. To remove carbon, the produced films were annealed in air at 800°C for 2h in a three-zone tube furnace for CVD using a heating rate of $10^oC\ min^{-1}$ after which they were left to cool in air.

PEC Measurements

To evaluate the PEC performance of the prepared photoanodes, the CompactStat. potentiostat by Ivium Technologies was used and photocurrents were measured under chopped simulated sunlight (AM 1.5G, 100 mW cm^{-2}) from a filtered 300W xenon lamp source (Lot Quantum Design). PEC cells were prepared using a three-electrode configuration consisting of a silver chloride reference electrode (Ag/AgCl 3.5M KCl), Pt as the counter electrode, and the photoanode as the working electrode. The pH of the aqueous NaOH electrolyte solution was 13.9. Front illumination was used, and photocurrent-potential curves were recorded at a scan rate of 20 mV s^{-1}. The Nernst equation (3) [31] was applied to convert the measured Ag/AgCl potentials ($E_{Ag/AgCl}$) to RHE potentials (E_{RHE}).

$$E_{RHE} = 0.1976\ V + 0.059\ pH + E_{Ag/AgCl} \quad (3)$$

Characterisation

Ultraviolet-visible (UV-vis) absorption and reflectance spectra were measured using Shimadzu IRS 2600PLUS UV-vis spectrophotometer for wavelengths of 200-800nm. X-ray diffraction (XRD) patterns were collected from 10-80° (2θ) Bragg-Brentano using Bruker AXS D8 Advance with Cu Kα (0.154 nm) radiation, 0.023°(2θ) step size and a total integration time of 1020s. Scanning electron microscopy (SEM) micrographs were obtained using LEO Gemini 1525 field emission gun scanning electron microscope (FEGSEM) using an acceleration voltage of 5kV.

Results and Discussion

SEM

Figure 2 displays SEM micrographs of pure TiO_2 and Fe^{3+} doped TiO_2 . It can be observed that porous morphology was synthesised for all samples. Pure TiO_2 (Figure 2a) was determined to have a similar structure to what is expected, (Figure 2b) according

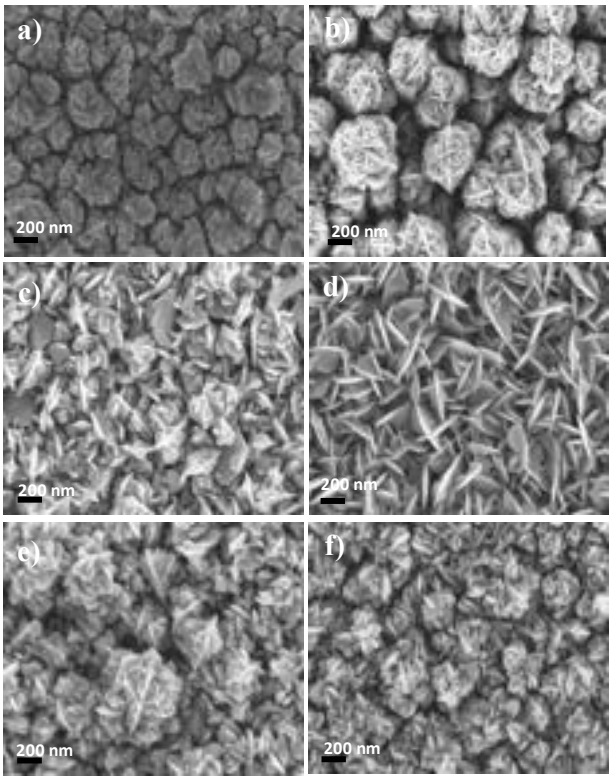


Figure 2: Scanning electron microscopy (SEM) images for (a) pure TiO_2 (b) expected pure TiO_2 desert rose morphology (c) 0.05 mol% Fe^{3+}-doped TiO_2 (d) 0.1 mol% Fe^{3+}-doped TiO_2 (e) 0.5 mol% Fe^{3+}-doped TiO_2 (f) 1 mol% Fe^{3+}-doped TiO_2

to findings by Regue et al [20] but does not exhibit the same unique desert rose morphology.

There are also general structural differences between the pure and doped samples in terms of size and shape. Factors such as carrier gas flowrate, precursor synthesis and temperature are known to influence morphology and therefore could explain this. The most likely cause of these variances is the precursor, as the addition of varying amounts of Fe^{3+} to obtain each concentration could have altered the reaction and decomposition paths which could affect morphology, as seen in previous studies. [9] There could also be slight variances between the precursor solutions as precursor synthesis was carried out multiple times.

Although it was aimed to produce and use identical precursors, this was not possible due to the dropwise addition of the ethanol-water mixture by hand as well as the time until use of the precursor solution each time, given its highly sensitive nature to air and water. Nevertheless, it was attempted to control this by staggering the dropwise addition of the ethanol-water mixture and creating a tight seal around the precursor bottle after each synthesis.

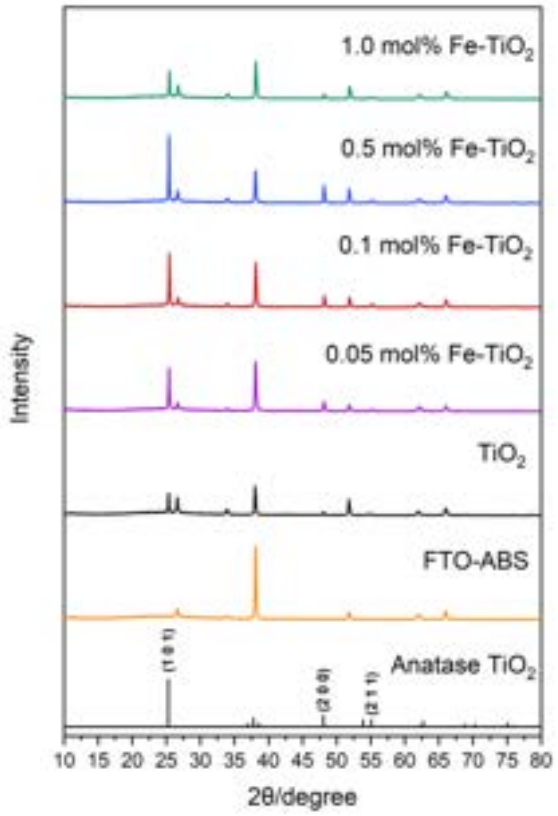


Figure 3 – X-ray diffraction (XRD) patterns of Fe^{3+} doped TiO_2, pure TiO_2, FTO-ABS substrate, and anatase TiO_2 (ICDD-JCPDS 71-1166)

XRD

Figure 3 shows the XRD patterns of the Fe^{3+} doped and pure TiO_2 films on FTO-ABS substrates, as well as the FTO-ABS substrate itself. No diffraction peaks associated with iron were observed for Fe^{3+} doped TiO_2 films, seemingly indicating that Fe^{3+} ions were successfully incorporated into the TiO_2 lattice structure without Fe_2O_3 formation on the surface of TiO_2 or any other impurity states. [32] Alternatively, the absence could be attributed to the very small iron content being undetected by XRD, which is very likely. All the diffraction peaks of TiO_2 can be assigned to the anatase phase of TiO_2 (ICDD-JCPDS 71-1166). Specifically, the diffraction peaks at 25.2°, 48.0°, and 55.1° (2θ) correspond to the (1 0 1), (2 0 0), and (2 1 1) diffraction planes. The other diffraction peaks can be assigned to the FTO-ABS substrate. The confirmation of anatase phase TiO_2 being the only present phase is important for maximal photocatalytic activity due to improved charge carrier migration to the surface. Doping with Fe^{3+} did not change the phase of TiO_2, as desired. To identify the high exposure facets, transmission electron microscopy (TEM) should be carried out.

UV-vis Spectroscopy

Figure 4a shows the relation between absorbance and doping concentration. Fe^{3+} doping did not have a significant effect on the visible light (400-700nm) absorption of TiO_2. However, a small increase was observed at 1.0 mol% Fe^{3+} doping, which is due to this sample possessing a smaller band gap than pure TiO_2.

In the UV region (200-400 nm), similarly, only 1.0 mol% Fe^{3+} doping resulted in increased absorbance in comparison to pure TiO_2. It can also be seen in Figure 4b that there is a red shift in absorbance for 1.0mol% Fe^{3+} doping. This red shift is likely to be a result of the successful inclusion of Fe^{3+} into the TiO_2 crystal structure, [32] and excitation of Fe^{3+} 3d electrons to the TiO_2 conduction band. [34] This effect has also been widely observed in previous instances of Fe^{3+} doping. [32,33,35,36]

The band gap was determined for each sample using data from reflectance and Tauc's relation:

$$\alpha h\nu = A(h\nu - E_g)^n \quad (4)$$

Where α is the absorption coefficient in m^{-1} , $h\nu$ is the photon energy in eV, A is a constant, E_g is bandgap in eV, and n is an index of value 0.5 and 2 for indirect and direct transitions, respectively. [33] In this case, n = 2 was chosen as anatase TiO_2 has an indirect band gap.

The Tauc plots obtained for each doping concentration can be seen in Figure 4d. The band gap was subsequently determined by drawing two tangents to the curve and taking the value at the point of their intersection. An example of this for pure TiO_2 can be seen in Figure 4e. The band gaps have been summarised in Table 1.

Table 1: Band gap values determined for pure TiO_2 and Fe^{3+} doped TiO_2 at different concentrations

Fe^{3+} concentration (mol %)	Bandgap (eV)
0	3.273
0.05	3.276
0.1	3.286
0.5	3.285
1	3.268

The value determined for pure TiO_2, 3.273 eV, has good agreement with the value found in literature as 3.2 eV. [37] 1.0 mol% Fe^{3+} doping resulted in a decreased band gap which agrees with the increased visible light absorbance observed. This is due to doping introducing new Fe^{3+}/Fe^{2+} energy levels which result in the excitation of 3d Fe^{3+} electrons to the TiO_2 conduction band. [38]

0.05 mol%, 0.1 mol%, and 0.5 mol% Fe^{3+} doping resulted in larger band gaps than pure TiO_2 which

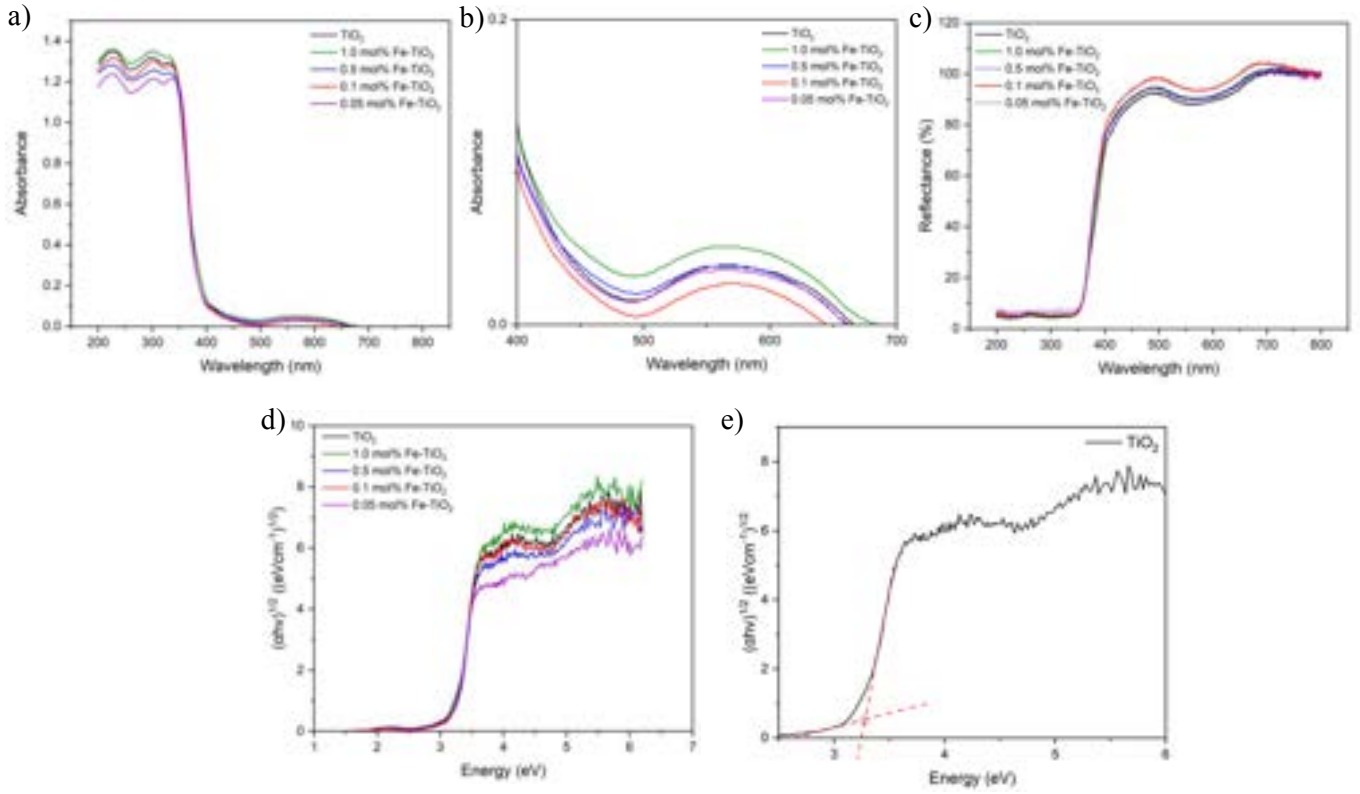


Figure 4: UV-Vis (ultraviolet -visible) (a) absorption spectra (b) absorption spectra in the visible light region (c) reflectance spectra (d) Tauc plot for TiO_2 and Fe^{3+} doped TiO_2 at different concentrations (e) Tauc plot for pure TiO_2 demonstrating how band gap was obtained

agrees with the lower absorbance observed in comparison. These results differ from trends in literature which generally show reduced band gaps and increased light absorbance for all Fe^{3+} concentrations. [32,33,35,36] This could potentially be attributed to Fe^{3+} acting as recombination centres, reducing photocatalytic activity and absorbance. This effect will be discussed in further detail in the PEC Measurements section.

The reason for reduced band gap not being observed for 0.05 mol%, 0.5 mol% and 0.1 mol% Fe^{3+} doping could potentially be Fe^{3+} ions not having an influence on the band structure due to the doping concentration being too low.

Finally, it is possible that the thickness of the deposition had an influence on absorption. The thickness of the deposition can change between experiments for many reasons. The dominant factor was determined to be cloudiness of the precursor solution caused by rapid addition of the ethanol-water mixture and was mitigated against as much as possible as described previously in the SEM section. The level of precursor liquid was also kept constant between experiments to ensure the flow of nitrogen was sufficient for adequate deposition to occur.

Photoelectrochemical Characterisation

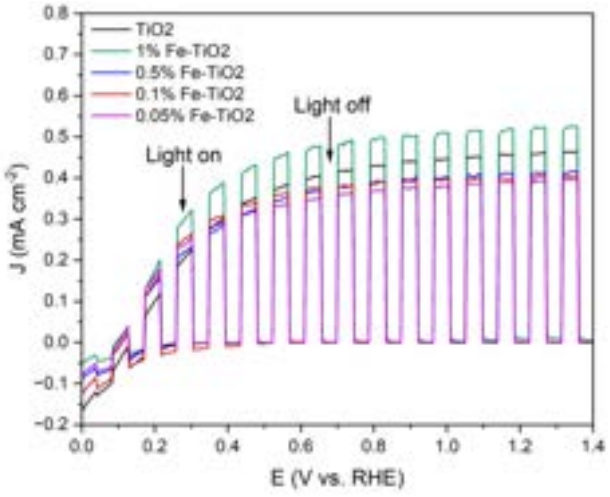


Figure 5: Photocurrent potential curves of TiO_2 and Fe^{3+} doped TiO_2 at different concentrations under 1 sun chopped illumination (AM 1.5G, 100mWcm^{-2})

Figure 5 presents the PEC performance of the samples. Fe^{3+} doping significantly affects the photocurrent density, with higher iron doping resulting in higher photocurrent density. This is because Fe^{3+} plays the role of an intermediate for the efficient separation of photogenerated electron-hole pairs. [39] Fe^{3+} can trap electrons due the Fe^{3+}/Fe^{2+} energy level being below the conduction band edge of TiO_2. Fe^{2+} can be oxidised to Fe^{3+} by transferring electrons to absorbed O_2 on the surface of TiO_2. Simultaneously, Fe^{3+} can trap holes due to the Fe^{4+}/Fe^{3+} energy level being above the valence band edge of TiO_2. The trapped holes in Fe^{4+} can migrate to the surface and absorb hydroxyl ions to produce hydroxyl radicals. [40] This inhibits the recombination of hole-electron pairs and improves photocatalytic activity. Fe^{3+} is a relatively stable ion due to the semi-full 3d electronic configuration, therefore, a charge trapped by Fe^{2+} or Fe^{4+} can easily return to the Fe^{3+} state and participate in the photocatalytic reactions. [40] However, when the concentration of Fe^{3+} becomes too large, Fe^{3+} can act as recombination centres for hole-electron pairs due to the decrease in distance between trapping sites. [41] It is unlikely that an excess Fe^{3+} concentration was reached in this study as there is no peak in photocurrent density.

Surprisingly, compared to pure TiO_2, lower concentrations of doping seemed to decrease the photocurrent density. There was only an improvement upon pure TiO_2 for Fe^{3+} doping of 1.0 mol% at which an optimum photocurrent of 0.52 mA cm^{-2} was achieved at +1.23 V_{RHE}. A potential explanation is Fe^{3+} doping having a net negative effect on the photocurrent density at lower concentrations, with the prominent effect of Fe^{3+} doping being charge recombination centres, rather than an increase in the concentration of holes and electrons available to react. While it should be noted that different fabrication techniques can lead to differences in photocatalytic activities, previous studies were consulted to interpret these results. Previously, increasing Fe^{3+} dopant concentration has led to increased photocurrent density until an excess Fe^{3+} concentration is reached. [36,39,42] However, decreasing photocurrent densities with increasing Fe^{3+} concentration have also been reported and attributed to defect states acting as recombination centres for charge carriers. [43] A decrease in photocatalytic activity with Fe^{3+} doping of TiO_2 was found by another study which reasoned this outcome with an unfavourable location of Fe^{3+} inside the interior matrix of TiO_2 rather than on the exterior surface, [44] where it could act as trap sites for eventual charge transfer at the interface. [45] Further structural characterisation techniques such as X-ray photoelectron spectroscopy (XPS) can be used to understand the nature of the Fe^{3+} incorporation into the TiO_2 structure.

However, the increasing photocurrent density with increasing doping concentration suggests there is a beneficial effect of Fe^{3+} doping, which dominates at 1.0 mol%.

Conclusion

It can be concluded that Fe^{3+} doping of nanostructured TiO_2 photoanodes has the potential to address the limitations of TiO_2. The maximum photocurrent density was obtained with the 1.0 mol% Fe^{3+} doped TiO_2 film which is higher than that achieved with the undoped TiO_2 film. At lower concentrations, a decrease in photocurrent was observed compared to pure TiO_2 which could be due to Fe^{3+} acting as hole-electron recombination centres.

The XRD results confirmed that anatase TiO_2 was successfully formed, the most photocatalytically active crystalline form of TiO_2. Fe^{3+} doping did not have an impact on the TiO_2 phase synthesised, and anatase phase was maintained, as desired. SEM micrographs verified the production of porous morphology in all samples. However, structural differences, in especially the structure size and shape were seen across the samples because of variances in the precursor used. UV-vis spectroscopy confirmed that Fe^{3+} doping reduced band gap at 1.0 mol%, with increased light absorption in the UV and visible ranges.

More experiments need to be conducted to confirm the effect of Fe^{3+} doping on the performance of TiO_2 photoanodes. By testing higher concentrations of Fe^{3+} doping, a peak in photocurrent density should be identified to be able to realise the full potential of Fe^{3+} doping in addressing the limitations of TiO_2. Further structural characterisation techniques such as XPS and TEM can be used to further understand the effect of Fe^{3+} doping on TiO_2.

Acknowledgements

The authors would like to express their sincere gratitude to Mengya Yang for her continued guidance and support throughout the entire duration of the project.

References

1. Development, O. f. E. C. a. *Solar energy perspectives;* International Energy Agency: Paris, 2011; , pp 228.

2. Napporn, T. W.; Holade, Y. *Metal oxide-based nanostructured electrocatalysts for fuel cells, electrolyzers, and metal-air batteries;* Elsevier: Amsterdam, 2021; .

3. Kim, C.; Choi, M.; Jang, J. Nitrogen-doped SiO2/TiO2 core/shell nanoparticles as highly efficient visible light photocatalyst. *Catalysis communications* 2010*, 11*, 378-382.

4. Colmenares, J. C.; Aramendía, M. A.; Marinas, A.; Marinas, J. M.; Urbano, F. J. Synthesis, characterization and photocatalytic activity of different metal-doped titania systems. *Applied catalysis. A, General* 2006*, 306*, 120-127.

5. Li, H.; Duan, X.; Liu, G.; Liu, X. Photochemical synthesis and characterization of Ag/TiO2 nanotube composites. *J Mater Sci* 2008*, 43*, 1669-1676.

6. Ismael, M. Enhanced photocatalytic hydrogen production and degradation of organic pollutants from Fe (III) doped TiO2 nanoparticles. *Journal of environmental chemical engineering* 2020*, 8*, 103676.

7. Šuligoj, A.; Arčon, I.; Mazaj, M.; Dražić, G.; Arčon, D.; Cool, P.; Štangar, U. L.; Tušar, N. N. Surface modified titanium dioxide using transition metals: nickel as a winning transition metal for solar light photocatalysis. *Journal of materials chemistry A : materials for energy and sustainability* 2018*, 6*, 9882-9892.

8. Reghunath, S.; Pinheiro, D.; KR, S. D. A review of hierarchical nanostructures of TiO2: Advances and applications. *Applied Surface Science Advances* 2021*, 3*, 100063.

9. Regue Grino, M. Development of Nanostructured Metal Oxides for Solar Fuels, ProQuest Dissertations Publishing, 2020.

10. Chen, X.; Mao, S. S. Titanium Dioxide Nanomaterials: Synthesis, Properties, Modifications, and Applications. *Chemical reviews* 2007*, 107*, 2891-2959.

11. Regue, M.; Ahmet, I. Y.; Bassi, P. S.; Johnson, A. L.; Fiechter, S.; van de Krol, R.; Abdi, F. F.; Eslava, S. Zn-Doped Fe 2 TiO 5 Pseudobrookite-Based Photoanodes Grown by Aerosol-Assisted Chemical Vapor Deposition. *ACS applied energy materials* 2020*, 3*, 12066-12077.

12. Sivula, K.; van de Krol, R. Semiconducting materials for photoelectrochemical energy conversion. *Nature reviews. Materials* 2016*, 1*.

13. Ahmed, S.; Rasul, M. G.; Martens, W. N.; Brown, R.; Hashib, M. A. Heterogeneous photocatalytic degradation of phenols in wastewater: A review on current status and developments. *Desalination* 2010*, 261*, 3-18.

14. Thangamuthu, M.; Ye, J.; Xiong, L. Solar Driven Water Splitting. *https://www.ucl.ac.uk/solar-energy-advanced-materials/solar-driven-water-splitting* (accessed 04/12/22) .

15. Decker, F.; Cattarin, S. PHOTOELECTROCHEMICAL CELLS | Overview. In *Encyclopedia of Electrochemical Power Sources*2009; Vol. 4, pp 1-9.

16. Landman, A.; Dotan, H.; Shter, G. E.; Wullenkord, M.; Houaijia, A.; Maljusch, A.; Grader, G. S.; Rothschild, A. Photoelectrochemical water splitting in separate oxygen and hydrogen cells. *Nature materials* 2017*, 16*, 646-651.

17. Kang, X.; Liu, S.; Dai, Z.; He, Y.; Song, X.; Tan, Z. Titanium Dioxide: From Engineering to Applications. *Catalysts* 2019*, 9*, 191.

18. Yu, J. M.; Lee, J.; Kim, Y. S.; Song, J.; Oh, J.; Lee, S. M.; Jeong, M.; Kim, Y.; Kwak, J. H.; Cho, S.; Yang, C.; Jang, J. High-performance and stable photoelectrochemical water splitting cell with organic-photoactive-layer-based photoanode. *Nature communications* 2020*, 11*, 5509.

19. Pan, J.; Liu, G.; Lu, G. Q. (.; Cheng, H. On the True Photoreactivity Order of {001}, {010}, and {101} Facets of Anatase TiO_2 Crystals. *Angewandte Chemie (International ed.)* 2011*, 50*, 2133-2137.

20. Regue, M.; Sibby, S.; Ahmet, I. Y.; Friedrich, D.; Abdi, F. F.; Johnson, A. L.; Eslava, S. TiO 2 photoanodes with exposed {0 1 0} facets grown by aerosol-assisted chemical vapor deposition of a titanium oxo/alkoxy cluster. *Journal of materials chemistry. A, Materials for energy and sustainability* 2019*, 7*, 19161-19172.

21. Choy, K. L. Chemical vapour deposition of coatings. *Progress in materials science* 2003*, 48*, 57-170.

22. Anonymous 8.5: Semiconductors- Band Gaps, Colors, Conductivity and Doping. *https://chem.libretexts.org/Courses/Barry_University/CHE360%3A_Inorganic_Chemistry/08%3A_The_Crystalline_Solid_State/8.05%3A_Semiconductors-_Band_Gaps_Colors_Conductivity_and_Doping* (accessed 12/12/22)

23. Mahmoud, M. S.; Ahmed, E.; Farghali, A. A.; Zaki, A. H.; Abdelghani, E. A. M.; Barakat, N. A. M. Influence of Mn, Cu, and Cd–doping for titanium oxide nanotubes on the photocatalytic activity toward water splitting under visible light irradiation. *Colloids and surfaces. A, Physicochemical and engineering aspects* 2018*, 554*, 100-109.

24. Miriam Regue; Katherine Armstrong; Dominic Walsh; Emma Richards; Andrew L Johnson; Salvador Eslava Mo-doped TiO2 photoanodes using [Ti4Mo2O8(OEt)10]2 bimetallic oxo cages as a single source precursor. *Sustainable energy & fuels* 2018*, 2*, 2674-2686.

25. Wang, C.; Hu, Q.; Huang, J.; Deng, Z.; Shi, H.; Wu, L.; Liu, Z.; Cao, Y. Effective water splitting using N-doped TiO2 films: Role of preferred orientation on hydrogen production. *International journal of hydrogen energy* 2014*, 39*, 1967-1971.

26. Valero-Romero, M. J.; Santaclara, J. G.; Oar-Arteta, L.; van Koppen, L.; Osadchii, D. Y.; Gascon, J.; Kapteijn, F. Photocatalytic properties of TiO2 and Fe-doped TiO2 prepared by metal organic framework-mediated synthesis. *Chemical engineering journal (Lausanne, Switzerland : 1996)* 2019*, 360*, 75-88.

27. Singh, A. P.; Kumari, S.; Shrivastav, R.; Dass, S.; Satsangi, V. R. Iron doped nanostructured TiO2 for photoelectrochemical generation of hydrogen. *International journal of hydrogen energy* 2008*, 33*, 5363-5368.

28. Dholam, R.; Patel, N.; Adami, M.; Miotello, A. Hydrogen production by photocatalytic water-splitting using Cr- or Fe-doped TiO2 composite thin films photocatalyst. *International journal of hydrogen energy* 2009*, 34*, 5337-5346.

29. Mohammad Ghorbanpour; Atabak Feizi Iron-doped TiO2 Catalysts with Photocatalytic Activity. *Journal of Water and Environmental Nanotechnology* 2019*, 4*, 60-66.

30. Eslava, S.; Goodwill, B. P. R.; McPartlin, M.; Wright, D. S. Extending the Family of Titanium Heterometallic–oxo–alkoxy Cages. *Inorganic chemistry* 2011*, 50*, 5655-5662.

31. Mohammadi, T.; Sharifi, S.; Ghayeb, Y.; Sharifi, T.; Mohamad Mohsen Momeni Photoelectrochemical Water Splitting and H2 Generation Enhancement Using an Effective Surface Modification of W-Doped TiO2 Nanotubes (WT) with Co-Deposition of Transition Metal Ions. *Sustainability (Basel, Switzerland)* 2022*, 14*, 13251.

32. Ali, T.; Tripathi, P.; Azam, A.; Raza, W.; Ahmed, A. S.; Ahmed, A.; Muneer, M. Photocatalytic performance of Fe-doped TiO2 nanoparticles under visible-light irradiation. *MRX* 2017*, 4*.

33. Shyniya, C. R.; Bhabu, K. A.; Rajasekaran, T. R. Enhanced electrochemical behavior of novel acceptor doped titanium dioxide catalysts for photocatalytic applications. *J Mater Sci: Mater Electron* 2017*, 28*, 6959-6970.

34. Einert, M.; Hartmann, P.; Smarsly, B.; Brezesinski, T. Quasi-homogenous photocatalysis of quantum-sized Fe-doped TiO2 in optically transparent aqueous dispersions. *Scientific reports* 2021*, 11*, 17687.

35. Tong, T.; Zhang, J.; Tian, B.; Chen, F.; He, D. Preparation of Fe3+-doped TiO2 catalysts by controlled hydrolysis of titanium alkoxide and study on their photocatalytic activity for methyl orange degradation. *Journal of hazardous materials* 2008*, 155*, 572-579.

36. Elghniji, K.; Atyaoui, A.; Livraghi, S.; Bousselmi, L.; Giamello, E.; Ksibi, M. Synthesis and characterization of Fe3+ doped TiO2 nanoparticles and films and their performance for photocurrent response under UV illumination. *Journal of alloys and compounds* 2012*, 541*, 421-427.

37. Zhu, T.; Gao, S. The Stability, Electronic Structure, and Optical Property of TiO2 Polymorphs. *Journal of physical chemistry. C* 2014*, 118*, 11385-11396.

38. Carneiro, J. O.; Azevedo, S.; Fernandes, F.; Freitas, E.; Pereira, M.; Tavares, C. J.; Lanceros-Méndez, S.; Teixeira, V. Synthesis of iron-doped TiO2 nanoparticles by ball-milling process: the influence of process parameters on the structural, optical, magnetic, and photocatalytic properties. *J Mater Sci* 2014*, 49*, 7476-7488.

39. Lei, J.; Li, X.; Li, W.; Sun, F.; Lu, D.; Yi, J. Arrayed porous iron-doped TiO2 as photoelectrocatalyst with controllable pore size. *International journal of hydrogen energy* 2011*, 36*, 8167-8172.

40. Sun, L.; LI, J.; Wang, C. L.; Li, S. F.; Chen, H. B.; Lin, C. J. An electrochemical strategy of doping Fe3+ into TiO2 nanotube array films for enhancement in photocatalytic activity. *Solar energy materials and solar cells* 2009*, 93*, 1875-1880.

41. Cong, Y.; Zhang, J.; Chen, F.; Anpo, M.; He, D. Preparation, Photocatalytic Activity, and Mechanism of Nano-TiO2 Co-Doped with Nitrogen and Iron (III). *Journal of physical chemistry. C* 2007*, 111*, 10618-10623.

42. Lee, T.; Ryu, H.; Lee, W. Photoelectrochemical properties of iron (III)-doped TiO2 nanorods. *Ceram. Int.* 2015*, 41*, 7582-7589.

43. Hamilton, J. W. J.; Byrne, J. A.; McCullagh, C.; Dunlop, P. S. M. Electrochemical Investigation of Doped Titanium Dioxide. *International Journal of Photoenergy* 2008*, 2008*, 1-8.

44. Othman, S. H.; Abdul Rashid, S.; Mohd Ghazi, T. I.; Abdullah, N. Fe-Doped TiO2 Nanoparticles Produced via MOCVD: Synthesis, Characterization, and Photocatalytic Activity. *Journal of Nanomaterials* 2010*, 2011*.

45. Li, Z.; Hou, B.; Xu, Y.; Wu, D.; Sun, Y. Studies of Fe-doped SiO2/TiO2 composite nanoparticles prepared by sol-gel-hydrothermal method. *Journal of materials science* 2005*, 40*, 3939-3943.

Flexible Poly(ethylene glycol) Diacrylate/Acrylamide Microneedle Patch for Non-invasive, Continuous Glucose Monitoring

Tai Xuan Tan and Marieke De Bock

Abstract Non-invasive, robust, flexible, and reversible detection systems that are easily manufactured are an unmet biomedical need for providing continuous monitoring of glucose levels in diabetic patients. In this study, the development and successful integration of a hydrogel from poly(ethylene glycol) diacrylate (PEGDA) and acrylamide (AM) as a microneedle (MN) base is demonstrated. MN sensors are emerging as a non-invasive, point-of-care technology with great potential to replace traditional sampling methods for continuous glucose detection in interstitial fluid (ISF), but material limitations prevent their widespread usage. The hydrogel was formulated with a differing PEGDA:AM ratio through photopolymerization, and subsequently, the dimensions of the MN base were characterised using a light microscope. The PEGDA:AM ratio in the formulation was optimised to obtain the desired mechanical properties. The compression modulus was recorded at 0.174 ± 0.023 MPa at PEGDA:AM 70:30, indicating high flexibility, whilst also reporting excellent compressive strength at 35.0 ± 0.7 MPa for human skin insertion. These optimal mechanical properties were compared to those of MN patches formulated from poly(vinyl alcohol) (PVA), a frequently reported MN base material. An insertion test in porcine skin confirmed insertion capabilities as a MN base, achieving an insertion depth of 159 ± 1 µm. The MN base did not impose any optical limitation on signal transmission when attached to a Förster resonance energy transfer (FRET) biosensor, where the device provided a linear response (R^2 = 0.984) to variable glucose concentration in artificial ISF. These attractive properties of this newly proposed hydrogel, such as ease of preparation, flexibility and excellent mechanical performance, indicate its potential towards mainstream treatment of diabetes mellitus.

Keywords microneedle, continuous glucose monitoring, poly(ethylene glycol) diacrylate, acrylamide, mechanical properties, fluorescence

1. Introduction

1.1 Diabetes mellitus

The prevalence of diabetes mellitus, a lifelong condition in which the pancreas produces no or insufficient insulin to control glucose levels, is expected to increase by 10.1% from 2020 to 2030 according to WHO data sources [1]. Today, it is estimated that over 530 million adults worldwide are affected by this disease [2].

Patients diagnosed with diabetes mellitus suffer from abnormal blood glucose levels, leading to hypo- or hyperglycaemia. These levels must be monitored regularly to effectively administer insulin injections. The lack of strict regulation puts patients at risk for blindness, heart disease, or liver disease [3]. The glucose enzyme biosensor was introduced by Clark and Lyons in 1962 [4] and many discoveries have been made in the field since then. However, challenges remain in clinically accurate tight glycaemic monitoring. The most widely used technique for diabetes management is capillary blood sampling using a finger prick device with a test strip. This method is costly and invasive, making the user more prone to infections, and must be performed daily, especially for type 1 patients [5].

1.2 Continuous glucose monitoring

Continuous glucose monitoring (CGM) has significant advantages over traditional intermittent monitoring, such as trend prediction capabilities and detection of unsuspected hypo- or hyperglycaemia [6]. In 2019, the U.S. Food and Drug Administration had approved nine CGM devices for commercial use, of which only one based on a fluorescence detection mechanism: Eversense (Senseonics, Germantown, MD, USA) [7]. Almost half of these devices analyse ISF, found in the surroundings of cells in the dermal layer. This body fluid is formed by capillary filtration through blood and therefore has a composition similar to that of plasma. The composition provides a reliable alternative to blood biomarker concentrations for a wide range of health-related parameters, including glucose [8]. Bruen et al. (2017) reported that the blood glucose concentration of a normal healthy adult ranges from 4.9 to 6.9 mM, while for diabetic patients it is reported to vary between 2.0 and 40.0 mM [9]. In comparison, the glucose concentration in ISF has a smaller range, between 3.9 and 6.6 mM for healthy adults and between 2.0 and 22.2 mM for diabetic patients. Furthermore, ISF composition is not influenced by fluctuating flow rates, lack of fluid replenishment, or sample dilution, which are issues that commonly affect other biofluids [8]. These interesting properties of ISF are utilized by microneedle (MN) technology to carry out non-invasive detection. Despite major advances in this field, many CGM wearables still face barriers such as the need for periodic replacement, re-calibrations, lack of clear interpretation of results, or discomfort of the device [10], which has resulted in limited implementation.

1.3 Microneedle patches

One of the most promising approaches for CGM is using a MN array. These can be applied directly to the skin or organs with applications in local drug delivery and detection [9]. The patches, which measure biomarkers in ISF, are attractive as micrometre-sized needles penetrate the

protective dermal barrier, but do not reach deep enough to initiate a painful sensation. Direct measurement occurs through mechanisms such as electrochemical, optical, magnetic, or colorimetric read-out [6]. MNs with fluorescent sensors have emerged in applications due to their high sensitivity and specificity, ease of use, low cost, and ability to measure biometrics. A typical technique, Förster resonance energy transfer, has reported a high signal-to-noise ratio [11] and is capable of elucidating dynamic interactions.

The arrays can be formulated from materials found in living organisms or from synthetic polymers for more customisable characteristics (Table 1). As a result of their biocompatibility and affinity with functional groups for fluorescence sensing, hydrogel MNs are discussed in this study. Initially manufactured in stainless steel with the intention of administering drugs, polymeric MNs have now shown great potential to measure various biomarkers including glucose [12]. Although the research field on MN sensor devices has been developing rapidly, the optimisation and comparison of biomaterials for general MN patch purposes has received considerably less attention.

1.4 Challenges and prospects

The fabrication of MNs as functional wearable for long-term use remains especially challenging. First, no flexible biosensor patch that allows continuous and reversible measurements has yet been reported. Fundamental aspects, such as durability and usability, are influenced by material selection and formulation. In addition, the fabrication of MNs must consider biocompatibility, cost, lifetime, and sensor accuracy [13]. However, there is little information at the moment on the suitability of the material for a MN sensor patch. This necessitates an exploration of materials and their characterisation to bring MN devices to the next level, which is the focus of this work.

This study introduced a hydrogel formulation from poly(ethylene glycol) diacrylate (PEGDA) and acrylamide (AM) and evaluates its ability as a MN base for continuous and reversible glucose sensing. First, the consistency of the dimensions of the material is discussed for different PEGDA:AM ratios and is compared to MNs fabricated using poly(vinyl alcohol). Second, the mechanical properties of the patch are characterised, which allows to quantify the flexibility, and the ability of the patch to pierce porcine skin is assessed. Finally, this report presents an assessment of the selected MN base material for FRET detection of glucose concentration when integrated with a biosensor layer.

2. Background

Hydrogels are an emerging class of polymeric materials that are broadly recognised to have potential in the biotechnology industry. These 3D-dimensional network materials can be found naturally, such as gelatine and chitosan, but also artificially synthesised, with prominent examples of PVA and perfluoromethyl vinyl ether, or a combination of both [14]. Hydrogels can be classified according to their cross-linking methods [15]. Physical cross-linking, by physical entanglement or interactions such as hydrogen bonding or van der Waals forces, reports good bioavailability. In chemically cross-linked hydrogels, the polymer chains are covalently linked, providing excellent mechanical strength. Photoinitiated chemical cross-linking occurs through radiation exposure, where photoinitiators 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 [16]. This cross-linking technique was applied in fabrication of PEGDA/AM hydrogels in this work.

The unique property of hydrogels, specifically non-solubility in water while hydrophilic [15], due to the presence of hydroxyl, carboxylic, amidic, sulfonic acid or primary amide group residues, offers many applications in the biomedical industry. Their matrix structure filled with water resembles living tissue and structures, making them suitable for applications to skin, such as MN patches or other wearables. The swelling capacity and other mechanical properties of this material can be tailored by changes in its chemical composition, length of the elastic chain, or cross-link density [17]. The variety of hydrogel materials offers great potential in achieving desirable characteristics of a MN base, such as flexibility, transparency, and durability.

2.1 Poly(ethylene glycol) diacrylate/Acrylamide

PEGDA is a monomer derivative of the biodegradable polymer polyethylene glycol (PEG) with two acrylate groups. Due to its ability to be cross-linked by various methods and its availability in a wide range of molecular sizes, it has tuneable mechanical properties [15]. Although it is a monomer, it acts as a crosslinker in this formulation, where it bonds linear chains together. AM is a commonly used hydrogel, having good hydrophilicity with its amide functional group ($-NH_2$). It is well known as a small molecular weight monomer [18]. However, because it is a monomer with only an acrylate group, it is only able to form linear polymer chains without a crosslinker present.

A mixture of AM and PEGDA forms a highly stretchable network, with the addition of a water-soluble photoinitiator, such as 2-hydroxy-2-methylpropiophenone [19]. Polyethylene glycol-bonded polyacrylamide hydrogel fibres have shown potential for continuous blood glucose detection, preserving their functionality for up to 140 days [20]. Only one MN has been reported that included both AM and PEGDA in its formulation [5], however, they were not combined instead incorporated in different layers of the patch. The PEGDA MN base was coated with colloidal crystals through the incorporation of a fluorophenylboronic acid-based matrix which included AM for attachment to the base layer. However, PEGDA alone as a MN base has been widely reported for transdermal drug delivery. Its precisely controlled structure and ability to cross-link in a short time were the deciding factors for PEGDA-based MN as it did not compromise the potency of the peptide [21]. Gao et al. have demonstrated the potential of film-coated PEGDA MNs for antibacterial applications [22], which is a useful

property for a CGM patch that is applied to the skin over a long period of time. However, there is a literature gap on reproducible fabrication methods and study of mechanical properties of the co-hydrogel formulated from PEGDA and AM, which will be addressed in this work.

2.2 Poly(vinyl alcohol)

The synthetic hydrogel PVA serves many applications in the biomedical field, amongst which in wearable devices. It is composed of linear polymer chains with a simple aliphatic [-C-C-] backbone and repeating hydroxyl (-OH) functional groups. This allows the formation of microcrystalline domains through hydrogen bonding, resulting in interesting properties, such as chemical resistance and high flexibility, while also being mechanically strong [23]. The simple chemical structure of PVA poses a minimal risk of toxicity to the human body [24]. Numerous projects such as those reported by Chen et al. (2021), He et al. (2020), and Coyne et al. (2017) have demonstrated that PVA is a suitable material with the necessary mechanical integrity for MN applications.

A key reason for the comparison of PVA to PEGDA/AM is its variety of fabrication methods [25]. Freezing and thawing is among the most widely used physical cross-linking methods in biotechnology. Through varying the freezing/thawing cycles, PVA can obtain viscoelastic properties that closely resemble those of natural human tissue. The drastic change in temperature encourages the formation of hydrogen bonds and hence increases the degree of cross-linking in the hydrogel. He et al. developed a MN sensor patch from a mixture of PVA and chitosan to extract ISF. The loading force required to break the needle tips was reported to be greater than 3.0 N per needle, indicating that the patch had the required mechanical strength to penetrate the stratum corneum and epidermis layer [26].

PVA can also be chemically cross-linked to fabricate MNs, as reported by Tekko et al. [27]. Cross-linking was performed through the addition of citric acid (CA), which acts as a nontoxic crosslinker. At high temperatures, CA converts to a cyclic anhydride, which esterifies with the hydroxyl groups present, forming chemical cross-links. The patches demonstrated sufficient mechanical strength and piercing ability in porcine skin, with a maximum force of 10.8 N for 15% PVA + CA 5% (w/w). Furthermore, PVA was selected as the patch material due to its transparency and ability to form fine films needed to observe the colour change [28]. In this study, the performance of the patches through both cross-linking methods will be evaluated.

2.3 Other polymeric materials

Despite copolymer poly(methylvinylether co. maleic acid) (PMVA/MA) having limited application due to its brittleness, when cross-linked with a plasticiser, such as PEG, it has excellent mechanical strength and antibacterial properties [29]. It was found that a patch with 121 needles consisting of 20% PMVA/MA and 7.5% PEG did not show deformation up to a force of 36.3 N, which is far sufficient for insertion into human skin [30].

Modifying natural hydrogel gelatine with methacrylic anhydride (MA) not only prevents thermal degradation, but also improves its elasticity and stiffness [31]. Gelatine methacrylate (GelMA) is optically transparent, indicating its suitability in signal transmission for fluorescence-based biosensors [32]. The highest compressive modulus of the MN patch designed by Zhu et al. (2020) was reported to be 7.23 MPa [33].

A third hydrogel reported for MN patches is methacrylated hyaluronic acid (MeHA). Hyaluronic acid (HA) is a nonsulfated glycosaminoglycan that can be found in human biofluid and almost all tissues [34]. Its linear chemical structure consists of repeating units of glucuronic acid and N-acetyl glucosamine [35]. This includes hydrophilic carboxyl and hydroxyl groups, allowing it to absorb water. Although HA alone provides a weak mechanical structure and quickly dissolves in aqueous solution, the covalent bonding with methacrylate groups increases rigidity and resistance to degradation while maintaining biocompatibility [36].

Table 1. Summary of MN materials for biochemical sensing and diagnosis

Material	Class	Young's modulus (MPa)	Ref.
PVA/PVP	Hydrogel	199-211	[37]
PVA	Hydrogel	10-200	[38]
MeHA	Hydrogel	0.175-0.218	[39]
PMVA/MA 5% : PEG 10,000	Hydrogel	71.1 – 115	[40]
Gelatine Methacrylate	Hydrogel	0.003-0.180	[41]
Clear Resin from Formlabs	Plastic	1600-2800	[42]

3. Methods

3.1 Materials

Poly(ethylene glycol) diacrylate (average M_W 700), acrylamide (purum, ≥ 98%), dimethyl sulfoxide, 2-hydroxy-2-methylpropiophenone (97%), poly(vinyl alcohol) (M_W 85000-124000, 87-89% hydrolysed), citric acid (anhydrous, ≥99.5%), 3-(acrylamido)phenylboronic acid (98%), acryloxyethyl thiocarbamoyl rhodamine B, N,N'-methylenebis(acrylamide) (99%) , calcium chloride (anhydrous, ≥93.0%), HEPES (≥ 99.5% (titration)), potassium chloride (ACS reagent, 99.0-100.5%), magnesium chloride (anhydrous, ≥98%), sodium chloride (100.0%), sodium phosphate monobasic (≥99.0%), sucrose (≥99.5%), D-(+)-glucose (anhydrous) and methylene blue were purchased from Sigma-Aldrich. Standard microneedle polydimethylsiloxane (PDMS) mould was obtained from Blueacre Technology (needle height 600 μm, needle base 300 μm, spacing 600 μm, array size 11x11). Ethanol absolute was obtained from VWR chemicals.

3.2 Equipment

An ultraviolet crosslinker (UVP CX-2000, Fisher Scientific) was used for the chemical cross-linking of PEGDA. Freezer (ES Series Combination, 363C-AEV-TS, Thermo Scientific) was used for freeze-thawing cycles. The EZ50 Universal Materials Testing Machine (Lloyd Instruments) with a 100N load cell was used to conduct

mechanical tests. The light microscope (Leica DM2700 P) was used to observe samples and capture pictures of experiment samples. The Microplate Reader (Varioskan LUX Multimode, ThermoFisher) was used to measure the intensities for different glucose concentrations.

3.3 Preparation of PEGDA/AM hydrogel

5 different PEGDA:AM ratios were formulated. These ratios were expressed as mole fractions, $\frac{mol_{PEGDA}}{mol_{PEGDA+mol_{AM}}}$ (% mol/mol of total monomers) and denoted as PEGDA/AM%, ranging from 50% to 90%. The concentration of total monomers used are kept constant as 7.5 M HMPP (11.45% v/v) was added as the photoinitiator for each solution. They were all dissolved in DMSO solvent at room temperature.

3.3.1 Preparation of PEGDA/AM MN

120 µL of each precursor solution was placed on the PDMS MN mould, covered with a glass slide with spacers (Figure 1). Glass slides were carefully lowered onto the PDMS mould to avoid trapping bubbles. They were UV cross-linked for 20 minutes, inverted, and then run for another 10 minutes to ensure that even cross-linking was achieved. The mould was carefully peeled off the MN patches at least 30 minutes after cross-linking. They were left to dry under the fume hood to detach themselves from the glass slides before being immersed in DI water overnight. This was done to prevent cracks or shattering due to rapid and uneven swelling. The dimensions of the MN were examined under the light microscope and measured to scale for 3 measurements per dimension per composition provided by the Leica computer software (LAS 14.2), connected to the microscope.

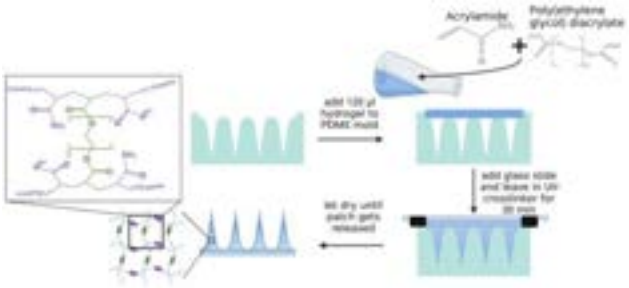


Fig. 1 Procedure of MN patch fabrication with PEGDA/AM hydrogel material via photoinitiated cross-linking

3.3.2 Preparation of PEGDA/AM hydrogel films

Thin films of the material were prepared for their consistency and ease of fabrication. 200 µL of each polymer solution was pipetted onto a flat polyester aluminised film, between tape strips attached acting as spacers, and then covered with glass slides. The setup was then UV cross-linked for 30 mins until solid hydrogel films were obtained. The films were peeled off the glass slides and stored for characterisation.

3.4 Preparation of PVA MN patches

3.4.1 Preparation of Freeze/Thawed PVA hydrogel (FT-PVA)

An aqueous 30% w/w PVA precursor solution was prepared by dissolving PVA in DI water at 90 °C for 4h. After being cooled to 37 °C, the solution was poured into the PDMS mould and covered with a glass slide. The freezing the/thawing cycle involves freezing hydrogel at -20 °C for 10h, then thawing at room temperature for 4h. The patch was removed from the mould and stored at room temperature.

3.4.2 Preparation of PVA + CA hydrogel

Aqueous 15% w/w PVA precursor solution was prepared through the same manner as FT-PVA. CA (1.5% w/w) was added and dissolved. The MN patches were prepared by pouring the precursor solution onto the mould, then placed in a vacuum chamber for 2h to remove trapped bubbles. They were dried in an oven at 60 °C for 1h, then cured at 130 ± 1 °C for 40min. The patch was then cooled to room temperature for storage.

3.5 Mechanical testing

3.5.1 Mechanical testing of films

3 sample films of each PEGDA:AM ratio were tested using the EZ50 universal testing machine in both compression and tensile mode. Their thickness was first measured using the light microscope, and surface areas were measured using image processing software (ImageJ) with a ruler as scale. The load exerted and the extension of the sample was measured as the upper platform was lowered onto the MN films to a maximum force of 50 N. All tests were run at 0.1 mm min^{-1}. The compressive modulus, E_c, of the films was calculated as the gradient of the linear region on the stress-strain plots of the films (Eq. 1), where τ is stress, the applied force per unit of cross-sectional area normal to the extension force, and ε is strain, the ratio to its original thickness of the sample.

$$E_c = \frac{\tau}{\varepsilon} \quad (1)$$

For tensile tests, the films were clamped and stretched to the point of fracture. The tensile modulus, E_t, was calculated in the same manner as E_c. The breaking stress, or the ultimate tensile stress, was also measured and plotted for comparison.

3.5.2 Comparison of MNs for different materials

MN patches fabricated from PEGDA/AM 70%, FT-PVA, and PVA/CA were observed under the microscope for comparison of dimensions, with 3 measurements per material per dimension. They were then tested in compression mode, in the same manner as the films. The patches were placed with needles facing upward for the tests. MN patches differ from films where their stress-strain curves are influenced by the shape and behaviour of needle structures under stress. To avoid dependencies on needle structure across different materials, the stress-strain curve was divided into two parts: prefracture and post-fracture. Assuming elastic behaviour and constant contact area, the postfracture region was identified as the first linear region encountered and was used to calculate the compression moduli. The compressive strengths of the patches were also measured, which is defined as the maximum stress withstood by the needles before the point of failure or fracture. The cross-sectional area is taken as the area at the

tip, A_{tip}. The strain at point of fracture, ε_{frac}, can be calculated using Eq. 2, where $\varepsilon_{0,linear}$ is the strain at the start of the height of the linear region, h is the needle of the sample and l_0 is the original thickness of the MN base.

$$\varepsilon_{frac} = \varepsilon_{0,linear} - \frac{h}{l_0} \quad (2)$$

The safety margin was then calculated using Eq. 3, which quantifies the ability of the MN to pierce through skin [43].

$$Safety\ margin = \frac{compressive\ strength}{insertion\ stress\ required} \quad (3)$$

3.5 MN Insertion tests

The PEGDA/AM 70% MN patches were inserted into porcine skin. Methylene blue dye was prepared by first dissolving the methylene blue powder (1.5% w/v) in 95% ethanol. The saturated solution was then diluted with DI water in a ratio of 3:10. For top-view insertion, the MN patches were inserted. Several drops of methylene dye were added immediately after the patch removal. After 5 seconds, excess dye was wiped from the skin surface. For the cross-sectional view, a piece of porcine skin was immersed in the dye for 5 seconds, then washed with DI water to remove excess dye. The MN insertion was carried out near the edge of the skin tissue. The tissue was then sliced to obtain a cross-sectional sample. Samples were observed under an optical microscope. Digital images were captured and compared. The depth of the insertion was measured in 5 samples.

3.6 Quantification of glucose concentration through FRET signalling

The fabrication of the biosensor layer has been optimised by Dr. Hu and demonstrated high sensitivity and reversibility to glucose in PBS buffer solutions.

3.6.1 Preparation of fluorescent biosensor

AM (19.0 % w/v), 3-APBA (12.73 % w/v), Fluorescent dye (0.27 % w/v), RhB (0.22 % w/v) and MBA crosslinker (0.13 % w/v) were dissolved in DMSO for the preparation of the fluorescent biosensor. HMPP photoinitiator (1% v/v) was added to the solution. 80µL of the solution was added to the PDMS mould and the MN base fabricated from Section 3.3 was carefully placed on top, covering the solution. The set-up was UV-cross-linked for 45 mins on one side, flipped, and cross-linked for another 30 mins. The MN patch was dried and immersed in DI water for storage.

3.6.2 Preparation of artificial interstitial fluid

Artificial ISF was prepared by adding 2.5 mM $CaCl_2$, 10 mM HEPES, 3.5 mM KCl, 0.7 mM $MgCl_2$, 1.5 mM NaH_2PO_4 and 7.4 mM sucrose to DI water. To compare the intensities of different glucose concentrations, D-(+)-glucose was added to artificial ISF in concentrations of 4 mM, 8 mM, 12 mM, 16 mM, 20 mM and 24 mM glucose.

3.6.3 Fluorescence intensity measurements

The intensities of MN with biosensor were recorded with emission spectra 500 - 700 nm by excitation at 470 nm in 1 ml of artificial ISF solution. The MN device was incubated in solution for 25 min and measurements were repeated three times to ensure the equilibrium of the glucose complexation.

4. Results & Discussion

4.1 Fabrication of PEGDA/AM MN patches

After photopolymerization of PEGDA/AM, the MN patches exhibited varying degrees of curling behaviour. This phenomenon could be explained by the surface tension of excess non-cross-linked solution on the glass slide being transferred to the polymer chains through an attractive interaction during the drying process, pulling them together and causing the patch to curl up [44]. This process occurred unevenly, often with local release of the patch from the slide, which initiated its breakage. Immediate immersion of the patch in water, however, did not resolve the issue because rapid swelling of the hydrogel when attached to the slide led to deformation and cracking. To address these fabrication difficulties, the PDMS mould was kept on the cross-linked patch for at least 30 min, providing a mechanical support for the patch to release itself evenly from the glass slide. To avoid deformation, caused by breakage or cracks during curling, the total patch area was manipulated to a minimal size of 70 mm^2.

Two key trends were observed in the dimensions of MNs across the different formulations. First, there were no large differences in the diameter of the MN base (Figure 2a) and the spacing (Figure 2b) with a changing PEGDA:AM ratio. A source of data variability was the curved surface of the curled patches, as the scale would only be accurate at the plane of focal length of the magnification lens [45]. Attempts to flatten the patch introduced cracks (lines in Figure 2d), especially at a lower PEGDA:AM ratio, and risked damaging the MN tip. A second observation was a parabolic trend in the measured needle height (Figure 2c), which corresponded to the trends of mechanical properties explained in Section 4.2. The overall difference in dimension was correlated with a trend in needle volume, where the swelling ratio plays a role. The swelling ratio has an inverse relationship with the cross-linking density [46]. Furthermore, it was confirmed that the fabrication method allowed reproducible MN patches with small standard deviation (Figure 2a, 2b, and 2c). This is an important factor in achieving reproducible skin insertion depth and surface area contact with the ISF, leading to reliable glucose concentration measurements across the entire surface of the patch.

A final observation was that the tip height was far from the needle depth at 600 µm of the negative mould. This suggests trapped air that prevented the hydrogel from filling the mould, which also greatly affected the fabrication of PVA patches and is further discussed in Section 4.3.

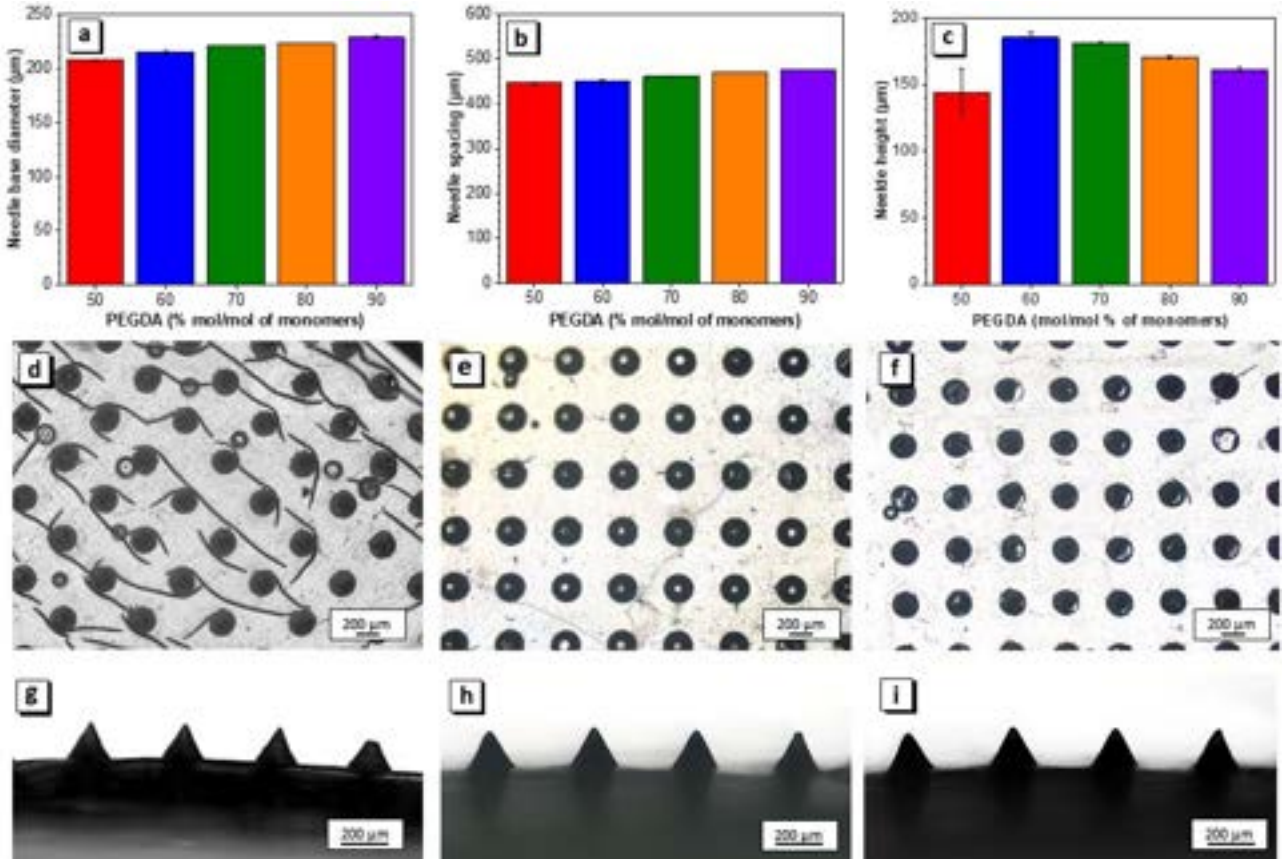


Fig. 2 Dimensions of MN base for 50%, 60%, 70%, 80% and 90% formulations of PEGDA/AM. (a) Comparison of MN base diameter for different PEGDA:AM ratios. (b) Comparison of tip-to-tip space for different PEGDA:AM ratios. (c) Comparison of needle height from base to tip for different PEGDA:AM ratios. (d,e,f) Top view of MN patch for (d) PEGDA/AM 50%, (e) PEGDA/AM 70% and (f) PEGDA/AM 90% at 2.5x magnification. (g,h,i) Side view of MN patch for (g) PEGDA/AM 50%, (h) PEGDA/AM 70% and (i) PEGDA/AM 90% at 5x magnification.

4.2 Mechanical testing of PEGDA/AM films

The PEGDA/AM films were measured to have an average thickness of 190 ± 5 µm and a surface area ranging from 212 mm^2 to 504 mm^2. During both the compression test (Figure 3a) and the tensile test (Figure 3b), the films demonstrated elastomer behaviour throughout the load range, where the modulus increased proportionally to strain. Thus, the moduli for both tests were obtained from the gradient of the approximated linear region of the stress-strain curve. The compressive modulus revealed a positive parabolic relationship with the PEGDA:AM ratio, whereby PEGDA/AM 70% with 0.174 ± 0.013 MPa showed the lowest compressive modulus of all ratios (Figure 3c). A similar trend was observed in Figure 3d with a tensile modulus of 15.6 ± 1.5 MPa for PEGDA/AM 70%.

The key factor in determining the mechanical properties of a hydrogel, derived from its response to stress, is the cross-linking density. The correlation between modulus (E) and cross-linking density can be captured from a thermodynamic standpoint in Eq. 4 where R is the universal gas constant, T is temperature, n refers to the cross-link density and $\overline{r_i^2}/\overline{r_0^2}$ is the ratio of the end-to-end distance of a polymer chain in a cross-linked and non-cross-linked state, which is assumed to be 1 [47]. For a given material at constant temperature, the modulus is directly proportional to the cross-linking density:

$$E = 3n\frac{\overline{r_i^2}}{\overline{r_0^2}}RT \quad (4)$$

The cross-linking density is greatly affected by two parameters: the crosslinker concentration and the mesh size. PEGDA acts as a crosslinker in this hydrogel network. Therefore, an increase in PEGDA concentration leads to more cross-links and improved chain entanglement, resulting in a higher cross-linking density within the hydrogel structure and increased stiffness [48]. However, a smaller mesh size, affected by the molecular size of the monomers, also increases the cross-linking density. The monomer AM (M_w 71.08) is significantly smaller in size than PEGDA (M_w 700). The increase in molar ratio of a low M_w monomer produces a hydrogel mesh with a denser cross-linking network, which simultaneously increases the stiffness of the hydrogel. According to Flory [49], long chains of polymers, such as PEGDA, contribute to the mobility of the internal polymer chain, which allows a greater degree of freedom and flexibility within the hydrogel network.

The U shape across the curves in Figure 3 accurately describes these two effects, while agreeing with the trend in dimension described in Section 4.1. Consequently, PEGDA/AM 70% was selected for subsequent investigations in this work, as it was the most flexible according to moduli and capable of conforming to the curvature of human skin.

The curve for tensile breaking stress (Figure 3e) closely followed this positive parabolic trend, with a minimum

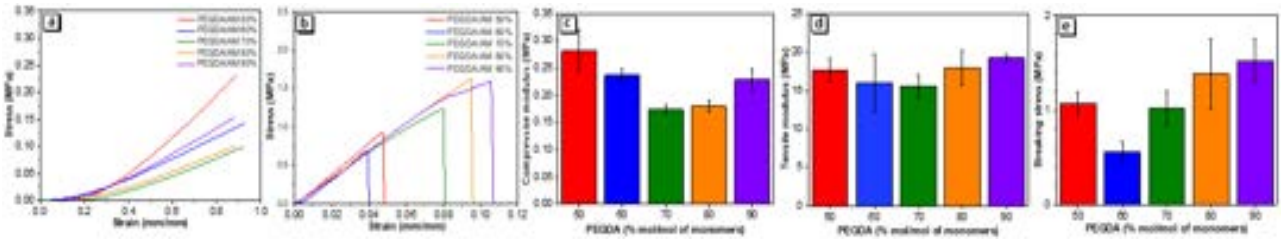


Fig. 3 Mechanical testing of PEGDA/AM films. (a,b) Stress-strain curves during compression and tensile tests, respectively, showing elastic behaviour across the test region. (c,d) The effect of the PEGDA: AM ratio on both compression and tension moduli, and (d) Variation of breaking stress measured during tensile tests to increase PEGDA concentration.

breaking stress recorded for PEGDA/AM 60% at 0.558 ± 0.110 MPa. Although on the lower end, PEGDA/AM 70% could withstand a tensile stress twice as high (1.02 ± 0.18 MPa) while stretched before breakage.

The standard errors of this investigation were quite significant as seen in Figure 3c, 3d and 3e. This variability may be due to the approximation of the modulus as a constant, which has been shown to vary with strain for elastomers [50]. Although the strain rates used are low, this was still insufficient to exclude the time-dependent behaviour of the hydrated hydrogels, which can be examined using rheology methods instead [51]. Methods such as thermal gravimetric analysis (TGA) can also further examine the internal cross-linking structure to arrive at a deeper understanding [52].

4.3 Comparison of PEGDA/AM MNs and PVA MNs

As PVA is a hydrogel widely reported for MN devices in literature, the performance of the PEGDA/AM patch (Figure 4f) was compared to that of the PVA patches manufactured by physical (FT-PVA) and chemical cross-linking (PVA / CA). From microscopic analysis, it was concluded that the PVA/CA patch failed to form needles; instead, it showed bubble-like structures (Figure 4d). FT-PVA was successful in needle formation; however, both its base (Figure 4c) and the needle itself (Figure 4e) showed defects and an inconsistent shape. This was also reflected in the large standard deviation of the dimensions relative to PEGDA/AM (Figure 4a). The PVA MNs showed a large disagreement between individual needles for several reasons. First, bubbles were easily trapped in the highly viscous precursor solution, resulting in MNs with a porous and irregular structure. The absence of MN structures of PVA/CA could be the result of moisture still present in the solution, which vaporises during the curing process at high temperature. Procedures such as centrifugation and extended vacuum drying may mitigate this issue, but this consumes more time and resources. On the contrary, the preparation of PEGDA/AM patches is straightforward and time-effective, taking only 30 minutes per MN patch. Apart from a UV crosslinker, all the equipment needed to fabricate the PEGDA/AM patches could be found in basic chemical laboratories, further suggesting the suitability of PEGDA/AM over other candidate MN base materials.

The compressive modulus and strength were calculated from mechanical stress testing. Both the FT-PVA and PVA/CA patches reported a significantly higher compressive modulus than PEGDA/AM 70% (Figure 4b). This was correlated with an increase in the stiffness of PVA, which is a less desirable property for wearable sensors. Furthermore, the compressive strength, defined as the uniaxial stress at the start of needle deformation, was recorded at 2.55 ± 1.78 MPa for FT-PVA and 35.0 ± 0.7 MPa for PEGDA/AM 70%. This suggested that PEGDA/AM was better able to penetrate the stratum corneum, which requires a stress of at least 3.18 MPa to overcome skin elasticity [53]. The MN had a safety margin of 11.0 (Eq. 3), indicating excellent insertion capabilities for human skin. This property, in addition to its low compressive modulus, confirmed the effectiveness of the material as an MN base.

These results also showed the influence of the cross-linking method on the mechanical strength. Although physically cross-linked FT-PVA was formulated with 30% PVA, its compressive modulus was similar in value to that of PVA/CA, which only had 15% PVA. As suggested in Section 2, these results confirmed the superior mechanical strength of chemically cross-linked hydrogels. In addition, the high uncertainty in the mechanical properties of FT-PVA was consistent with the variability in dimensions and porosity of the MNs, further supporting our observations.

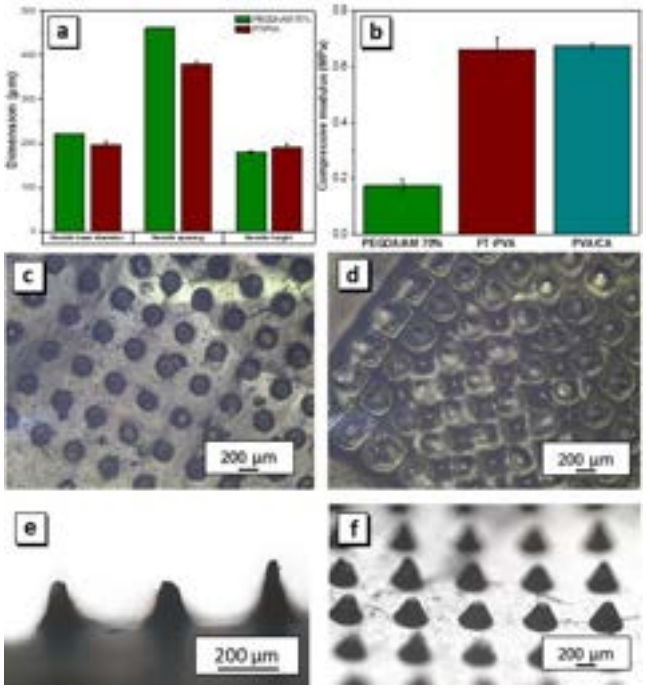


Fig. 4 Comparison of PEGDA/AM patches with freeze-thawed (FT) PVA and PVA/CA formulation. (a) Comparison of MN base diameter, tip-to-tip needle spacing and MN height for PEGDA/AM 70% and freeze-thawed PVA patch. (b) Top view of PVA/CA MN patch at 2.5x magnification, showing needleless, bubble-like structures. (c) Top view of FT PVA MN array at 2.5x magnification with bubbles trapped. (d) Side view of FT PVA MN array at 5x magnification demonstrates irregularity in the needles. (e) Angled view of PEGDA/AM 70% patch at 2.5x magnification.

4.4 Skin insertion

From the analysis in Sections 4.1 and 4.2, PEGDA/AM 70% showed the most promising properties for successful skin insertion. The porcine skin was successfully penetrated by this patch (Figure 5), with a measured insertion depth of 159 ± 1 µm. However, the challenge of keeping the MN patch flat remained, as the curled surface led to less efficient penetration near the edge of the patch.

Porcine skin has been widely used and well researched as a skin model due to its histological similarity to human skin [43], sharing a similar thickness of the epidermal layer and stratum corneum, and a similar elastin ratio. This validated that the compressive strength of PEGDA/AM 70% is sufficient to penetrate human skin and interact with ISF for accurate monitoring of glucose levels.

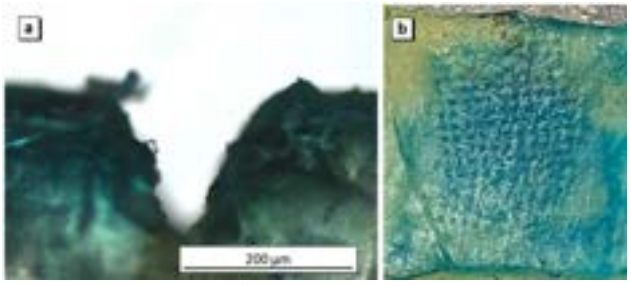


Fig. 5 Insertion of PEGDA/AM 70% microneedle in porcine skin stained with methylene blue. (a) Cross-sectional view of porcine skin after insertion of MN at 20x magnification. (b) Top view of porcine skin after insertion.

4.5 Feasibility of the FRET fluorescence process with MN base of detecting glucose

To further explore its performance as MN sensors, replicas of PEGDA/AM 70% patches adhered to FRET biosensor through hydrogen bonds (Figure 7). Förster resonance energy transfer is a mechanism where the excited donor group transfers energy to the acceptor group of a different energy level through dipole-dipole coupling interactions, allowing the emission of photons of a different wavelength [54]. In the presence of glucose, the diol functions of glucose bind to the boronic acid groups of 3-APBA in the glucose-specific moiety. Upon this binding, an intermolecular dissociation of the fluorescein / rhodamine B pair occurs, inhibiting the FRET process (Figure 6). This mechanism is at the core of the glucose-responsive capability of the patch. Fluorescence spectra were acquired with an excitation wavelength of 470 nm, corresponding to the absorbance maxima of the FRET donor. Two emission peaks were observed, one at 520 nm corresponding to the green emissive donor fluorophore fluorescein, and the second at 578 nm, attributed to the emission of the orange emissive fluorophore rhodamine B (Figure 7a). The presence of the peak at 578 nm is considered evidence of energy transfer from excited fluorescein molecules to the rhodamine B subunit in the biosensor layer and is consistent with the orange colour of the MN device observed by the naked eye upon excitation by blue light (Figure 7c, 7g). For PEGDA with molecular weight 700, optical transparency is achieved above a concentration of 40% [55].

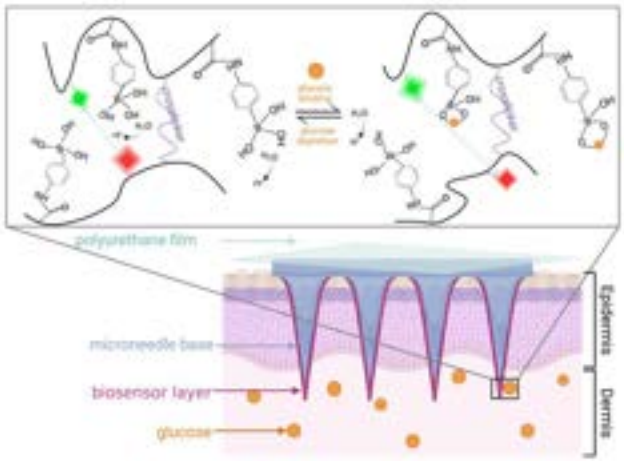


Fig. 6 Illustration of the glucose monitoring microneedle device inserted into human skin. The MN consists of a MN base, providing the rigid structure and biosensing layer. A transparent polyurethane film is attached on top to allow adhesion to the skin. The figure in the top illustrates the FRET mechanism for the fluorescence response to change in ISF glucose concentration, where the biosensor interacts with glucose to inhibit energy transfer between donor and acceptor.

Furthermore, the fluorescent quenching ability of glucose, illustrated by an increase in donor emission accompanied by a reduction in acceptor emission as the concentration increases, was confirmed (Figure 7a). On this basis, glucose levels can be tracked by the MN device through the ratiometric fluorescence method. The emission intensity ratio ($^{I_{578}}/_{I_{520}}$) displayed a quasi-linear dose-response towards glucose (from 4 to 24 mM) (Figure 7b). The equation of linear regression was $^{I_{578}}/_{I_{520}} = -14.0C + 0.545$ ($R^2 = 0.984$) where C represents the glucose concentration in molar units. This linear response supports the consistent measurement of glucose levels. Therefore, it can be concluded that integration of the PEGDA/AM 70% MN base did not hinder the signalling efficiency of the biosensor layer.

5. Conclusion

This work presented an end-to-end investigation on the fabrication, optimisation, and characterisation of the PEGDA/AM hydrogel MN base for non-invasive, CGM. After limitations in fabrication, such as curling and breakage, were mitigated, it can be concluded that this report proposes a strong and flexible hydrogel material for next step fabrication of MN sensors. Further improvements in its shape, especially a longer needle tip using vacuum or centrifugation methods on the precursor solution, can be investigated. The optimised PEGDA/AM 70% was characterized to have a modulus of 0.174 ± 0.023 MPa and a compressive strength of 35.0 ± 0.7MPa, demonstrating improved mechanical properties over PVA, which is a widely studied MN material. More detailed mechanical analysis using frequency-based dynamic testing could be used to study its time-dependent viscoelasticity behaviour. For the first time, the performance of a PEGDA/AM MN base was validated by successful penetration into porcine skin, achieving an insertion depth of 159 ± 1 µm.

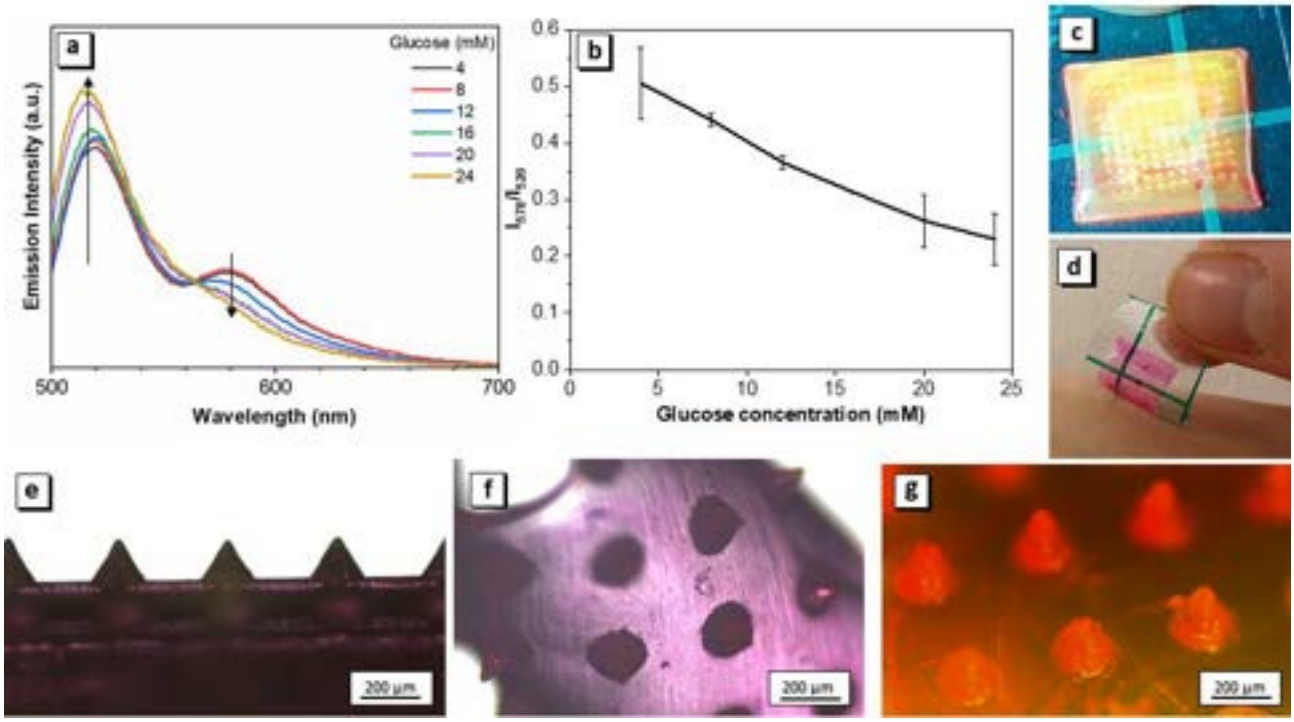


Fig. 7 Response of the photonic MN biosensor to glucose. (a) Glucose concentration dependence of the normalized fluorescence of the biosensor-integrated MN immersed in artificial ISF across wavelengths 500-700 nm. (b) The peak intensity, I_{578}/I_{520}, shows a linearly decreasing trend versus glucose concentration from 4 to 24 mM. (c) Fluorescence of the MN device under blue light. (d) Application of MN device with polyurethane film on human skin. (e) Side view of MN device with biosensor layer at 5x magnification. (f) Angled view of the MN device with biosensor layer at 5x magnification. (g) Fluorescent microscope image of the MN device excited by blue light at 5x magnification.

When a biosensing layer was integrated, a CGM device was produced and characterised that could quantitatively monitor glucose concentrations. Fluorescence spectroscopy showed that it was successful in providing a linear response ($R^2 = 0.984$) to the nominal glucose level in artificial ISF reversibly. The reliability of the quantification can be further improved by using new techniques that allow for a consistent thickness in the biosensor layer. Before adaptation in real life, the stability and degradation of the biosensor in long-term use and at various pHs should be tested. Finally, a user-friendly mobile application system capable of quantifying the fluorescence response can be developed to accurately monitor ISF glucose concentrations, to achieve the ultimate objective of providing an all-round solution for patients diagnosed with diabetes.

Acknowledgements

Tai Tan and Marieke De Bock express their sincere gratitude to Dr. Hu Yubing for her constant guidance throughout the entire duration of project and to everyone in the Yetisen group for their encouragements and support.

Reference

[1] Ampofo, A.G., Boateng, E.B. (2020) Beyond 2020: Modelling obesity and diabetes prevalence. *Diabetes Research and Clinical Practice*. **167**.
[2] Sun, H., et al. (2022) IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Research and Clinical Practice*. **183**, 109119.
[3] Chien, M.N., et al. (2020) Continuous Glucose Monitoring System Based on Percutaneous Microneedle Array. *Micromachines*. **13**(3).
[4] Clark, L.C.J., Lyons, C. (1962) Electrode systems for continuous monitoring in cardiovascular surgery. *Ann. N. Y. Acad. Sci.*, 1962. **102**, 29-45.
[5] Zeng, Y., et al. (2020) Colloidal crystal microneedle patch for glucose monitoring. *Nano Today*. **35,** 100984.
[6] Klonoff, D.C. (2005) Continuous Glucose Monitoring. *Diabetes care*. **28**(5), 1231-1239.
[7] Kim, J., et al. (2019) Wearable biosensors for healthcare monitoring. *Nature Biotechnology*. **37**(4), 389-406.
[8] Tehrani, F., et al. (2022) An integrated wearable microneedle array for the continuous monitoring of multiple biomarkers in interstitial fluid. *Nature Biomedical Engineering*.
[9] Bruen, D., et al. (2017) Glucose Sensing for Diabetes Monitoring: Recent Developments. *Sensors*. **17**(8), 1866.
[10] Iqbal, S.M.A., et al. (2021) Advances in healthcare wearable devices. *npj Flexible Electronics*. **5**(1).
[11] Shrestha, D., et al. (2015) Understanding FRET as a Research Tool for Cellular Studies. *International Journal of Molecular Sciences*, **16**(12), 6718-6756.
[12] Kulkarni, D., et al. (2022) Recent Advancements in Microneedle Technology for Multifaceted Biomedical Applications. *Pharmaceutic*. **14**(5).
[13] Rodbard, D. (2016) Continuous Glucose Monitoring: A Review of Successes, Challenges, and Opportunities. *Diabetes Technology & Therapeutics*. **18**(S2), 3-13.
[14] Wichterle, O., Lím, D. (1960) *Hydrophilic Gels for Biological Use.* Nature,. **185**(4706), 117-118.
[15] Choi, J.R., et al. (2019) Recent advances in photo-crosslinkable hydrogels for biomedical applications. *BioTechniques*. **66**(1), 40-53.
[16] Ahmed, E.M. (2015) Hydrogel: Preparation, characterization, and applications: A review. *Journal of Advanced Research*. **6**(2), 105-121.

[17] Li, X., et al. (2018) Functional Hydrogels With Tunable Structures and Properties for Tissue Engineering Applications. *Frontiers in Chemistry*. **6**.
[18] Sennakesavan, G., et al. (2020) Acrylic acid/acrylamide based hydrogels and its properties - A review. *Polymer Degradation and Stability*. **180**, 109308.
[19] Zhang, B., et al. (2018) Highly stretchable hydrogels for UV curing based high-resolution multimaterial 3D printing. *Journal of Materials Chemistry B*,. **6**(20), 3246-3253.
[20] Heo, Y.J., et al. (2011) Long-term in vivo glucose monitoring using fluorescent hydrogel fibers. *Proceedings of the National Academy of Sciences*. **108**(33), 13399-13403.
[21] Liu, S., et al. (2017) Peptide delivery with poly(ethylene glycol) diacrylate microneedles through swelling effect. *Bioengineering; Translational Medicine*. **2**(3), 258-267.
[22] Gao, Y., et al. (2021) Intradermal administration of green synthesized nanosilver (NS) through film-coated PEGDA microneedles for potential antibacterial applications. *Biomaterials Science*. **9**, 2244-2254.
[23] Yang, S., et al. (2015) Phase-Transition Microneedle Patches for Efficient and Accurate Transdermal Delivery of Insulin. *Advanced Functional Materials*. **25**(29), 4633-4641.
[24] DeMerlis, C.C., Schoneker, D.R. (2003), Review of the oral toxicity of polyvinyl alcohol (PVA). *Food and Chemical Toxicology*. **41**(3), 319-326.
[25] Kodavaty, J. (2022) Poly (vinyl alcohol) and hyaluronic acid hydrogels as potential biomaterial systems - A comprehensive review. *Journal of Drug Delivery Science and Technology*. **71**.
[26] He, R., et al. (2020) A Hydrogel Microneedle Patch for Point-of-Care Testing Based on Skin Interstitial Fluid. *Advanced Healthcare Materials*. **9**(4), 1901201.
[27] Tekko, I.A., et al. (2020) Development and characterisation of novel poly (vinyl alcohol)/poly (vinyl pyrrolidone)-based hydrogel-forming microneedle arrays for enhanced and sustained transdermal delivery of methotrexate. *International Journal of Pharmaceutics*. **586**, 119580.
[28] Wang, Z., et al. (2020) Transdermal colorimetric patch for hyperglycemia sensing in diabetic mice. *Biomaterials*. **237**.
[29] Chandran, R., et al. (2022) Factors influencing the swelling behaviour of polymethyl vinyl ether-co-maleic acid hydrogels crosslinked by polyethylene glycol. *Journal of Drug Delivery Science and Technology*. (68).
[30] Donnelly, R.F., et al. (2012) Hydrogel-Forming Microneedle Arrays for Enhanced Transdermal Drug Delivery. *Advanced Functional Materials*. **22**(23),4879-4890.
[31] Zhu, M., et al. (2019) Gelatin methacryloyl and its hydrogels with an exceptional degree of controllability and batch-to-batch consistency. *Scientific Reports*. **9**(1).
[32] Sharifi, S., et al. (2021) Systematic optimization of visible light-induced crosslinking conditions of gelatin methacryloyl (GelMA). *Scientific Reports*. **11**(1).
[33] Zhu, J., et al. (2020) Gelatin Methacryloyl Microneedle Patches for Minimally Invasive Extraction of Skin Interstitial Fluid. *Small*. **16**(16), 1905910.
[34] Spearman, B.S., et al. (2020) Tunable methacrylated hyaluronic acid-based hydrogels as scaffolds for soft tissue engineering applications. *Journal of Biomedical Materials Research Part A*. **108**(2), 279-291.
[35] Bencherif, S.A., et al. (2008) Influence of the degree of methacrylation on hyaluronic acid hydrogels properties. *Biomaterials*. **29**(12), 1739-1749.
[36] Burdick, J.A., et al. (2005) Controlled Degradation and Mechanical Behavior of Photopolymerized Hyaluronic Acid Networks. *Biomacromolecules*. **6**(1), 386-391.
[37] Sheik, S., Nagaraja, G.K., Prashantha, K. (2018) Effect of silk fiber on the structural, thermal, and mechanical properties of PVA/PVP composite films. *Polymer Engineering & Science*. **58**(11), 1923-1930.
[38] Cha, W.I., et al. (1996) Mechanical and wear properties of poly(vinyl alcohol) hydrogels. *Macromolecular Symposia*. **109**(1), 115-126.
[39] Tavsanli, B., Okay, O. (2017) Mechanically strong hyaluronic acid hydrogels with an interpenetrating network structure. *European Polymer Journal*. **94**, 185-195.
[40] Singh, T.R.R., et al. (2009) Physicochemical characterization of poly(ethylene glycol) plasticized poly(methyl vinyl ether-co-maleic acid) films. *Journal of Applied Polymer Science*. **112**(5), 2792-2799.
[41] Gan, D., et al. (2019) Mussel-inspired dopamine oligomer intercalated tough and resilient gelatin methacryloyl (GelMA) hydrogels for cartilage regeneration. *Journal of Materials Chemistry B*. **7**(10), 1716-1725.
[42] Formlabs, I. (2017) *Material Data Sheet Standard*.
[43] Makvandi, P., et al. (2021) Engineering Microneedle Patches for Improved Penetration: Analysis, Skin Models and Factors Affecting Needle Insertion. *Nano-Micro Letters*. **13**(1).
[44] Chavda, H., et al. (2012) Preparation and characterization of superporous hydrogel based on different polymers. I*nternational Journal of Pharmaceutical Investigation.* **2**(3).
[45] Bi, G., et al. (2018) The measuring method for actual total magnification of metallographic microscope — digital image method. IOP Conference Series: Materials Science and Engineering. 397, 012148.
[46] Mohammed, A.H., et al. (2018) Effect of crosslinking concentration on properties of 3-(trimethoxysilyl) propyl methacrylate/N-vinyl pyrrolidone gels. *Chemistry Central Journal*. **12**(1).
[47] Sperling, L.H. (1986) *Introduction to physical polymer science*.
[48] Chavda, H., Patel, C. (2011) Effect of crosslinker concentration on characteristics of superporous hydrogel. *International Journal of Pharmaceutical Investigation.* **1**(1).
[49] Flory, P.J. (2003) *Principles of polymer chemistry*. 7th ed. Ithaca, NY, Cornell University Press.
[50] Arnold, F.J., Maran, F.S. (2019) Young's modulus determination ofelastomeric materials using capacitancemeasurement. *European Journal of Physics,* **40**(5).
[51] Lee, D., Zhang, H., Ryu, S. (2019) Elastic Modulus Measurement of Hydrogels, in Polymers and Polymeric Composites: A Reference Series. *Springer International Publishing.* **4***,* 865-884.
[52] Xia, Z., et al. (2013) Determination of crosslinking density of hydrogels prepared from microcrystalline cellulose. *Journal of Applied Polymer Science*. **127**(6), 4537-4541.
[53] Abser, M.N., Gaffar, M., Islam, M.S. (2010) Mechanical feasibility analysis of process optimized silicon microneedle for biomedical applications. *International Conference on Electrical & Computer Engineering (ICECE 2010)*. 222-225.
[54] Jones, G.A., Bradshaw, D.S. (2019) Resonance Energy Transfer: From Fundamental Theory to Recent Applications. *Frontiers in Physics.* **7**
[55] Torres-Mapa, M.L., et al. (2019) Fabrication of a Monolithic Lab-on-a-Chip Platform with Integrated Hydrogel Waveguides for Chemical Sensing. *Sensors*. **19**(19), 4333.

Modelling A Perfusion Bioreactor for IgG Antibody Production Using CHO Cell Lining

Rayan, Mohammad and Erinjogunola, Abdullah
Department of Chemical Engineering, Imperial College London, U.K.

Abstract

Immunoglobulin G (IgG), a biopharmaceutical, is made by cultivating recombinant animal cells, particularly Chinese hamster ovary (CHO) cells. Historically, large fed batch cultures are employed however a perfusion culture offers unmatched performance and savings over the former. Perfusion is a complex process as fresh media is continuously fed and spent media is removed while keeping cells in culture. Therefore, modelling the process in a simulation software such as gPROMS can offer a greater understanding of the process and underlying principles as well as offer estimates for various important process parameters. In this research, a perfusion bioreactor was modelled using gPROMS and was used to carry out parameter estimation for process parameters. The performance was compared with that of a batch reactor and the estimated parameters were used to rerun the simulation to compare with the experimental results. The results for the fed batch model are in good agreement and follow an expected trend. However, the perfusion model struggles to provide reliable results which may be due to the model not being sufficient for perfusion or that initial parameter guesses, and their bounds are not able to provide reliable parameters which provide a good fit.

Background

When the body feels under attack, it makes special proteins called antibodies. These antibodies are made by plasma cells and are released throughout the body to kill bacteria, viruses, and other germs. Immunoglobulin G (IgG) is a very common type of antibody which is made by plasma B cells and makes about 75% of the plasma serum. IgG deficiency is a health condition in which the body does not produce enough immunoglobulin G (IgG). People with IgG deficiency are more likely to get infections. IgG has four different subclasses, IgG1-4. IgG is always present to prevent infection. They are also ready to multiply and attack when a foreign object enters the body.[1]

Biopharmaceuticals such as immunoglobulin G (IgG) are manufactured by culturing recombinant animal cells, especially Chinese hamster ovary (CHO) cells. Due to their similarity to human cell lines, CHO cells are used in scientific and medical research, especially in genomic and chromosomal, toxicity, nutrition, and gene expression studies. Production of recombinant proteins in bioreactors is one of his main goals for CHO cells, as CHO cells represent more than 70% of the entire biopharmaceutical industry. They can produce on the order of 3-10 grams of recombinant protein per litre of CHO cell culture.[2]

A bottleneck in biopharmaceutical manufacturing is believed to be the intracellular IgG secretion mechanism. Improved productivity has been demonstrated in numerous studies of regulation of expression levels of endogenous secretory proteins. However, not all proteins performed better as a result of these efforts. Based on the understanding of the secretion mechanism in IgG-producing CHO cells, more rational and effective design of high-producing cells is required. [3]

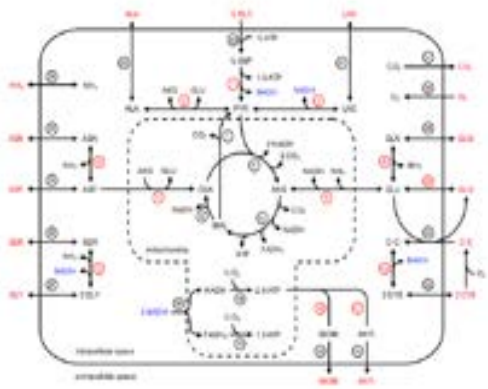


Figure 1 Metabolism of CHO cell [4]

Despite their practical and commercial relevance, there are few reports on the growth and production kinetics of Chinese Hamster Ovary (CHO) cells. Currently, there are over 200 pharmaceutical companies on the market and the economic value of these medicines continues to grow, with sales increasing by US$30 billion in 2003 and reaching US$100 billion by 2012. Among biopharmaceutical compounds, monoclonal antibodies (mAbs) are an increasingly accepted class of therapeutics, especially in the fields of oncology, immunology, and organ transplantation. Since their introduction in 1986, mAbs have become the dominant products in the biopharmaceutical market. Mammalian cell cultures are used because therapeutic proteins require complex post-translational modifications and mammalian cells (including CHO) are capable of carrying out these manipulations. CHO cells grow to very high densities in suspension culture in bioreactors of up to 10,000 L, making them suitable for large-scale culture. They are relatively stable in heterologous gene expression over time. [5]

Throughout the industry fed-batch cultures are mostly used for cell cultures. Fed-batch culture is an operating technique in biotechnology processes in which substrates are supplied to a bioreactor during cultivation and the product remains in the bioreactor until the end. In general, fed-batch culture is superior to traditional batch culture when controlling the level of a nutrient affects yield. Another understanding for this technique is that "the basal medium supports the initial cell culture and the nutrient medium is added to prevent nutrient starvation".[6] This is also a kind of "semi-batch culture". In some cases, all nutrients are fed into the bioreactor. Fed batch cultures offer the advantage that the concentration of the fed-batch substrate in the culture medium can be controlled at any level.[7]

An alternative to the fed batch culture is the perfusion bioreactor. Perfusion is a continuous culture process in which cells are retained in a bioreactor or returned to the bioreactor. Therefore, the harvested medium is cell-free, resulting in higher cell concentrations and product yields in the reactor. This also avoids the risk of cell washout due to excessive dilution. Perfusion cell culture uses a cell retention device and continuous medium exchange to achieve and maintain high cell densities and viability for extended periods of time, typically weeks. The cell retainer holds the cells in the bioreactor while fresh medium is added and products of interest, waste, and spent (or depleted) medium are continuously removed. Fresh medium is provided at the same rate as product and spent medium are removed from the bioreactor. Hollow fibre-based membrane filters are the most reliable and commonly used membrane type. Long-term perfusion is just one application of perfusion, in contrast to short-term-based perfusion applications such as High Productivity Harvest (HPH), which focus on enhancing the fed batch process over several days. Perfusion can also be used to enhance the seed train, reducing steps and reducing overall time to production.[8]

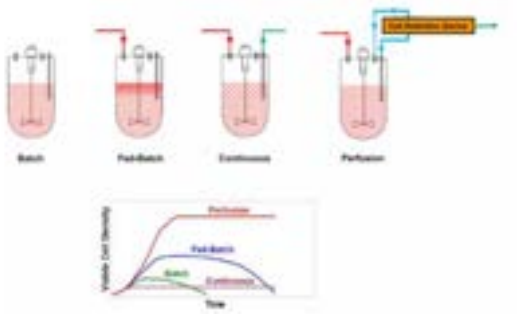


Figure 2 Different reactor concentrations and their performance [9]

Currently, perfusion is not common on large scale processes due to laborious nature of the process and incomplete understanding of the process. Modelling the process provides understanding of the system and how it was to behave in different conditions. Modelling provides estimates of parameters which are relevant to the process such as the yields. The model may be extrapolated to large scale production. In the development of all pharmaceutical manufacturing processes, including those using hMAbs produced by CHO cells, optimal process parameters and methods are determined based on cost, time and potency comparisons. Multiple scalable platforms are often considered before moving the final process to a pilot or scale-up lab. Significant R&D time and money are invested to increase yields, reduce costs, and improve current bioreactor and bioprocess technology.[10]

Methods

To build the model and estimate reliable parameters, a general the mass balance was carried out for the process. The mass balances take into account the cell culture as well as the substrates/ products. The mass balances were obtained from literature.[11] The relevant equations are shown in in Table 1.

$\frac{d[VCD]}{dt} = (\mu_g - \mu_d) \times [VCD]$
$\mu_g = \mu_{max} \times \frac{[GLC]}{K_{glc} + [GLC]} \times \frac{[GLN]}{K_{gln} + [GLN]} \times \frac{K_{Ilac}}{K_{Ilac} + [LAC]} \times \frac{K_{Iamm}}{K_{Iamm} + [AMM]}$
$\mu_d = k_d \times \frac{[LAC]}{K_{Dlac} + [LAC]} \times \frac{[AMM]}{K_{Damm} + [AMM]}$
$\frac{d[GLC]}{dt} = -\left(\frac{\mu_g - \mu_d}{Y_{VCD/glc}} + m_{glc}\right) \times [VCD] + V[(F_{in} \times C_{in,GLC}) - (F_{out} \times C_{GLC})]$
$\frac{d[LAC]}{dt} = Y_{lac/glc} \times \left(\frac{\mu_g - \mu_d}{Y_{VCD/glc}}\right) \times [VCD] + V[(F_{in} \times C_{in,LAC}) - (F_{out} \times C_{LAC})]$
$\frac{d[GLN]}{dt} = -\left(\frac{\mu_g - \mu_d}{Y_{VCD/gln}} + m_{gln}\right) \times [VCD] + V[(F_{in} \times C_{in,GLN}) - (F_{out} \times C_{GLN})]$
$m_{gln} = \frac{a_1 \times [GLN]}{a_2 + [GLN]}$
$\frac{d[AMM]}{dt} = Y_{amm/gln} \times \left(\frac{\mu_g - \mu_d}{Y_{VCD/gln}}\right) \times [VCD] + V[(F_{in} \times C_{in,AMM}) - (F_{out} \times C_{AMM})]$
$\frac{d[mAb]}{dt} = Q_{anti} \times [VCD]$

Table 1 Mass balance equations

After constructing the mass balance, literature [12] was used to obtain initial guesses for the parameter values which will be used to do parameter estimation. Finally, a model was built in gPROMS with the aforementioned balances and parameter estimation was run with a schedule to include the different phases of the process. These parameters were in turn used to estimate the results which were then compared to obtained experimental values. The table below shows values for the parameters obtained.

Parameter	Description	Value	Unit
$VCD_{initial}$	Starting viable cell concentration	0.5000	$X10^6$ cells mL^{-1}
$GLC_{initial}$	Starting glucose concentration	50.95	mM
$GLN_{initial}$	Starting glutamine concentration	9.950	mM
$LAC_{initial}$	Starting lactate concentration	0.0000	mM
$AMM_{initial}$	Starting ammonium concentration	0.0198	mM
μ_{max}	Maximum growth rate	0.0290	h^{-1}
k_d	Maximum death rate	0.016	h^{-1}
$Y_{VCD/glc}$	Yield coefficient cell conc./glucose	$X10^9$ cells $mmol^{-1}$	
$Y_{VCD/gln}$	Yield coefficient cell conc./glutamine	$X10^9$ cells $mmol^{-1}$	
$Y_{lac/glc}$	Yield coefficient lactate/glucose	1.23	mmol $mmol^{-1}$
$Y_{amm/gln}$	Yield coefficient ammonium/glutamine	0.67	mmol $mmol^{-1}$
Q_{anti}	Specific production rate	$X10^{-14}$ mmol $cells^{-1}$ h^{-1}	
m_{glc}	Glucose maintenance coefficient	$X10^{-14}$ mmol $cells^{-1}$ h^{-1}	
a_1	Coefficient for mgln	$X10^{-12}$ mmol $cells^{-1}$ h^{-1}	
a_2	Coefficient for mgln	2.000	mM
K_{glc}	Monod constant glucose	0.1500	mM
K_{gln}	Monod constant glutamine	0.2200	mM
K_{Ilac}	Monod constant lactate for inhibition	45.00	mM
K_{Iamm}	Monod constant ammonium for inhibition	9.500	mM
K_{Dlac}	Monod constant lactate for death	45.8	mM
K_{Damm}	Monod constant ammonium for death	6.51	mM

Table 2 Parameter description and literature values [12][14]

The parameter estimation was performed using the Maximum Likelihood Estimation (MLE) in gPROMS. The goal of maximum likelihood estimation is to determine the parameters for which the observed data have the highest joint probability. gPROMS assumes independent, normally distributed measurement errors, with zero means and standard deviations. The mathematical equation used for MLE and the corresponding symbols shown below:

$$\Phi = \frac{N}{2}\ln(2\pi) + \frac{1}{2}\min_{\theta}\left\{\sum_{i=1}^{NE}\sum_{j=1}^{NV}\sum_{k=1}^{NM_{ij}}\left[\ln\left(\sigma_{ijk}^2\right) + \frac{\left(z_{ijk}-z_{ijk}\right)^2}{\sigma_{ijk}^2}\right]\right\}$$

Objective function symbol definitions

N	Total number of measurements taken during all the experiments
θ	Set of model parameters to be estimated. The acceptable values may be subject to given lower and upper bounds, i.e. $\theta^L \leq \theta \leq \theta^U$
NE	Number of experiments performed
NV_i	Number of variables measured in the i^{th} experiment
NM_{ij}	Number of measurements of the j^{th} variable in the i^{th} experiment
σ_{ijk}^2	Variance of the k^{th} measurement of variable j in experiment i. This is determined by the measured variable's variance model
z_{ijk}	k^{th} measured value of variable j in experiment i
$\hat{z}_{ijk}$	k^{th} model-predicted value of variable j in experiment i

Table 3 Symbols for MLE equation

The experimental data for comparison was provided by MERCK. The experiment was performed in a 6.3 L glass reactor with a working volume of 4.2 L at 36.8 C with 40% dissolved oxygen. The ATF flowrate was 1 L/ min and perfusion was started at day 3. The experimental results are plotted below.

To simulate the experiment the model reactor has a volume of 4.2L. Perfusion began at day 3 with a rate of 2vvd (total medium exchange rate) and the bleed was 0.14vvd. The feed during perfusion was the mean experimental concentrations for the reactants and waste products during steady state.

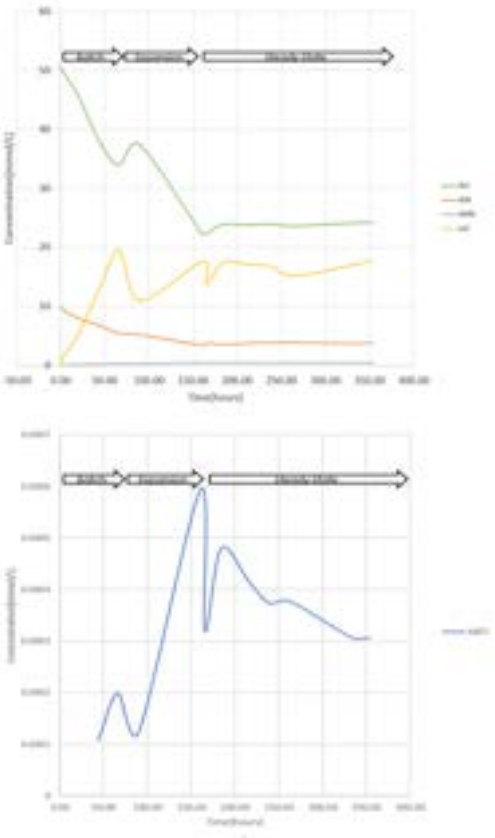


Figure 3 Concentration vs time for various components

The timeline can be divided into three main steps.

- **Batch phase** has a fed batch configuration in which the cell culture kept inside the reactor.
- **Perfusion expansion phase** has perfusion turned on and the system tries to reach the steady state.
- **Steady state phase** has the system in a stable steady state condition with constant concentrations.

From the graphs it can be observed that the steady state concentration of lactate and ammonia are around 17 and 0.3 mM. However, the threshold values for the two reagents to have an inhibition effect is 15 and 5 mM [13]. Therefore, lactate is expected to have an inhibition effect on cell growth as it has a higher concentration than the threshold. In contrast, ammonia is not expected to affect the growth due to the small concentration.

Furthermore, an analysis was performed on glutamine and glucose to check if they are he limiting reagents as suggested by vast literature. This was done by plotting the concentration gradient against the cell growth. The graphs for the two are shown below.

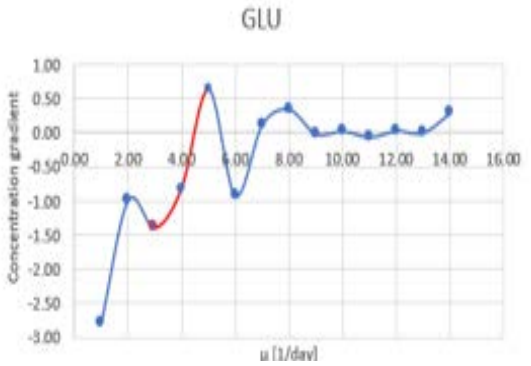


Figure 55 Concentration gradient vs growth graph for glutamine

Figure 44 Concentration gradient vs growth graph for glucose

The graphs show a positive trend of concentration gradient vs the cell growth. This signifies a direct relation which implies that the two are limiting reagents.

Results and Discussion

Whole data set

Using the whole data set provided by MERCK. The parameters for the whole model were estimated using MLE. Table 7 shows the parameter values estimated and their bounds. Parameters a_2, K_{Dlac}, $K_{glc}$and Q_{anti} were the parameters not estimated to be at their bounds. Table 8 shows these parameters had an individual 95% t-value greater than the reference t-value which suggest that the data available was sufficient to estimate the parameters precisely.

Graphs from figure 9 show, the VCD Whole values predicted by the model is initially close to the measured data. After 90 hours there is a spike in VCD measured values, the model failed to predict this and reached steady state at a significantly lower value than the measured values.

The reactants are predicted by the model similarly as shown by graphs GLC Whole and GLN Whole. Initially the prediction for glutamine concentration became more inaccurate as time passed, while for glucose concentration remained close to the measured values. When perfusion at day 3 started the model predicted the steady-state values accurately.

Before perfusion started, graph AMM Whole shows the ammonia predicted values were significantly greater than the measured values and peaked at 3 mmol/L. Once perfusion started the model predicts the measured values. The lactate predicted values showed a similar pattern as seen in graph LAC Whole, however pre-prefusion the model is close to the measured values but peaked when prefusion started, then it decreased to a value slightly greater than the measured value at steady state.

As shown by graph ANTI Whole. The antibody predicted values pre-perfusion were close to the

measured values but fell to a low of 4.81x10-6 mmol/L at steady state.

Batch phase data set

To develop a more accurate model. The dataset was split to reflect the different cell growth phases. This section focuses on the batch phase.

The parameters a_2,, $K_{glc}$and Q_{anti} was not estimated at their bounds. Table 8 shows the individual 95% t-values are greater than the reference t-value which indicates that the data available was sufficient to estimate the parameters precisely.

From figure 9, graphs VCD Batch, GLC Batch, LAC Batch, GLN batch, and ANTI Batch show the model predicted similar values to the measured values in the batch phase for glucose, lactate, glutamine, antibody concentrations and VCD. For ammonia concentrations the model predicted higher values compared to the measured values as seen in graph AMM Batch.

Perfusion phase

The parameters a_2, K_{Dlac}, $K_{glc}$and Q_{anti} was not estimated at their bounds. Table 8 shows the individual 95% t-values were greater than the reference t-value which indicates that the data available was sufficient to estimate the parameters precisely.

The perfusion graphs from figure 9 show the model failed to predict the perfusion expansion phase (89.77- 160.71 hours) for every component. At steady state (184.17-354 hours) the model predicted values were close to the measured values for ammonia, glutamine, glucose and lactate concentrations. However, for VCD and antibody concentrations the predicted values were significantly lower than the measured values.

Perfusion steady state phase.

For the perfusion steady state model, the parameters a_2, K_{Dlac}, $K_{glc}$and Q_{anti} was not estimated at their bounds. Table 8 shows the individual 95% t-values are greater than the reference t-value which indicates that the data available was sufficient to estimate the parameters precisely except for Q_{anti}were the 95% t-value is lower than the reference t-value.

The perfusion steady state graphs from figure 9 show that the model formed using the parameter estimated from the perfusion steady state data set; VCD predicted values were the closest to the measured values of any model that simulated perfusion. However, the antibody concentration predicted values were significantly lower than the measured values. The other component concentrations predicted resembled the perfusion model predicted values.

Parameter data

Parameter	Whole data set	Batch data set	Perfusion data set	Perfusion Steady State data set
a_2	2.056	2.130	2.026	2.101
K_{Dlac}	286.7	311.0*	277.5	309.3
Q_{anti}	1.443 ×10^{-15}	1.862× 10^{-15}	1.000x10^{-15}	1.000x10^{-15}**

All other parameters were constant for every model

Model Parameter	**Final Value**	**Lower Bound**	**Upper Bound**
a_1	1.000×10^{-13}	1.000×10^{-13}*	1.500×10^{-11}
a_2		0.200	4.000
k_d	0.0160	0.0160*	0.0160*
K_{Damm}	10	1.44	10.00*
K_{Dlac}		15.00	311.0
K_{glc}	0.1500	0.1500	1.000
K_{gln}	0.0600	0.0600*	0.8000
K_{Iamm}	20.00	1.000	20.00*
K_{Ilac}	140.0	8.000	140.0*
m_{glc}	0.0000	0.0000*	2.000×10^{-10}
Q_{anti}		1.000×10^{-15}	4.000×10^{-14}
μ_{max}	0.0290	0.0290*	0.0290*
$Y_{amm/gln}$	0.6700	0.6700*	0.6700*
$Y_{VCD/glc}$	1.690×10^{8}	1.690×10^{8}*	1.690×10^{8}*
$Y_{VCD/gln}$	9.740×10^{8}	9.740×10^{8}*	9.740×10^{8}*
$Y_{lac/glc}$	1.230	1.230*	1.230*

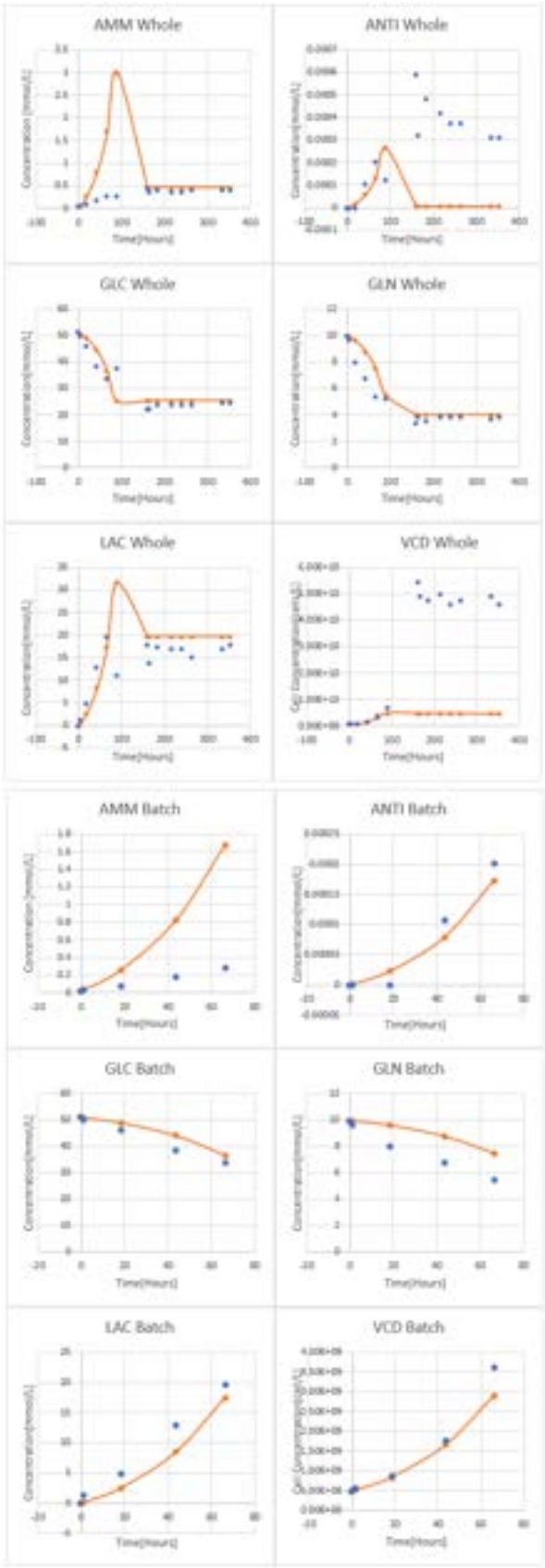
*a parameter that lies on one of its bounds

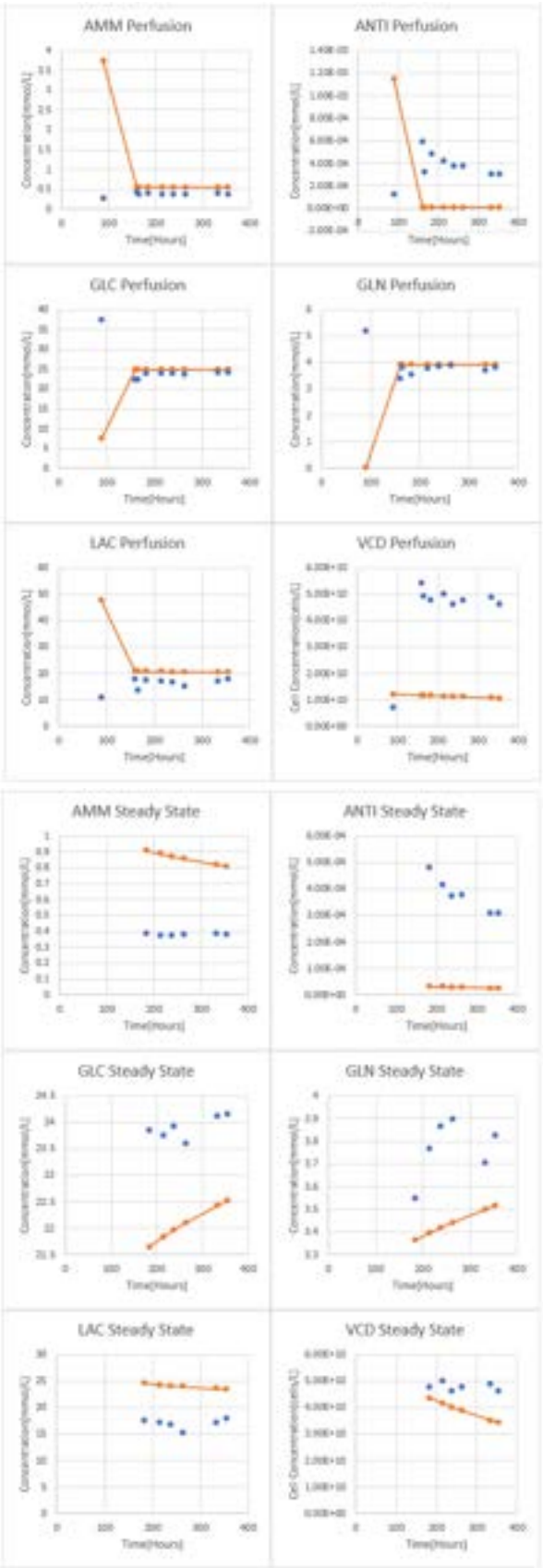
**Data from experiment may not be sufficient to estimate the parameter precisely.

Tables 7 Model parameters, Literature values for bounds [14]

	95% t-value				t-reference
	a_2	K_{Dlac}	K_{glc}	Q_{anti}	
Whole	4.831x 10^4	1.172x 10^8	6.313x 10^8	1.524 x10^1	1.666
Batch	2.053 x10^4		1.647 x10^8	9.199	1.725
Perfusion	1.687 x10^8	1.584 x10^8	4.260 x10^8	5.069	1.680
Steady state	2.207 x10^7	3.622 x10^8	6.843 x10^8	3.402 x10^{-1}	1.706

Table 8 Parameter t-values and t-references





Orange line: Predicted values
Blue dots: Measured values
Figure 9 Set of graphs comparing model predicted values to measured values.

Discussion

The model that used parameter estimation with the whole data set predicted a low antibody concentration once perfusion started. This may have been caused by the low predicted VCD since the antibody production is proportional to the VCD. The predicted antibody values when perfusion began decreased and the rate of antibody production was constant because VCD was constant, hence the model predicts the rate of antibody production was lower than the flowrate of antibodies out of the reactor.

Overall, the model was sufficient at predicting the bioreactor in the batch phase, except for ammonia concentration values. This may be due to a change in the ammonia mass balance equation than ignored the parameter for ammonia removal rate since the bioreactor does not have an ammonia removal system. This suggest that parameters relating directly to ammonia concentration for example yields were not transferable to the model.

Due the model failed to predict perfusion expansion phase, the data set for parameter estimation was focused on perfusion steady state. The steady state graphs from figure_ show the model did not reach true steady state. Compared to the perfusion phase model the difference between the predicted values and measured values for ammonia, glucose and glutamine concentration is greater for the perfusion steady state model. Lactate concentration predicted values are similar for both perfusion models. The perfusion model and perfusion steady-state model results indicate that the model's failure to reach VCD values after the expansion phase is not the sole reason for low antibody concentration predictions.

Comparing the parameters in the models that were different. Across all models the predicted value for glutamine concentration was close to the measured values; this suggests that the slight difference in a_2 values does not affect the model's predictive capability. K_{Dlac} is a key parameter in the cell death rate equation. Considering that every model except the batch model failed to predict VCD, the changes in K_{Dlac} values could be significant during perfusion phase. The poor prediction of antibody concentration by models that simulate perfusion also suggests changes in Q_{anti} values could be significant during perfusion.

The models that included perfusion data values in its estimation all failed to predict VCD and antibody concentration. This suggests that initial parameter guesses, and their bounds are preventing the model from finding parameter values that fit the measured values.

Conclusion.

The model to simulate the perfusion bioreactor did not replicate the data from Merck, using the whole data for parameter estimation by maximum likelihood estimation. The model failed to predict

perfusion expansion phase and, the viable cell density and antibody concentration in the perfusion phase. The investigation into how the different bioreactor phases affect the parameter estimation showed that the model could predict closely the measured bioreactor data in the batch phase. However, in the perfusion phase the model continued to be unsuccessful with viable cell density and antibody concentration predictions.

To conclude, differences in the parameter values across all models had a minimal effect on predicted values. The model's prediction in the batch phase was close to the measured data and the perfusion phase predictions was poor. This suggest that the model is either not sufficient for perfusion since it sufficient for the batch phase or that initial parameter guesses, and their bounds are preventing the model from finding parameters that fit the measured values. To improve the model, we suggest a change in the cell mass balance in the perfusion phase because VCD was poorly predicted repeatedly, in addition using parameters that are more specified for the reactor configuration rather than using yields and maximum growth and death rate for a batch CHO bioreactor.

References

1. IGG deficiencies (no date) IgG Deficiencies - Health Encyclopedia - University of Rochester Medical Center. Available at: https://www.urmc.rochester.edu/encyclopedia/content.aspx?ContentTypeID=134&ContentID=109 (Accessed: December 15, 2022).
2. *Cho cells - 7 facts about the Chinese hamster ovary cell* (2022) *evitria*. Available at: https://www.evitria.com/journal/cho-cells/cho-cells/#:~:text=Due%20to%20their%20similarity%20to,express%20recombinant%20proteins%20in%20bioreactors. (Accessed: December 15, 2022).
3. Kaneyoshi, K. *et al.* (2019) *Analysis of the immunoglobulin G (IGG) secretion efficiency in recombinant Chinese hamster ovary (CHO) cells by using citrine-fusion IGG - cytotechnology*, *SpringerLink*. Springer Netherlands. Available at: https://link.springer.com/article/10.1007/s10616-018-0276-7 (Accessed: December 15, 2022).
4. Mariano Monterio, "Integration of mechanistic and data-driven models to support the transition to continuous biomanufacturing", 2022
5. López-Meza, J. *et al.* (2015) *Using simple models to describe the kinetics of growth, glucose consumption, and monoclonal antibody formation in naive and infliximab producer Cho cells - cytotechnology*, *SpringerLink*. Springer Netherlands. Available at: https://link.springer.com/article/10.1007/s10616-015-9889-2 (Accessed: December 15, 2022).
6. *How single-use, mini bioreactors could revolutionize bioprocess scale-up* (no date) *Pharmaceutical Processing*. Available at: https://web.archive.org/web/20151020005950/http://www.pharmpro.com/articles/2014/11/how-single-use-mini-bioreactors-could-revolutionize-bioprocess-scale (Accessed: December 15, 2022).
7. *Fed-batch culture* (2022) *Wikipedia*. Wikimedia Foundation. Available at: https://en.wikipedia.org/wiki/Fed-batch_culture (Accessed: December 15, 2022).
8. *Perfusion cell culture: Upstream process intensification* (no date) *Repligen*. Available at: https://www.repligen.com/applications/perfusion (Accessed: December 15, 2022).
9. KBI Biopharma Follow (no date) *Debottlenecking manufacturing capacity: Initiating cell culture manuf...*, *Debottlenecking Manufacturing Capacity: Initiating cell culture manuf...* Available at: https://www.slideshare.net/kbibiopharma/debottlenecking-manufacturing-capacity-initiating-cell-culture-manufacturing-campaigns-using-seed-train-cryopreserved-in-a-disposable-bag (Accessed: December 15, 2022).
10. Michelet Dorceus, S.S.W. (2017) *Comparing culture methods in monoclonal antibody production: Batch, fed-batch, and perfusion*, *BioProcess International*. Available at: https://bioprocessintl.com/analytical/upstream-development/comparing-culture-methods-monoclonal-antibody-production-batch-fed-batch-perfusion/#:~:text=In%20the%20batch%20method%2C%20all%20nutrients%20are%20supplied,waste%2C%20supply%20of%20nutrients%2C%20and%20harvesting%20of%20product. (Accessed: December 15, 2022).
11. Author links open overlay panelCleoKontoravdiaSteven P.AspreybEfstratios N.PistikopoulosaAthanasiosMantalarisaPe

rsonEnvelope *et al.* (2006) *Development of a dynamic model of monoclonal antibody production and glycosylation for product quality monitoring*, *Computers & Chemical Engineering*. Pergamon. Available at: https://www.sciencedirect.com/science/article/pii/S0098135406000871?casa_token=IxsiqN6dNK0AAAAA%3AB_EXxDKVd8xilUZ6uI1aITx0iqJua9ggtT5x0czJdzfObjBr0fihPC-iskZfPYcGl8Urmfea4uo#bib2 (Accessed: December 15, 2022).

12. Kornecki, M. and Strube, J. (2019) *Accelerating biologics manufacturing by upstream process modelling*, *MDPI*. Multidisciplinary Digital Publishing Institute. Available at: https://www.mdpi.com/2227-9717/7/3/166/htm#B33-processes-07-00166 (Accessed: December 15, 2022).
13. Lao, M.-S. and Toth, D. (1997), Effects of Ammonium and Lactate on Growth and Metabolism of a Recombinant Chinese Hamster Ovary Cell Culture. Biotechnol Progress, 13: 688-691. https://doi.org/10.1021/bp9602360
14. Xing, Z., Bishop, N., Leister, K. and Li, Z.J. (2010), Modeling kinetics of a large-scale fed-batch CHO cell culture by Markov chain Monte Carlo method. Biotechnol Progress, 26: 208-219. https://doi.org/10.1002/btpr.284

Investigation on PEG-PCL Nanoparticles for Intracellular Drug Delivery

Minzhi Chen and Yichen Zhang

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

Abstract: Poly(ethylene glycol)-Poly(ε-caprolactone) (abbreviated as PEG-PCL) copolymers have demonstrated strong potential as the building block for an intracellular drug delivery system. Using nanoprecipitation, PEG-PCL nanoparticles were used to encapsulate a negatively charged payload (calcein) to examine the copolymer's ability to encapsulate payload without lipids and to explore whether the charged nature of the payload would bring any differences to the performance of the nanoparticles. The synthesized nanoparticles had an averaged size of around 200 to 300nm and an average zeta potential of around -17mV. The synthesized nanoparticles showed stability in deionized water and rapid releases of payload when exposed under human endosomal pH between 5.5 and 6.5. The synthesized particles were also resilient to external disturbances and exhibited a leakage of around 10% under 40 rpm shaking for 20 hours. However, the synthesized nanoparticle showed instability when exposed to physiological pH of 7.4 and a low encapsulation efficiency between 10% to 20%. Once the two problems are dealt with, single-component PEG-PCL nanoparticles could be favorable for an efficient intracellular drug delivery vector.

Keywords: *pH-responsive, drug delivery, nanoprecipitation, solvent evaporation, PEG-PCL copolymers, encapsulation efficiency*

1. Introduction

1.1. Background

A lot of effort has been put into the field of intracellular drug delivery systems in recent years. These systems are key to treating non-infectious diseases such as cancer and tumor, where drugs would not be effective unless the drug can be delivered into the specific organ and mutated cells. The process faces multiple delivery barriers including organ-level barriers, sub-organ level barriers, and subcellular barriers.[1]

A popular technique to overcome these barriers would be to synthesize bio-compatible drug delivery vectors. These vectors are nanoparticles that encapsulate the payload and have diameters in the scale 100 nm.[2] Ideally, such vectors are stable during in vitro storage and in vivo transport and quickly release the payload once they reach target cells. Among these materials, PEG-PCL copolymers have demonstrated superior stability in physiological conditions as well as fast payload release performance upon entry in target cells.[3] Hence, this study focuses on evaluating the performance of PEG_x-PCL_y nanoparticles as drug delivery vectors with calcein as payload, where x and y would stand for the average molecular weight of PEG and PCL in the copolymer, respectively.

1.2. Motivation

Extensive research has been done on PEG-PCL nanoparticles as a drug delivery agent. Some of these study have nanoparticles consisting of two components – a copolymer and a lipid.[4] It is of interest to see how well PEG-PCL alone can protect the payload during transportation and release the payload at the designated site. In addition, most payloads studied so far has been charge neutral, such as tetrandrine, anastrozole, and doxorubicin.[4-6] Using a payload that is negatively charged would be a good simulation of

how good PEG-PCL nanoparticles are at delivering negatively charged drugs, such as nucleic acids.

2. Methodology

2.1. Materials

Tetrahydrofuran (THF), acetone, sodium chloride, and ethyl alcohol were purchased from VWR Chemicals (Leicestershire, UK). Calcein, potassium chloride, and sodium phosphate dibasic were purchased from Sigma-Aldrich (Dorset, UK). Anhydrous citric acid, sodium citrate dihydrate, and potassium dihydrogen orthophosphate were purchased from Fisher Scientific (Leicestershire, UK). Hydrochloric acid solution (0.1M, 1M and 2M) and sodium hydroxide solution (0.2M and 2M) were prepared by members of Chen Research Group. PEG-PCL copolymers (PEG_{2k}-PCL_{5k}, PEG_{5k}-PCL_{10k}, PEG_{5k}-$PCL_{11.5k}$, PEG_{5k}-PCL_{13k}) were synthesized and characterized by Yifan Liu and Xinyu Lu (members of Chen Research Group).

2.2. Preparation of Solutions

To make calcein solution, first, solid calcein was added to water in a 10ml or 20 ml glass vial. The initial concentration of calcein was set to be around 5 mM (3mg ml^{-1}), well below the self-quenching concentration of calcein. After shaking, a small amount of 2M sodium hydroxide solution was added to the solution to help the calcein dissolve. Finally, 2M hydrochloric acid solution was added to the solution to calibrate the pH of the solution to be between 7.0 and 7.4. This solution was further diluted to different concentrations based on the need of the study.

Phosphate buffered saline (PBS) was used to make buffers of pH 7.4 and pH 6.5. PBS buffers contained 137mM sodium chloride, 2.7mM potassium chloride, 10mM of sodium phosphate dibasic and 1.8mM of potassium dihydrogen orthophosphate. Additional 2M hydrochloric acid solution and 2M sodium hydroxide solution was used to calibrate the solution to its target pH.

Citrate buffer was used to make buffers of pH 5.5. This buffer contains 70.3mM sodium citrate dihydrate and 29.7mM of anhydrous citric acid. Additional 2M hydrochloric acid solution and 2M sodium hydroxide solution was used to calibrate the solution to its target pH.

2.3. Measurement of calcein concentration

Earlier studies reported that the relationship between calcein concentration and fluorescence reading is proportional below 4mM.[7] Following this relationship, a calibration curve was made as the standard for calcein concentration for later experiments.

Different concentrations of calcein solutions from 10^{-5}mM to 0.01mM were prepared using the calcein solution mentioned in 2.2. To measure concentration of calcein in a sample, three copies of 200µl solution from the sample were extracted and pipetted into a 96-well plate for calcein florescence reading in a plate reader (GloMax Multi-Detection System, Promega, USA). The calibration curve was then plotted for future reference,

2.4. Synthesis of PEG-PCL Nanoparticles

Firstly, a certain amount of copolymer (between 8mg and 10mg) was dissolved in corresponding volume of THF or acetone to make a solution of 10mg/ml in a 1.5ml plastic vial, called the organic phase. Secondly, calcein was extracted from calcein solution made from 2.2. and diluted with deionized water to make an 800 µl 2mg/ml solution in a 10ml glass vial, called the liquid phase. Then, 800 µl of the organic phase was pipetted quickly into the aqueous phase. The glass vial containing the mixture was then left on a magnetic stirrer (Topolino, IKA, Germany) for continuous stirring to fully evaporate the organic phase. This process is referred to as nanoprecipitation. Another sample with no payload and THF as the organic phase was also

synthesized which worked as a comparison group.

2.5. Characterization of Nanoparticles

The average size, polydispersity index (abbreviated as PDI), and average zeta potential of synthesized PEG-PCL nanoparticles were measured. As these properties were inherent to the nanoparticles, they would be referred as "inherent properties". All nanoparticle-containing suspensions were diluted to above 1.5ml for accurate measurement in a particle size analyser (Litesizer 500, Anton Paar, Austria). Size and PDI were measured using the particle size series mode for five repetitive measurements per sample, and zeta potential was measured using the zeta potential series mode for three repetitive measurements per sample.

2.6. Purification and Encapsulation Efficiency

After nanoprecipitation, 600 µl of the nanoparticle-containing suspension was extracted from the glass vial and diluted to around 1ml to 1.5ml. This diluted suspension was then pipetted into a dialysis tube (Float-A-Lyzer G2 100kD 1ml, Spectra/Por, USA). Dialysis tube was then assembled with a floater and immersed in 200ml of deionized water in a sanitized 250ml beaker. The dialysis would be run for some time, during which the water was replaced hourly until the concentration of calcein in water dropped below a threshold value (5×10^{-5} mM).* For each replacement of water, the concentration of calcein inside the replaced water was recorded using the plate reader. All was summed up to calculate the overall amount of calcein escaped from the dialysis tube, which equals the amount of unencapsulated calcein. The encapsulation efficiency (EE) is then calculated by:

$$EE\ \% = \frac{V_{n,extracted} C_{c,n} - V_{water} \sum_i C_{c,i}}{V_{n,extracted} C_{c,n}} * 100\%$$

where $V_{n,extracted}$ is the volume extracted from the glass vial after nanoprecipitation (600 µl), $C_{c,n}$ is the concentration of calcein inside the glass vial after nanoprecipitation (2mg/ml assuming complete evaporation of organic phase), V_{water} is the volume of water (200ml) and C_c is the concentration of calcein in water.

2.7. Storage and Measurement of Leakage

All synthesized nanoparticles were stored in tightly sealed glass vials at 4 degrees with no agitation and no light irradiation for 1, 3, or 5 consecutive days before a leakage study was carried out. Leakage measurement is similar to that of encapsulation efficiency measurement. The only difference would be the amount leaked would be measured only once at 2 hours into the test. Leakage percentage was calculated by:

$$Leakage\ \% = \frac{V_{n,extracted} C_{c,n} - V_{water} C_c}{V_{n,extracted} C_{c,n}} * 100\%$$

2.8. pH-Dependent Payload Release

After complete purification (described in 2.6) to remove the unencapsulated calcein, the volume of nanoparticle-containing suspension in the dialysis tube was recorded and calibrated to 3 ml. The nanoparticle-containing suspension was then separated into three equal portions. Two of them were transferred into two new dialysis tubes, one in 200ml pH 6.5 buffer and another in 200ml pH 5.5 buffer with the third portion left for leakage test, or characterization measurements as discussed in 2.5, or 200ml pH 7.4 buffer for another payload release. The concentration of calcein in the buffers were measured at of 2, 5, 15, 30, 60, 90, 120, 150, and 180 minutes into payload release. The concentration measurement took the same protocol as described in 2.3. For each measurement, an equal volume of buffer was refilled into the beaker to account for the plate reading loss. After 4 hours (or sometimes overnight), plate reading was carried out for both the nanoparticle-containing suspension inside the dialysis tube and outer buffer. The release

* Dialysis was sometimes conducted overnight, during which the water in the beaker was not changed.

percentage is calculated by:

$$Release\ \% = \frac{C_{buffer} * V_{buffer}}{C_{buffer} * V_{buffer} + C_{tube} * V_{tube}} * 100\%$$

where C_{buffer} is the concentration of calcein in the buffer, V_{buffer} is the buffer volume (200ml), C_{tube} is the concentration of calcein in dialysis tube and V_{tube} is the volume of nanoparticle-containing suspension in the dialysis tube at the end of payload release.

2.9. Shaking Test

After purification, some samples were extracted from the dialysis tube, diluted to 1 ml if the volume left in the dialysis tube was less than 1 ml, and transferred into 1.5ml plastic vials. The plastic vials were covered with aluminum foil to avoid light irradiance. The vials were then clipped on to a shaker (Stuart SB3, VWR, UK) and rotation speed was set to 40 rpm. Samples underwent horizontal or vertical shaking with a duration of 2 or 20 hours. After shaking, the samples go through the purification step mentioned in 2.7. again to determine additional leakage.

2.10. Statistical Analysis

All data reported was the average of at least three repetitive measurements. Error bars or error margins were calculated for 95% confidence interval. A student's *t* test was employed to evaluate whether comparable values from different data sets were statistically different with a value of $P < 0.05$ as an indication of statistical difference. Values without error margins were usually generalizations of the data. Should two sets of data turned out to be statistically different, the magnitude of difference was also usually discussed.

3. Results and Discussion

3.1. Inherent Properties: Size, PDI and Zeta Potential

3.1.1. General description

Size indicates the feasibility of the nanoparticles to enter the cells. PDI represents the homogeneity of nanoparticles synthesized. Zeta potential measures the surface charge and may also indicate the stability of the nanoparticles in a suspension. Nanoparticles need to 'pass' (has a similar magnitude with what other researchers reported) these three basic criteria before further experiments are carried out.

3.1.2. Trends across copolymers and solvents

Figure 3.1 shows how the average size change across various copolymers synthesized with THF. PEG_{2k}-PCL_{5k} nanoparticles showed a bigger standard deviation in average size compared to other nanoparticles. Additionally, size distribution diagram shows that PEG_{2k}-PCL_{5k} nanoparticles had a large PDI with two peak intensities. The lack of homogeneity and variation from batch to batch makes PEG_{2k}-PCL_{5k} unsuitable as a drug delivery vector and is hence excluded from further analysis. Figure 3.2 shows the size distribution of PEG_{2k}-PCL_{5k} compared to that of an 'acceptable size distribution'. Nanoparticles synthesized with PEG_{5k}-PCL_{10k} PEG_{5k}-$PCL_{11.5k}$ and of PEG_{5k}-PCL_{13k} do not show statistical difference in terms of average sizes.

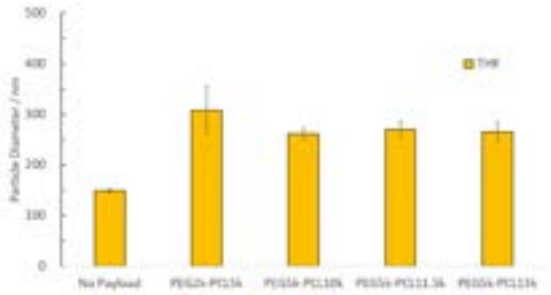
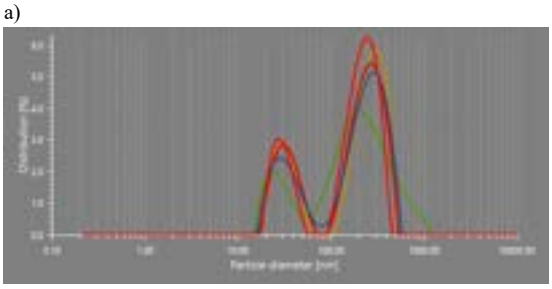


Figure 3.1 Average sizes of PEG-PCL nanoparticles synthesized with THF. 'No payload' indicates a trial synthesized with PEG_{5k}-PCL_{10k} without calcein.

a)



b)

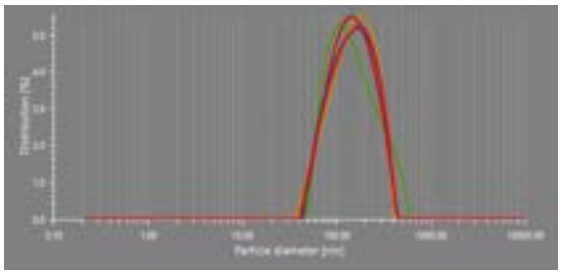


Figure 3.2 Particle size distribution of a) PEG_{2k}-PCL_{5k} nanoparticles synthesized with THF compared to b) PEG_{5k}-$PCL_{11.5k}$ nanoparticles synthesized with THF

Figure 3.3 shows the average nanoparticle sizes of all the copolymers that were considered 'qualified' for further payload release testing. Similar to THF, nanoparticles synthesized with acetone do not show statistical differences in average size for different copolymers. Average sizes of nanoparticles synthesized with acetone were smaller than those synthesized with THF, although this difference is not statistically significant for nanoparticles synthesized from PEG_{5k}-PCL_{10k}. The average of the averaged sizes of PEG-PCL nanoparticles synthesized with THF were around 270nm while those synthesized with acetone were around 200nm.

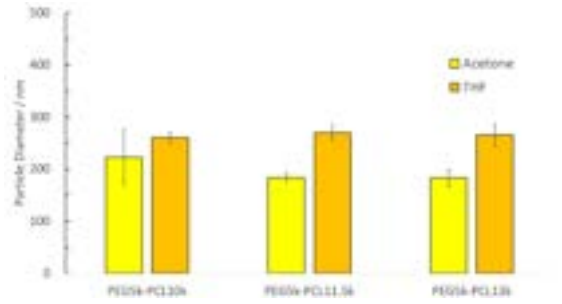


Figure 3.3 Size comparison between nanoparticles synthesized with acetone and nanoparticles synthesized with THF.

Figure 3.4 shows the zeta potentials of the nanoparticles synthesized across different copolymers. The zeta potentials of the nanoparticles are all negative. Nanoparticles synthesized with PEG_{5k}-$PCL_{11.5k}$ tend to have a statistically significant more negative zeta and potential regardless of the choice of organic solvent.

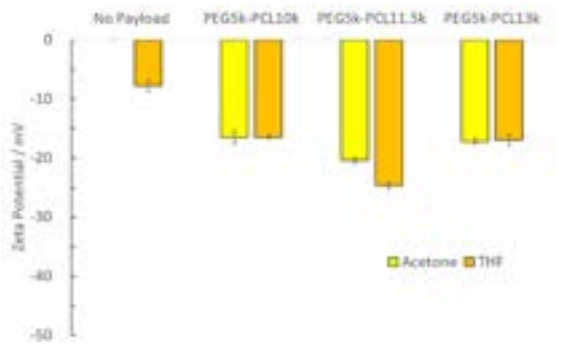


Figure 3.4 Different average zeta potentials of various PEG-PCL copolymers synthesized with THF and acetone.

3.1.3. Comparison with literature

Overall, average sizes of the nanoparticles are within the ranges of what other researchers have reported which generally varies between 80nm to 285nm.[4,6] The differences in the average sizes of the nanoparticles arising from the organic solvent was also reported by other researchers using a similar synthesizing method. A study on PCL-PEG-PCL nanoparticles claimed that while some organic solvents are completely miscible with water, the difference in solvent water miscibility led to difference in nanoparticle size.[8] The average size of PCL-PEG-PCL nanoparticles synthesized with THF was 40% larger than that synthesized with acetone which matches the result of this study.

The averaged zeta potentials of PEG-PCL nanoparticles in this study, which ranges from 16.48 ± 0.31mV to -24.66 ± 0.58mV, are more negative than what other studies report, which lie between -0.04 ± 0.04mV and -11.47 ± 0.64mV.[6] This difference in zeta potential could be contributed by the payload. Calcein, having 4 of its 6 pKas below 5.5, would readily dissociate into calcein anion at pH 7 where zeta potential measurement took place.[9] Encapsulated calcein anions can contribute to negative charges during the zeta potential measurements. The fact that zeta potentials of the unloaded nanoparticle being -7.79 ± 1.04mV, within the range of what other researchers had reported and significantly less negative than loaded PEG-PCL nanoparticles,

offers further validation of calcein contributing to the negative zeta potential.

The PDI of the nanoparticles varied between 0.25 to 0.30, which is higher than what other researchers have obtained which were between 0.05 and 0.18.[6] One potential explanation would be passive coating methods are difficult to control, as it is very difficult to determine the exact stop point of nanoprecipitation (i.e., full evaporation of organic phase) and control the exact stirring speed,

3.2. Encapsulation Efficiency

3.2.1. Trends across copolymers and solvents

Figure 3.5 depicts the encapsulation efficiencies of nanoparticles made from different copolymers and different solvents. The encapsulation efficiencies across different copolymers are not statistically different from each other for both solvents. On the other hand, the encapsulation efficiency of nanoparticles synthesized with acetone is higher than that synthesized with THF, which is statistically significant across all three copolymer nanoparticles.

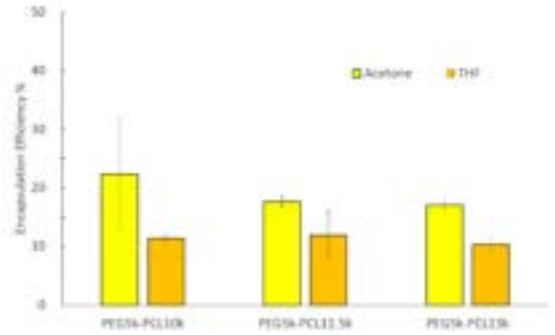


Figure 3.5 Different average zeta potentials of various PEG-PCL copolymers synthesized with THF and acetone.

3.2.2. Comparison with literature

The measured encapsulation efficiency has a range between 10.37±1.28% and 22.34±9.42%. This is much lower than some studies. One study pointed out that the encapsulation efficiencies of PEG-PCL with nobiletin as the payload could be as high as 76.34±3.25% using DMSO as the organic phase solvent for nanoprecipitation.[10] Besides the different choice of organic phase solvent, which had already proven its ability to influence different encapsulation efficiencies in figure 3.5, the study had the payload and the copolymer dissolved in the same organic phase which may lead to a different encapsulation process. In addition, the charged nature of calcein may also lead to low encapsulation efficiency as the anion, together with negatively charged PEG-PCL, created electrostatic repulsion that makes assemble of nanoparticle more difficult.

There are currently a limited number of studies on how the choice of organic phase solvent may influence encapsulation efficiency. It is plausible that the difference in solvent water miscibility, as discussed in 3.1.3, caused this discrepancy.[8] Another study reported that different solubilities of organic solvent in water may lead to encapsulation efficiency difference.[11] Based on this, it is likely that the differences in size and encapsulation efficiency were caused by the interactions between different organic solvent molecules and water molecules. The key properties that need to be investigated may include viscosity and surface tension coefficient.

3.3. Payload Release

3.3.1. Payload Release Profile

Figure 3.6 depicts payload release over time for a single batch of nanoparticle at four different pH. The presence of ions (in buffers) would lead to significant release of the payload. All release profiles show decreasing releasing rate with time. Also, from the graph, PEG-PCL nanoparticles are stable under DI water storage conditions.

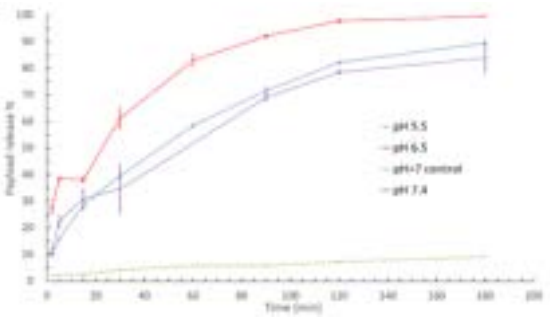


Figure 3.6 A payload release profile of PEG_{5k}-PCL_{13k} nanoparticles synthesized with THF across four pHs.

3.3.2. Trends across different variables

Figure 3.6 shows that the payload release rate at pH of 6.5 is quicker compared to that at pH 5.5 and pH 7.4. At pH of 6.5, 82.96±2.12% of the payload were released after 1 hour compared to a release of and 58.48±0.94% and 51.82±0.02% at pH 5.5 and pH 7.4, respectively. The difference between payload release at pH 5.5 and that at pH 7.4 are not significant.

The trend of payload release at pH 6.5 faster than that at pH 5.5 holds for all nanoparticles synthesized with acetone. Specifically, 1 hour into the experiment see approximately 45% of payload released at pH 5.5 and 60% of payload release at pH 6.5. For some nanoparticles synthesized with THF, however, there is no difference in payload release rate between pH 6.5 and pH 5.5. Figure 3.7 provides a good example.

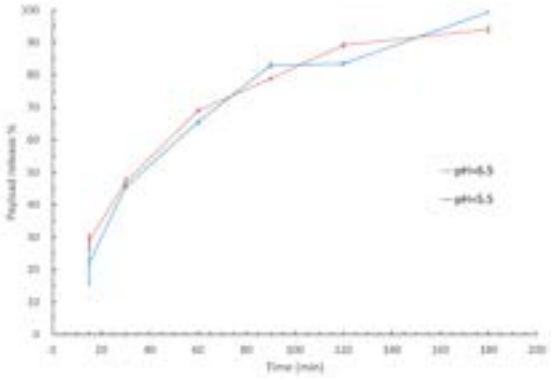


Figure 3.7 A payload release profile for PEG_{5k}-PCL_{13k} nanoparticles synthesized with THF at pH 6.5 and pH 5.5. The difference between two profiles is not significant in magnitude.

3.3.3. Effect of payload release on Inherent Properties

The average size of nanoparticles did not show significant differences. For example, the average size of PEG_{5k}-$PCL_{11.5k}$ nanoparticles synthesized with THF after a 24-hour payload release at pH 6.5 was 264.55±11.35nm, which does not deviate heavily from the average size for the same nanoparticle before dialysis, which was 285.56 ± 6.45nm. However, the average zeta potential showed an increase. The averaged zeta potential for PEG_{5k}-$PCL_{11.5k}$ nanoparticles after payload release was -10.92±0.38mV. As the zeta potential for the same nanoparticles without payload release was -24.66±0.58mV, this is a strong validation that encapsulated calcein escaped from PEG-PCL nanoparticles during payload release. Additionally, as the average size did not change much during payload release, it could be reasonably inferred that an ionic buffer environment created gaps in PEG-PCL nanoparticles that are larger than the size of calcein molecules, allowing calcein to escape. The overall structure of PEG-PCL nanoparticles, however, remained intact.

3.3.4. Comparison with literature

Various researchers have claimed that PEG-PCL nanoparticles are stable when stored under pH 7.4 environment for months, which the data from this study strongly disagrees.[3] One study went further and reported that PEG-PCL nanoparticles released only around 20% of the payload within 10h under pH 7.4 with similar solvent evaporation preparation method and doxorubicin as the payload.[12] Two key differences may explain the difference in payload release rate. The first would be that the study did not publish the molecular weight of PEG-PCL used for synthesis and it is evident (from the case of PEG_{2k}-PCL_{5k} discussed in 3.1) that the molecular weight of copolymer could affect properties of the synthesized nanoparticles. Secondly, doxorubicin was not highly charged at pH 7.4 while calcein was negatively charged at pH 7.4. The difference in charge may affect the rate which the payload escapes the nanoparticle should there be a hole that is similar in size to the payload. In addition, other studies also pointed out that while PEG-PCL nanoparticles helped reduce the payload release rate at pH 7.4 compared to free, unencapsulated payload, the magnitude of reduction was rather moderate.[10] The study, which uses PEG-PCL nanoparticles to encapsulate nobiletin, observed a 90% release of free nobiletin released compared to a 65% released of PEG-PCL encapsulated nobiletin under pH 7.4 environment for 12 hours.

3.3.5. Real-life implications of the results

Comparing the ideal scenario (as discussed in 1.1) to the results of this study, the biggest problem of single-component PEG-PCL nanoparticle would be instability under pH 7.4. While payload release rate at pH 6.5 and pH 5.5 may differ, all samples record a release of more than 45% within 1 hour and more than 70% within 3 hours. Such results are acceptable and further optimization is not a priority.

3.4. Leakage under static condition

3.4.1 Leakage Profile across copolymers

Figure 3.8 shows leakage profiles of PEG-PCL nanoparticles synthesized with acetone.

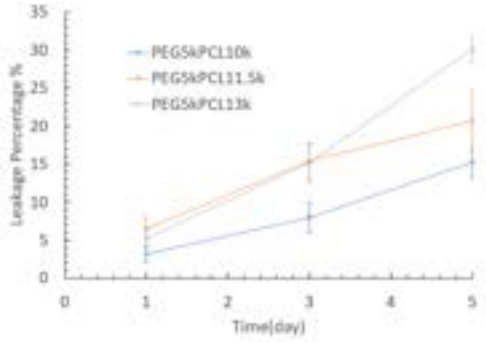


Figure 3.8 Leakage profile over time for nanoparticles made from three copolymers with acetone

PEG_{5k}-PCL_{13k} nanoparticle saw the largest leakage over the 5 days, reaching around 30% at day 5. All three copolymer nanoparticles exhibited similar leakage amount around 5% at day 1, while the difference became more significant with increasing time. It could be concluded from the figure that larger molecular mass of PEG-PCL monomers would result in higher leakage within a certain amount of time. However, leakage percentages for different copolymer nanoparticles were sometimes not statistically different from each other.

3.4.2 Comparison with Literature

The leakage profile over days is in line with what other researchers reported. When varying the coating material, the payload leakage is approximately increasing from 10% at day 1 to 30% at day 5.[13]

3.5. Effect of Shaking

3.5.1. Effect of Short Shaking on Leakage

Shakings for 2 hours simulates external disturbances experienced by the nanoparticles during intra-city transports. Figure 3.9 displays examples of the leakage of nanoparticles after 2 hours of shaking. The additional leakage was small in magnitude (less than 10%) and sometimes smaller than that of the control group.

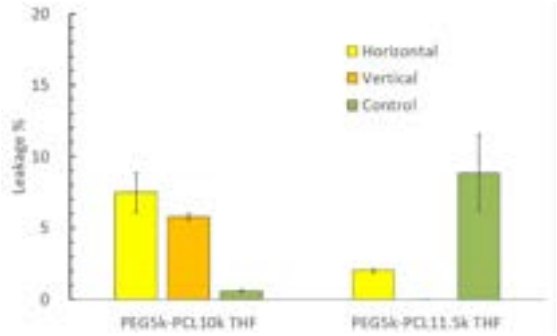


Figure 3.9 Leakage after 2 hours of shaking

3.5.2. Effect of Prolonged Shaking on Leakage

Shakings for 20 hours simulates of external disturbances experienced by the nanoparticles during inter-city or international transports. Figure 3.10 shows the leakage of various nanoparticles over 20 hours. Additional leakage was statistically significant at around 10%, indicating that the amount of additional leakage is positively correlated with the duration of shaking.

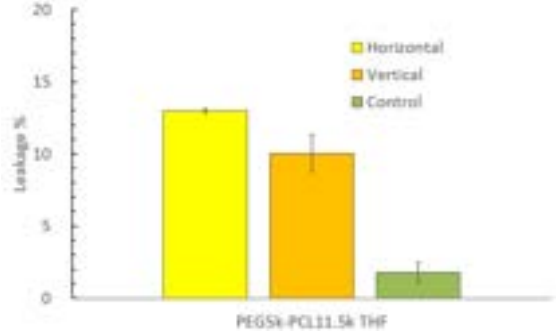


Figure 3.10 Leakage after 20 hours of shaking

3.5.3. Effect of Shaking on Inherent Properties

Inherent properties of PEG-PCL nanoparticles did not undergo significant changes after shaking. The effect of short-time shakings on these properties were statistically insignificant. For example, the average size of PEG_{5k}-$PCL_{11.5k}$ nanoparticles (synthesized with acetone) after horizontal shaking for 1 hour was 223.77 ± 7.32nm which is similar to 222.34 ± 7.45nm in the control group.* The average zeta potential of the same nanoparticles after shaking were -20.86±0.63mV compared to -24.66±0.58mV in the control group. The results met with leakage results as small or statistically insignificant leakage should not lead to changes in inherent properties. For longer shaking, the average size of nanoparticles also remained unchanged. PEG_{5k}-$PCL_{11.5k}$ nanoparticles (synthesized with THF) returned an average size of 266.88±10.20nm after a 20-hour horizontal shaking, which is still comparable to the control group's average size of 270.10 ± 15.66nm. However, a slight increase in average zeta potential was observed. The average zeta potential of the same nanoparticles after shaking was -21.18±0.52mV, which is statistically higher than -24.66±0.30mV in the control group. These observations were coherent with the leakage results, as longer shakings incur a statistically significant leakage. Average zeta potential of the nanoparticles became less negative as a result of calcein anions escaping the nanoparticle.

4. Conclusions

The study of PEG-PCL nanoparticles proves that conventional nanoprecipitation techniques can be used to encapsulate negatively charged drugs just as to encapsulate charge neutral drugs. PEG-PCL nanoparticles also showed superior payload release performance in early and late endosomal pH environments. Based on in vitro analysis results, acetone was determined to be the superior organic solvent for PEG-PCL nanoparticle synthesis as nanoparticles synthesized with acetone had smaller average sizes which enabled better cell wall penetration and higher encapsulation efficiencies which means more drugs could be loaded to the nanoparticle.

Meanwhile, this study also shows that simple nanoprecipitation process led to low encapsulation efficiencies in the range of 10% to 20%. In addition, a single layer of PEG-PCL was not sufficient to achieve stability in physiological pH conditions. Further analysis could be done on measuring the behavior of the nanoparticles under actual blood and cell conditions, and further work needs to be done on improving the stability of single-layer PEG-PCL nanoparticles in physiological pH conditions and improving the encapsulation efficiencies. Despite the challenges, PEG-PCL nanoparticles have demonstrated good potential towards a successful drug delivery vector.

■ **Acknowledgements**

The authors would like to express their sincere gratitude to Prof. Rongjun Chen and Yifan Liu for their consistent support, feedback, and insights throughout the research. The authors would also like to thank Xinyu Lu, Yifan Ding, and Dr. Apanpreet Kaur for their assistance and expertise during the research.

■ **References**

(1) Poon, W.; Kingston, B. R.; Ouyang, B.; Ngo, W.; Chan, W. C. W., A framework for designing delivery systems. *Nat Nanotechnol.* **2020**;15(10):819-829.

(2) De Jong, W.H.; Borm P.J., Drug delivery and nanoparticles: applications and hazards. *Int J Nanomedicine*. **2008**;3(2):133-149.

* Size measurements were statistically different from that reported in figure 3.x. Such differences could arise from simple batch to batch difference or due to the fact that these data were measured 4 days after synthesis of the nanoparticles.

(3) Grossen, P.; Witzigmann, D.; Sieber, S.; Huwyler, J., PEG-PCL-based nanomedicines: A biodegradable drug delivery system and its application. *J Control Release.* **2017**;260:46-60.

(4) Massadeh, S.; Omer, M. E.; Alterawi, A.; Ali, R.; Alanazi, F. H., Almutairi, F., Almotairi, W., Alobaidi, F. F.; Alhelal, K.; Almutairi, M. S.; Almalik, A.; Obaidat, A. A.; Alaamery, M.; Yassin, A. E., Optimized Polyethylene Glycolylated Polymer-Lipid Hybrid Nanoparticles as a Potential Breast Cancer Treatment. *Pharmaceutics*. **2020**;12(7):666.

(5) Diao, Y. Y.; Li, H. Y.; Fu, Y. H.; Han, M.; Hu, Y. L., Jiang, H. L.; Tsutsumi, Y.; Wei, Q. C.; Chen, D. W.; Gao, J. Q., Doxorubicin-loaded PEG-PCL copolymer micelles enhance cytotoxicity and intracellular accumulation of doxorubicin in adriamycin-resistant tumor cells. *Int J Nanomedicine.* **2011**;6:1955-1962.

(6) Li, R.; Li, X., Xie, L.; Ding, D.; Hu, Y., Qian, X.; Yu, L., Ding, Y.; Jiang, X.; Liu, B., Preparation and evaluation of PEG-PCL nanoparticles for local tetradrine delivery. *Int J Pharm*. **2009**;379(1):158-166.

(7) Hamann, S.; Kiilgaard, J.F.; Litman, T.; Alvarez-Leefmans, F.J.; Winther, B.R.; Zeuthen, T., Measurement of Cell Volume Changes by Fluorescence Self-Quenching. *J. Fluoresc.*, **2002**;12:139-145.

(8) Anh Nguyen, T.H.; Nguyen, V.C., Formation of nanoparticles in aqueous solution from poly(ε-caprolactone)–poly(ethylene glycol)–poly(ε-caprolactone). *Adv. Nat. Sci.: Nanosci. Nanotechnol.,* **2010**;1.

(9) Dojindo EU GmbH 2022, *Calcein*, accessed 15 December 2022, https://www.dojindo.eu.com/store/p/623-Calcein.aspx

(10) Wang, Y.; Xie, J.; Ai, Z.; Su, J., Nobiletin-loaded micelles reduce ovariectomy-induced bone loss by suppressing osteoclastogenesis. *Int J Nanomedicine.* **2019**;14:7839-7849.

(11) Ravi, S.; Peh, K. K.; Darwis, Y.; Murthy, B. K.; Singh, T. R.; Mallikarjun, C., Development and characterization of polymeric microspheres for controlled release protein loaded drug delivery system. *Indian J Pharm Sci*. **2008**;70(3): 303–309.

(12) Sun, H.; Guo, B.; Cheng, R.; Meng, F.; Liu, H.; Zhong, Z., Biodegradable micelles with sheddable poly(ethylene glycol) shells for triggered intracellular release of doxorubicin. *Biomaterials*, **2009**;30(31):6358–6366.

(13) Chiu, H. T.; Su, C. K.; Sun, Y. C.; Chiang, C. S.; Huang, Y. F., Albumin-Gold Nanorod Nanoplatform for Cell-Mediated Tumoritropic Delivery with Homogenous ChemoDrug Distribution and Enhanced Retention Ability. *Theranostics*, **2017**;7(12): 3034–3052.

Forecasting Building Energy Usage to Drive Net Zero Investments for a Major Food Retailer

Kamila Tohlukov *(CID: 01776494)* and Christophorus Gilbert *(CID: 01741325)*

Department of Chemical Engineering, Imperial College London, South Kensington Campus, London SW7 2AZ, United Kingdom

REPORT INFO

General information:

Word count: 7134
Number of pages: 10
Session: 14/12/2022

Key words:

Energy
Environment
Economics
Mathematical Modelling
Optimisation
Numerical Analysis

ABSTRACT

This work develops regression models to investigate the factors affecting gas and electricity usage of food retail buildings with the objective of developing a net-zero investment strategy, in partnership with Sainsbury's. The relationship between store area and weather factors on energy usage was implemented on linear and multi-variate linear regression (MLR) models and their performance was compared. It was concluded that the MLR model gave a better fit for the data sets used, yielding a more accurate ranking of the stores' performance. Energy optimisation was performed by converting the gas boilers to GSHPs and implementing electricity-saving technologies. While installing GSHPs led to a substantial reduction in carbon emissions, it was not attractive from a financial standpoint due to its negative NPV. An investment scenario, Scenario 2, where energy optimisation would be implemented on the worst 5 performing stores, was deemed to make the most financial sense as it had the greatest NPV per store. This scenario was also predicted to reduce the carbon emitted in 5 years by 5,090 tonnes, aligning with Sainsbury's net-zero targets. Finally, this work highlights the need for further research on energy-saving technologies.

1 Introduction

In 2021, the operation of buildings accounted for 30% of the global final energy consumption [1]. The increase in energy use in the building sector is therefore responsible for one-third of global energy-related CO2 emissions. Moreover, the outline made by the Intergovernmental Panel on Climate Change (IPCC) in their last 2022 climate report highlights the continuing rise of total net anthropogenic greenhouse gas (GHG) emissions [2]. Therefore, to minimize the temperature rise in the upcoming years, the IPCC believes in the reduction of carbon emissions through performance and efficiency improvements, allowing major cost savings. The retail sector, in particular supermarkets, accounts for approximately 1% of the UK's annual GHG emissions. The high energy demand for food retail buildings is mainly due to the presence of gas and electricity features, refrigeration, ventilation, and air conditioning [3]. The necessity of reducing these emissions is faced by many operators, including Sainsbury's, the UK's second largest grocery chain, with a 15.1% market share in 2022 [4]. Their commitment to aligning to a 1.5C trajectory across all scopes, therefore reducing GHG emissions from their operations by 2035, has already been put in place. Reaching their net-zero target is unhesitatingly combined with the improvement of the emissions outline of their portfolio.

2 Background

There is a significant amount of literature on the analysis of building energy performance and efficiency. To identify the best emissions and cost reduction solutions, understanding the complex load requirements of power, heating, and cooling of energy-intensive buildings is required. While some studies on energy demand drivers in food retailers have shown that evaluating annual energy demands can be done by knowing only a few characteristics of food retail stores [1], others have developed data-driven methods to predict this demand [4]. These methods correspond to benchmarking techniques, allowing a comparison of the energy performance of similar commercial buildings [8, 9, 10]. Benchmarking consistently measures buildings' energy use in relation to their size and other core characteristics. It is essential to understand which buildings are the most inefficient and what design improvements could be made to address them. Current energy benchmarking methods can be categorized into black, gre,y and white box methods [11]. A number of these methods have been developed based on project requirements and available monitoring data.

Multiple regression methods have been developed with various applications including linear regression (LR), multi-linear regression (MLR) [8], and decision trees (DT). While these methods are conceptually simple and thus easy to use, other methods that show a

different approach can be more challenging due to the embedded statistical knowledge and concepts. This includes artificial neural networks (ANN), used by Kalogirou in his 2000's work to predict the energy consumption of a passive solar building. [6]. While these models can predict a building's electricity and gas consumption, they cannot inform on the actions needed to be taken to achieve a better performance of this building. A few literature gaps were identified, mostly around the prediction of building energy performance considering weather and carbon factors. Moreover, other limitations were found regarding general investment strategy roadmaps for particular national organizations [5]. Strategic financial planning in the sector's large organizations, such as Sainsbury's, requires knowledge of future demands for gas and electricity consumption.

By partnering with Imperial, Sainsbury's benefits from the university's research and expertise to assess state-of-the-art technologies and make decisions on energy savings and carbon reduction strategies. The second largest food retailer in the UK has provided access to data on energy consumption, including gas and electricity usage, and information on store characteristics for a large set of stores.

The main aim of this research is to deliver a potential investment strategy specifically tailored for Sainsbury's case. This work principally explores two existing benchmarking methods to develop a rigorous forecast of energy consumption, with a principal focus on the implementation of weather and carbon features, including store characteristics such as store area. Further, this research presents a financial analysis based on the evaluation of costs included in retrofitting buildings with different heating and electricity technologies.

This paper is composed of four sections. The current section has provided the background, motivations, and scope of the problem. The second section informs of the methodology along with the mathematical conceptualization of the models used to categorize the estate and the means through which energy forecasting was performed. This section also describes the investment strategy scenarios based on energy and financial considerations. The third section provides the results from the two regression models, as well as the energy consumption forecast and cost insights. Finally, the last section of this work provides concluding marks, including a potential investment strategy based on the earlier sections' results.

3 Methodology

3.1 Data Collection and Treatment

To develop regression models, data was first collected, cleaned, and processed.

3.1.1 Building Data

A database containing building characteristics was provided by Sainsbury's. Data was given for 1927 assets from the financial year 2017 to 2020. This data set included all stores such as supermarkets, convenience stores and click & collect grocery markets. To have uniformity in building types and to only consider operationally intense assets, only Sainsbury's supermarkets were considered, which narrowed down the number of assets to 601. A few building characteristics were gathered, including gas and electricity consumption by four-week periods, sales area, and building technology features. The stores that contained missing data or anomalies were then removed, leaving 183 stores for gas consumption data and 174 stores for electricity consumption data.

3.1.2 Weather Data

UK hourly raw weather dataset was taken from the CEDA Weather Database [12]. An SQL database was built to store the raw data, which could then be used to easily query the weather data. This data was gathered by coordinating the postcodes of each store to the closest weather station. The two weather features processed, namely the average heating degree days (HDD) and cooling degree days (CDD) [13], were calculated based on equations 1 and 2. These two features provided a more accurate interpretation of the effects of temperature on the heating and cooling demand of the stores [14].

$$HDD = \sum_{i=day} v_i(T_i - T_b) \quad v_i \begin{cases} 1 & T_i < T_b \\ 0 & Else \end{cases} \tag{1}$$

$$CDD = \sum_{i=day} v_i(T_i - T_b) \quad v_i \begin{cases} 1 & T_i > T_b \\ 0 & Else \end{cases} \tag{2}$$

T_i as daily average temperature, T_b as baseline temperature.

Based on literature, and to maximise correlation between energy consumption and temperature, the baseline temperature was selected to be 15°C for CDD and HDD [15].

3.1.3 Cost and Carbon Emission Factors

In this case, cost factors represented the cost of electricity and natural gas per unit energy consumed, with units of $£/kWh$. Similarly, the carbon emission factor described the amount of CO_2 released per unit energy consumed with units of $kgCO_2/kWh$. The dataset provided by Sainsbury's included forecasted annual cost factor data for all stores from FY-21/22 to FY-24/25. The cost factor values are shown in Table 5. The carbon emission factor was taken as a constant at 0.184 $kgCO_2/kWh$. To evaluate the Net Present Value (NPV) of the two chosen scenarios, carbon abatement

costs for different technologies was taken from the literature [15]. These factors were considered to be more representative to develop an investment strategy.

3.2 Electricity and Gas Demand Forecast

Electricity and gas demand was forecasted by performing linear regression where the year was the independent variable, whereas the electricity and gas demand represented the dependent variables. The forecasting was performed by taking the average of the year-on-year (y-o-y) growth of electricity and gas demand for FY 17/18 to FY 19/20. The average y-o-y growth was found to be -5.6% for electricity demand and 0.13% for gas demand. These two values were then used to forecast the energy consumption in the next 5 years by taking the assumption that the y-o-y growth would remain constant at these two values. These findings will be further described in Section 4.1.

3.3 Building Energy Performance

The building energy performance was analysed with the purpose of understanding the main factors that affect a building's gas and electricity consumption and exploring possible methods to optimise the overall performance. For the purpose of this study, we have narrowed down the factors that affects the building energy performance to heating degree days (HDD), cooling degree days (CDD) and store area.

The dataset provided by Sainsbury's contained the building characteristics showing the different heating and cooling technologies of each store. As these technologies would greatly affect the efficiency and performance of the building energy usage, the building energy performance analysis was conducted separately for the different types of technologies. Regarding gas consumption, only the stores with gas boilers were considered for this study as they represented the majority of the stores. For electricity consumption, the stores were segmented into two categories of buildings: those with Ground Source Heat Pumps (GSHP) and those without, as this technology was found to have a substantial effect on the electricity consumption of a store.

The study used here may have several limitations since it did not consider any other weather parameters such as wind or humidity, or any other technical advances and building improvements. Furthermore, change in footfall was taken to be negligible. Despite that, this method yielded substantial results in an easily realisable way.

3.3.1 Linear Regression

Linear regression method was used to analyse the impact of weather and store area on building energy performance and to determine whether a linear relationship between these parameters could be observed. The linear regression was performed based on two key equations, one for electricity and one for gas consumption. The two equations are shown below:

$$G = \alpha \times HDD \times SA_{incl.} \quad (3)$$

$$E = \beta \times CDD \times SA_{excl.} \quad (4)$$

G is the gas annual consumption, E is the electricity annual consumption and SA is the store area. In this equation α and β are the regression coefficients to be calculated by plotting the dependent variable against the independent variable and taking the gradient at that point. The store area, $SA_{incl.}$, used to calculate α includes sales area and checkouts, while the $SA_{excl.}$ to determine β discards checkouts. This is due to electricity, which is mainly used for refrigeration systems, only including the sales area, while gas provides for the heating system in the whole building.

The calculated α and β coefficients could then be used to rank the store based on their performance. The lower the α and β values, the more efficient the store was in utilizing energy. Based on these coefficients, the stores were ranked, from best to worst-performing buildings. This information was then used to inform investment decisions.

3.3.2 Multi-variate Linear Regression (MLR)

Analogous to Linear regression, multi-variate linear regression (MLR) was used to analyse the impact of weather and store area on the building energy performance. MLR was performed as it was predicted to provide a better fit in comparison to the linear regression. The multi-linear regression was achieved using the two equations below:

$$G = A \times SA_{incl.} + B \times HDD \times SA_{incl.} \quad (5)$$

$$E = C \times SA_{excl.} + D \times CDD \times SA_{excl.} \quad (6)$$

G is the gas annual consumption, E is the electricity annual consumption and SA is the store area. In the equation above, A, B, C and D are regression coefficients that can be determined by performing the multi-variate regression across the stores for the different heating and cooling technologies. This regression was performed by using the data analysis tool in Microsoft Excel. For example, to calculate A and B, G was set to be the dependent variable while $SA_{incl.}$ and $HDD \times SA_{incl.}$ were considered as the independent variables.

The regression coefficients were then used to calculate the baseline gas/electricity annual consumption for each store by using equation 5 and 6. Baseline consumption represents the gas/electricity consumption of the stores performing at the average level. Stores could then be ranked based on how far they were performing from the baseline consumption. This was done using the equation below:

$$\theta = \frac{[Actual\ Cons. - Baseline\ Cons.]}{Actual\ Cons.} \quad (7)$$

θ represents the deviation of the actual gas/electricity consumption of the stores from the baseline consumption. A more negative value of θ yields a better performance of the store. This ranking system will then be used to execute data-driven investment decisions.

3.4 Investment Strategy: Cost and Carbon Savings

Financial analysis was performed under two investment scenarios to determine the potential cost-savings that Sainsbury's will be able to attain. Each of the scenarios were run using datasets from both linear regression and multi-linear regression separately. Carbon emission analysis for the two scenarios was also conducted to examine the reduction in carbon emissions from optimising the store energy performance.

3.4.1 Scenario 1: Optimising all stores

The impact of optimising all the stores was quantified by calculating the electricity consumption, if all the stores were to perform as efficiently as the stores at the 75th percentile and the additional electricity consumption, if all the gas boilers were converted to Ground Source Heat Pumps (GSHP).

For linear regression, the β-coefficient was obtained for the 75th percentile store and used to calculate the optimised electricity annual consumption, as described in equation 4. For multi-linear regression, the regression coefficients C and D were calculated with simultaneous equations by taking the 75th and 80th percentile store weather and store size data. The coefficients were then used to calculate the optimised electricity annual consumption, i.e. if all stores were to perform at the 75th percentile, using equation 6.

Assuming that all the stores were to convert their gas boilers to GSHPs, the gas annual consumption would be reduced to zero. However, the operation of these heat pumps would lead to additional electricity consumption. For every 1 kWh of electricity used, 3 kWh of natural gas is required to yield the same heating power. Therefore, the additional electricity required annually is one-third of the forecasted gas annual consumption.

Total potential cost-savings could therefore be calculated by multiplying Gs_{GSHP}, $E_{Additional}$, $Es_{Optimisation}$ by the cost factors in Table 5. Gs_{GSHP}, $E_{Additional}$, $Es_{Optimisation}$ were calculated using the equation below:

$$Gs_{GSHP} = G_{Forecasted} \quad (8)$$

$$E_{Additional} = \frac{1}{3} \times G_{Forecasted} \quad (9)$$

$$Es_{Optimisation} = E_{Forecasted} - E_{Optimised} \quad (10)$$

$$CE_{reduced} = Gs_{GSHP} \times C_{factor} \quad (11)$$

Gs_{GSHP} is the gas annual consumption saved from implementing GSHP, $G_{Forecasted}$ is the forecasted gas annual consumption, $E_{Additional}$ is the additional electricity annual consumption required from GSHP, $Es_{Optimisation}$ is the electricity annual consumption saved from optimising the stores to perform at 75th percentile, $E_{Forecasted}$ is the forecasted electricity annual consumption, $E_{Optimised}$ is the optimised electricity annual consumption, if the stores were to perform at 75th percentile, $CE_{reduced}$ is the annual reduction in carbon emission and C_{factor} is the carbon factor, taken as 0.184 $kgCO_2/kWh$ [13].

To further analyse the scenario, Net Present Value (NPV) of the project was calculated by taking the investment timeline to be 5 years as requested by Sainsbury's. The total NPV was calculated by summing the two NPVs below:

NPV-G: NPV of implementing GSHP to reduce gas consumption [16]

NPV-E: NPV of implementing electricity-saving technologies for all the stores to perform at the 75th percentile

$$NPV - G = \sum_{t=1}^{T} \frac{C_t}{(1+r)^t} - IC \quad (12)$$

NPV-G was calculated using the equation above where t is the time period in years, T is the total number of time period taken to be 5 years, C_t is the cost-saving at year t, r is the discount rate taken to be 6% from the United Kingdom's forecasted interest rate in the next 5 years [17], and lastly IC is the investment cost for the GSHP. The investment cost and cost-saving can be calculated using the two equations below.

$$IC = HP_{Capacity} \times HP_{Cost} \quad (13)$$

$$HP_{Capacity} = \frac{G}{Hours\ in\ a\ year} \quad (14)$$

$$C_t = C_{Gs} + C_{CE} + C_{E\ Add} \quad (15)$$

$HP_{Capacity}$ is the heat pump capacity, HP_{Cost} is the heat pump cost is taken to be £$633/kW$ [18], $Hours\ in\ a\ year$ is taken as the total number of hours in a year assuming that heating is required all year round, C_{Gs} is the annual gas cost-saving from GSHP, C_{CE} is the annual carbon emission cost-saving from GSHP and $C_{E\ Add}$ is the additional cost of electricity from GSHP.

$$NPV - E = AC_{Energy} \times E_{Saved} \quad (16)$$

NPV-E was calculated using the equation above where AC_{Energy} is the average energy abatement cost in £$/kWh$ derived from multiplying carbon abatement cost in £$/kgCO_2$ with the carbon factor. E_{Saved} is the total electricity saved from implementing electricity-saving technologies. These technologies are listed in Table 6.

The total NPV can therefore be calculated by summing NPV-G and NPV-E. If the NPV is positive, the investment should logically increase Sainsbury's earnings, making the investment attractive.

Normalization was performed to compare the two scenarios, where the Total NPV per store was calculated using the equation below.

$$\frac{Total\ NPV}{per\ store} = \frac{NPV - G}{No.of\ Stores} + \frac{NPV - E}{No.of\ Stores} \qquad (17)$$

3.4.2 Scenario 2: Optimising worst 5 stores in gas and electricity consumption

For Scenario 2, the same methodology was applied to attain the optimised electricity annual consumption based on the 75th percentile store and additional GSHP electricity consumption. However, instead of performing the optimisation to all the stores, this scenario only implemented the improvements on the worst 5 stores in terms of gas consumption and worst 5 stores in terms of electricity consumption. The reduction in carbon emissions was obtained following the same approach as in Scenario 1. This scenario was predicted to be more ideal in the short term as it required a smaller investment, yet still leading to substantial results.

4 Results

4.1 Electricity and Gas Demand Forecast - Output

Figures 1 and 2 show the plot of the linear regression performed to forecast the energy demand for the next 5 years.

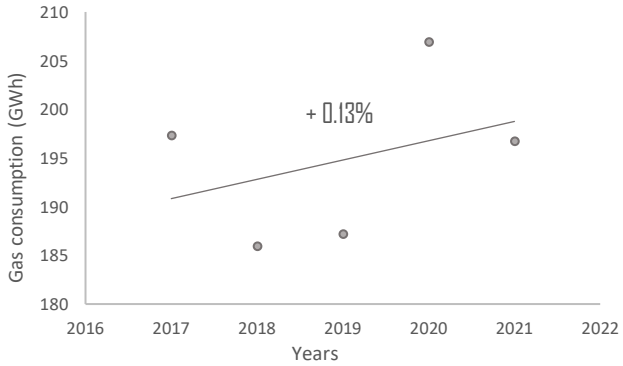


Figure 1: Forecast of Gas Annual Consumption for the next 5 years

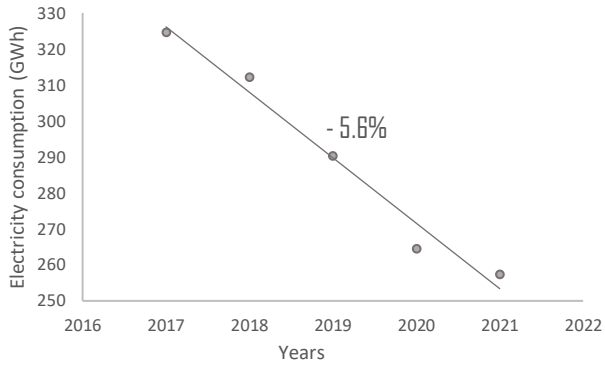


Figure 2: Forecast of Electricity Annual Consumption in the next 5 years

While the electricity consumption has been clearly declining throughout the last 5 years, from 2017 to 2021, the consumption of natural gas has remained relatively constant, as highlighted by the linear trendline. Assuming a constant y-o-y growth, the demand in electricity was forecasted to follow a constant decline of 5.6% on average, while the demand in gas would be relatively stable in the next few years at a constant growth rate of 0.13% on average.

4.2 Building Energy Performance - Output

Understanding the models' outputs is essential to this research since it allows an effective benchmarking of buildings. The outcomes of this analysis, applied to different categories of buildings, led to the ranking of supermarkets by performance efficiency.

4.2.1 Linear Regression for all the stores

To assess the impact of the two variables considered, i.e the area of each building as well as the CDD and HDD, on the buildings' energy demand, a variation of both natural gas and electricity consumption was analysed as a function of these two components, as shown in Figure 3.

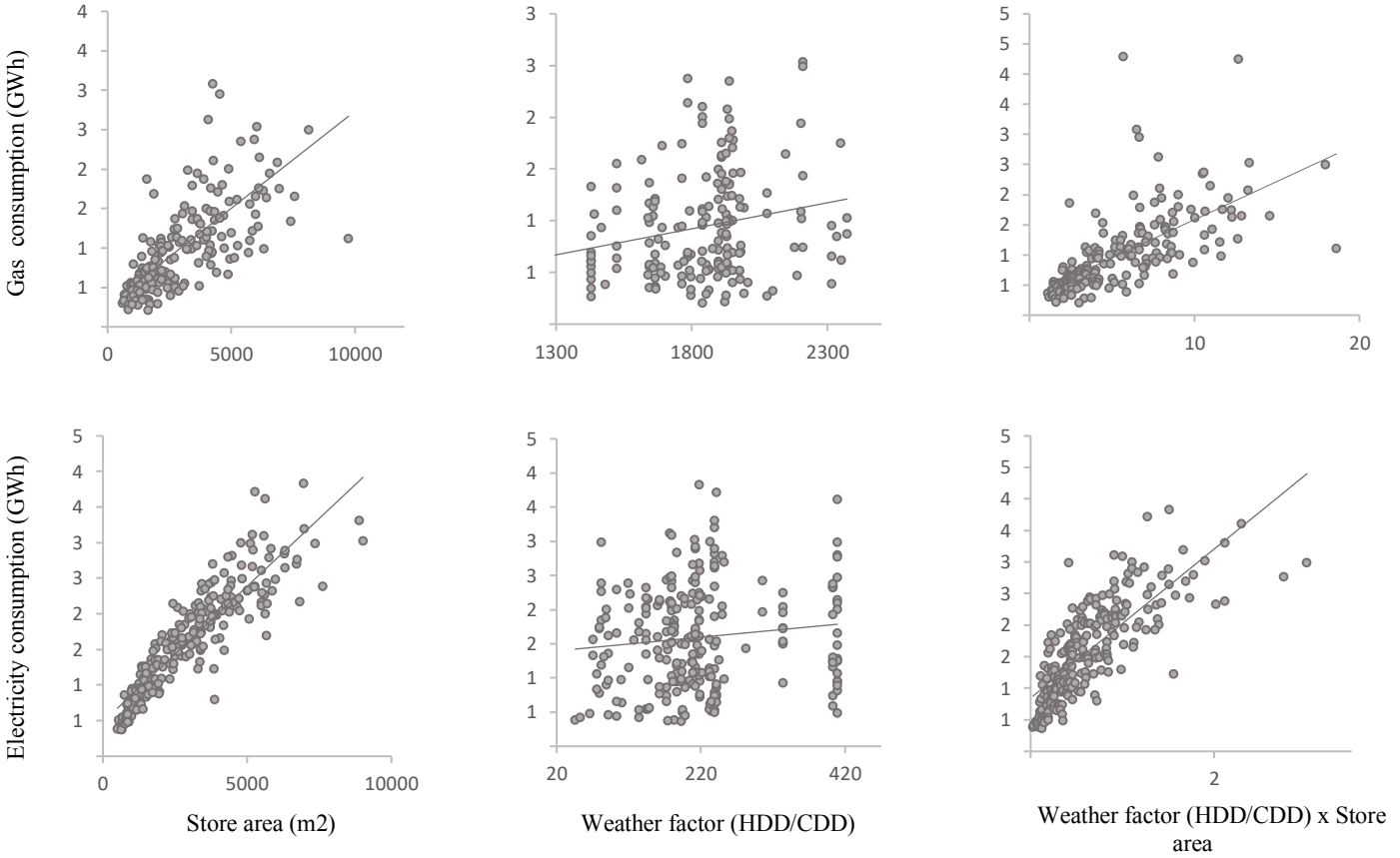


Figure 3: Linear plots of 2019 Gas and Electricity consumption against Store area, Weather factor, and Weather Factor x Store Area

A few data points, considered as anomalies due to abnormally high energy consumption, were removed from these graphs since it altered the overall fit of the linear trendline. Although the size of the store exhibited a relatively clear linear relationship with the overall gas and electricity consumption, the heating and cooling degree day indexes did not show any clear correlation with the energy consumption as seen in Figure 3. This can also be seen by the coefficient of determination, R^2, which not only indicate the goodness of fit, but also can be interpreted as the amount of variation of the dependent variable explained by the regression equation [8]. Table 1 summarizes the R-square term for the three correlations plotted for both gas and electricity.

Table 1: R-square values describing the plot of Gas and Electricity Consumption as functions of Store area, Weather factor and Weather factor x Store area

R-square values 2019	
Gas consumption VS Store Area	0.4697
Gas consumption VS HDD	0.0486
Gas consumption VS HDD x Store Area	0.495
Electricity consumption VS Store Area	0.8074
Electricity consumption VS CDD	0.0147
Electricity consumption VS CDD x Store Area	0.5984

It can be observed that most of the values are reaching 50%, which means that half of the variance in the gas and electricity consumption is explained by the linear model. This is due to some variability in the data that cannot be accounted for by this model. The R-square coefficient was therefore found to be insignificant, which indicates that linear regression was not appropriate to this case.

To further understand the impact of store size and weather factors on energy demand, a sensitivity analysis was attempted on SobolGSA, a tool for global sensitivity analysis. However, no sensible output was produced due to the data input being erroneous.

4.2.2 Linear Regression for the three categories of stores

As mentioned previously, most of the supermarket buildings selected comprise a gas boiler for heating. The majority of the remaining stores lacked ground source heat pumps (GSHPs) and only a few of them would possess this electricity – demanding technology.

The linear regression of each category of stores for both natural gas and electricity was computed.

Figure 4: Variation of α and β coefficients for the different categories throughout the years, from 2017 until 2020

Figure 4 displays the variation of α and β coefficients throughout the years, from 2017 until 2020. While the gas consumption coefficient, exclusively applied to stores that possess gas boilers, shows a relatively stagnant value throughout the four years, the electricity demand coefficient for both stores with and without GSHPs displays a significant change, principally a decrease of 4.4% from 2017 to 2018 for the stores lacking a GSHP system, and a 6.6% decrease for those owning one. This means that these stores were performing better from 2018 since their electricity consumption per area and CDD was decreasing. This could be due to the addition of more efficient technologies such as LED lighting, reducing the electricity consumption by around 58% for the stores in 2018 [19].

The main goal of this linear regression analysis was to suggest a ranking of the top 5 and worst 5 performing stores for the different categories. This can be viewed in the Appendix, Tables 3 and 4.

While the Sainsbury's supermarket from Slough revealed to be the least efficient building regarding its gas consumption from gas boilers throughout the four years, Chichester's supermarket seemed to be the best-performing building with a very low α – coefficient of 0.03% on average. On the other hand, Newton Stewart's supermarket ranked as the worst-performing store in terms of electricity demand, whereas Leatherhead proved to be the most efficient one in 2020. An interesting observation to make is that Slough's Sainsbury's store ranked first in the best-performing supermarkets classification for two consecutive years, from 2017 to 2019.

Stores that included a heat pump had already an optimized gas consumption, thus only their electricity demand was considered. From the initial number of stores considered for this research, only 7 of them used a GSHP. Therefore, it was not necessary to evaluate the top and bottom 5 stores, as it was done for the other two categories.

4.2.3 Multi-variate linear regression

The second step was to look at the multi-variate linear regression, which was applied directly to the three categories of buildings.

The ranking outcome of this regression technique required a different approach since MLR was based on the calculation of two regression coefficients for both natural gas and electricity consumption.

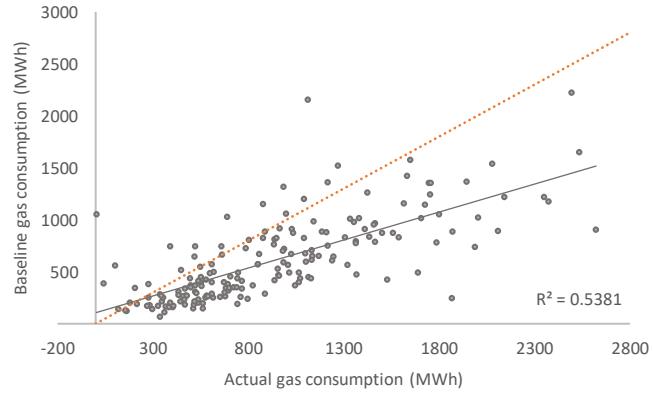


Figure 5: 2019 Baseline gas annual consumption against Actual gas annual consumption

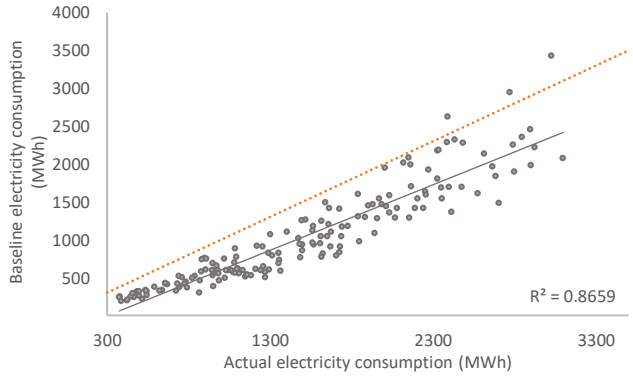


Figure 6: 2019 Baseline electricity annual consumption against Actual electricity annual consumption

Figures 5 and 6 exhibit the predicted gas and electricity consumption of the multi-variate regression model as a function of the actual energy consumption in

2019. The correlation between the electricity consumption and the area of the building including the HDD is better described by this regression model than the gas consumption. This can be easily noticed by the R-square value of 0.87 for electricity, which is far greater than the R-square of the gas consumption graph, being around 0.54. This indicates that MLR might not be fully suitable to accurately benchmark the supermarkets that own gas boilers, and other methods should be considered.

These two plots only account for the data from 2019, the rest of the analysis for the other three years can be found in the Appendix, Figures 7, 8, and 9.

Tables 3 and 4 in the Appendix display the ranking of the best and worst-performing stores. The MLR analysis revealed a similar ranking order to the LR analysis, where Sainsbury's supermarkets owning a gas boiler in the region of Slough remained to be the worst-performing buildings for the last two consecutive years. On the other hand, Chichester was ranked as the region that had the best-performing stores in the same category. An interesting observation to make is that there was no overlap between the classifications of the worst-performing stores without GSHPs generated by the two regression methods. MLR showed that buildings from the Guildford region had the worst electricity performance, while the first place was attributed to the region of Leatherhead in 2021. This distinction can be explained by the effectiveness of the regression fit to the data. MLR clearly exhibits a better fit than the LR, thus the ranking that this regression method generated is more accurate.

4.3 Investment Strategy

Two different investment strategy scenarios were compared, as discussed in sections 3.4.1 and 3.4.2. In Scenario 1, optimisation was performed on all stores while Scenario 2 described the same optimisation strategy, but applied to the 5 worst-performing stores in terms of electrical and gas annual consumption.

The results obtained from the two scenarios for both linear regression and multi-variate regression models are summarized in Table 2, in terms of NPV from implementing GSHP, NPV from implementing the optimisation in accordance with the 75th percentile store and the Total reduction in carbon emitted in 5 years. Analysis on the results from NPV calculations in exhaustive detail can be found in the Appendix, Tables 7, 8, 9, 10.

The outcome of NPV-G was negative for all scenarios, and this implied that the investment on GSHP was not attractive from a financial standpoint as it would result in a negative impact on Sainsbury's overall earnings. However, it should be noted that the total reduction in carbon emissions in 5 years for all scenarios was relatively substantial. Scenario 1 would allow Sainsbury's to reach their net-zero carbon emission target. In this scenario, all gas boilers were converted to GSHPs, which used renewable energy source to produce heating power. In Scenario 2, the total reduction in carbon emitted in 5 years represented 14% of that from Scenario 1. This is due to the reduction in the number of stores considered in Scenario 2, accounting for only 3% of the total number of stores in Scenario 1.

NPV-E for all scenarios yielded positive values. Thus, the investment made on optimising the building electricity usage by implementing several technologies such as PVs or night blinds, as listed in Table 6, would be beneficial for Sainsbury's. The positive value of the Total NPV for all the scenarios is due to the absolute value of NPV-E being greater than that of NPV-G. Therefore, investing in either Scenarios 1 or 2 will impact positively on Sainsbury's earnings. Lastly, to compare the different scenarios and to evaluate the impact of the overall investment per store, the Total NPV per store was determined. Scenario 2 yielded a higher Total NPV per store, suggesting that this scenario was more optimal than Scenario 1. Scenario 2 should therefore be considered for Sainsbury's investment strategy.

Table 2: Summary of the cost and carbon analysis of the two scenarios in terms of Net Present Value per store and Reduction in Carbon Emitted in 5 years.

	Units	Scenario 1 - LR	Scenario 1 - MLR	Scenario 2 - LR	Scenario 2 - MLR
Total GSHP Capacity	(kW)	22,200	33,300	733	653
Investment Cost - GSHP	(£ x 1000)	14,000	21,100	464	413
Total Reduction in Carbon Emitted in 5 years	(tCO2)	173,000	173,000	5,710	5,090
NPV - G	(£ x 1000)	-1,487	-1,487	-223	-198
Total Electricity Saved in 5 years	(MWh)	147,000	105,000	20,500	14,300

Average Energy Abatement Cost	(£/kWh)	-0.271	-0.271	-0.271	-0.271
NPV - E	(£ x 1000)	39,900	28,600	5,570	3,870
Total NPV	(£ x 1000)	38,400	27,100	5,340	3,670
Total NPV/Store	(£ x 1000)	221.2	156.0	1,068.4	734.9

5 Discussion

In order to understand the energy trends that would affect Sainsbury's supermarkets in the next few years, linear regression of the energy demand was computed. Comparing the electricity and natural gas consumption forecasts, the observed decline in the electricity demand could be explained by the implementation of energy-saving technologies such as LED lights. Sustainable operations, for instance dimming, also reduce the lighting energy consumption by an average of 70% [20]. Regarding the prediction of natural gas consumption, the results suggested an overall stability in this energy demand. This was assumed to be consistent for the next 5 years at least.

The impact of the two independent variables, i.e. the area of the store and the two weather factors, on the energy usage was carefully analysed. It was concluded that the area of the store had a greater influence on the energy demand than the HDD and the CDD. The previous results suggest that the size of the building would have a linear relationship with the electricity consumption, which means that the electricity demand increases proportionally with the area of the supermarket. On the other hand, the absence of a clear correlation between the two weather factors and the energy demand could be due to the lack of uniformity of store sizes for the same HDD or CDD value.

Linear and multi-variate linear regressions were then performed on the three categories of stores introduced earlier in this paper. From a general perspective, MLR has generated a better fit of the observed energy data, leading to a more accurate ranking of the stores.

Two investment strategy scenarios were proposed: one that would be implemented in all stores and the other focusing on the worst-performing supermarkets. While it was decided to optimize both electricity and natural gas demands, the choice between the two scenarios was made based on the impact of this optimization on these two cases. Looking at the total NPV per store, Scenario 2 was concluded to be the most attractive from a financial point of view, and therefore a more suitable investment strategy for Sainsbury's.

While this research focused mainly on providing an investment strategy as part of Sainsbury's sustainability roadmap, further considerations affecting the energy demand should be considered. Firstly, the energy consumption could be predicted more effectively using other regression methods such as the decision tree (DT) model, as mentioned in [13]. Furthermore, additional investigations should be made on different weather factors, including humidity, which was not evaluated in this work. As part of the investment strategy, the quality of the building should be considered, comprising its age and the insulation technology used, as well as the quality of the refrigeration system [21]. Finally, this energy benchmark analysis was developed only for a certain number and type of buildings. Further research on different building categories should be therefore considered. Lacking data should also be recovered in order to obtain more accurate results.

6 Conclusion

This paper investigated the factors affecting building energy usage and developed several net-zero investment scenarios for Sainsbury, the second major food retailer in the UK. In this work, an investment scenario, Scenario 2, aiming to implement electricity and heat-saving technologies on the worst 5 performing stores, has been chosen to be the most ideal. This scenario was determined to have a higher total NPV per store, hence yielding a greater financial impact on Sainsbury's. This was achieved by developing linear and multi-variate linear regression models, which were then integrated into a carbon and cost analysis model comparing the two scenarios.

The LR and MLR models were performed to analyse the impact of weather and store area on building energy consumption and were used to rank the stores based on the output of these models. The MLR model was concluded to be more effective in predicting the energy consumption as it provided a better fit. This model was applied to different categories of buildings, leading to the ranking of stores by performance efficiency.

In the carbon and cost analysis, NPV was derived from the financial impact of optimising the building energy usage by implementing electricity-saving technologies, such as PVs and biofuels, but also gas-saving technologies, including GSHPs. The total NPV per store for Scenario 2 – MLR was found to be £734,900, almost five times greater than that of Scenario 1 – MLR. Scenario 2 – MLR would also reduce carbon emissions by 5,090 tonnes of carbon dioxide in 5 years. This decrease aligns with Sainsbury's net-zero target.

Though this research has touched upon numerous factors that would affect the net-zero investment strategy, there is still a great research space yet to be

explored. The impact of other factors influencing the energy demand model, such as humidity and quality of the building environment, should be analysed. Since the implementation of GSHPs requires a high investment cost, other gas-saving technologies should be examined. Lastly, this work was conducted with a limited number and type of buildings, such as only supermarkets with certain technologies were considered. Therefore, research on other building types and characteristics can be considered to further develop a more robust investment strategy.

7 References

[1] Chiara Delmastro ,Tanguy De Bienassis ,Timothy Goodson, Kevin Lane, Jean-Baptiste Le Marois, Rafael Martinez-Gordon, Martin Husek. (2022) Buildings. https://www.iea.org/reports/buildings .

[2] Jim Skea et al. (2022) Climate Change (2022): Mitigation of Climate Change. https://www.ipcc.ch/report/ar6/wg3/ .

[3] M. Hart, W. Austin, S. Acha, N. LeBrun, C.N Markides & N. Shah. (2020) A roadmap investment strategy to reduce carbon intensive refrigerants in the food retail industry. Journal of Cleaner Production.

[4] Fraser McKevitt. (2022) Inflation starting to drive grocery behaviour as pandemic loosens hold on Brits.https://www.kantar.com/inspiration/inflation/2022-wp-inflation-starting-to-drive-grocery-behaviour-as-pandemic-loosens-hold-on-brits.

[5] M.S. Spyrou, K. Shanks, M. Cook, J. Pitcher & R. Lee. (2014) An empirical study of ele ctricity and gas demand drivers in large food retail buildings of a national organisation. Energy and Buildings. 68 172-182. https://doi.org/10.1016/j.enbuild.2013.09.015.

[6] S.A. Kalogirou & M. Bojic. (2000) Artificial neural networks for the prediction of the energy consumption of a passive solar building. Energy. 25 (5), 479-491. https://doi.org/10.1016/S0360-5442(99)00086-9.

[7] G. Mavromatidis, S. Acha & N. Shah. (2013a) Diagnostic tools of energy performance for supermarkets using Artificial Neural Network algorithms. Energy and Buildings. 62 304-314. https://doi.org/10.1016/j.enbuild.2013.03.020.

[8] G.K.F. Tso & K.W.Yau. (2007) Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. Energy. 32 (9), 1761-1768. https://doi.org/10.1016/j.energy.2006.11.010.

[9] X. Gao & A. Malkawi. (2014) A new methodology for building energy performance benchmarking: An approach based on intelligent clustering algorithm. Energy and Buildings. 84 607-616.

[10] W. Chung, Y.V. Hui & M. Lam. (2006) Benchmarking the energy efficiency of commercial buildings. Applied Energy. 83 (1), 1-14. https://doi.org/10.1016/j.apenergy.2004.11.003.

[11] Z. Li, Y. Han & P.Xu. (2014) Methods for benchmarking building energy consumption against its past or intended performance: An overview. Applied Energy. 124 325-334. https://doi.org/10.1016/j.apenergy.2014.03.020.

[12] The Natural Environment Research Council's Data Repository for Atmospheric Science and Earth Observation. United Kingdom, CEDA.

[13] M.J.S. Gulliford, R.H.Orlebar, M.H.Bird, S. Acha & N.Shah. (2022) Developing a dynamic carbon benchmarking method for large building property estates. Energy and Buildings. 256 111683. https://doi.org/10.1016/j.enbuild.2021.111683.

[14] M. Yalcintas. (2006) An energy benchmarking model based on artificial neural network method with a case example for tropical climates. International Journal of Energy Research. 30 (14), 1158-1174. https://doi.org/10.1002/er.1212.

[15] V. Caritte, S. Acha, N. Shah & N.Ekins-Daukes. (2012) Developing a low carbon roadmap for a food retail chain in the UK: targets & challenges. Imperial College London.

[16] How to Calculate Net Present Value (NPV). (2022) https://www.investopedia.com/ask/answers/032615/what-formula-calculating-net-present-value-npv.asp#toc-npv-formula.

[17] United Kingdom Interest Rate. (2022) https://tradingeconomics.com/united-kingdom/interest-rate .

[18] E.J.S Escriva, M. Hart, S. Acha, V.S. Frances, N. Shah, C.N. Markides (2021) Techno-economic evaluation of integrated energy systems for heat recovery applications in food retail buildings.

[19] (2018) Sustainability Report 2019. United Kingdom, Sainsbury's.

[20] P. Dunne. (2022) Why Sainsbury's is investing in green technology start-ups. https://www.thegrocer.co.uk/sainsburys/why-sainsburys-is-investing-in-green-technology-startups/669814.article .

[21] A. Mota-Babiloni, J. Navarro-Esbri, A. Barragan-Cervera, F. Moles, B. Peris & G. Verdu. (2015) Commercial refrigeration – An overview of current status. International Journal of Refrigeration. 57 186-196. https://doi.org/10.1016/j.ijrefrig.2015.04.013.

[22] D.R. Anderson, D.J. Sweeney, T.A. Williams, J.D. Camm, J.J. Cochran, M.J. Fry & J.W. Ohlmann. (2019) Essentials of Modern Business Statistics with Microsoft Excel . 8th edition.

[23] Lariviere & G. Lafrance. (1999) Modelling the electricity consumption of cities: effect of urban density. Energy Economics. 21 (1), 53-66. https://doi.org/10.1016/S0140-9883(98)00007-3.

[24] M. Bourdeau, X.Q. Zhai, E. Nefzaoui, X. Guo & P. Chatellier. (2019) Modeling and forecasting building energy consumption: A review of data-driven techniques. Sustainable Cities and Society. 48 https://doi.org/10.1016/j.scs.2019.101533.

The importance of international support for the sustainable development of Zambia's power sector

Giulio Fiorani and Maxime Van Eyseren
Department of Chemical Engineering, Imperial College London, U.K.

Zambia's compliance to the Paris Agreement aligns with its desire to meet its Nationally Determined Contribution targets of carbon dioxide emission reduction. Equally, through its Integrated Resource Plan, Zambia aims to ensure electricity access across the country, currently at 45% of the population. This paper developed long-term power system expansions until 2070 and flexibility assessments on Zambia's power grid in 2030, using the OSeMOSYS and FlexTool modelling software. A scenario-based approach was investigated. The Business as Usual (BAU) scenario replicated the inclusion of all current and scheduled technologies and was not constrained to carbon dioxide emissions. Scenarios 1 and 2 aligned with Zambia's NDC targets and both supported net zero by 2060. Scenario 1 (SC1) was committed to reducing greenhouse gas emissions by 25% in 2030 compared to 2010 and assumed limited international support. On the other hand, scenario 2 (SC2) assumed substantial international support and emission reductions of 47%. It could be observed that a considerable increase in solar capacity is necessary for Zambia, with most other VRE generation sources at maximum capacity. Nuclear was also present in the energy mix in 2070 because of the ever-increasing demand particularly at times where solar generation is low. The flexibility assessments of SC1 and SC2, in 2030, highlighted the importance of international support and identified the need for investments in transmissions and battery capacities for the power grid. In fact, SC2 eliminated the loss and excess load from 11.47% & 52.27% of the annual electricity demand and reduced the curtailment by 70.2% compared to SC1. Overall, to achieve its NDC targets, Zambia would need to invest in at least 44GW of VRE generation capacity by 2050, along with significant investments in storage and transmission capacity, highlighting the importance of international support.

1. Introduction

Zambia has a goal of becoming a world leader in copper production, with the aim of tripling its production to 3 million tonnes in the next ten years [1]. The mining sector is therefore crucial to the country's economy, accounting for more than three-quarters of the export earnings [2]. Energy consumption is mainly concentrated in the industrial sector because of this, accounting for more than half of the country's electricity consumed. Consequently, electricity is a key driver for the economic development of Zambia.

As of 2021, Zambia had 3.2 GW of installed electricity generation capacity, with hydro representing the vast majority, at 85% of the total capacity [3]. However, the main problem in Zambia is the access to electricity. The national average is 45% with 82% of urban and 8% of rural areas having access to power [4]. This poses an enormous challenge for Zambia's ability to increase its development.

The government of Zambia has set through their Integrated Resource Plan (IRP) a pledge to "ensure access to reliable, clean, and affordable electricity across the country at the lowest economic and financial costs consistent with local, national and regional development goals" [5]. Zambia's Nationally Determined Contribution (NDC) has also recently been updated with a pledge to reduce greenhouse gas emissions by 2030, through two scenarios: with and without international support [6].

However, as mentioned previously there are numerous challenges that impact Zambia's ability to attain these targets. The heavy reliance on storage-less hydro energy has proved problematic with fluctuations in rainfall patterns because of climate change which has, in turn, provoked deficits in hydropower generation. In addition, the high population growth rate of 3% [7] poses an issue with meeting an increasing demand by a grid that currently only reaches 45% of the country. Transmission and distribution losses are equally substantial due to the high temperatures and theft. In fact, theft directly from the grid and vandalism highlight the need for Zambia to increase security and control of its energy grid. Recent changes have been implemented to overcome these issues, such as the shift from using copper cables to aluminium [8], but more drastic measures are required.

Equally, since Zambia's electricity consumption comes principally from its mining activity, the country's goal of becoming a world leader will cause a sustained increase in the electricity demand, posing another challenge for the power grid.

There is also a hurdle with international investments, a key driver in the IRP and NDC targets. There was a reduction of 42% in investments projects in Zambia in 2021 compared to 2020 [9], largely due to the pandemic. International support depends heavily on the mining of copper, with large investments from Canada, Australia, United Kingdom, China, and the United States. Moreover, large infrastructures and other projects have been funded entirely by Chinese companies, with China owning 69 percent of the construction industry in Zambia [10]. Consequently, Zambia is heavily dependent on international support for its development, particularly in advancing its infrastructure such as the road network, railway, and the construction of power plants.

This highlights Zambia's potential challenge of facing insufficient international support, which could result in negative consequences for the improvements of their

power grid and affect their ability to meet their development and climate goals.

This study, therefore, has the aim of assessing the possibility of Zambia to achieve its objectives as set out by the IRP and NDC, within various modelling scenarios. The varied influences of international support, along with the limits on carbon emissions, were driving factors in the development of the scenarios, with the aim of demonstrating the importance of such investments on developing countries and their power grid.

2. Background

The challenges for the development of Zambia's power grid, mentioned in the previous section, drive the need for accurate power system planning and are the reasons for the ongoing research in this area. As a result, there are numerous studies investigating the expansion of the power grid in Zambia, but there is a clear gap regarding the combination of long-term power system expansion and flexibility assessments. For example, numerous papers investigated the feasibility of integrating VRE in Zambia's power grid and the continuous power system expansion until 2030 such as the IRP [5] or the report on integration of VRE sources in Zambia by the RES4Africa Foundation and Enel Foundation [11]. However, they do not address the main challenges regarding the long-term sustained increase in electricity demand after 2030, arising from the high population growth, high mining activity, and the electrification of the transport industry, which are of crucial importance for Zambia's development goals. Equally, some studies do not cover the scope of the country, such as McPherson et al. [12], whose paper highlights the long-term regional planning in Lusaka, Zambia's capital, for variable renewable energy and electric vehicle integration.

Hence the purpose of this study was to fill in gaps in existing research and help Zambia identify both potential development pathways up to 2070 and flexibility challenges for the power grid. The development pathways were achieved with the Open-Source modelling software OSeMOSYS, which has the unique feature of developing intertemporal cost optimization pathways for long-term energy planning. To be precise, it is a linear bottom-up energy optimization software that focuses on the detailed representation of flows and technologies in the energy system, which includes their cost parameters, performance, and environmental impacts. In simpler terms, the bottom-up characteristic of this software enables the model to identify the optimal solution for meeting the defined demand while minimizing costs, completely aligning with the demand driven challenges in Zambia's power grid.

Additionally, the flexibility of this long-term energy planning was further investigated with the IRENA FlexTool, a linear optimization program. According to the International Energy Agency, the flexibility of a power system is "the ability of a power grid to reliably and cost-effectively manage the variety and uncertainty of demand and supply that VRE generation introduces across all time scales" [13], which aligns with Zambia's NDC goals, and justifies the significance of this assessment. This linear program modelling software can solve the hourly capacity expansion problem for a one-year horizon. It determines the optimal flexibility solution for the power systems by simulating technical constraints on energy balances, reserve requirements & others. The linear program solution uses GNU MathProg, with this solver ensuring greater certainty in determining a global optimum compared to integer programs, due to the convexity of the optimization.

These modelling approaches have been researched in studies that collaborated with governments ranging from national to continental analysis, such as Cyprus, Costa Rica, and Africa [14] [15] [16]. They have not yet been utilized to investigate Zambia's electricity grid, with previous research using models with different optimization characteristics. The IRP made use of Antares [5], which does not produce a development pathway but generates capacity planning at selected years, and the RES4Africa Foundation used GRARE [11], a Monte Carlo analysis, to assess the reliability of the power grid. Hence, this study provided both long-term capacity expansion and flexibility assessments with a different approach and solution to the problem, highlighting the generation of novel material in this research.

3. Methods

3.1. Overview:

OSeMOSYS and FlexTool require varied techno-economic input data to match the electricity demand of a predefined region with an energy supply mix. The inputs of data were found directly from sources readily available online or computed through different correlations and data-driven assumptions. For OSeMOSYS, projections were required until the year 2070 to develop long-term power expansion planning. The output of the power grid in 2030 was subsequently used in FlexTool for flexibility assessments. Different scenarios were developed within OSeMOSYS using different constraints on total capacity of a technology, new capacity investments, carbon emission limits and cost of new technologies. For a more detailed explanation of these modelling software, the following resources are helpful [17] [18].

3.2. Scenario development.

- **Business as usual (BAU)**

This scenario was characterized by the inclusion of all current and scheduled technologies used for electricity generation in Zambia. No emissions constraints were placed on the model.

- **Scenario 1: achieving net zero without external investments (SC1)**

A constraint on carbon emissions was added to the BAU to create SC1, which had the aim of meeting the

emissions target set out by Zambia's NDC [6]. This involved reducing GHG emissions[1]by 25% from 2010 to 2030. A linear reduction in emissions was computed to achieve net zero by 2060. Although Zambia is currently in discussion about the year by which it can reach net zero, an estimate of 2060 was used for this scenario [19].

- **Scenario 2: achieving net zero with external investments (SC2)**

This scenario had the aim of meeting the emissions target set out by Zambia's NDC with substantial international support. This involved reducing GHG emissions by 47% from 2010 to 2030 and achieving net zero by 2060. Similarly, to SC1, a linear reduction in emissions was computed to meet net zero by 2060.

3.3. Development of OSeMOSYS models

An OSeMOSYS model was developed for each scenario. To do this, the Climate Compatible Growth starter data kit for Zambia [20] a pre-made model for the Zambia electricity grid, was utilised, and modified, as detailed in sections 3.3.1. and 3.3.2.

3.3.1. Technical data input

- **Power Plants**

To remain technology neutral, a wide variety of power generation units were included in the models. For the different technologies, residual capacities and capacity factors were computed for all installed power plants and projects in development [21]. The potential of all technologies in Zambia were equally set as constraints [22] [23] [24] [25], to highlight when Zambia had maximized its capacity in these technologies.

For solar PV and CSP plants, constraints were added to prevent the software from forecasting unrealistic amounts of annually added capacity. For SC1 and BAU, a power curve was computed starting at Zambia's current level of 0.2 GW and reaching 4GW of annual investments, a level resembling developed countries like Germany and the Netherlands [26]. For SC2, due to the increase in investment in renewable energy sources because of international funding, the same approach was implemented but increased by 20%. For CSP, annual capacity increases have been less significant, so an annual limit of 100 MW of added capacity was set from 2024. This was increased by 100 MW every 10 years.

With regards to nuclear, Zambia has considered its use in the past and planned to install a 2.4GW capacity power plant by 2035 [27]. A constraint of 1 power plant of 2.4GW built every 5 years until 2050, and 2 every 5 years beyond this point was added to the model.

- **Transmission and distribution**

Losses due to transmission and distribution stood at 7% and 11 % respectively, in 2015 [28], as a results of high heat and theft. Zambia aims to reduce these losses, with plans in place to improve the electricity transmission and distribution infrastructures, as set out by the ERB [21] in 2021. For the BAU and SC1, an aim of achieving 10% total losses by 2070 (comparable to Russia and Portugal) was set. For SC2, this value was set to 8% (comparable to the UK and South Africa), assuming greater control measures are implemented to prevent disruptions.

- **Fuel availability, imports, and exports**

Data for the reserves of coal present in the country was gathered and included in the models [29]. Raw materials, such as oil and gas, are not present in Zambia, so would only be imported.

Electricity imports and exports were obtained from the latest ERB report, with exact values up until 2021. Beyond that point, projections were made by considering Zambia's long-term contracts and trade relations with countries in SAPP [29].

- **Electricity demand and projections**

The ERB provided exact values for the electricity demand in Zambia until 2021, and projections of subsequent years were performed for each scenario.
Future demand was divided by sector and determined by using the GDP, population growth and income per capita data [31] [32] [7].

The **industrial** projections were correlated with the income per capita, GDP growth, investments from international support and the production targets of copper mines, representing 87% of the industrial activity in Zambia. This resulted in an increase by 124% and 160% in 2030; and 1200% and 1400% in 2070 for scenarios 1 and 2 respectively from 2015. The high difference for the scenarios is due to the increased investments from international support towards the country's development, hence achieving higher production targets and a greater demand.
The projections for the **commercial** and **residential** sector considered the GDP, income per capita and population growth. The increase in access to the grid was also included, with 100% of the population expected to have access to electricity in 2035 and 2030 for scenario 1 and 2 respectively. This resulted in an increase of 189% and 363% in 2030 and of 1240% and 1290% by 2070 for scenario 1 and 2 respectively.
The **transport** energy sector was projected in the short term until 2026 using the population growth and income per capita, as from their IRP targets, Zambia is aiming to maintain a supply of petroleum products for the transport industry until 2030. The long-term projections, after 2026, were calculated using Zambia's sustainable development goals (phase out fossil fuel cars by 2050), their near-future investments in EV charging ports and Zambia's agreement with DRC for the manufacturing of EV batteries [33] [34]. Consequently, the increase in demand in this sector was considerably different for each scenario as international support resulted in earlier implementation of EV infrastructures. This increased demand by 63% (SC1), 204% (SC2) by 2030 and a factor of 180 (SC1) and 300 (SC2) in 2050.

[1] In OSeMOSYS it was assumed that all GHG emissions were from carbon dioxide emissions.

This resulted in an average distribution for the electricity demand of 58.6%, 31.5% and 9.9% between the industrial, residential, and commercial, and transport sectors respectively, aligning with projections from the starter data kit and population projections [20] [7].

3.3.2. Economic data

- **Technology costs**

For SC2, due to the increase in investments in renewables, the capital cost and fixed costs were reduced by 40% and 25%. This aligned with the Announced Pledges Scenario (APS) from the World Energy Outlook 2022 report [35], which adjusted renewable energy technologies costs to meet the government targets.

- **Fuel Costs**

The cost (in $/GJ) for extraction and imports of coal, crude oil and natural gas were added to the models. Forecasts and recent volatility experienced in 2022 were considered due the impact on commodity prices [36].

3.4. FlexTool

Flexibility assessments were computed on the electricity grid in 2030 for both scenarios 1 & 2 using the IRENA FlexTool software. The year 2030 was analysed as it aligns with the NDC targets, ensures greater accuracy for the projections, and gives the ability to identify near-future challenges in the power grid. The model was set in the dispatch operational mode, linearly optimising the hourly ability of the power system to have sufficient flexibility over a one-year horizon. IRENA FlexTool uses four fundamental sets: Grid (1), Node (2), Unit (3) & Time (4). An energy grid is a product in the power system and within this study only the electricity grid was investigated.

3.4.1. Nodes

The nodes are used to separate the country into regions and allocate their generation capacity, demand, transmission, and reserve requirements. As shown in Fig.1, Zambia was split into 4 different nodes (A, B, C & D) and the relative electricity demand (including imports and exports) used in OSeMOSYS was distributed within these nodes for each demand sector, shown in Table 1, where the numbers in the first column represents percentages.

Figure 1: Representation of Zambia separated into four nodes, where A and B represent urban areas and C and D rural areas [37].

Table 1: Distribution of the Electricity per demand sector for each node, and with the explanation.

Distribution	Explanation
Industry: A:87, B:2, C:5.5, D:5.5	Mining industry accounts for 87% of industrial activity, located at node A, the remainder is agricultural activity at node C&D
Commercial: A: 40, B:40, C:10, D:10	Correlated with urbanisation and commercial activity of each node
Residential: A:29, B:30, C:14, D:27	Correlated with the current population distribution, percentage of the population having access to the grid, and urbanisation from C&D to A&B
Transport: A:45, B:35, C:10, D:10	Aligned with Zambia's EV development plan and urbanisation

3.4.2. Transmission of Electricity between Nodes

The nodes required information on their transmission capacities, illustrated in Fig.1. These were evaluated at different capacity levels for each scenario. Scenario 1 used the existing transmission capacities, and the development expansion plans up to 2030 reported by ZESCO (state-owned power company in Zambia) [38], totalling 5016 MW. For Scenario 2, additional transmission capacities were incorporated from international investments[39] (total of 12949 MW). The transmission and distribution losses between these nodes were also determined for each scenario using the values from OSeMOSYS.

3.4.3. Imports and Exports of Electricity within South African Power Pool

As Zambia is part of the Southern African Power Pool (SAPP), imports and exports of electricity are present at each node. The total imports and exports OSeMOSYS were distributed between the nodes using the current distribution in 2022, long-term contracts and trade relations with neighbouring countries such as their DRC power agreement with Congo [40]. It also aligned with the expansion of transmission capacities.

3.4.4. Reserves

Both static and dynamic reserves were set as constraints in the model due to the high share of VRE. The static reserves were set at 15% of the electricity demand as modelled in the OSeMOSYS program. They were distributed between nodes by correlating them to the ratio of generation of electricity to the demand of electricity at each node. The dynamic reserves were set at a reserve increase ratio of 0.1 for all VRE generation units, supported by the following study [41].

3.4.5. Minimum Inertia Limit

A minimum inertia limit[2] was incorporated to ensure that there is sufficient energy stored in the large rotating generators and motors. This is to prevent a loss of load when large power plants fail and maintain the predefined RoCoF (rate of change of frequency in the system). The computation of this value required Zambia's System Frequency of 50 Hz [42], the worst-case size of the largest credible multiple contingency which corresponds to the Kafue Gorge Hydro Plant of 1740 MW, and an additional 10% was added to ensure sufficient inertia is present. Zambia does not currently have a predefined value for the RoCoF, but it was deemed reasonable to assume the same value as South Africa, 1.5Hz/s, as they are both part of SAPP, resulting in a limit of 31900 MW.s.

3.4.6. Unit capacity distribution

The third fundamental set is the unit set, representing the generation and storage of energy sources. The following information was required for the input options of each unit: the fuel (price and emissions) or the use of a capacity factor profile. Also, the inertia constants, storage capacity, ramp up/down[3], costs, efficiency and max reserves were required for each unit type, determined from various resources [21] [43] [44]. Then the generation capacity of each unit was allocated to their corresponding nodes. As this study was evaluated for 2030, OSeMOSYS results were essential to determine the total generation capacity for each energy source in 2030. The existing capacity in 2022 was distributed according to its known respective locations, and the additional capacity until 2030 was distributed by correlating it with the total technology potential within each node, considering the localisation of near-future investments [23] [24] [45].

3.4.7. Projection for The Storage Capacity Expansion

The OSeMOSYS tool did not compute battery capacities within the model. An estimation was required to calculate the additional battery capacity in 2030 for each scenario. This was achieved by analysing the development projects in Zambia. As all the investments were achieved through international support [46], it seemed reasonable to allocate 25% of SC2's capacity to SC1. The total battery capacity by 2030 was of 2400 MW for SC2.

3.4.8. Time series

Time series corresponding to each hour in the year were implemented for the capacity factors of VRE generation sources, such as PV and wind, as they varied throughout the day. With regards to variation in demand, a load curve for the hourly electricity demand in sub-Saharan African countries was used [47].

4. Results

To meet the decarbonisation constraints set out by the NDC, an energy supply dominated by renewables was necessary.

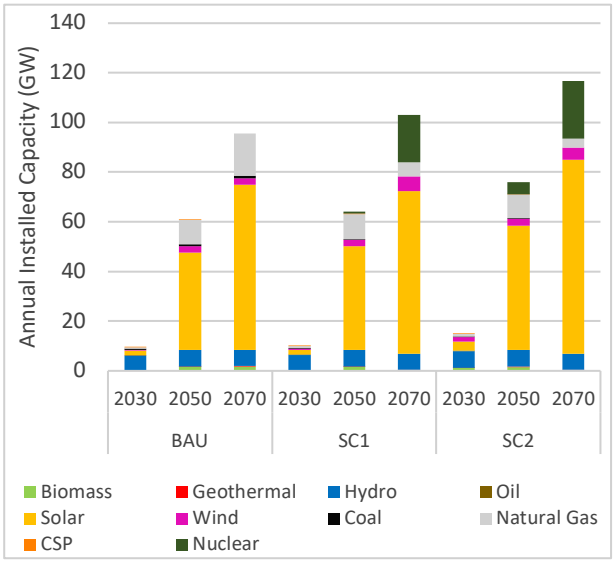


Figure 2: Comparison of the annual capacity of installed electricity generation technologies for different scenarios.

Fig.2 illustrates the installed capacity forecasts for Zambia. In the BAU scenario, the electricity demand is met through an energy supply mix largely dominated by hydro, solar and natural gas. Hydro power generation occupies most of the production until 2030, beyond which this technology is at max capacity. To meet the increasing demand there is heavy investment in solar and natural gas. Only a small amount of biomass and wind is used throughout, due to the limited potential of the country in these resources.

Both SC1 and SC2 highlight the decrease in use of fossil fuels, particularly natural gas, with lower carbon and renewable alternatives implemented instead. The electricity mix varies over time, with hydro the main source up until 2030, beyond which the capacity is at its maximum. As result, solar generation increased analogously to the BAU scenario. However, rather than utilising natural gas, nuclear energy was implemented from 2050 onwards to meet the increasing demand and reduce carbon emissions.

Due to the different carbon constraints, the uptake of fossil fuels varied between SC1 and SC2. With stricter decarbonisation, SC2 had a lower share of fossil fuels in the energy grid, causing increases in renewable investments. This can be observed in Fig.2 for the year 2050, where the installed capacity of natural gas represented 12% of the total capacity for SC2, down from 15% for SC1 and 18% for BAU. Equally, a greater demand in SC2 resulted in a greater total capacity each year when compared to the BAU and SC1.

Although Fig.2 highlights that there is natural gas capacity in 2070, this technology is no longer used for electricity generation from 2060 onwards. Therefore, SC1 and SC2 successfully achieve net zero by 2060 (see

[2] Inertia: refers to the energy stored in spinning motors and generators, giving the ability to maintain reliable generation.

[3] Ramping rate is the speed at which generators can change their output, increasing (ramp up) or decreasing (ramp down) generation.

supplementary material for electricity production graph).

A flexibility assessment on the IRENA FlexTool dispatch operational mode was successfully computed for both scenarios, as shown in Table 2.

Table 2: Flexibility Assessment of Scenario 1&2

Flexibility Criteria	SC1	SC2
VRE share (% of annual demand)	96.1	96.9
Loss of load (% of annual demand)	11.4	0.0
Ramp up constrained (% of annual demand)	0.0	0.0
Excess load (% of annual demand)	52.2	0.0
Insufficient reserves (% of reserve demand)	6.1	4.4
Insufficient inertia (% of inertia demand)	0.3	0.3
Curtailment (% of VRE gen.)[4]	20.1	6.0
Ramp down constrained (% of VRE gen.)	0.0	0.0
Peak load (MW)	3901.2	5428.9
Peak net load (MW)	-626.8	-1022.6

The comparison between the two scenarios regarding the flexibility criteria, particularly the loss and excess of load (Fig.3), and the curtailment (Fig.4) highlight the importance of international support. In fact, loss of load and excess load are completely removed in SC2.

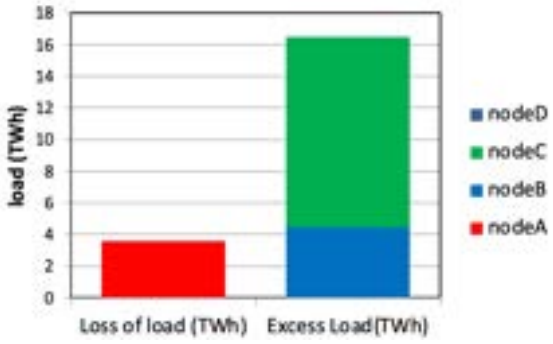


Figure 3: Distribution of the Loss and Excess Load in Scenario 1

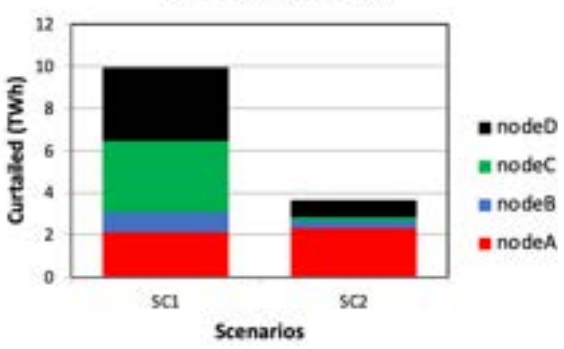


Figure 4: Distribution of the VRE Curtailment for Scenario 1&2

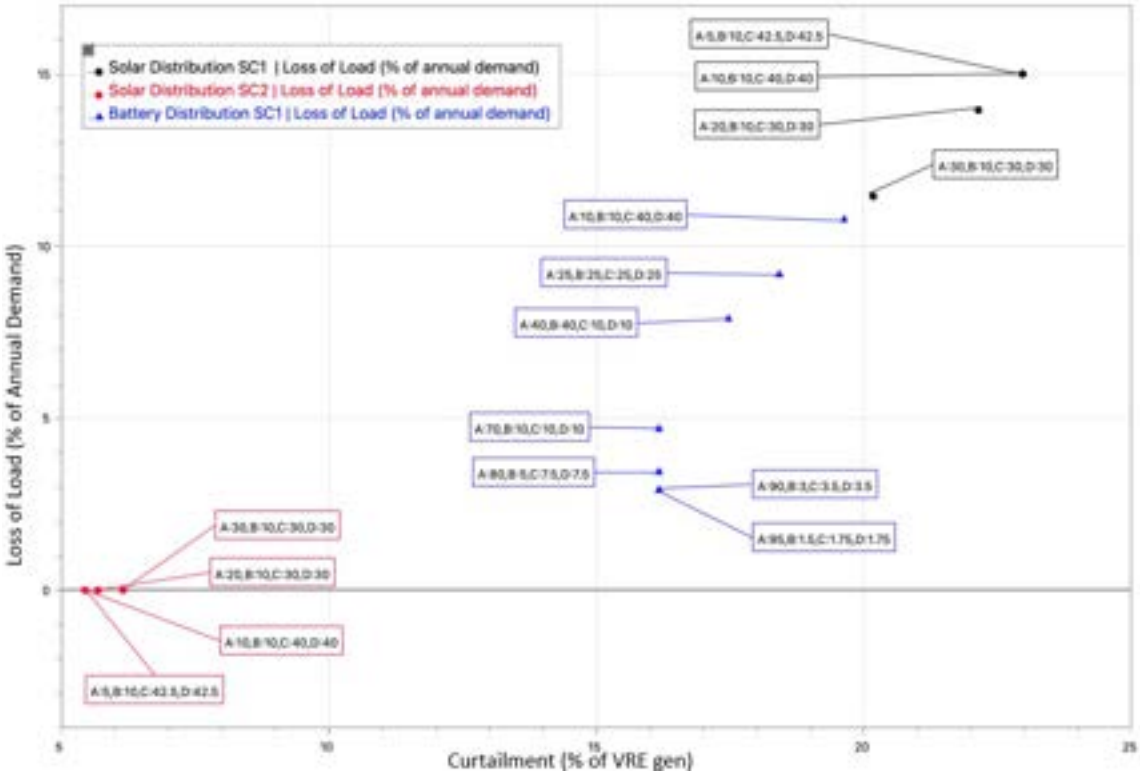


Figure 5: Sensitivity analyses for: the distribution of the battery capacity in SC1 (blue triangles), the distribution of solar capacity in SC1 (black circles) and SC2 (red circles).

[4] VRE Curtailment is a reduction in the capacity factors of renewable generators due to the lack of demand flexibility.

Additionally, two sensitivity analyses were performed (Fig.5), on the distribution of battery storage in Zambia for SC1 and on the decrease in the distribution of solar capacity at node A for both scenarios. These were important to determine the impact of possible changes on the flexibility of the power grid.

5. Discussion

As illustrated by Fig.2 Zambia must invest heavily in solar power to meet the increasing electricity demand, regardless of the scenario. The results highlight that the decarbonisation scenarios require the installation of 44GW for SC1 and 56GW for SC2 of renewable technologies by 2050 for Zambia to be on track to reach net zero in 2060. Although this seems ambitious, the country has the potential and has already begun installation of solar power units, with the cost being significantly lower than fossil-intensive power plants.

However, there is a caveat. To meet the increasing demand from the residential, commercial, industrial and transport sectors, solar alone will not suffice. This is highlighted by the inability of solar to meet the demand during periods of low sunshine, such as the early morning or at night. As a result, SC1 and SC2 turned to nuclear power to meet the remaining demand. The government has considered this option, going as far as developing a National Nuclear Policy in 2020 [48], but the costs of implementing nuclear power are immense. It was estimated that a 2.4GW plant was expected to cost around $30 billion [49]. With Zambia's annual budget at around $8 billion this implementation is impossible without the aid of international investment. There is equally a great scepticism from the Zambian population towards this energy source regarding disposal of radioactive waste.

Nevertheless, when nuclear power was excluded from the OSeMOSYS models, an infeasibility occurred due to hydro, biomass and geothermal being at maximum capacity, and solar unable to meet the demand at night. To overcome this issue, carbon capture systems could have been implemented in the model, to maintain the use of natural gas. However, its feasibility in the model would greatly depend on international support and innovations in this technology due to its high cost compared to solar technologies [50]. Equally storage facilities could have been added. This would have allowed an increase in solar capacity and electricity generation from solar technologies, that would then be stored and utilised at a later stage when there is a demand. Zambia would therefore have to invest heavily in storage capacities instead of nuclear power plants. This further highlights the importance of external funding to enable Zambia to achieve its development and climate goals.

However, the model does have limitations. Certain assumptions made have a great effect on the output of the scenarios. For instance, a linear decrease of the CO2 emissions was computed, restricting year on year the amount of fossil fuels used by scenario 1 and 2. An alternative narrative could have been used to postpone the reduction in GHG emissions as the country gets richer due to greater ease of reducing emissions, which would align with Zambia's vision of becoming a middle-income country by 2030 [5]. This would change the energy generation mix outputted by the model as a result. Equally, assumptions were made regarding the predictions of future demand for the different sectors, affecting the energy production of the grid. Due to these assumptions, the viability of the proposed scenarios should be scrutinised and assessed in further detail.

The Main Flexibility Issues

From Table 2, scenario 1 has significant flexibility challenges for the power grid in 2030 compared to scenario 2, due to the high loss and excess load and high curtailment. The combination of the excess load at node B&C and loss of load at node A (Fig.3), highlight the main challenge of geographical disparity between electricity demand at node A and the generation at node C. In fact, node C has 48% of the generation and 11% of the demand, compared to node A with 14% of the generation and 52% of the demand. This arises from Zambia's high dependence on storage-less hydro energy in 2030 at 85.4% (SC1) and 75.1% (SC2), shown in Fig.2. This generation is principally located at node C.

The comparison with the significantly improved flexibility in scenario 2, justifies the lack of transmission capacity between the nodes to eliminate the loss of load at node A. It is constrained by the lack of transmission capacity from node C to A, as transmissions to node A are only possible from node C (shown in Fig.1), and since node C has a significant excess load (Fig.3). Equally, there was a lack of battery capacity in SC1 to eliminate this excess load. The transmission and battery capacities were lower in SC1 by 62.5 % and 75% respectively.

This lack of capacity was further confirmed by the high curtailment in Scenario 1 (Fig.4), found in node C&D, 3.2 times greater than SC2. The curtailment in Scenario 2 is present but is inevitably due to the high share of VRE (96.9%) in the power grid. It was determined that to eliminate the 6.0 % curtailment, found mainly in node A, 1705 MW of battery capacity would need to be added to the system. From an economic standpoint, this was not beneficial, with a total project cost of $2.47 Billion [51]. Hence It was regarded as of insignificant magnitude compared to problems experienced in SC1.

Insufficient Inertia Flexibility challenge

A challenge that was observed for both scenarios was insufficient inertia added to the system by the electricity grid to match the limit set (Table 2). This was due to the high share of VRE in the system. However, since this infeasibility is of a low magnitude for both scenarios, it can be considered an insignificant flexibility challenge for the power grid. This was further justified with the elimination of this flexibility issue when the inertia limit was recomputed without the 10% reserves added to the worst-case size of the largest credible multiple

contingencies, originally computed in **section 3.4.5**. Also, it is important to acknowledge that Zambia does not currently have an official RoCoF value for their grid. The assumption to use the RoCoF value of South Africa's highlights the potential inaccuracy with this flexibility issue.

Insufficient Reserves Flexibility challenge
The flexibility issue regarding insufficient reserves was experienced in both scenarios (Table 1). From the analysis of the results, instead of achieving a 15 % static reserve, only approximately 10% of the electricity demand was achieved. This confirms the importance of investing in storage unit capacities to ensure flexibility in the energy grid and eliminate the insufficient reserves, principally caused by the dependence on storage-less hydro units in 2030 shown in Fig.2.

Sensitivity Analysis
The results for SC1 emphasised the main challenges of important loss, excess load, and curtailment due to the low transmission and battery storage capacities in the power grid. Therefore, a sensitivity analysis on the distribution of battery storage in Zambia was computed to determine its importance on the flexibility of SC1. As shown in Fig.5, a greater distribution of battery storage in node A significantly decreased the loss of load and the curtailment. This correlation is due to the loss of load only being present at node A, and hence a greater battery capacity in that node eliminated the curtailment in node A and reduced the loss of load. Likewise, it was computed for the excess load, but it stayed constant at 52.27%, due to the lack of transmission capacity from node C to A. This analysis validates the importance of battery distribution, within Zambia's power grid and the need for high battery storage at node A to minimize the loss of load and curtailment.

The solar capacity was distributed with the following proportionality (A:30, B:10, C:30, D:30), using the rationale mentioned in **section 3.4.6**. However, high distribution in node A may be unfeasible due to the important urbanization aligned with the projected increase in mining production targets from 2030, as shown in Fig.1. Therefore, a sensitivity analysis on the decrease in the distribution of solar capacity at node A was evaluated for both Scenarios. As shown with Fig.5, a higher battery and transmission capacity for the power grid in Zambia considerably affects the impact of the change in distribution on the grid's flexibility. Scenario 1's loss of load and curtailment increase with a lower solar capacity in node A, justified by the lower generation present at node A and a greater increase in curtailment at node D&C. Contrarily, Scenario 2's curtailment decreases due to greater decrease in curtailment at node A compared to the increase at the nodes D&C which is counteracted by the higher battery storage.

An evaluation was also performed on the cost difference of the batteries and transmission capacities for SC1 and SC2. The total investment cost difference was $429 million for transmission capacities [52], and $2.65 billion for the battery capacity [51]. Therefore, increasing the battery and transmission capacity to the same level as in scenario 2 to eliminate the flexibility issues may be a challenge for Zambia. This underlines the importance of attracting international support and of developing strategies to minimize transmission and distribution losses from heat dissipation and theft, for Zambia to reach its goals. Also, it is important to acknowledge that this flexibility assessment was only performed for 2030, investigating for different years until 2070 would be an additional factor to consider when comparing both development pathways.

6. Conclusion

This research aimed to contribute to the body of knowledge and generate novel material regarding Zambia's power grid, which lacks literature on its long-term development. This study successfully investigated different decarbonization pathways for the electricity grid in Zambia, in line with their IRP and NDC targets and development goals. It was achieved with the computation of long-term power system expansion and flexibility assessments on the electricity grid. These models required varied techno-economic input data, which were generated from sources available online and through data-driven assumptions. The study involved a direct comparison of three scenarios, which enabled the identification of challenges in the Zambian power grid and highlighted the importance of international support.

To meet the increasing demand, Zambia must invest heavily in solar power. The country rapidly exhausted its capacities in hydro, wind and geothermal so an alternative source of power was required. The OSeMOSYS model outputted nuclear energy as a substitute to natural gas, which was heavily present in the BAU scenario. However, this technology presents a lot of constraints with regards to safety and costs. As a results, a more detailed assessment of the possibility for Zambia to incorporate nuclear into its energy generation mix should be performed.

The evaluation of the flexibility assessments on Scenario 1 & 2 in 2030, identified the main challenges in Zambia's power grid. To mitigate the geographical disparity of electric demand at node A and generation at node C arising from the high dependence on storage-less hydro energy generation (85.4% for SC1 and 75.1% for SC2), investments in transmissions from C to A and battery capacities are essential. This was validated with the combination of the loss (11.46%) and excess (52.27%) load and the high curtailment (20.17%) in SC1 compared to the insignificant flexibility challenges in SC2. Therefore, for Zambia to achieve its targets of increasing access to secure and reliable electricity, along with achieving its desired mining production, international investments and strategies to minimize transmission and distribution losses are essential.

Nevertheless, with a rapidly increasing population and energy demand, and droughts which have directly affected hydropower electricity generation, solar energy

appears to be one of the most effective solutions to produce sustainable and clean energy in Zambia.

As mentioned previously, there are limitations with the accuracy of this study, as for certain projections the models required data-driven assumptions. The uncertainty of these assumptions was investigated and somewhat minimized, through comparisons with various sources and sensitivity analyses. The linear characteristic of the modelling software resulted in uncertainty, as the dispatch for electricity generated from a unit increased linearly instead of being directly switched on as with an integer optimization model, which could present inaccuracies for solar unity.

Regarding this research, it can be deduced that there are numerous innovative pathways to expand the analysis. The OSeMOSYS model did not incorporate long-term modelling for storage capacities, as illustrated by the paper by Howells et al. [18], which would provide a more accurate projection. Equally, alternative energy sources and technologies could have been explored and added to the model, such as hydrogen [53] and carbon capture.

Regarding potential extensions on the flexibility assessment, an increase in nodes in the FlexTool model would help assess issues at a micro level, beneficial for the identification of challenges regarding Zambia's goal of ensuring access to electricity in rural areas. Also, sector coupling with the heating system and the electrification of the transport sector with electric vehicles could be implemented in the FlexTool model. The use of smart strategies in sector coupling would provide demand-side flexibility in the system, extremely relevant to the flexibility assessments as there is an important implementation of EV from 2030 in Zambia.

Finally, another extension to this study would be the use of the investment operational mode within FlexTool, to determine the exact required investments in battery and transmission capacities in scenario 1 to eliminate the loss and excess load and to reduce the curtailment at an economically beneficial level.

This novel research provides a reliable reference for studies looking to assess the ability of developing countries to meet their long-term decarbonization goals, without hindering their development. The methods used in this paper could be applied to other developing countries and provide an assessment of their energy grid.

7. References

[1] Lusaka Times, *Zambia: Zambia's target to increase copper production to 3 million tonnes in the next ten years is attainable*, 2022.
URL: https://www.lusakatimes.com/2022/11/01/zambias-target-to-increase-copper-production-to-3-million-tonnes-in-the-next-ten-years-is-attainable/

[2] British Geological Survey, *Zambia: The Copper Mining Powerhouse looking towards a safer, low-carbon future,* 2022. URL: https://www.bgs.ac.uk/news/zambia-the-copper-mining-powerhouse-looking-towards-a-safer-low-carbon-future/

[3] IRENA, Energy Profile Zambia, 2022.
URL: https://www.irena.org/-/media/Files/IRENA/Agency/Statistics/Statistical_Profiles/Africa/Zambia_Africa_RE_SP.pdf

[4] IEA, *Tracking SDG7: The Energy Progress Report,* 2022.
URL: https://www.iea.org/reports/tracking-sdg7-the-energy-progress-report-2022.

[5] Ministry of Energy, *Integrated Resource Plan for the Power Sector in Zambia*, 2021. URL: https://www.moe.gov.zm/irp/?wpdmpro=integrated-resource-plan-inception-report

[6] *Nationally Determined Contribution (NDC) of Zambia*, 2021. URL: https://unfccc.int/sites/default/files/NDC/2022-06/Final%20Zambia_Revised%20and%20Updated_NDC_2021_.pdf.

[7] United Nations Population Division, *World Population Prospects 2022,* 2022. URL: https://population.un.org/wpp/

[8] Lusaka Times, Zambia: Increased acts of vandalism and theft force ZESCO to shift from using copper cables to aluminium, 2022. URL: https://www.lusakatimes.com/2022/03/25/increased-acts-of-vandalism-and-theft-force-zesco-to-shift-from-using-copper-cables-to-aluminum/

[9] Lloyds Bank, *Foreign Direct Investment (FDI) in Zambia*¸ 2022. URL: https://www.lloydsbanktrade.com/en/market-potential/zambia/investment.

[10] Hsiang, E. *Chinese investment in Africa: A re-examination of the Zambian debt crisis*. Harvard International Review, 2022. URL: https://hir.harvard.edu/chinese-investment-in-africa-a-reexamination-of-the-zambian-debt-crisis/.

[11] RES4Africa Foundation and Enel Foundation, *Integration of Variable Renewable Energy Sources in the National Electric System of Zambia*, 2019. URL: https://static1.squarespace.com/static/609a53264723031eccc12e99/t/60ed4dee4182611181b4e7bc/1626164746091/Integration-of-Variable-Renewable-Energy-Sources-in-the-National-Electric-System-of-Zambia.pdf

[12] McPherson et al., *Planning for variable renewable energy and electric vehicle integration under varying degrees of decentralization: A case study in Lusaka, Zambia,* 2018. URL: https://doi.org/10.1016/j.energy.2018.03.073

[13] IEA, *Status of Power System Transformation 2019 - Power System Flexibility,* 2019. URL: https://www.iea.org/reports/status-of-power-system-transformation-2019

[14] Godínez-Zamora et al., *Decarbonising the transport and energy sectors: Technical feasibility and socioeconomic impacts in Costa Rica*, 2020. URL: https://doi.org/10.1016/j.esr.2020.100573

[15] Taliotis et al., *Natural gas in Cyprus: The need for consolidated planning*, 2017. URL: https://doi.org/10.1016/j.enpol.2017.04.047

[16] Taliotis et al., *An indicative analysis of investment opportunities in the African electricity supply sector — Using TEMBA (The Electricity Model Base for Africa),* 2016. URL: https://doi.org/10.1016/j.esd.2015.12.001

[17] *Power system flexibility for the Energy Transition (2018) IRENA.* URL: https://www.irena.org/publications/2018/Nov/Power-system-flexibility-for-the-energy-transition

[18] Howells et al. *OSeMOSYS: The Open-Source Energy Modelling System: An introduction to its ethos, structure, and development*, 2011. URL: https://doi.org/10.1016/j.enpol.2011.06.033

[19] Ariemu, O. *COP27: Achieving net-zero in developing countries possible by 2060*, 2022. URL: https://dailypost.ng/2022/11/07/cop27-achieving-net-zero-in-developing-countries-possible-by-2060-un/

[20] Allington et al*., CCG starter data kit: Zambia, Zenodo, 2022*. URL: https://zenodo.org/record/6542410

[21] *The Energy Sector Report 2021, Energy Regulation Board*. URL: https://www.erb.org.zm/document/the-energy-sector-report-2021

[22] *Investment incentives for renewable energy in Southern Africa, 2012*. URL: https://www.iisd.org/system/files/publications/investment_incentives_zambia.pdf

[23] DVN.GL*, Wind resource mapping in Zambia 12 Month Site Resource Report - World Bank*, 2018. URL: https://documents1.worldbank.org/curated/en/528711526549758961/pdf/Renewable-energy-wind-mapping-for-Zambia-12-month-site-resource-report.pdf

[24] Mwanza, M. *Assessment of solar energy source distribution and potential in Zambia*, 2017. URL: https://www.researchgate.net/publication/318118164_Assessment_of_Solar_Energy_Source_Distribution_and_Potential_in_Zambia

[25] *Renewables readiness assessment: Zambia*, 2013. URL: https://cms.irena.org/-/media/Files/IRENA/Agency/Publication/2013/RRA_Zambia.ashx

[26] *Solar power by country Wikipedia. Wikimedia Foundation*, 2022. URL: https://en.wikipedia.org/wiki/Solar_power_by_country.

[27] Rosatom *Nuclear is the technology of Zambia's future*, 2018. URL: https://www.rusatom-overseas.com/media/mass-media-about-us/nuclear-is-the-technology-of-zambia-s-future-msiska.html

[28] USAID*. Education Data Activity Annual Performance Report,* 2018. URL: https://pdf.usaid.gov/pdf_docs/PA00ZQC1.pdf

[29] *Zambia coal Worldometer*, 2018. URL: https://www.worldometers.info/coal/zambia-coal/

[30] ESMAP, *The potential of regional power sector integration*, 2009. URL: https://www.esmap.org/sites/esmap.org/files/BN004-10_REISPCD_The%20Potential%20of%20Regional%20Power%20Sector%20Integration-Literature%20Review.pdf

[31] IMF, *World Economic Outlook Database,* 2022. URL: https://www.imf.org/en/Publications/WEO/weo-database/2022/October

[32] Mfula, C. *Zambia expects economy to grow 4% in medium-term, president says*, 2022. URL: https://www.reuters.com/world/africa/zambia-expects-economy-grow-4-medium-term-president-says-2022-09-09/

[33] *COP26 declaration on accelerating the transition to 100% zero emission cars and Vans,* 2022. URL: https://www.gov.uk/government/publications/cop26-declaration-zero-emission-cars-and-vans/cop26-declaration-on-accelerating-the-transition-to-100-zero-emission-cars-and-vans

[34] *Phase-out of fossil fuel vehicles,* 2022. URL: https://en.wikipedia.org/wiki/Phase-out_of_fossil_fuel_vehicles#cite_note-Glasgow-41

[35] IEA, *World energy outlook 2022 – analysis, IEA,* 2022. URL: https://www.iea.org/reports/world-energy-outlook-2022

[36] *Price forecast series – Capital,* 2022.
URL: https://capital.com/oil-price-forecast-2030-2050
URL: https://capital.com/coal-price-forecast-2030-2050
URL: https://capital.com/natural-gas-prices-forecast-2030-2050

[37] GISGeography. *Zambia map, GIS Geography*, 2020. URL: https://gisgeography.com/zambia-map/

[38] *REMP for Zambia final report 01 – JICA*, 2006. Available at: https://openjicareport.jica.go.jp/pdf/11871027_01.pdf

[39] EIB *Lusaka Power Transmission and distribution network,2015.* URL*:* https://www.eib.org/en/projects/pipelines/all/20120602

[40] UNECA. *Trade ties: Zambia and DRC sign cooperation agreement to manufacture electric batteries, create jobs,* 2022. URL: https://www.un.org/africarenewal/magazine/may-2022/trade-ties-zambia-and-drc-sign-cooperation-agreement-manufacture-electric

[41] IRENA, *Abu Dhabi, FlexTool Model*, 2021

[42] ESMAP. *Zambian Distribution Grid Code*, 2016. URL: https://esmap.org/grid-integration-requirements-for-variable-renewable-energy

[43] Chown, G. *System inertia and rate of change of frequency (rocof)*, 2018. URL: https://www.researchgate.net/publication/324280415_System_inertia_and_Rate_of_Change_of_Frequency_RoCoF_with_increasing_non-synchronous_renewable_energy_penetration

[44] Joshi et al. *Ramping up the ramping capability,* 2020. URL: https://www.osti.gov/biblio/1665866-ramping-up-ramping-capability-india-power-system-transition

[45] Shane et al. *Bioenergy Resource Assessment for zambia*, 2014. URL: https://www.researchgate.net/publication/282792504_Bioenergy_resource_assessment_for_Zambia

[46] *BESS EOI, 2022*. URL: https://www.africagreenco.com/bess/

[47] McQuillen, J. *Estimating electricity demand of Sub-Saharan Africa using AI*, 2021. URL: https://omdena.com/blog/electricity-demand/

[48] *National Nuclear Policy 2020*, 2020. URL: https://inis.iaea.org/collection/NCLCollectionStore/_Public/52/011/52011593.pdf?r=1

[49] Lusaka Times, *Zambia : Zambia's nuclear power plant to cost around US$30 billion-expert*, 2018. URL: https://www.lusakatimes.com/2018/05/23/zambias-nuclear-power-plant-to-cost-around-us30-billion-expert/

[50]. IEA. *Is carbon capture too expensive? – analysis*, 2021. URL: https://www.iea.org/commentaries/is-carbon-capture-too-expensive

[51] iCLIMA, 2022 URL: https://www.iclima.earth/article/forecasting-demand-for-batteries-until-2030-and-considerations-on-supply-different-applications-technologies-minerals-and-costs

[52]. POWERGRID International Directors C.E.C. *Underground vs. overhead: Power Line installation-cost comparison and mitigation,* 2022. URL: https://www.power-grid.com/td/underground-vs-overhead-power-line-installation-cost-comparison/#gref

[53] Lusaka Times, *Zambia : Why green hydrogen should be part of the energy mix in Zambia*, *LusakaTimes.com*. URL: https://www.lusakatimes.com/2022/06/29/why-green-hydrogen-should-be-part-of-the-energy-mix-in-zambia/

Exploring SAFT-Focused Deep Learning for Interfacial Tension Modelling

Anujan Kirupakaran and Keerthanan Rajan
Department of Chemical Engineering, Imperial College London, U.K.

Abstract: Machine learning (ML) has emerged as a powerful predictive tool for describing complex thermodynamic relationships that far surpasses the capacity of more time-consuming traditional models. This report details the ability of a model to replicate a correlation for the interfacial tension (IFT) of pure fluids using a statistical associating fluid theory. A feedforward neural network, trained using the Adam optimiser, was employed on a synthetic database produced from SGTPy, to obtain a result that mirrored Garrido's relationship. Upon achieving this, the practicality of the model was tested by computing the IFT of real fluids using molecular parameters. The deep learning model developed a relationship for interfacial tension with an average absolute deviation (AAD) of 1.4%, surpassing Garrido's AAD of 2.2%, in a significantly shorter time. The computational model for experimental data provided an accurate correlation with an AAD of 3-6%. Whilst noting a substantial increase relative to the synthetic set, this can be attributed to underlying errors in the experimental dataset. Ultimately, this shows that ML can be used in thermodynamics to supplement existing relationships and offer superior correlations, much faster than any human approach.

Keywords: Deep Learning, Interfacial Tension, Thermodynamics, Neural Networks, Quantitative Structure-Property Relationship

1. Introduction

Since the 20th century, measuring interfacial tension (IFT) of fluids has rapidly gained interest, due to its significance in product development, as it characterises prominent system behaviours. This key system property dictates the interfacial phenomena of liquid-liquid systems that influence heat and mass transfers between phases. This is observed countlessly in the world today, from the emulsifiability of different phases to tertiary oil recovery on offshore platforms, and multiphase microfluidic flow.[1] Evidently, IFT is central to the pharmaceutical, medical, and other major industries, thus, understanding and modelling IFT is extremely beneficial.

However, despite its abundance of usages, current thermodynamics models used to measure IFT utilise empirical correlations that took years to fabricate. These models are centred around a scarce amount of raw data that are only applicable to a few pure compounds at set system properties. As such, many theory-based estimation methods have also been developed, but these lack accuracy or are hard to parametrise.[2] Despite this, a particular theoretical model by Garrido et al., to predict the IFT of pure fluids mapped from Mie fluids, obtained an average absolute deviation of 2.2%.[3] Not only is this very accurate but it applies to a wide range of fluids. Yet, it was still time-consuming taking months to develop. To better these laborious processes to develop a model, machine learning (ML) can be deployed to produce a correlation for IFT in a significantly reduced time. This method will not only yield a model similar in accuracy to Garrido's model but will account for non-ideal behaviour as well. Non-ideality has conspicuous effects on IFT, yet few models can truly account for this, simply due to its complexity.[2]

Machine learning techniques have been a new tool utilised across research and industry in all disciplines, from finance to medicine. The ability to identify and form relationships between multiple variables to a high complexity makes it very useful in tackling large-scale problems. These same techniques have the potential to overcome non-ideality in molecular thermodynamics, making modelling more accurate over wide ranges of materials and properties, valuable information in today's industry as highlighted before.[4]

This research project aims to make a proof of concept for the application of ML, specifically deep learning, in IFT modelling. This would be achieved using data generated by the model used in Garrido et al. during training and testing, to see if an average absolute deviation lower than 2.2% can be achieved.[3] If this is successful, further modelling will be conducted utilising a small set of experimental data to test the model's application to real fluids.

2. Background

2.1 Interfacial Tension correlations

In 1876, Gibbs introduced IFT in his paper on composite-system thermodynamics and as described before, this term compensates for the excess molecules and energy at intermolecular interface interactions.[5] This allows for a complete system to be defined. Scientists exploited this excess energy to directly measure IFT because it causes interfaces to minimise interfacial area. This drive towards minimised geometry can be interpreted as a tensile force per unit length applied in the interfacial plane, which numerically equals IFT.[6]

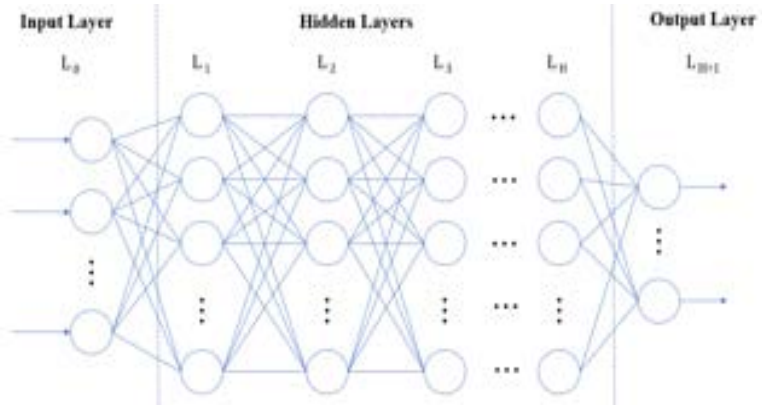


Figure 1: A visualisation of a fully connected neural network. There n_{L0} input nodes and $n_{L[H+1]}$ output nodes. In between these layers there are H hidden layers, with each layer containing n_h nodes.

Experimentally, this property is used to measure IFT indirectly, using a combination of Wilhelmy plates and du Noüy rings for force calculations or through pendant drop shape analysis. For further information on IFT measuring equipment and techniques, the reader is advised to study source.[6] However, although having high accuracy, these methods have a very limited application range, due to the strict requirements and numerous correction factors. For example, the rings must remain parallel to the surface, where the ring must experience perfect wettability.[6] Also, rings are extremely prone to deformation during handling and cleaning, which will cause a very large error in the measurement.

Consequently, the popularity of theoretically driven simulations has risen in the past years, in particular, quantitative structure-property relationship (QSPR) models. Based on the relationship of molecular descriptors to model different chemical properties, this approach offers an efficient methodology for predicting critical parameters and complex correlations for pure fluids from scratch.[7] This has been utilised alongside other mathematical theories, such as multiple linear regression or square gradient theory, to obtain very accurate results for physio-chemical properties, that surpass any experimental-based correlations.

For this computation, the QSPR model proposed by Lafitte et al. 2013, the statistical associating fluid theory for variable range interactions through the Mie potential (SAFT-VR-Mie) equation of state (EoS), was applied.[8] This semiempirical EoS characterizes molecules using spherical fragments linked in a chain, symbolising beads on a pearl necklace. These 'beads' are quantified by the spherical radius [σ], the number of Mie 'beads' [m], the repulsive component for the bonds [λ_r], the attractive component of the bonds [λ_a], and the energy well depth[ε]. It was selected because of its advantage over the previous renditions, resulting from its exploitation of the Baker and Henderson high-temperature perturbation theory and the radial distribution function of the reference monomer fluids.[8] For further information on the development of this theory, the reader is advised to review source 8. This application of various theories fabricated an EoS that can be applied to a much broader range of molecular fluids and enhanced accuracy in near-critical regions. This coupled with accurate modelling of second-derivative thermodynamic properties (heat capacities, Joule-Thomson coefficients, speed of sound, etc) from previous SAFT approaches, facilitates a significantly enriched global representation of system properties and phase equilibria of fluids.[8]

2.2. Deep Learning for Thermodynamics

A lot of attraction has been received by a certain subsection of ML, deep learning. This is because of its ability to process information like the human brain whilst being capable of sifting through much more information. The models are made of multiple layers of interconnected nodes, called artificial neural networks (ANN), where an input layer feeds the data into the subsequent layer nodes, known as hidden layers, until the output layer is reached. There are different types, such as convolutional networks, that serve to fit specific use cases, like computer vision. From a regression standpoint, it is much simpler to train a feedforward neural network (FNN).[9] A visualisation of an FNN can be seen in Figure 1. The nodes in an FNN take a weighted sum of the outputs of the previous layer and a bias term before being permuted further by an activation function and sent to the subsequent node. FNNs can model complex functions thanks to the non-linearity of the activation function. These include rectified linear unit, tanh and sigmoid, and each function will transform the subsequent inputs differently. For a single node of input size n and step function h(x):

$$y = \sum_{i=1}^{n} [w_i x_i] + w_0 \qquad \textit{Eqn. 1}$$

$$z = h[y] \qquad \textit{Eqn. 2}$$

Since the same activation function is applied to the whole layer, a single layer with m nodes can be represented as:

$$\boldsymbol{z} = h\left(\begin{bmatrix} w_{1,0} + \cdots + w_{1,n}x_n \\ \vdots \\ w_{m,0} + \cdots + w_{m,n}x_n \end{bmatrix}\right) \qquad \textit{Eqn. 3}$$

$$= h\left([1 \quad \dots \quad x_n]\begin{bmatrix} w_{1,0} & \cdots & w_{m,0} \\ \vdots & \ddots & \vdots \\ w_{1,n} & \cdots & w_{m,n} \end{bmatrix}\right)$$

$$= h[\boldsymbol{x}^T\boldsymbol{W}] \qquad \textit{Eqn. 4}$$

To find these optimal values for **W**, the parameters are initialised with random values and initial predictions from the training set are made. This is utilised in a loss function, with the true values, to find an error that is to be minimised through an optimization process known as backpropagation. Backpropagation is a simple function that, for each parameter, calculates the gradient of the loss algorithm from the training set using automatic differentiation and adjusts that value proportional to the learning-rate, α. Thus, after every complete pass of the training set, the parameters converge closer to the optimal values. To reduce the memory requirement, the data set is randomly split into mini-batches and calculates average adjustment for each batch. The number of times the algorithm goes through the whole set is known as the epoch number and will continue to run until the stopping value or maximum epoch number is reached.

Epoch number and batch size are known as the hyperparameters (HPs) of the model. Other HPs include the learning rate, minibatch size, activation function as well as layer number and size. Like the parameters of the network, the hyperparameters must be optimised so that an optimal solution can be reached. This is formally done in a stage called hyperparameter tuning to optimise for the best values utilising algorithms such as Hyperband and Bayesian Optimization (BO).

3. Methodology

For this experiment, a method that combines ML with QSPR was exploited, due to its distinct advantages over published literature models: high predictive accuracy; sufficient interpolation and extrapolation ability; and ease of interpretation of very large datasets.[10]

To test the viability of training a model to predict IFT, a large amount of data had to be collated to cover the training and validation portions of the workflow. Consequently, a database of artificial data was constructed to ensure better data quality and quantity, following the methodology proposed by Zhu and Muller.[9] This was completed by utilising a Python package called SGTPy which allowed for inbuilt functions for calculating interfacial properties through the square gradient theory. [11] A pipeline was constructed where the SAFT-VR Mie EoS parameters were randomly generated within a realistic domain, to avoid computational error, and using the sgtpy.component function to obtain state properties for this simulated molecule. IFT values were generated from a saturated temperature range of 0.5-0.95 of the critical temperature (T_c) through the sgtpy.sgt_pure function. The database was compiled where each data point referred to a single IFT value for specific saturated temperature and SAFT parameters. This method had the advantage of being able to control the size of the dataset, and not relying on experiment data meaning that the data was not restricted to revolving around room temperature and/or pressure. This allowed for much more general models to be investigated, leading to more conclusive results.

A reduced temperature range from 0.5-0.95 was selected because outside these bounds the data becomes unreliable and cannot be computed accurately. At the upper extreme, errors arise from the system reaching a critical point. IFT evolves from the unbalanced adhesive forces of one phase at the boundary and is related to the molecule's energy. At the critical point, this force vanishes as the two phases merge into one supercritical phase.[12] As IFT approaches zero, the molecules are in a state of constant fluctuations as they are very mobile near the surface, causing continuous adjustments to the surface tension.[12] At the lower bound of the reduced temperature, this is the general region just before fluids reach their freezing point. Similarly, the IFT drops as the molecules become more organised, reducing the net cohesion forces at the interface.[13] However, this bound is an estimate and is heavily dependent on the fluid. Hence, gaining accurate readings for these unstable regions for IFT are difficult. This problem also appears in molecular simulations, as models are unable to produce an accurate relationship in this area.[13]

Mirroring Zhu and Muller, the problem dimensionality was reduced by non-dimensionalising T and IFT(γ). This data pre-processing reduces the scale of the respective parameters to a common scale to enhance the model's accuracy. These equations are based on prior knowledge of thermodynamic relations. This

$$T^* = \frac{T.k_B}{\varepsilon} \qquad \textit{Eqn. 5}$$

$$T^*_{red} = \frac{T}{T_c} \qquad \textit{Eqn. 6}$$

$$\gamma^* = \frac{\gamma.\sigma^2}{\varepsilon} \qquad \textit{Eqn. 7}$$

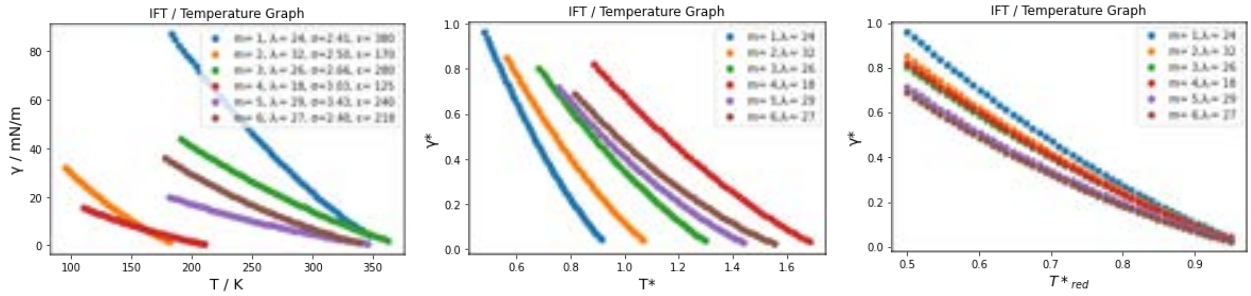


Figure 2: The graph on the left plots 5 IFT series for different SAFT parameters, each having 20 points between 0.5-0.95 Tc. The curves are far apart and have widely differing ranges. Middle plot shows the same series but non-dimensionalised using the SAFT parameters. Now the IFT values are set between a fixed similar range, but the temperature scale still differs significantly. However, when plotting non-dimensionalized IFT against T/Tc* instead we see a much better scaling where the values differ slightly, meaning the model only needs to correlate fewer relationships to predict the values and reducing problem complexity.*

can be taken further by reducing non-dimensionalised T by normalising it to the critical temperature. This is shown by the graphs in Figure 2, which visualises the further reduction in the value ranges and complexity.

In this project, an FNN was trained using the Adam optimiser due to its computational efficiency and suitability to large datasets and problems and is generally viewed as the recommended default option.[14] BO was selected for the hyperparameter optimisation due to its fast convergence speeds, in conjunction with babysitting to allow for initialising hyperparameters and small tweaks near an optimum.[15] K-Fold Cross-validation, K=10, was done to prevent data leakage and overfitting.

It was determined that no more than 4 hidden layers of size 50 would be utilised, with the aim of reducing that to <20 to avoid overfitting to the training set. The exact number in each layer was determined in the HP optimization stage. The hyperbolic tanh was chosen as the activation function due to its continuous nature resulting in an infinite number of continuous derivatives. For the loss function, Mean Squared Error (MSE) was selected but other metrics such as R^2, Root Mean Squared Error (RMSE) and Mean Average Percentage Error (MAPE) were the main metric monitored due to ease of interpretability and comparison between models. For all modelling pipelines, the databases were split 80:20 training/validation split, meaning 80% of data would be used to train the model whilst 20% would be used to test its accuracy. TensorFlow python package, in particular the Keras application programming interface, was utilised in the coding due to the ease of developing and training models with large datasets for high performance.[16]

For training and validation, different databases were generated to take a methodological process to get to a complete model. Initially, m was set to 1. This was done to further reduce the dimensionality of the problem to only 2 independent variables: T and λ_r. Then the BO HP optimizer determines the remaining hyperparameters, which are used alongside the dataset to train the model. The model is trained until a MAPE under the tolerance level of 2% is achieved, after which the validation set is utilised to assess performance. This is repeated until an optimal value is reached below the minimum threshold. A hold-out set is generated from a collection of IFT series to test the capability of the model on unseen data points.

The problem complexity is further intensified by increasing the value range of m in fixed increments of 1, utilising the same training methodology. This process is halted when a model for $1 \leq m \leq 6$ is obtained as this matches the initial criteria of Garrido's model and for most real pure fluids. Upon completion, the ML algorithm would provide proof of concept that deep learning can be utilised to predict interfacial tension. From here, the computational predictive tool was tested against existing literature values to truly appreciate the accuracy of the model.

4. Results and Discussion

Table 1: Overview of Synthetic Model Architecture

ANN Specifications	
Inputs	T_{red}*, λ_r, m
Output	γ*
Hidden Layers	[11,8,6]
Activation Function	Tanh

As aforementioned in section 3, multiple ANNs were developed throughout the model, each with increasing complexity. The goal of accurately predicting IFT using the SAFT-VR Mie EoS model consists of six Mie beads and a λ_r range between 8

to 44, paralleling Garrido's theory and real fluids. From an overview of the 3 models based on synthetic data, very little changed with the model architecture. An overview can be seen in Table 1, where each network had 3 inputs and 1 output, tanh activation function and hidden layer shape of [11,8,6]. During training and validation, a common trend was found with the hidden layer shape where the much larger networks with ≤50 nodes/layer would perform noticeably better, with roughly a 0.5% decrease in MAPE. However, the smaller models were chosen to avoid the effects of overfitting to the dataset and provide a concrete example of machine learning outperforming theoretical models. The following sections go into detail about the results of each model.

4.1 Case 1: m = 1

Table 2: m=1 Model Test Performance Metrics

Performance Specifications			
	Training	Validation	Testing
MSE	1.88×10^{-6}	1.89×10^{-6}	1.89×10^{-6}
R^2	≈ 1	≈ 1	≈ 1
RMSE	0.00137	0.00137	0.00137
MAPE	0.319	0.317	0.372

This case was the first to be modelled to ease into the complexity of this problem. The performance metrics, seen in Table 2, show the model reached an MSE of 1.88×10^{-6} in training and 1.89×10^{-6} in validation, with the very low values implying a strong convergence to a true model. This was backed up further by the coefficients of determination (R^2) being near identical, signifying that the data fit due to the regression model being high and replicable. This was shown very clearly in the parity plot in Figure 3, where the points were almost exactly on the true line. The low MAPE values in training and validation further emphasised the accuracy of this model. During the testing phase, similar results were achieved with there being a slight increase in the metrics, which can be attributed to the absence of bias in the learning process. A few cases can be seen in Figure 4, where the IFT series are plotted against reduced non-dimensionalized T (T_{red}*) for different values of λ_r. The circles represent the predicted values by the neural network whilst the true curve was plotted for different values of λ_r, visualising the 0.398% MAPE achieved in testing. Although performing highly, other factors like the large dataset and low dimensionality of the inputs could have led the model to just memorise the values, since dimension reduction had limited the scale that was required for predicting.

Figure 3: Parity Plot for Model m=1

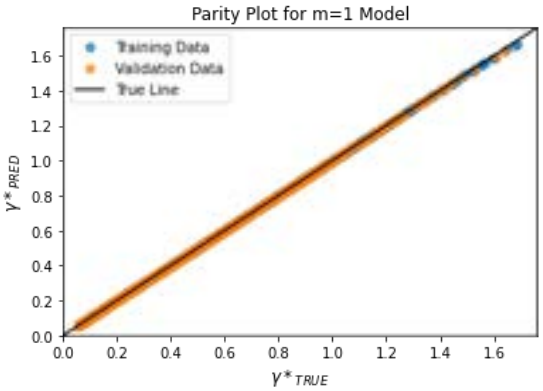
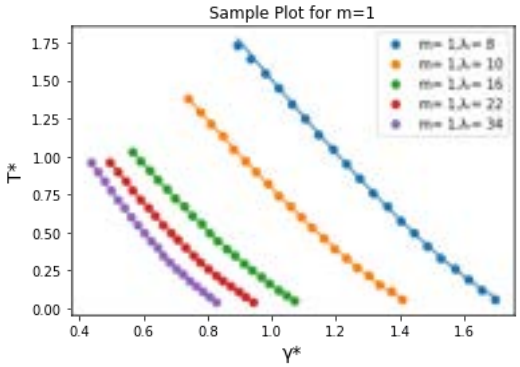


Figure 4: Value Plots for Model m=1



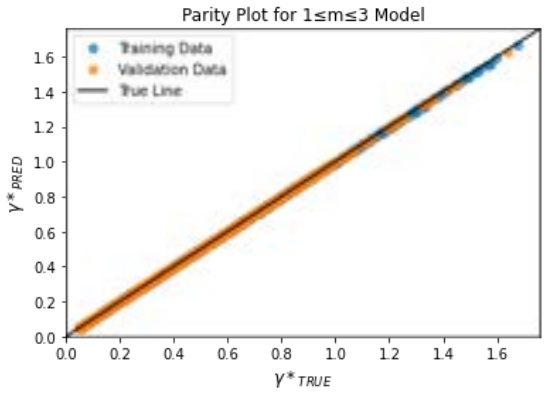
4.2 Case 2: 1 ≤ m ≤ 3

During this second case, the range for the number of beads, m, was increased to add some complexity and move towards a more useful, general solution. Similar results were yielded with almost perfect R^2 values and negligible MSE values for both validation and training, as shown by the performance specifications in Table 3. This suggested that the ANN comfortably produced a replicable regression model with a strong

Table 3: 1 ≤ m ≤ 3 Model Test Performance Metrics

Performance Specifications			
	Training	Validation	Testing
MSE	8.60×10^{-6}	8.76×10^{-6}	1.02×10^{-5}
R^2	≈ 1	≈ 1	≈ 1
RMSE	0.00293	0.00296	0.00320
MAPE	0.779	0.781	1.08

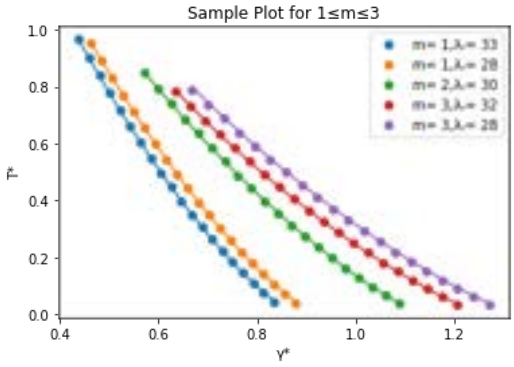
Figure 5: Parity Plot for Model $1 \leq m \leq 3$



convergence towards predicting true values. This can be seen in Figure 5, where the parity plot again plotted an almost perfect true-line. This sustained high performance, despite the increased difficulty, indicated that the model was not simply memorising the values, but the model has obtained an extremely accurate relationship for these parameters.

Hold-out testing performed on the finalised model cemented the accuracy of the model, highlighted by the low test metrics mirrored by the training & validation stages. The small increase in error between the two suggests a lack of overfitting within the model. The γ^*-T_{red}^* plot, in Figure 6, demonstrated the low error margins achieved in testing, as little to no deviation is seen.

Figure 6: Example Plot for *$1 \leq m \leq 3$ Case*



Nevertheless, when compared to the previous renditions, the error margins have increased by a few folds. This was anticipated because of the increased difficulty of the new dataset, as an additional variable, m, was considered. Yet, this error was negligible as it was still in the order of 10^{-3} or smaller for the RSME and MSE.

4.3 Case 3: $1 \leq m \leq 6$

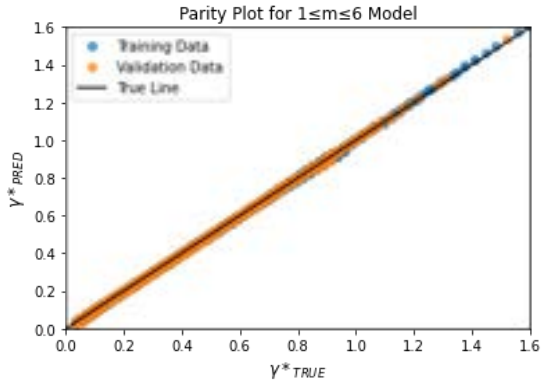
For the final theoretical case, the parameters are expanded to a real range of chemicals that matched the theoretical case. Table 4 shows the training and validation performance for the optimal model during the learning process. The MSE for this case was very low, implying that the model had converged close to

Table 4: $1 \leq m \leq 6$ Model Test Performance Metrics

Performance Specifications			
	Training	Validation	Testing
MSE	1.70×10^{-5}	1.68×10^{-5}	2.15×10^{-5}
R^2	≈ 1	≈ 1	≈ 1
RMSE	0.00412	0.00410	0.00464
MAPE	1.34	1.40	1.58

the true parameters. As with the previous case, we saw a trend where the MSE was increased with the complexity of the inputs, however, it was not significant enough to affect the convergence. An R^2 approximately equal to 1 was seen, indicating that the model could consistently predict IFT values, with Figure 7 further highlighting this. Notably, the distribution of data was not large enough across the validation data as there were very few values above 1.2 in this set. This could be assumed because the majority of high γ^* occurred at low temperatures, and very few existed at 0.5 T_c. This made it difficult to assess this region properly. However, the rest of the set performed well with the values lying on the true line reiterating the low MAPE.

Figure 7: Parity Plot for Model $1 \leq m \leq 6$

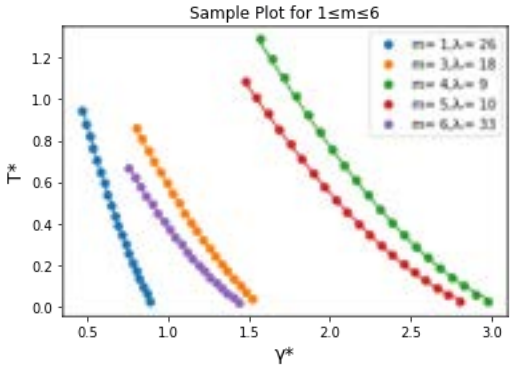


In testing with the hold-out set, there was a large jump from 1.08% in the previous case to 1.58% in the current one. This can be explained by both models having the same network size, but the current one was training with a more complex dataset. This was proven in previous stages of the modelling process where for much bigger neural networks, a MAPE <1% was achieved. In comparison to the 2.2% AAD in the SGT model, it's clear FNNs have the potential to form engineering correlations without the need for scientific rigour.

Figure 8 plots non-dimensionalised IFT against non-dimensionalised temperature for some of the series in the hold-out set. Comparing the neural network predictions, represented by the dots,

and the true curves of the same parameters, reinforce the performance metrics' low error rate.

Figure 8: Example Plot for Model $1 \leq m \leq 6$



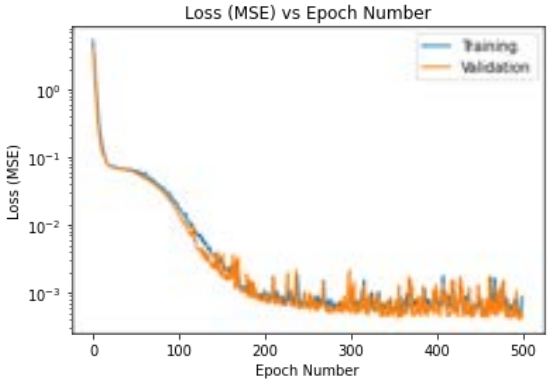
4.4 Case 4: Experimental Models

The ANN has been shown to excel in replicating a relationship for the IFT of pure fluids, based on theoretical data. To validate the proof-of-concept for real pure fluids, the practicality of the ANN was tested against a collection of experimental values sourced from Christian Wohlfarth and SAFT molecular parameters taken from Herdes et al.[17,18] For each chemical, the most up-to-date IFT values were selected, and any data older than eight years was disregarded. This ensured the most accurate dataset with no conflicting values, and for each chemical, the data was sourced from the same experiment to maintain some consistency in conditions. For the reduction calculations, the critical temperatures were taken from the NIST chemistry database.[19] In this new database, there were 25 different chemicals including inorganic nonmetals, hydrocarbons, and aromatics, with Appendix A going into further detail. In total there were 219 individual data points each characterized by T_{red}*, λ_r, m and γ*.

An overview of the development of the experimental-based model can be found in Table 5, alongside the performance of the model. It follows a similar approach to the previous cases, with the hidden layer size reducing further to [4,2] to prevent overfitting to the smaller dataset. The train-test ratio for this set was 70:30, with the test size increasing to gain a broader perspective of the model's ability. It was also vital that the distributions across the training and testing sets were equal, to ensure a good reflection of performance. Due to the lack of data, the training set was split 70:30 into a new training set and validation set, so to replicate the previous methodology of an unseen test set. With these alterations, the ML algorithm was tuned and trained until an optimal model was reached.

In training, the model converged to an MSE of 3.78×10^{-4}, and this is illustrated by Figure 9 which plots the loss against the epoch number, with the validation loss following behind at 6.29×10^{-4}, being almost double that of the training metric. Initially, the losses started at high values, with little deviation between training and testing. Over each epoch, it approaches a solution but struggles to converge properly as there is some noise, oscillating around the 10^{-3} value. During final testing, a smaller MSE of 5.63×10^{-4} was achieved, further showcasing the model's accuracy.

Figure 9: Loss Plot for Experimental Model



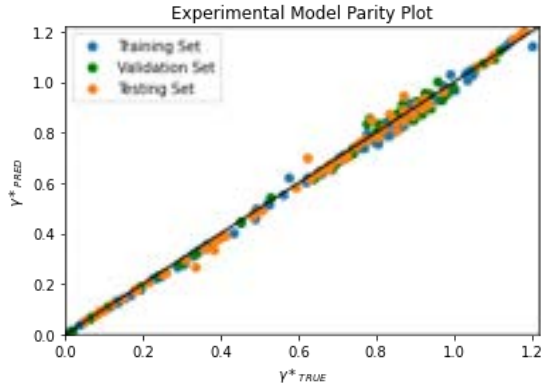
The experimental model successfully achieved a strong regression correlation to the experimental data, with an R^2 value of over 99% across every set. This is illustrated in Figure 10, highlighting the ability to perform consistently when interpolating data and proving that the model has little bias towards training set values.

Looking at the percentage errors for this case, the MAPE for validation seemed abnormally high in comparison to the training metric achieved. However, upon inspecting the testing metric the value appears to agree much more with the training

Table 5: Overview of Experimental Model

ANN Specifications			
Inputs	T_{red}*, λ_r, m		
Output	γ*		
Hidden Layers	[4,2]		
Activation Function	Tanh		
Performance Specifications			
	Training	Validation	Testing
MSE	3.78×10^{-4}	6.29×10^{-4}	5.63×10^{-4}
R^2	0.996	0.994	0.994
RMSE	0.0194	0.0251	0.0237
MAPE	2.99	5.58	3.12

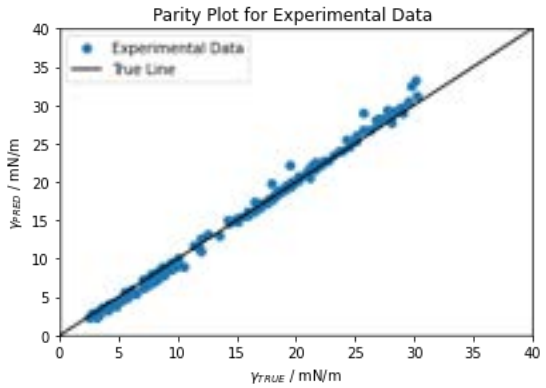
Figure 10: Parity Plot for Experimental Model



value, implying that the validation percentage may be inflated. This is further validated by the RMSE of each subset being not that different from the other. One factor that could have caused this inflation is the validation dataset containing much smaller values close to zero. Errors in these values have more significant contributions as the error would be much larger than the true value itself, leading to a larger overall percentage error. Although an attempt was made to make it uniform across the non-dimensionalised IFT values, this was unlikely to occur in every input parameter range for the model. This slight disagreement could have caused the values to be distributed this way.

In comparison to the realistic artificial model's performance, there was a significant increase in the error. This was expected as a small dataset is attributed to an increase in error, due to there being fewer data to learn from and hence predict from. Additionally, the experimental values encompassed a much larger range for temperature, compared to that of the artificial model, where it was fixed between 0.5-0.95. Henceforth, the real data explored the extremities of the physical region: near the critical and freezing points. The model was expected to struggle in these zones, due to the aforementioned IFT fluctuations, but it overcame these by developing relatively accurate predictions in these regions. Although unable to beat the 2.2% theoretical benchmark, a 3.12% error in testing is quite accurate and provides an impressive performance relative to the small dataset utilised.

Figure 11: Parity Plot for Artificial Model using Experimental Data



To get a better picture of the quality of performance from this dataset, the artificial model was further validated by utilising this dataset as another hold-out testing set. As shown by the parity plot in Figure 11, the model could predict IFT accurately and illustrated the $R^2 = 0.994$ calculated for this set. The percentage error of 3.12% obtained again exemplified that SAFT-focused neural network as the concept can be utilised as a high accuracy alternative for molecular thermodynamic modelling.

4.5 Training Deep Learning Model vs. Developing a Theory

In this section, some of the advantages and disadvantages of the two modelling methods will be discussed, as there are other factors in deciding between the two approaches.

Deep learning methods need to perform many iterations and calculations for every training cycle. However, this can be done utilising computational graphs and is relatively fast compared to the time taken to comprehensively develop a theory, where it is common to spend years perfecting it. The ability to bypass a conventional time constraint allows correlations to be investigated and discovered much more quickly. The ability to tackle high-complexity problems with little outside input is any factor in deep learning's superiority, as coupled with the computational time a larger range of relationships can investigate between many different variables. This is especially pertinent to engineering applications, as accelerating the modelling process would more breakthroughs for real applications in chemical engineering throughout research and industry.

The ability to go around a scientific background can be seen as an advantage. Many models utilise solely data-driven methodologies as they can reduce the complexity to a simple correlation problem. However, regarding scientific progress, it does little to advance the understanding and knowledge of the mechanics behind real phenomena. ML with QSPR modelling tries to bridge this gap with more molecule-focused inputs. Yet, due to the black-box nature of most ML techniques, it blurs any perception towards understanding the model's method. Inputs from QSPR models cannot always match those of the real values, such as the critical temperature literature values differing from those forecasted by the SAFT-VR Mie model. This could have led to inconsistencies between T_{red} and the corresponding IFT, causing some bias.

One of the biggest factors that limited exploration in this project is the availability of consistent, experimental data. Training using neural

networks requires large amounts of data, as shown by the artificial model requiring 1000s of data points in the training set. Whilst this size is normal in data science, getting enough empirical data from experiments would be challenging. If it's not possible to fill this requirement, there will be a significant drop in accuracy and generality as seen by the attempt at a model correlated purely from experimental data.

On the flip side, having an unnecessarily large database leaves the model open to overfitting, increasing bias toward values seen more often. This goes hand in hand with the requirement for good distributions across the input and output parameters as failure to do so would result in sub-par predictions and worsening extrapolations that may be attempted.

The ability to extrapolate far outside the training data is something deep learning fails to do. Theoretical-based approaches have the upper hand in both data requirement and extrapolation power as they can be formed without the need for large amounts of data and can have far better results when extrapolating outside of these ranges. This is exemplified by Garrido's work as the database utilised is far smaller and achieves similar performance benchmarks to FNNs compiled in this project.

Regardless of the disadvantages discussed here, the performance metrics witnessed across the modelling stage surpassed that of any theoretical approach.

Conclusion

The models developed in this investigation have showcased the potential for SAFT-focused deep learning to replace the conventional theoretical modelling practice. Trained using artificial datasets, models were able to effectively correlate interfacial tension from the SAFT-Mie parameters, m and λ_r, for a specific temperature, illustrated by metrics such as high coefficient of determination and RMSE. The artificial models were also able to outperform the average absolute deviation benchmark of 2.2%, obtained by the SGT theory and utilised to test model viability. With this successfully achieved, an experimental database was generated for a variety of different industrial pure fluids so that further analysis could be conducted on the viability of real fluids. The metrics obtained from the model trained with this proved again the ability of the neural networks to perform well even with experimental real data, with the artificial model validating this with a low percentage when the experimental data was utilised as a hold-out testing set. Many shortfalls in these methods were addressed, like the inability to extrapolate far outside the training set, lack of scientific foundation, and the hindrance by a small experimental dataset. However, they did not outweigh the result, showcasing the computational speed and accuracy achieved by SAFT-focused neural networks. The disadvantages found can also be addressed by further avenues of research which came to move towards concrete proof of deep learning being the superior methodology.

Whilst deep learning was the focus of this research project, there are many other machine learning algorithms and techniques which utilise less data-heavy approaches and enhance extrapolative power. Transfer learning is the idea of taking a previously trained model and further fitting it with new data. This idea can be utilised to reduce the requirement for the dataset as models trained using theoretical models, such as the one compiled in this project, can be further retrained using a small real database to move predictions towards more accurate results. Ensemble learning models can be utilised as they train multiple weak learners and combine them to create much stronger models. Algorithms such as random forest and XGBoost which have already proven to be accurate in modelling thermodynamic properties, can allow for much wider predictions to be made using small datasets.[9,20] Future SAFT-focused modelling can incorporate either of these algorithms to close the gap in accuracy and produce a better model overall.

Another major factor which limited our analysis of the two modelling approaches was the metric utilised to numerically compare the two. Whilst mean average percentage error, synonymous with average absolute deviation, is used commonly in comparison of models it failed to be truly consistent in concretely proving method superiority, as seen by inflations in values for sets including close to zero values. Finding a more optimal solution is important as this can lead to the dismissal of high-accuracy models. Alternatives such as root mean squared scaled error or mean average scaled error can be employed for evaluating accuracy.[21] Utilising a standardized training set, such as the one from the theory, in conjunction with standardised hold-out sets, would robustly determine which methodology is more viable for optimal molecular modelling.

The scope of this project was focused on pure fluids, allowing for simpler systems to be modelled. However, industrial chemical process modelling focuses more on mixtures of inhomogeneous fluids, such as aqueous-organic emulsions and separation systems for tertiary oil recovery. The logical solution to this is to expand the scope of the project to modelling schemes such as binary mixtures and others more complex. This would include assessing model viability as well as investigating correlations from experimental data, which may be more common due to its necessity. Overall, this exploration would allow for a more beneficial real-industry application.[22]

Acknowledgements

We would like to thank Professor Erich Muller for his guidance and wisdom, and especially thank Gustavo Chaparro Maldonado, for debugging any mistakes and queries we had along the way. Without their help, the project wouldn't have been possible; it was truly appreciated.

References

1. D'Apolito, R. et al. Measuring Interfacial Tension of Emulsions in Situ by Microfluidics. Langmuir 34, 4991–4997 (2018).
2. Yang, Y. L., Tsao, H. K. & Sheng, Y. J. Molecular structure incorporated deep learning approach for the accurate interfacial tension predictions. J Mol Liq 323, 114571 (2021).
3. Garrido, J. M., Mejía, A., Piñeiro, M. M., Blas, F. J. & Müller, E. A. Interfacial tensions of industrial fluids from a molecular-based square gradient theory. AIChE Journal 62, 1781–1794 (2016).
4. Ding, J. et al. Machine learning for molecular thermodynamics. Chin J Chem Eng 31, 227–239 (2021).
5. GIBBS, J. W. ART. LII.--On the Equilibrium of Heterogeneous Substances; American Journal of Science and Arts (1820-1879) 16, 441 (1878).
6. Drelich, J., Fang, C. & White, C. Measurement of interfacial tension in Fluid-Fluid Systems. in Encyclopedia of Surface and Colloid Science 3152–3166 (2002)
7. Ghasemi, F., Mehridehnavi, A., Pérez-Garrido, A. & Pérez-Sánchez, H. Neural network and deep-learning algorithms used in QSAR studies: merits and drawbacks. Drug Discov Today 23, 1784–1790 (2018).
8. Lafitte, T. et al. Accurate statistical associating fluid theory for chain molecules formed from Mie segments. J Chem Phys 139, 154504 (2013).
9. Zhu, K. & Müller, E. A. Generating a Machine-Learned Equation of State for Fluid Properties. J Phys Chem B 124, 8628–8639 (2020).
10. Li, R., Herreros, J. M., Tsolakis, A. & Yang, W. Machine learning-quantitative structure property relationship (ML-QSPR) method for fuel physicochemical properties prediction of multiple fuel types. Fuel 304, 121437 (2021).
11. Mejía, A., Müller, E. A. & Chaparro Maldonado, G. SGTPy: A Python Code for Calculating the Interfacial Properties of Fluids Based on the Square Gradient Theory Using the SAFT-VR Mie Equation of State. J Chem Inf Model 61, 1244–1250 (2021).
12. Winkler, C. A. & Maass, O. AN INVESTIGATION OF THE SURFACE TENSION OF LIQUIDS NEAR THE CRITICAL TEMPERATURE. 9, 65–79 (2011).
13. Haji-Akbari, A. & Debenedetti, P. G. Computational investigation of surface freezing in a molecular model of water. Proc Natl Acad Sci U S A 114, 3316–3321 (2017).
14. Ruder, S. An overview of gradient descent optimization algorithms. (2016) doi:10.48550/arxiv.1609.04747.
15. Yang, L. & Shami, A. On hyperparameter optimization of machine learning algorithms: Theory and practice. Neurocomputing 415, 295–316 (2020).
16. Chollet, F. & others. Keras. Preprint at (2015).
17. Wohlfarth, C. Surface Tension of Pure Liquids and Binary Liquid Mixtures. Surface Tension of Pure Liquids and Binary Liquid Mixtures (2016) doi:10.1007/978-3-662-48336-7.
18. Herdes, C., Totton, T. S. & Müller, E. A. Coarse grained force field for the molecular simulation of natural gases and condensates. Fluid Phase Equilib 406, 91–100 (2015).
19. P.J. Linstrom & W.G. Mallard (Eds.). NIST Chemistry WebBook, NIST Standard Reference Database Number 69. (National Institute of Standards and Technology, 2022).
20. Zhang, J. et al. A unified intelligent model for estimating the (gas + n-alkane) interfacial tension based on the eXtreme gradient boosting (XGBoost) trees. Fuel 282, 118783 (2020).
21. Hyndman, R. J. & Koehler, A. B. Another look at measures of forecast accuracy. Int J Forecast 22, 679–688 (2006).
22. Papavasileiou, K. D., Moultos, O. A. & Economou, I. G. Predictions of water/oil interfacial tension at elevated temperatures and pressures: A molecular dynamics simulation study with biomolecular force fields. Fluid Phase Equilibrium 476, 30–38 (2018).

Extracting the Biomarker Potential of N-glycans with a Machine Learning Framework Applied to Colorectal Cancer

Joseph Davies, Shoh Nakai
Department of Chemical Engineering, Imperial College London, U.K.

Abstract Aberrant glycosylation is considered to be a hallmark of colorectal cancer. It remains a challenge, however, to effectively diagnose colorectal cancer patients and to identify well-validated and clinically useful biomarkers that can be exploited in the development of advanced therapies. Here, a framework for a machine learning approach that considers multiple algorithms and scaling methods for colorectal cancer diagnosis and patient stratification is proposed and potential biomarkers are identified. Using data on the glycomic profiles, age, gender and BMI for 1413 colorectal cancer patients and 538 controls, a Soft-Voting Ensemble binary classifier was trained and achieved a mean Area Under the Curve (AUC) of 0.723. An XGBoost model trained on the dataset augmented using the Synthetic Minority Oversampling Technique (SMOTE) achieved a mean AUC of 0.920, with a sensitivity of 99.9% and specificity of 43.4%. Further, a Random Forest multiclass model trained on non-augmented data classified controls and early-stage cancer patients with an AUC of 0.701, while classifying controls and late-stage cancer patients with an AUC of 0.729. Permutation feature importance analysis indicated that changes in the core-fucosylation of IgG glycans are a potential biomarker, corroborating earlier work into the glycosylation traits associated with colorectal cancer.

1. Introduction

On a global basis, colorectal cancer (CRC) is the third most commonly diagnosed cancer (10.2% of total cases in 2018) and the second most common cause of cancer-related deaths (9.2% of total cancer deaths in 2018) (1). The incidence rate of CRC is increasing year-on-year and in the coming years, its mortality rate is projected to overcome that of heart diseases, which are the leading cause of death globally (2, 3). Current screening methods for CRC have limitations in terms of invasiveness, low sensitivity and high costs (4, 5). The success of conventional treatments, such as surgery, chemotherapy and radiation therapy, has been limited by the failure to diagnose CRC at an early stage, indicating the need to identify specific molecular targets for the development of more effective diagnostic procedures and therapies (4).

Understanding of the molecular basis of CRC has been improved by advancements in genomic and proteomic studies. Even so, the identification of well-validated and clinically useful biomarkers for CRC has been relatively scarce (5). The emerging field of glycomics is a promising direction of study and has gained traction in cancer research. While it appears that glycans play a role in tumour proliferation, the understanding of the mechanisms that drive this lags significantly behind that of other key cell components, namely genes and proteins (10). Aberrant glycosylation is considered to be a hallmark of CRC, and it has been suggested that human serum N-glycans may serve as an important biomarker for diagnosis and the development of advanced therapeutic intervention (6). Of particular interest are N-glycans found on immunoglobulin G (IgG), a glycoprotein abundantly present in human serum (7). In fact, in healthy adults, IgG constitutes approximately 75% of total serum immunoglobulins (8). The ease of sample collection, coupled with the indications of CRC's associations with aberrant glycosylation, makes IgG N-glycans an excellent candidate to exploit in diagnosis and the discovery of biomarkers (7, 9).

In this study, the dataset obtained from the Study of Colorectal Cancer in Scotland (SOCCS) study, which includes the glycome profiles of cancer patients and controls, is used to propose a framework for a machine learning approach to CRC diagnosis and patient stratification. The framework proposed here presents a supplementary tool to existing screening methods for CRC. The hope is that the framework proposed is general and can be extended to the diagnosis of diseases beyond CRC using similar glycomic data. Further, by investigating the features that drive classification, this study aims to identify potential biomarkers for CRC to motivate future research into the pathophysiology of CRC and the development of more effective diagnostic procedures and advanced therapies.

2. Background

Since 2017, the field of machine-learning-based disease diagnosis has seen considerable growth in the number of journal publications and holds great promise towards developing inexpensive and time-efficient diagnostic procedures (11). Supervised machine learning algorithms use optimization with statistical and probabilistic methods to detect patterns in labelled data to make predictions on new unlabeled datasets (9). In the past, machine learning has been demonstrated to successfully diagnose various diseases, including breast cancer using image data (12) and diabetes using clinical and gene expression data (13). In the clinical setting, prostate cancer is among the diseases that are routinely diagnosed using the help of machine learning with image data from CT scans (14, 15). More recently, the oxygen needs of COVID-19 patients have been predicted using machine learning at over 20 hospitals (16). Regarding the algorithms used in machine-learning-based disease diagnosis studies, Uddin et al (17) found that Support Vector Machines are applied most frequently, while Random Forest showed superior accuracy in 53% of the studies where it was applied.

Recent advancements into the glycosylation traits associated with specific diseases have opened up avenues for disease diagnosis using glycomic data – specifically, the relative abundances of N-glycans that decorate the immunoglobulin G (IgG) antibody present in human serum. Hepatocellular carcinoma (18), urological diseases (19), liver fibrosis (20) and gastrointestinal cancers (esophageal, gastric and colorectal cancer) (21) are examples of diseases that have been demonstrated to be identifiable using machine learning models trained on glycomic data.

Previous studies have exploited the indications that glycosylation is aberrant in colorectal cancer patients. Vučković et al (9) built a regularized Logistic Regression model using data from the Study of Colorectal Cancer in Scotland (SOCCS) study (1999-2006), a case-control study performed by clinicians in Scotland. Vučković et al (9) used the Area Under the Receiver Operating Characteristic Curve, or AUC, as the main performance metric. This is commonly used to estimate the predictability of a classifier, whereby an AUC of 1 corresponds to a model that predicts class labels perfectly, while an AUC of 0.5 corresponds to a model that has no discriminative power (17). While a model based only on age and sex did not show success in discriminating between cancer patients and controls (AUC = 0.499), adding the glycome profiles of samples to the training set increased the discriminative power of the model considerably (AUC = 0.755). Further, the study found that CRC associates with a decrease in IgG galactosylation, IgG sialylation and an increase in core-fucosylation of neutral glycans with a concurrent decrease of core-fucosylation of sialylated glycans.

The present study builds on the earlier work of Vučković et al (9) by applying a wider range of algorithms to a modified version of their dataset – our version of this dataset has 653 more cancer patients. This study also provides an investigation into the impact of various scaling methods on model performance, which previous studies into machine-learning-based disease diagnosis have largely failed to do (22). Further, this study investigates the potential of patient stratification using glycomic data. Finally, while univariate statistical tests are commonly used for biomarker discovery, a model-agnostic version of permutation feature importance is used here to capture multivariate interactions of the glycan features.

3. Methods

In this study, a dataset from the Study of Colorectal Cancer in Scotland (SOCCS) study (1999-2006) was used. This includes the composition of 24 glycan profiles normalized by summing all peak areas to 100%, body mass index (BMI), as well as the known glycan covariates, age and gender, for 1413 patients with pathologically confirmed colorectal adenocarcinoma and 538 matching controls (9). Limitations of the dataset include missing BMI data for 238 samples and a significant imbalance in the number of controls and cancer patients for ages over 60 (719 cancer patients vs. 4 controls). All computational work was conducted using Python, with the scikit-learn library being used extensively (23).

3.1 Pre-processing

For binary classification, the class labels were encoded into numerical values based on whether the sample was from a control or a cancer patient. For multiclass classification, the class labels were encoded into three numerical values based on whether the sample was from a control, an early-stage patient (stage 1 or 2 CRC) or an advanced-stage (stage 3 or 4 CRC) patient, as identified by the SOCCS dataset. Similarly, data on the gender of each sample were encoded into two numerical values. To overcome the problem of missing BMI data for certain samples, data imputation was employed. Specifically, an iterative imputer that incorporates all features to create a regression model in a round-robin fashion was used to achieve this.

Imbalanced datasets are known to pose a challenge to machine learning algorithms when learning minority class concepts (24). In this regard, data from samples over the age of 60 were initially discarded, yielding 694 cancer patients and 534 controls. Upon identifying that more data would improve model performance, data augmentation was employed for the binary classification to address the class imbalance observed among the samples for ages over 60. In doing so, the Synthetic Minority Over-Sampling Technique (SMOTE), one of the most prominent methods in the literature to address imbalanced classification problems (25), was used. SMOTE works by first selecting a minority class instance, a, at random and finds the k-nearest neighbours to a ($k = 5$ for this study). One of the k-nearest neighbours, b, is then selected at random and a synthetic instance is generated as a linear combination of a and b. The SOCCS dataset contains no control samples over the age of 74, and to avoid extrapolation, samples over the age of 74 were discarded before data augmentation was performed. Age, gender and, to a lesser extent, BMI, are known covariates of human glycome profiles (26, 27), and to ensure that model classification decisions were attributable to differences in glycome profiles, as opposed to differences in the covariates, these three features were controlled in applying SMOTE. This was achieved by firstly keeping the proportion of males to females approximately constant. Secondly, the BMIs of the synthetic instances were not generated by SMOTE, but rather, the median of the BMI of the non-augmented dataset (after BMI imputation) was added. For the 61 to 71 age range, the data was augmented to match exactly the number of controls and cancer patients for each age. For the 72 to 74 age range, the data was augmented so that the number of controls for each age would be half the number of cancer patients. This sampling strategy was employed in

an attempt to have no statistically significant differences in the underlying distributions for age, gender and BMI across cancer patients and controls.

Statistical hypothesis testing was performed before and after data augmentation to assess whether there was a statistically significant difference in the underlying distributions for age, gender and BMI across cancer patients and controls. The null hypothesis for each test was that there was no statistically significant difference in the underlying distributions of the features between controls and cancer patients at a 95% confidence interval. For all features except for gender, the Mann-Whitney U test (28) was carried out. For gender, a categorical variable, the Fisher's exact test (29) was used. On all datasets used, the null hypothesis was accepted for the known covariates, age and gender.

Data scaling methods are known to greatly influence the performance of machine learning algorithms (22, 30). To that end, this study evaluates the performance of each machine learning algorithm using 3 scaling methods: Min-Max, Standard and Robust scaling. These scaling methods were selected as they have been demonstrated to perform well with the machine learning algorithms used in this study (22). The Min-Max scaling method scales each feature so that all values lie between 0 and 1 (31). Min-Max scaling has been demonstrated to show particularly strong performance when used with the Support Vector Machines algorithm (32). The Standard scaling method involves subtracting each value from the mean and dividing by the standard deviation (31). Standard scaling has been noted to work especially well with the Random Forest algorithm (33). In the presence of outliers, the mean and standard deviation calculated using the Standard scaling method may be skewed. An approach to remove the effect of outliers in the scaling process is to use the Robust scaling method, which scales the data by subtracting the median of a feature and dividing by the interquartile range (31).

3.2 Machine Learning Algorithms

In this study, five machine learning algorithms were explored: Random Forest (RF), Support Vector Machines (SVMs), Logistic Regression (LR), XGBoost (XGB) and Soft-Voting Ensemble (SVE). RF and SVMs are robust and widely used algorithms for disease diagnosis (34), while LR has shown success with the original SOCCS dataset (9). XGB was implemented in this study as previous studies have shown that tree-based ensemble algorithms, such as XGB, are widely accepted as the recommended option for real-life tabular data (35). XGB specifically has generated increased interest due to its success in recent machine learning challenges (36).

The machine learning algorithms used in this study are largely very different but share some similarities. Firstly, RF is a bagging-based ensemble algorithm consisting of decision tree classifiers, where each decision tree casts a unit vote based only on a randomly-selected subset of features, and a new instance is classified based on the majority vote (37). For binary classification, the SVMs algorithm works by finding a hyperplane which separates the d-dimensional data into its two classes. For data that is not linearly separable, the SVMs algorithm casts the data into a higher dimensional space via a kernel function where the data is separable (38). LR is an extension of Linear Regression. The linear equation used to describe the relationship between features and outcomes in Linear Regression is wrapped into the exponential of the logistic function so that the output of the model is a number between 0 and 1 (39). By selecting a threshold between 0 and 1, the LR algorithm can be used for classification problems. XGB is similar to the RF algorithm as it is also a decision-tree-based ensemble algorithm. Unlike RF, however, the algorithm is based on an extreme gradient boosting approach, wherein trees are added to the ensemble one at a time and a gradient descent algorithm is used to minimise errors in subsequent models (40, 41). To balance out the weaknesses of individual classifiers, such as overfitting, the SVE was explored. Much like RF and XGB, SVE is an ensemble algorithm that comprises more than one base classifier. The probabilities of the class predictions from each base classifier are used to calculate a weighted average and produce a final class prediction (42). In this study, for binary classification, base models of LR with Robust scaling and XGB with Min-Max scaling were used to generate an equally-weighted SVE. Similarly, for multiclass classification, an equally-weighted SVE comprising of RF with Min-Max scaling and LR with Robust Scaling base classifiers was built.

3.3 Model Tuning and Evaluation

Given the limited data available from the SOCCS study, a separate test set was not created. Instead, nested cross-validation (NCV) was employed. Using non-nested cross-validation, hyperparameter tuning is performed to maximize a model's performance on a given validation set. The evaluation of the model's performance is then performed on the same validation set. In this way, information about the validation set may 'leak' to the model during hyperparameter tuning leading to overfitting. As such, a model selection process using non-nested cross-validation will often lead to overly optimistic estimates of a model's generalization ability on unseen data (43). This can be avoided without a test set by using nested cross-validation. While nested cross-validation is a computationally expensive procedure, Tsamardinos et al (44) demonstrated that the AUC bias using non-nested cross-validation is more apparent with smaller sample sizes. In order to propose a generalized framework for future applications that can be extended to diseases other than colorectal cancer, where data availability may be even more limited, it was concluded that nested cross-validation was worthwhile.

Nested cross-validation incorporates tuning of hyperparameters in an inner loop and model training and evaluation in an outer loop (43). In the inner loop, a stratified 5-fold cross-validation was used, whereby the full data is shuffled and split into five folds, with one fold held out as the validation set. The model would train on four folds and validate on one fold. In the following evaluation, a different fold is used as the validation set. The procedure is repeated until all folds have been used as the validation set. With each iteration in the inner loop, a particular configuration of hyperparameters is selected, and the mean AUC across five folds is used as the performance metric in determining the best configuration. The configured model is then fed into the outer loop, which involved a stratified 10-fold cross-validation procedure. The outer loop follows a similar procedure to the inner loop, but ten folds are used and only the best hyperparameter configuration is used. For model selection, the mean AUC score across the ten folds is used as the primary model evaluation metric. Stratified folds were used to ensure that a similar proportion of observations with a given categorical value are present in each fold (45).

In the inner loop, a random search or an exhaustive grid search is often performed over a pre-defined hyperparameter search space. In this study, three inner loops were used. Firstly, the relevant hyperparameters and a range of hyperparameter configurations were identified. The best model configuration in the first inner loop was used as the initial estimate of the best configuration. In the second inner loop, a grid search was employed over a search space that ranged ± 15% around the initial estimate for each hyperparameter. In the third inner loop, a final grid search was employed over a search space that ranged ± 10% around the best hyperparameter values from the previous inner loop. For the sake of runtime, non-numerical hyperparameters determined in the first inner loop were not investigated in subsequent inner loops. By using this method, a greater range of model configurations was tested in a time-efficient manner.

The Receiver Operating Characteristic (ROC) curve captures the trade-off between the true positive rate and false positive rate that exists as the classification threshold is varied. As such, the AUC score quantifies a model's performance across different classification thresholds. For a model to be used in clinical practice, a classification threshold must be specified. This threshold was selected using an adjusted method of Youden's J statistic approach. In its raw form, Youden's J statistic is calculated at various classification thresholds, where each threshold corresponds to a point along the ROC curve. The classification threshold corresponding to the maximum J statistic is selected to represent the classifier's optimal operating threshold (46). An adjusted version of this method was

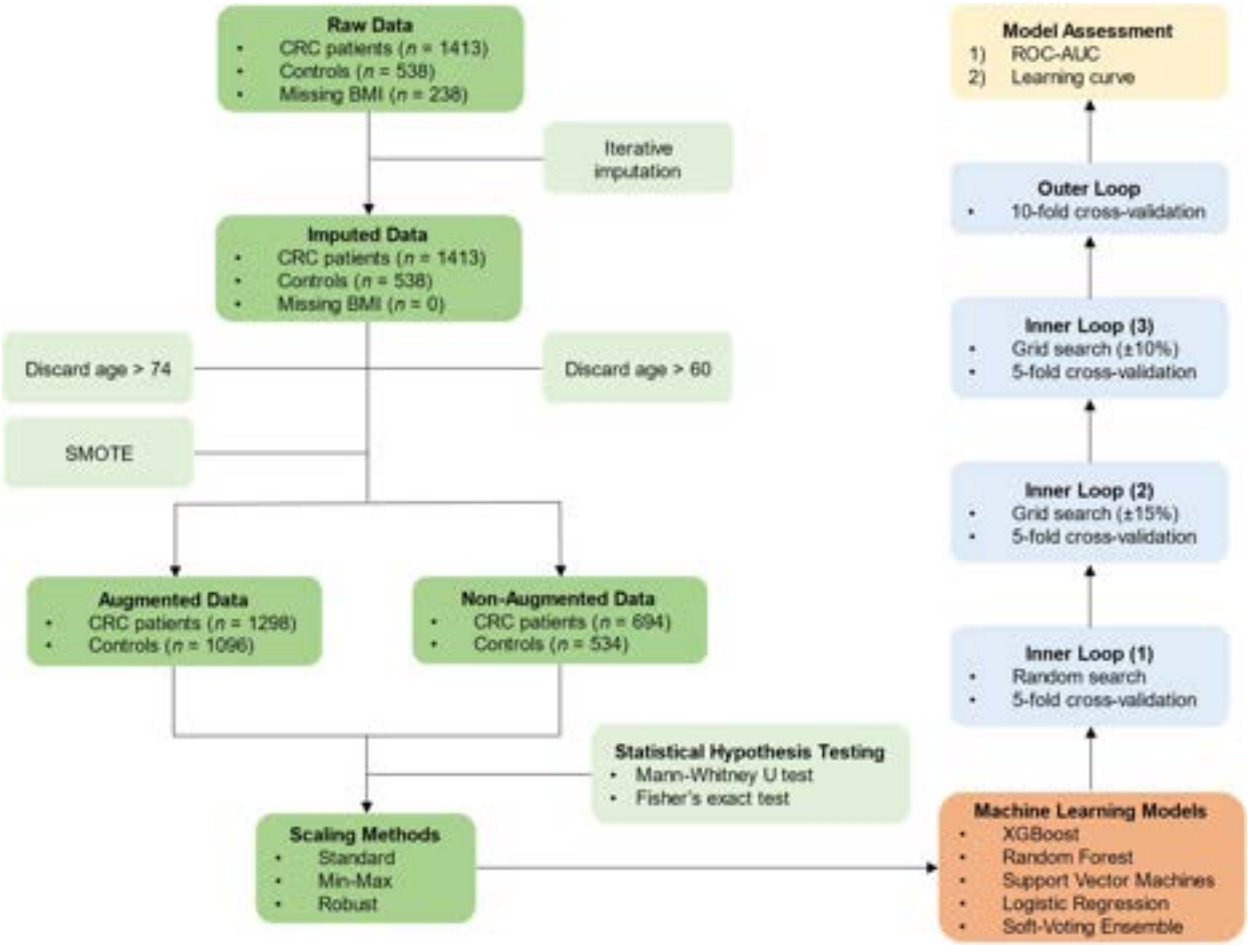


Figure 1. Schematic of the machine learning framework specific to the data from the SOCCS study.

incorporated by setting a minimum true positive rate, and so a maximum false negative rate, for which to calculate the J statistic within. Favouring high true positive rate is generally desired in a clinical setting (47) as incorrectly classifying a person as negative would limit a patient's access to treatment. The adjusted method used here accounts for this.

3.4 Feature Importance

Univariate hypothesis testing, which is widely used for biomarker discovery in bioinformatics, fails to capture the complex variable interactions inherent to biological processes, and can only identify features important to the output variable in isolation from the other inputs (48). In this study, a model-agnostic version of the permutation feature importance measurement, capable of capturing multivariate interactive effects between features, is used to identify features that are 'important' for the SVE model's predictions on the non-augmented dataset (49). The importance of a feature is measured by calculating the increase in the model's prediction error upon permuting the feature. A feature is deemed 'important' if shuffling its values increases the model error, and 'unimportant' if shuffling its values does not change the model error (49). The feature importance is quantified as the model prediction error upon permuting less the original model prediction error.

4. Results

Data on the features (age, gender, BMI and glycome compositions) of 1413 patients with pathologically confirmed colorectal adenocarcinoma and 538 matching controls were used to train four base machine learning algorithms and one SVE. Three scaling methods were tested for each base algorithm and SMOTE data augmentation was used for binary classification only. Further, the 'importance' of an input feature was quantified with permutation feature importance.

4.1 Binary Classification

The best scaling methods for each of the machine learning algorithms tested, and the corresponding mean AUC scores are summarized in Table 1. Among the models trained on the non-augmented dataset, excluding the SVE, XGB had the highest mean AUC score of 0.723. However, as illustrated in Figure 2 (b), the learning and validation curves show poor convergence, which is indicative of overfitting on the training dataset and poor generalization ability on unseen data. While LR had a worse mean AUC score of 0.696, the learning curve in Figure 2 (a) for LR demonstrates relatively good convergence between the training and validation curves compared to XGB. These observations motivated the decision to build an SVE comprising LR and XGB base classifiers. Among the combinations of machine learning algorithms and scaling methods tested, the SVE using Min-Max scaling for the XGB base model and Robust scaling for the LR base model showed the highest mean AUC of 0.727. Additionally, the learning curve for the SVE in Figure 2 (c) showed improved convergence. For the models trained on the augmented dataset, the RF model using the Standard scaling method showed the highest mean AUC score of 0.921. The XGB model using the Robust scaling method closely followed, with a mean AUC score of 0.920. While the difference in mean AUC scores is marginal, the learning curves for XGB showed far better convergence for the training and validation curves, as illustrated in Figure 2 (d).

The ROC-AUC curves shown in Figure 3 (a) and (b) illustrate the performance of the SVE and XGB models, trained on the non-augmented and augmented datasets respectively, at different classification

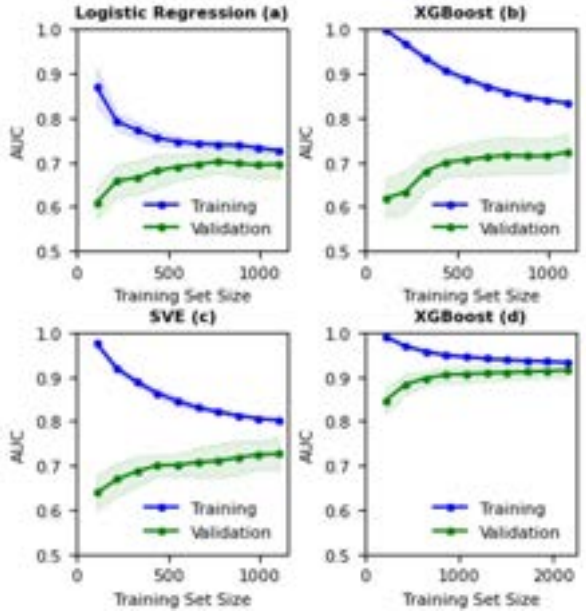


Figure 2. Learning curves: (a) LR with non-augmented data, (b) XGB with non-augmented data, (c) SVE with non-augmented data and, (d) XGB with augmented data, where the shaded region around each curve corresponds to one standard deviation away from the plotted mean (of 10 cross-validation folds).

Table 1. Mean AUC and standard deviation (std) results for the best-performing scaling method for each algorithm and the SVE model (comprising of XGB and LR) for augmented and non-augmented input binary datasets. Note that the scaling method of the SVE is determined by each base classifier.

	Non-Augmented			Augmented		
Algorithm	**Scaling Method**	**Mean AUC**	**Std**	**Scaling Method**	**Mean AUC**	**Std**
XGB	Min-Max	0.723	0.044	Robust	0.920	0.016
RF	Robust	0.722	0.043	Standard	0.921	0.017
SVM	Robust	0.702	0.030	Robust	0.861	0.018
LR	Robust	0.696	0.033	Standard	0.724	0.033
SVE	-	0.727	0.038	-	0.900	0.038

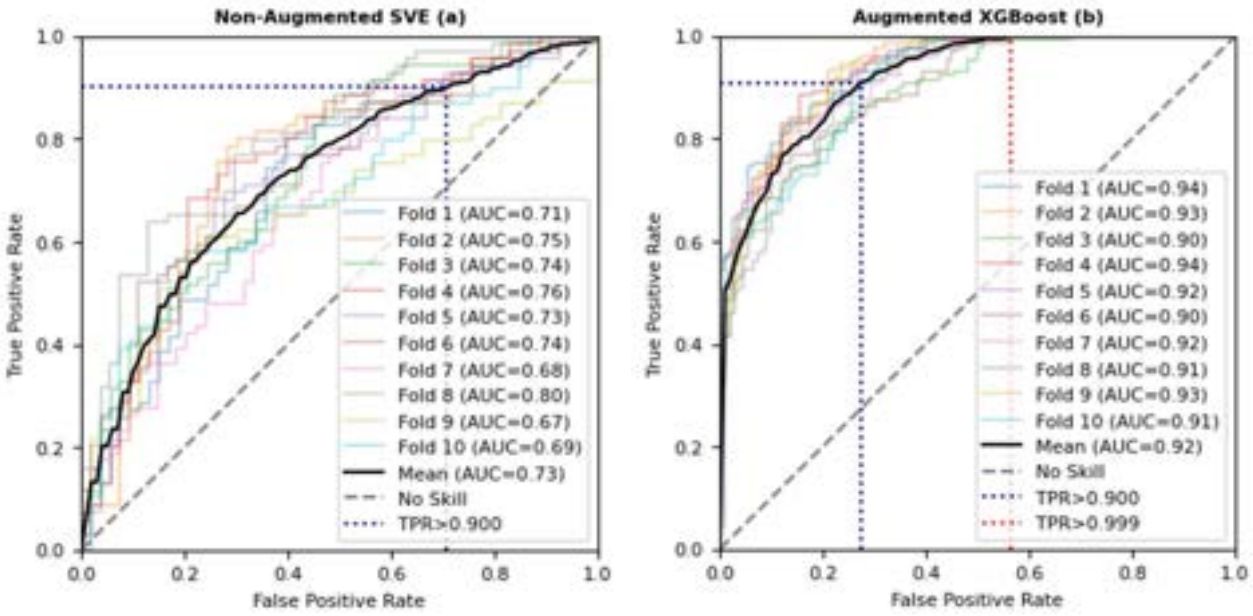


Figure 3. ROC-AUC curves for each outer loop cross-validation fold for: (a) the SVE model trained on non-augmented data and, (b) the XGB model trained on augmented data (as per Table 1), where "TPR" corresponds to "True Positive Rate".

thresholds. These curves were useful in determining potential operating points for the classifier, i.e., what thresholds might be desirable given their sensitivity and specificity. For the SVE model trained on the non-augmented dataset, a minimum 90.0% sensitivity threshold was selected to give a sensitivity of 90.1% and a specificity of 29.3%, at a classification threshold of 46.1%. The strong performance of the XGB model trained on the augmented dataset enabled a minimum 99.9% sensitivity threshold to be specified. This yielded a sensitivity of 99.9% and a specificity of 43.4% at a classification threshold of 51.0%.

4.2 Multiclass Classification

For multiclass classification, three mean AUC scores are calculated to illustrate a model's ability to discern between controls and early-stage patients, controls and late-stage patients, and early-stage and late-stage patients. The best combinations of algorithms and scaling methods were identified by calculating the average of the three mean AUC scores (average mean AUC), and are summarized in Table 2. The SVE model used for multiclass classification comprises two of the classifiers that had the highest average mean AUC scores, namely the RF and LR models. Although the SVE achieved the best mean AUC for controls and late-stage patients, overall it still performed worse than the base RF classifier. Notably, due to a statistically significant difference in the age distributions between early-stage and late-stage patients upon data augmentation, this augmented dataset was not used.

4.3 Feature Importance

The permutation feature importances derived from the SVE's predictions on the non-augmented dataset are shown in Table 3. The SVE model was used to calculate feature importances as it was the best-performing model on the non-augmented data. The top five 'important' features are core-fucosylated, supporting past indications that increased core-fucosylation of N-glycans on IgG is a hallmark of colorectal cancer.

Table 3. Top 4 and bottom 4 ranked features based on their mean importance score, where "std" corresponds to "standard deviation". The core-fucosylated glycan structures were identified using Figure 1 in the study by Vučković et al (9).

		Fucosylated	Importance	
Rank	Feature	Glycan?	Mean	Std
1	GP14	Yes	0.286	0.009
2	GP18	Yes	0.263	0.009
3	GP10	Yes	0.235	0.012
4	GP9	Yes	0.232	0.007
24	GP5	No	0.060	0.010
25	GP22	No	0.049	0.009
26	Age	N/A	0.016	0.003
27	Gender	N/A	0.001	0.001

Table 2. Mean AUC and standard deviation (std) results for the best-performing scaling method for each algorithm and the SVE model (comprising of RF and LR) for the non-augmented input multiclass datasets. Note that the scaling method of the SVE is determined by each base classifier.

		Control and Early-Stage		Control and Late-Stage		Early and Late-Stage		Average Mean
Algorithm	Scaling Method	Mean AUC	Std	Mean AUC	Std	Mean AUC	Std	AUC
XGB	Robust	0.667	0.068	0.701	0.042	0.521	0.077	0.630
RF	Min Max	0.701	0.045	0.729	0.034	0.637	0.061	0.689
SVMs	Min Max	0.671	0.070	0.691	0.054	0.478	0.034	0.613
LR	Robust	0.671	0.057	0.700	0.043	0.527	0.072	0.633
SVE	-	0.699	0.048	0.731	0.022	0.593	0.062	0.674

5. Discussion

5.1 Binary Classification

The best-performing model trained on the non-augmented dataset was the equally-weighted SVE, comprising LR and XGB models. This model has a mean AUC of 0.727, which, at a mean classification threshold of 46.1%, has a sensitivity of 90.1% and a specificity of 29.3%. The mean AUC for the SVE is worse than that achieved by Vučković et al (9) (AUC = 0.755), wherein an LR model was trained on a dataset obtained from the SOCCS study. The LR model trained in the present study has a mean AUC score of 0.696. The inferior performance of the LR model in this study may be attributed to the differences in the dataset that was used. Namely, our non-augmented dataset did not include samples over the age of 60, upon observing a significant class imbalance over this age. This yielded 694 cancer patients and 534 controls. Vučković et al (9) made no mention of discarding data and instead trained an LR model on a larger dataset of 760 CRC patients and 538 controls. So while the AUC score is higher for the model built by Vučković et al (9), it may be that the model will be less successful in classifying the minority class (controls) from unseen data. Most notably, the study by Vučković et al (9) used a 10-fold non-nested cross-validation procedure, as opposed to a nested cross-validation procedure. It is known that a non-nested cross-validation procedure may result in overly optimistic AUC scores (43). As such, it may be that the AUC score of 0.755 reported by Vučković et al (9) is a slight overestimate of the model's ability to generalise on unseen data.

Discarding samples over the age of 74 and performing data augmentation using SMOTE resulted in a dataset with 1298 cancer patients and 1096 controls. The XGB model, trained on the augmented dataset, using the Robust scaling method has a mean AUC score of 0.920. Similarly, high mean AUC scores have been achieved in past studies using glycomic data and a 5-fold non-nested cross-validation to predict diseases. In a study conducted by Iwamura et al (19), a neural network model, applied to the diagnosis of urological diseases, achieved a mean AUC score of 0.970, while in a study by Huang et al (18), an LR model achieved a mean AUC of 0.860 in predicting alpha-fetoprotein negative hepatocellular carcinoma. Outside of glycobiology, Dayan et al (16) achieved a mean AUC of 0.920 which crucially was considered sufficient to implement in a clinical setting to predict the oxygen requirement of symptomatic COVID-19 patients. At a mean classification threshold of 51.0%, our XGB model trained on the augmented dataset gives a sensitivity of 99.9% and a specificity of 43.4%. In contrast, colonoscopy, considered to be the gold standard for colorectal cancer screening, has a lower sensitivity but higher specificity of 92.5% and 73.2% respectively (50). Colonoscopy is generally performed under general anaesthesia and by trained professionals, and while uncommon, serious complications such as colon perforation and bleeding may occur. Meanwhile, the proposed test relies on a minimally invasive blood test. The ability of the XGB model to achieve such high sensitivity is significant in a clinical setting, as erroneously classifying a sample as negative may lead to severe consequences. This is especially significant for colorectal cancer where a prompt diagnosis is critical. The augmented dataset used to train the XGB model is indeed synthetic and thus, the framework proposed here cannot directly proceed to clinical trials. However, the high mean AUC score (AUC = 0.920) achieved using the augmented dataset is a hopeful indication that, given more samples to train and validate on, the machine learning framework proposed in this study can be used in conjunction with traditional diagnostic tests. By doing so, the framework would offer an inexpensive and time-efficient means to increase the reliability of diagnoses.

Significant variation in mean AUC is observed in both the augmented and non-augmented results. As illustrated in Table 1, the best-performing scaling method depends on the machine learning algorithm being used, confirming the earlier work of Ahsan et al (22), who demonstrated that a model's performance varies depending on the data scaling method. Furthermore, for the non-augmented dataset, three out of the four machine learning algorithms (excluding the SVE) tested favoured the Robust scaling method. Given the Robust scaling method is an effective way to remove the impact of outliers on classification, this may be an indication of significant outliers in the non-augmented dataset. While the XGB model has been overlooked in previous glycomic studies, the strong performance with both the non-augmented and augmented datasets indicates the model should continue to be investigated in future studies. Further to this, of the five models presented in Table 1, the three models with the highest mean AUC (XGB, RF and SVE) are ensemble models for both the non-augmented and augmented datasets. This is somewhat to be expected, as ensemble methods are designed to consider the outputs of several individual models, and combine to improve overall classification performance (51). Even so, the XGB model was observed to overfit the non-augmented dataset in Figure 2 (a). It is speculated that the increased complexity associated with the extreme gradient boosting approach makes the XGB model more susceptible to overfitting compared to the non-ensemble-based models tested here. Machine learning models are known to overfit small datasets (52), and this is reflected by the fact that the XGB model does not seem to overfit the larger augmented dataset, as can be seen in Figure 2 (d).

5.2 Multiclass Classification

Across the three classification categories presented in Table 2, the RF model showed the best performance with an average mean AUC of 0.689. Past studies using

the RF classifier for multiclass classification have also shown success. Specifically, a study into the patient stratification for lymph diseases recorded an average mean AUC score of 0.954 across four classification categories using the RF model (53). XGB showed strong performance with binary classification and in past multiclass classification studies (54). However, here, the XGB model showed poor performance with multiclass classification. Admittedly, the inclusion of additional techniques that helped achieve its success in literature, such as transfer learning, was not explored in our methodology. The comparative success of RF could be attributed to the fact that SVMs and LR models can only handle binary outputs natively. Therefore, the native SVMs and LR models require settings to be adapted to address multiclass classification problems (55).

It can be observed that the mean AUC scores shown for the multiclass problem in Table 2 are, in general, lower than those for the binary problem using the non-augmented dataset in Table 1. This is because it is widely acknowledged that classification problems become more challenging as the number of classes increases (55). Moral et al (55) argue that the additional complexity in multiclass classification comes from the heterogeneity of decision boundaries which gives rise to the interaction of decision boundaries.

The results in Table 2 also highlight the somewhat intuitive point that discerning between controls and late-stage patients is easier than discerning between controls and early-stage patients or early-stage and late-stage patients. These observations are consistent with the fact that poorer CRC prognosis (i.e. late-stage patients) is associated with increased IgG pro-inflammatory activity (56), which in turn is associated with the presence of certain glycan structures (7). Namely, poor prognosis has been associated with decreased galactosylation and decreased sialylation of fucosylated IgG glycans (56). We speculate that the multiclass models have exploited these glycosylation traits associated with late-stage patients to make better predictions with the control and late-stage classification category.

5.3 Feature Importance

The relative abundances of GP14, GP18, GP10 and GP9 were found by the SVE model trained on non-augmented data to be the most important. Beyond the top 4 shown in Table 3, the top 11 most 'important' features for the SVE classifier in making predictions contain at least one core-fucosylated glycan structure. This is particularly notable given that 15 of the 24 glycan features in this dataset contain core-fucosylated structures. Indeed, several cancer types are known to be associated with increased core-fucosylation on serum proteins (6). Further, the result here corroborates the earlier work of Vučković et al (9), which found that colorectal cancer is associated with an increase in core-fucosylation of neutral glycans. Interestingly, non-core-fucosylated glycans were deemed among the least 'important', meaning the SVE model finds non-core-fucosylated glycans less informative when making classification predictions. As such, we speculate that the changes in the relative abundance of core-fucosylated IgG N-glycans may serve as a potential biomarker for colorectal cancer. For clinical use, potential biomarkers must undergo validation processes. To that end, high-throughput methods for quantitative clinical glycan biomarker validation exist (57).

We further speculate that using data on the relative abundances of core-fucosylated glycans only may increase the discriminative power of the SVE model. In past studies, it has been demonstrated that various machine learning models, including RF and XGB, show improved accuracy by using only the highest-ranked features from the feature importance technique (58).

6. Conclusions

Key findings from this study support the existing potential for a machine-learning-based diagnostic tool for colorectal cancer using glycomic data, involving a minimally invasive procedure. The XGB model trained on the augmented dataset, in particular, achieves a high mean AUC score of 0.920. While this result is limited by the use of data augmentation, it provides a proof of concept for a supplementary diagnostic tool that can be developed given a clinically valid dataset larger than what is currently available. Therefore, future research on the diagnostic potential of glycomic data would involve acquiring more data points on controls and colorectal cancer patients. Such glycomic data should be accompanied by clinically relevant features, including age, gender and BMI. Indeed, the framework proposed here is designed to be extendable for other diseases where glycomic data may be relevant. Moreover, given aberrant glycosylation is observed with multiple cancer types, this framework has potential to be used as a universal cancer screening tool. While data augmentation was not explored for the multiclass classification, modifying the method used in this study to account for the statistical significance observed in the underlying distribution for age across early-stage and late-stage patients would enable a similar proof of concept study for patient stratification to be performed.

By investigating the features that drive classification decisions for the SVE model trained on the non-augmented dataset, it is speculated that changes in the core-fucosylation of glycans present on immunoglobulin G may serve as a potential biomarker for colorectal cancer. Further understanding of the molecular mechanism underlying core-fucosylation of glycans may unlock the ability to develop advanced therapies that target specific glycosylation pathways. Overall, this study supports the existing biomarker potential of N-glycans for colorectal cancer and directs future studies to further evaluate the feasibility of developing a machine-learning-based disease diagnostic tool for colorectal cancer.

7. Supplementary Information

The Python code of our framework and accompanying data for this work will be available upon request.

8. Acknowledgements

We would like to thank Konstantinos Flevaris for his invaluable guidance and insight throughout this project.

9. References

(1) Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;68(6):394-424.

(2) Mármol I, Sánchez-de-Diego C, Dieste AP, Cerrada E, Yoldi MJR. Colorectal carcinoma: A general overview and future perspectives in colorectal cancer. *Int J Mol Sci*. 2017;18(1):197.

(3) Juan José Granados-Romero, Valderrama-Treviño AI, Contreras-Flores EH, Barrera-Mera B, Enríquez MH, Uriarte-Ruíz K et al. Colorectal cancer: a review. *Int J Res Med Sci Technol*. 2017;5(11):4667-76.

(4) Wanga D, Madunićab K, Guinevere TZ, Lageveen-Kammeijera S, Wuhrer M. Profound diversity of the N-glycome from microdissected regions of colorectal cancer, stroma, and normal colon mucosa. *Engineering*. 2022.

(5) Balog CIA, Stavenhagen K, Fung WLJ, Koeleman CA, McDonnell LA, Verhoeven A et al. N-glycosylation of colorectal cancer tissues. *Mol Cell Proteomics*. 2012;11(9):571–85.

(6) Pinho SS, Reis CA. Glycosylation in cancer: Mechanisms and clinical implications. *Nat Rev Cancer*. 2015;15(9):540-55.

(7) Flevaris K, Kontoravdi C. Immunoglobulin G N-glycan biomarkers for autoimmune diseases: Current state and a glycoinformatics perspective. *Int J Mol Sci*. 2022;23(9):5180.

(8) Kdimati S, Mullins CS, Linnebacher M. Cancer-cell-derived IgG and its potential role in tumor development. *Int J Mol Sci*. 2021;22(21):11597.

(9) Vučković F, Theodoratou E, Thaçi K, Timofeeva M, Vojta A, Štambuk J et al. IgG glycome in colorectal cancer. *Clin Cancer Res*. 2016;22(12):3078-86.

(10) Lauc G, Pezer M, Rudan I, Campbell H. Mechanisms of disease: The human N-glycome. *Biochim Biophys Acta*. 2016;1860(8):1574-82.

(11) Ahsan MM, Luna SA, Siddique Z. Machine-learning-based disease diagnosis: A comprehensive review. *Healthcare*. 2022;10(3):541.

(12) Yao D, Yang J, Zhan X. A novel method for disease prediction: Hybrid of random forest and multivariate adaptive regression splines. *J Comput*. 2013;8(1):170-7.

(13) Yang J, Yao D, Zhan X, Zhan X. Predicting disease risks using feature selection based on random forest and support vector machine. In: Basu M, Pan Y, Wang J, editors. *Bioinformatics Research and Applications*; *Zhangjiajie, China*. Cham: Springer; 2014. p. 1-11.

(14) Microsoft. *A Microsoft AI tool is helping to speed up cancer treatment – and Addenbrooke's will be the first hospital in the world to use it* 2020 [cited 10th December 2022]. Available from: https://news.microsoft.com/en-gb/2020/12/09/a-microsoft-ai-tool-is-helping-to-speed-up-cancer-treatment-and-addenbrookes-will-be-the-first-hospital-in-the-world-to-use-it/.

(15) Jena R, Parkes M, Stranks A. *Medicine for members: Artificial intelligence in healthcare - improving patient outcomes.* [Presentation] Cambridge University Hospitals. 21-Jul 2021 [cited 10th December 2022]. Available from: https://www.cuh.nhs.uk/news/medicine-members-artificial-intelligence-in-healthcare/.

(16) Dayan I, Roth HR, Zhong A, Harouni A, Gentili A, Abidin AZ et al. Federated learning for predicting clinical outcomes in patients with COVID-19. *Nat Med*. 2021;27:1735-43.

(17) Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Inform Decis Mak*. 2019;19:281.

(18) Huang C, Fang M, Feng H, Liu L, Li Y, Xu X et al. N-glycan fingerprint predicts alpha-fetoprotein negative hepatocellular carcinoma: A large-scale multicenter study. *Int J Mol Sci*. 2020;149(3):717-27.

(19) Iwamura H, Mizuno K, Akamatsu S, Hatakeyama S, Tobisawa Y, Narita S et al. Machine learning diagnosis by immunoglobulin N-glycan signatures for precision diagnosis of urological diseases. *Cancer Sci*. 2022;113(7):2434-45.

(20) Scott DA, Wang M, Grauzam S, Pippin S, Black A, Angel PM et al. GlycoFibroTyper: A novel method for the glycan analysis of IgG and the development of a biomarker signature of liver fibrosis. *Front Immunol*. 2022;13:797460.

(21) Liu S, Liu Y, Lin J, Wang Y, Li D, Xie G-Y et al. Three major gastrointestinal cancers could be distinguished through subclass specific IgG glycosylation. *J Proteome Res*. 2022;21(11):2771-82.

(22) Ahsan MM, Mahmud MAP, Saha PK, Gupta KD, Siddique Z. Effect of data scaling methods on machine learning algorithms and model performance. *Technologies*. 2021;9(3):52.

(23) Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B. Scikit-learn: Machine learning in python. *J Mach Learn Res*. 2011;12:2825-30.

(24) Barua S, Islam MM, Murase K. A novel synthetic minority oversampling technique for imbalanced data set learning. In: Lu B-L, Zhang L, Kwok J, editors. *Neural Information Processing*; *Shanghai, China*. Berlin: Springer; 2011. p. 735-44.

(25) Elreedy D, Atiya AF. A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. *Inf Sci*. 2019;405:32-64.

(26) Perkovic MN, Bakovic MP, Kristic J, Novokmet M, Huffman JE, Vitart V et al. the association between galactosylation of immunoglobulin G and body mass index. *Prog Neuropsychopharmacol Biol Psychiatry*. 2014;48:20-5.

(27) Ding N, Nie H, Sun X, Sun W, Qu Y, Liu X et al. Human serum N-glycan profiles are age and sex dependent. *Age Ageing*. 2011;40(5):568-75.
(28) Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Statist*. 1947;18(1):50-60.
(29) Fisher RA. *Statistical methods for research workers*. Oliver and Boyd: Edinburgh; 1934.
(30) Shahriyari L. Effect of normalization methods on the performance of supervised learning algorithms applied to HTSeq-FPKM-UQ data sets: 7SK RNA expression as a predictor of survival in patients with colon adenocarcinoma. *Brief Bioinformatics*. 2019;20(3):985-94.
(31) scikit-learn. *Compare the effect of different scalers on data with outliers* 2022 [cited 12th December 2022]. Available from: https://scikit-learn.org/stable/auto_examples/preprocessing/plot_all_scaling.html.
(32) Ambarwari A, Adrian QJ, Herdiyeni Y. Analysis of the effect of data scaling on the performance of the machine learning algorithm for plant identification. *Jurnal RESTI*. 2020;4(1):117-22.
(33) Balabaeva K, Kovalchuk S. Comparison of temporal and non-temporal features effect on machine learning models quality and interpretability for chronic heart failure patients. *Procedia Comput Sci*. 2019;156:87-96.
(34) Pan Y, Zhang L, Zhang R, Han J, Qin W, Gu Y et al. Screening and diagnosis of colorectal cancer and advanced adenoma by bionic glycome method and machine learning. *Am J Cancer Res*. 2021;11(6):3002-20.
(35) Shwartz-Ziv R, Armon A. Tabular data: Deep learning is not all you need. *Inf Fusion*. 2022;81:84-90.
(36) Distributed (Deep) Machine Learning Community. *Awesome XGBoost* 2022 [cited 8th December 2022]. Available from: https://github.com/dmlc/xgboost/tree/master/demo#machine-learning-challenge-winning-solutions.
(37) Oshiro TM, Perez PS, Baranauskas JA. How many trees in a random forest? *Lect Notes Comput Sci*. 2012;7376:154-68.
(38) Boswell D. Introduction to support vector machines. *Department of Computer Science and Engineering University of California San Diego*. 2002.
(39) Stoltzfus JC. Logistic regression: A brief primer. *J Acad Emerg Med*. 2011;18(10):1099-104.
(40) Bentéjac C, Csörgő A, Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms. *Artif Intell Rev*. 2021;54:1937-67.
(41) Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Krishnapuram B, editor. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; *San Francisco, USA*. New York: Association for Computing Machinery; 2016. p. 785-94.
(42) Wang H, Yang Y, Wang H, Chen D. Soft-voting clustering ensemble. In: Zhou Z-H, Roli F, Kittler J, editors. *International Workshop on Multiple Classifier Systems*; *Nanjing, China*. Berlin: Springer; 2013. p. 307-18.
(43) Tsamardinos I, Greasidou E, Borboudakis G. Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation. *Mach Learn*. 2018;107:1895-922.
(44) Tsamardinos I, Rakhshani A, Lagani V. Performance-estimation properties of cross-validation-based protocols with simultaneous hyper-parameter optimization. In: Likas A, Blekas K, Kalles D, editors. *Artificial Intelligence: Methods and Applications*; *Ioannina, Greece*. Cham: Springer; 2014. p. 1-14.
(45) Zeng X, Martinez TR. Distribution-balanced stratified cross-validation for accuracy estimation. *J Exp Theor Artif Intell*. 2010;12(1):680-90.
(46) Kallner A. Interpretation of the elements of the ROC analysis. In: Fedor J, editor. *Laboratory statistics: Methods in chemistry and health sciences*. 2nd ed. Amsterdam: Elsevier; 2018.
(47) Ruopp MD, Perkins NJ, Whitcomb BW, Schisterman EF. Youden index and optimal cut-point estimated from observations affected by a lower limit of detection. *Biom J*. 2008;50(3):419-30.
(48) Huynh-Thu VA, Saeys Y, Wehenkel L, Geurts P. Statistical interpretation of machine learning-based feature importance scores for biomarker discovery. *Bioinformatics*. 2012;28(13):1766-74.
(49) Molnar C. Permutation feature importance. *Interpretable machine learning*. 2nd ed. Munich: Bookdown; 2022.
(50) Martín-López JE, Beltrán-Calvo C, Rodríguez-López R, Molina-López T. Comparison of the accuracy of CT colonography and colonoscopy in the diagnosis of colorectal cancer. *Colorectal Dis*. 2014;16(3):O82-O9.
(51) Sagi O, Rokach L. Ensemble learning: A survey. *Data Min Knowl Discov*. 2018;8(4):e1249.
(52) Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. *J Mach Learn Res*. 2014;15(56):1929-58.
(53) Azar AT, Elshazly HI, Hassanien AE, Elkorany AM. A random forest classifier for lymph diseases. *Comput Methods Programs Biomed*. 2014;113(2):465-73.
(54) Liew XY, Hameed N, Clos J. An investigation of XGBoost-based algorithm for breast cancer classification. *Mach Learn Appl*. 2021;6:100154.
(55) Moral PD, Nowaczyk S, Pashami S. Why is multiclass classification hard? *IEEE Access*. 2022;10:80448-62.
(56) Theodoratou E, Thaçi K, Agakov F, Timofeeva MN, Štambuk J, Pučić-Baković M et al. Glycosylation of plasma IgG in colorectal cancer prognosis. *Sci Rep*. 2016;6:28098.
(57) Shipman JT, Nguyen HT, Desaire* H. So you discovered a potential glycan-based biomarker; now what? We developed a high-throughput method for quantitative clinical glycan biomarker validation. *ACS Omega*. 2020;5(12):6270-6.
(58) Khan NM, C NM, Negi A, Thaseen IS. Analysis on improving the performance of machine learning models using feature selection technique. In: Abraham A, Cherukuri AK, Melin P, Gandhi N, editors. *Intelligent Systems Design and Applications*; *Vellore, India*. Cham: Springer; 2018. p. 69–77.

Development of an integrated system for tear ascorbic acid fluorescent detection

Wenhao Le and Guohui Liu

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

Abstract

Ascorbic acid (AA) is an imperative antioxidant, which is indicative of the state of human health. AA in tears has been suggested as a potential biomarker for diagnosing dry eye disease (DED). A fluorescent sensor, established on the contact lens for sensitive turn-on detection of AA, was invented based on the BSA-Au nanoclusters (NCs) because of its easy operation, low cost, high stability, selectivity and sensitivity. With the increasing demand for the point-of-care (POC) diagnosis of ocular diseases, a readout platform containing light-emitting lights (LEDs), a magnifying lens and an optical filter was developed for quantitative measurements and image acquisition by employing a smartphone camera. The use of a smartphone algorithm and user interface was aiming to process measurements from the fluorescent sensor and export quantitative diagnostic data. The fluorescent sensor was examined and exhibited a highly sensitive ($R^2 = 0.96$) AA detection (0 - 1.2 mmol L^{-1}). In this study, a smartphone application for tear AA fluorescent sensors was also developed with some optimisations of image-processing algorithms. Through this research, continuous monitoring of tear AA concentrations in POC settings was achieved. With the future use of artificial intelligence algorithms, cloud server data of DED patients in the UK would be acquired, allowing for accurate prediction of disease severity stages and giving appropriate suggestions.

Keywords: Fluorescent sensing, Contact lens sensors, Ascorbic acid, Tear monitoring, Ophthalmic diagnostics, Dry eye disease image-processing algorithms, Smartphone readout device, Ophthalmic health track application

1 Introduction

Ascorbic acid (AA), commonly well-known as vitamin C, found in many fruits and vegetables, has been treated as an important antioxidant in a variety of fields, including food, beverages, animal feed, cosmetics, nutraceutical and pharmaceutical formulations [1]. Meanwhile, AA, as a cofactor of enzymes [2], a reducing agent [3], and an essential nutritional factor [4], is critical to many physiological processes in human bodies. The level of AA in the human body is indicative of the state of human health. Variations in the amount of AA in the body can lead to a change in physiological conditions as well as severe diseases [5]. It has been shown that the deficiency of AA in the human body leads to immunity reduction, anemia, and scurvy [6]. Other disorders such as stomach convulsion, diarrhoea, and urinary stones [7] can be attributed to the elevation of AA. Besides, in order to maintain the necessary AA content in the human body, AA must be obtained from food intake since it is an exogenous [8] chemical substance. A human eye's polymorphonuclear leukocytes (PMNs) infiltration will increase the concentration of free radicals in tears during inflammation. Due to the presence of toxic free radicals released by early-stage corneal disorder [9] and alkali burns [10], the concentration of AA in tears will increase. Hence, it would be beneficial to analyse the variation of tear AA levels in order to monitor ocular health and arrive at a diagnosis.

In 2021, the global prevalence of dry eye disease (DED) was estimated at 11.59% [11] of people worldwide, while about 9.6% [12] of the population in the UK was affected by DED. NHS dispensing service filled over 6.4 million prescriptions for DED such as artificial tears, ocular lubricants and astringents, at the cost of more than £27 million [13]. Lack of tears or an excessive loss of tears causes dry eye and hyperosmolarity, which results in symptoms such as abrasions and discomfort to the cornea [14]. Although the majority of dry eye treatments have focused on a lack of tear production and inflammation, meibomian gland dysfunction (MGD) has recently been identified as the major cause of DED [15]. Lacrimal hyposecretion is another common cause of DED, which decreases with age and results in aqueous tear deficiency [16]. Diagnostic approaches for DED are mainly based on identifying symptoms such as burning, gritty, or sandy sensations, burning, and red eyes [17]. Slit lamp examinations and symptom surveys are the most common diagnostic methods. In tests for tear hyperosmolarity, such as those conducted with osmometers (TearLab), results are difficult to interpret due to large standard errors [18]. LipiView (TearScience), an interferometry imaging device that can be used on a benchtop, is used to monitor the condition of the lipid layer of an eye. Nevertheless, a patient's blinking rate may be inconsistent, which causes this high-cost device to produce inaccurate results [19].

The fluorescent detection for AA concentrations in the tear is considered one of the desirable alternatives to existing diagnosis methods of DED because of its easy operation, low cost, high sensitivity and selectivity. A number of methods have been developed for the fluorescent detection of AA, including electrochemistry [20], the titration with an oxidizing agent [21], spectrophotometry [22], chromatography [23] and chemiluminescence [24], enzymology [25]

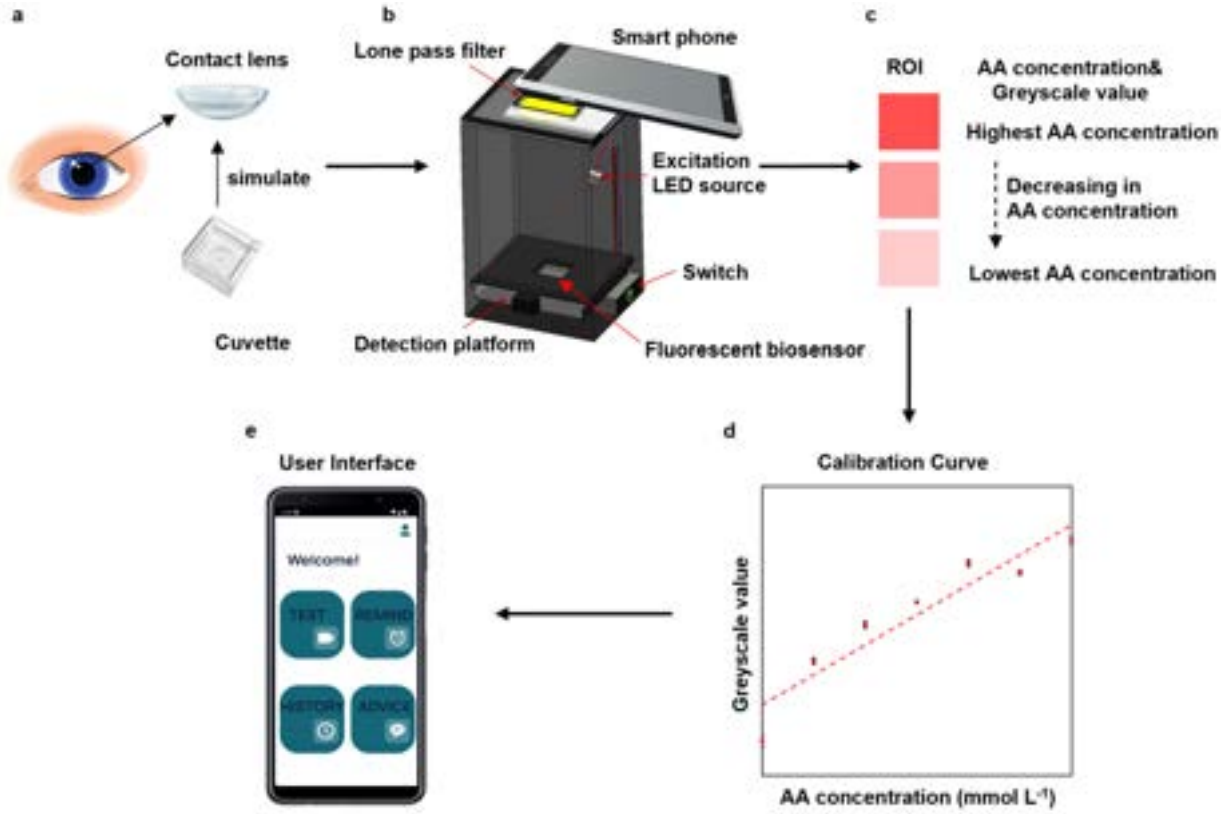


Figure 1: The overall experimental design and an integrated POC system for the fluorescent detection of tear AA (a) The use of cuvettes to simulate contact lens platforms for the sample collection. (b) The readout device used to reduce the noise during the photo-taking procedure [26] (c) The identified region of interest (ROI), the fluorescent intensity increases with the tear AA increments (d) Predicted calibration curve based on linear regression used for the smartphone application (e) The user interface design of the smartphone application

and capillary electrophoresis [27]. It remains urgently necessary to develop less complicated and expensive, but more sensitive and effective methods in order to determine the level of AA within the human body. For example, AA can be detected with high sensitivity using a fluorescent probe composed of CdTe/CdS/ZnS quantum dots (QDs) [28]. Although CdTe/CdS/ZnS QDs have excellent properties, the synthesis of a CdTe/CdS/ZnS QDs probe took a long time, and the use of Te power and $CdCl_2$ was environmentally unfriendly and toxic as a fluorescent sensor. A promising field of new fluorescent probes has opened up as a result of noble metal NCs. NCs become molecular species in the small size regime, and discrete states with significant fluorescence are observed [29]. In particular, BSA-Au NCs have been found to be highly desirable for biolabeling and bioimaging applications due to their non-toxic nature, which means they pose little adverse effect on biological systems. With the addition of $KMnO_4$, the oxidation status of BSA-Au NCs will be perturbed, resulting in fluorescence quenching. After introducing reductive AA into the quenched solution, added $KMnO_4$ was reduced by AA, then the fluorescence of the system was recovered [30]. A novel fluorescent probe for AA can therefore be created using BSA-Au NCs. A preliminary study of the proposed method was conducted on human tear fluid samples with satisfactory results.

Some patients would not be able to visit the hospital to obtain a real-time diagnosis of their ocular conditions. Therefore, recent studies have developed personalised POC [31] diagnosis that allows patients to self-check their eye conditions [32]. In order to meet the increasing demand for POC diagnostics, it is imperative to identify and monitor the biological and chemical molecules in tears under minimal concentrations within physiological conditions through rapid and accurate detection methods [33]. Consequently, biosensors based on fluorescence provide rapid detection of AA levels in tears. To achieve real-time monitoring and continuous sampling, the reversibility of fluorescent biosensors is also an imperative requirement for POC settings. Real-time monitoring, diagnosis of specific ocular diseases, and an understanding of physiological conditions within the eye systems would be achievable with the employment of specific fluorescent sensing technologies [34]. With a smartphone application and a readout device (Figure 1.b & 1.e), the quantitative POC measurement can be obtained from fluorescent sensors and further processed to export quantitative diagnostic data. The developed fluorescent sensor is highly sensitive (R^2 =

0.96) between 0 and 1.2 mmol L^{-1}. Cuvettes and artificial tear fluids (ATF) were used to simulate the contact lens platform and human tear compositions (Figure 1.a). The integrated ophthalmic system is demonstrated for quantitatively analysing and reporting DED severity stages through smartphone image-processing algorithms and a calibration curve of fluorescent intensity versus tear AA levels (Figure 1.c & 1.d). The present work allows diagnosis of DED severity stages in POC settings [17] and can be integrated with AI models in the future.

Basic regression algorithms are applied to predict values of AA concentrations in tears. The commercialisation of the integrated detection system including a readout device, a smartphone application and internal algorithms will be a breakthrough in POC and personalised ocular diagnosis, as many patients could now receive medical advice without attending specialist appointments. This could potentially improve the efficiency of medical resource allocation, shorten the waiting time of patients with more critical conditions and optimise the government expenditure on ophthalmic health care.

2 Materials and Method

2.1 Chemicals and instrument

Tris-Hydrochloride (Tris acid) purchased from EMD Millipore Corpration; Tris-2-amino-2-(hydroxymethyl)-1,3 propanediol (Tris base), chloroauric acid ($HAuCl_4$), bovine serum albumin (BSA), sodium hydroxide (NaOH), hydrochloric acid (HCl), nitric acid (HNO_3), ascorbic acid, potassium chloride (KCl), calcium chloride ($CaCl_2$), magnesium chloride ($MgCl_2$), lysozyme, glucose and DI water purchased from Sigma-Aldrich; Potassium permanganate ($KMnO_4$) and sodium chloride (NaCl) purchased from VWR International. In order to obtain solutions containing different concentrations of AA ranging from 0 to 1.2 mmol L^{-1}, serial dilutions of the maximum concentration of AA solution were performed. The pH levels of buffer solutions were adjusted and monitored using a pH meter (FiveEasy F20, Mettler Toledo). With the use of a pH meter, Tris acid and Tris base were used to adjust the pH values of the $KMnO_4$ solution, the Tris buffer solution and ATF to 6.5, 7.4 and 7.4 respectively. Cuvettes and ATF were used to simulate the contact lens platform and human tear compositions. A 3-D printer (Formlabs Form 3 SLA 3D Printer, Imperial College London) was used to fabricate the designed black readout device for data collection and analysis. Samsung Galaxy A7 was employed to capture and process images with the aid of the readout device, Oasis (health diagnosis application) and image algorithms developed by java and C++ on Android Studio.

2.2 Synthesis of BSA-Au NCs and solution preparations

Synthesis of BSA-Au NCs: All glassware used in the experiment was cleaned three times by freshly prepared aqua regia (HCl and HNO_3 in the volumetric ratio of 3:1). Subsequently, ethanol was required to rinse the glassware three times followed by distilled deionised water. $HAuCI_4$ solution (15 mL, 10 mmol L^{-1}) was added to the BSA solution (15 mL, 50 mg mL^{-1}) under magnetic stirring. After 3 minutes, NaOH solution (1.5 mL, 1 mol L^{-1}) was added to catalyse the synthesis. Next, the mixture was incubated at 37^oC for 48 hours, as the solution colour changed from light yellow to deep brown, the mixture was dialysed in DI water for another 48 hours to remove all the excess precursors such as unreacted $HAuCI_4$, BSA and NaCl. Finally, the prepared BSA-Au NCs could be stored in the refrigerator at 4^oC.

Preparation of ion solutions With the aid of a pH meter, the Tris buffer solutions of 0 and 1.2 mmol L^{-1} AA were adjusted to pH = 7.4 by mixing the solutions of Tris acid and base of 0 and 1.2 mmol L^{-1} AA respectively. As an example, 121.1 mg of Tris base and 157.6 mg of Tris acid were each dissolved in 100 mL of DI water. In order to obtain a Tris buffer solution with pH = 7.4, these two stock solutions were mixed and monitored using a pH meter. With the addition of 1.2 mmol L^{-1} AA, another Tris buffer solution was also prepared with a pH of 7.4. These two Tris buffer solutions (0 and 1.2 mmol L^{-1} AA) were used in the serial dilutions to obtain solutions of varying AA concentrations.

Preparations of ATF: ATF were used to simulate the ocular environment, in order to observe the influences on the fluorescent intensity carried by co-existing substances in tear, NaCl (150 mmol L^{-1}) KCl (20 mmol L^{-1}), $CaCl_2$ (1 mmol L^{-1}), $MgCl_2$ (0.6 mmol L^{-1}) lysozyme (2.36 mg mL^{-1}), BSA (50 μg mL^{-1}) and glucose (0.14 mmol L^{-1}) were dissolved into the stock Tris buffer solution of 0 mmol L^{-1} AA and 1.2 mmol L^{-1} AA with a pH value of 7.4. Serial dilutions were carried out to obtain the required concentrations.

2.3 Apparatus

A portable readout device (Figure 2) was developed for the data collection procedure. It consists of three parts: a lid, a main body and a detection platform. The lid contains a long pass filter allows only red light between 620 nm and 750 nm to pass through. A magnifying glass is used to reduce the smartphone camera natural focusing distance from 7 cm to 4 cm. The readout device's main body was designed to be a black cuboid (height is 4 cm) with one LED used to emit ex-

citing lights at a wavelength of 390 nm. The detection platform is movable and can be used to place the sample in its groove. The readout device was developed for a more unified data collection procedure by reducing environmental noises. In detail, the fluorescent intensity is easily interfered by some external factors, including the excitation light intensity, the reflection light of the surface where the sample is placed, the distance between the sample and the smartphone camera, and the shooting angle of the smartphone. Moreover, 3-D printing materials were also matt black which can effectively reduce the noise caused by the light reflection problem. The latter two challenges were solved by fixing the smartphone camera relative to the filter on the lid to ensure that, the shooting angle is always perpendicular to the sample and the shooting distance is 4 cm throughout the entire data collection procedure.

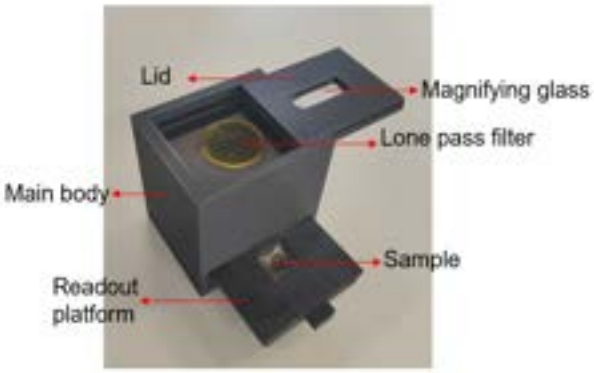


Figure 2: Readout device for image-taking procedure

2.4 Data Collection Procedure

After BSA-Au NCs were quenched by the $KMnO_4$ solution, a series of concentrations of AA were used to recover the fluorescence (the ratio of BSA-Au NCs : $KMnO_4$: AA = 2 : 1: 2), The fluorescent information of the samples are collected using the readout device and a smartphone, each prepared sample would be transferred to a cuvette, and this cuvette would be placed further at the groove of the mobile platform. The distance between the smartphone camera and the sample is controlled to be 4 cm at a constant ambient temperature (25^oC) to ensure that, the shooting environment is identical for each sample throughout the entire data collection procedure. With excitation lights (wavelength = 390 nm), the smartphone was used to capture the image of the excited-sample through the lid. Three samples were prepared to reduce intensity errors of each AA concentration due to some external factors mentioned in Apparatus. For each sample, three pictures were taken to find the average fluorescent intensity, aiming to reduce systematic errors caused by the shooting environment. Finally, image-processing algorithms developed by C++ could obtain the fluorescent information for a series of AA concentrations from these images captured with excitation lights.

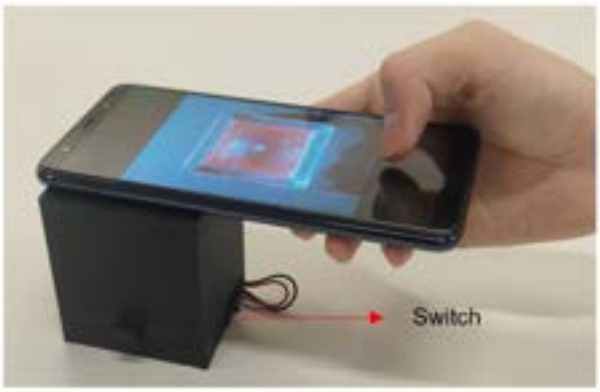


Figure 3: Operation of capturing images using the readout optical device. The magnified photographs on the top indicates the developed smartphone app for data collection and data analysis

2.5 Image Processing Algorithms

In order to obtain an equation of calibration curves which can be further applied in the smartphone health track software, image-processing algorithms were developed by C++ on VS Code 2022 using the OpenCV library. Upon acquiring images of samples, the image-processing algorithms are articulated into three steps: image segmentation, filtration of noises and conversion of RGB images into greyscale.

Foremost, the region of interest (ROI) of each image is required to be obtained. Since ROI is the red light of a wavelength = 690 nm (Figure 4.a), the regional-based segmentation splits the image into green, blue, and red channels. Since the region of interest is the red region, ROI is extracted, and others are transferred to black. (Figure 4.b)

Following this, erode method is used to denoise the image. To be specific, sample pictures may contain white dots in the middle due to the reflection of the light (Figure 4.a), the algorithms will assign the value of pure white dots to 255, which will pull up the average greyscale value, making it inaccurate. As a result, addition algorithms are used to remove the white dot and this is done by setting a thresh of 200. With the implementation of algorithms, white dots have been removed and transferred to black which will not affect the greyscale value anymore. Besides, a minimum thresh is set to filter the noise around the kernel. This is done by trial and error and found that when the thresh value is great than 120, the noise can

be effectively filtered.

The reason why this method can be used to describe the fluorescent intensity can be found in the following equation of greyscale value.

$$Greyscale\ value = 0.30R + 0.59G + 0.11B, \quad (1)$$

where R, G, and B represent the luminous intensity for red, green and blue channels, respectively. To be specific, the luminous intensity of the red channel is been calculated by equation 2:

$$R = \int_0^\infty S(\lambda)\overline{r}(\lambda)d\lambda, \quad (2)$$

where $S(\lambda)$ is the spectral power distribution, $\overline{r}$ is the RGB colour matching functions at standardised wavelengths of 700 nm (red light), and λ is the wavelength.

It is clear that the value of R is proportional to the spectral power, indicating that a greater value of R represents higher luminous intensity. Since algorithms developed in this research only extract the R channel, the greyscale value of the ROI can directly represent the fluorescent intensity of the sample (Figure 4.c). Lastly, algorithms will read out the greyscale value of ROI based on calculations of average greyscale values. This process is then followed by data processing, the fluorescence can be quantitatively related to the AA concentration in the tear fluid, a higher AA concentration can enhance the fluorescent intensity of the sensor.

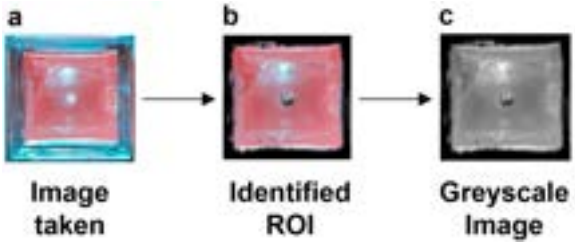


Figure 4: Image processing procedure (a) The image was taken from the readout device (b) The identified region of interest by using the algorithms. (c) The greyscale format of the identified ROI which will be used to calculate the average intensity

3 Results and discussion

This section includes the design of a smartphone application user interface, all observations and calibration curves gathered through the variations in sample volumes and their response to changes in temperature, storage time, and operation time in ATF

3.1 User interface design

With the purpose of POC ophthalmic diagnosis, an android smartphone software, Oasis, was developed. As shown in Figure 5, the application users are guided step by step to detect the AA concentration in the tear fluid based on the fluorescent sensing of the image capture. The welcome page (Figure 5.b) with four clickable buttons was designed for navigating different activities. The top left button, ***test*** navigates the user to where the test image can be imported by taking a photo or selecting it from the album. When the photo is captured by the smartphone camera, ***start*** button should be clicked to invoke the image-processing algorithm. Finally, in ***result*** page, the AA concentration is computed based on the average of regional greyscale values. As long as the image is successfully processed, the user can select to save test results and be navigated to ***advice*** page where proper suggestions would be provided. Nevertheless, errors may appear due to the failure of the camera focus or the extremely strong environmental light, then the user can return to ***test*** page to import or retake the image again. Moreover, the user can click the bottom left button, ***history***, to access all the previous test results by searching for the date.

3.2 Sample volume analysis

The effect of change in tear volume was investigated by four sets of experiments with different volumes of ATF (10 µL, 120 µL, 240 µL, and 360 µL). For each set of experiments, the fluorescent information of each sample at different AA concentrations was collected, and a corresponding calibration curve was plotted with the interested volume (Figure 6.a - d). The results showed that at higher ATF volumes, the gradients would be greater. However, for different volumes, the lowest and highest greyscale values were similar which were around 135 and 150 respectively, this indicated that the greyscale value was mainly related to the AA concentration. The key difference is the rate of reaching the maximum greyscale value which is greater at higher AA concentrations. This might be because the fluorescent effect exhibited better within the higher sample volume. As in the synthesis process, the BSA was able to trap at least one Au^+ ion and the clusters might not be on average bonded within the mixture, higher sample volumes (360 µL) hence would obtain a higher fluorescent intensity outcome. However, if the tear volume is under 120 µL, the gradient would be similar and around 13. In reality, the average human tear

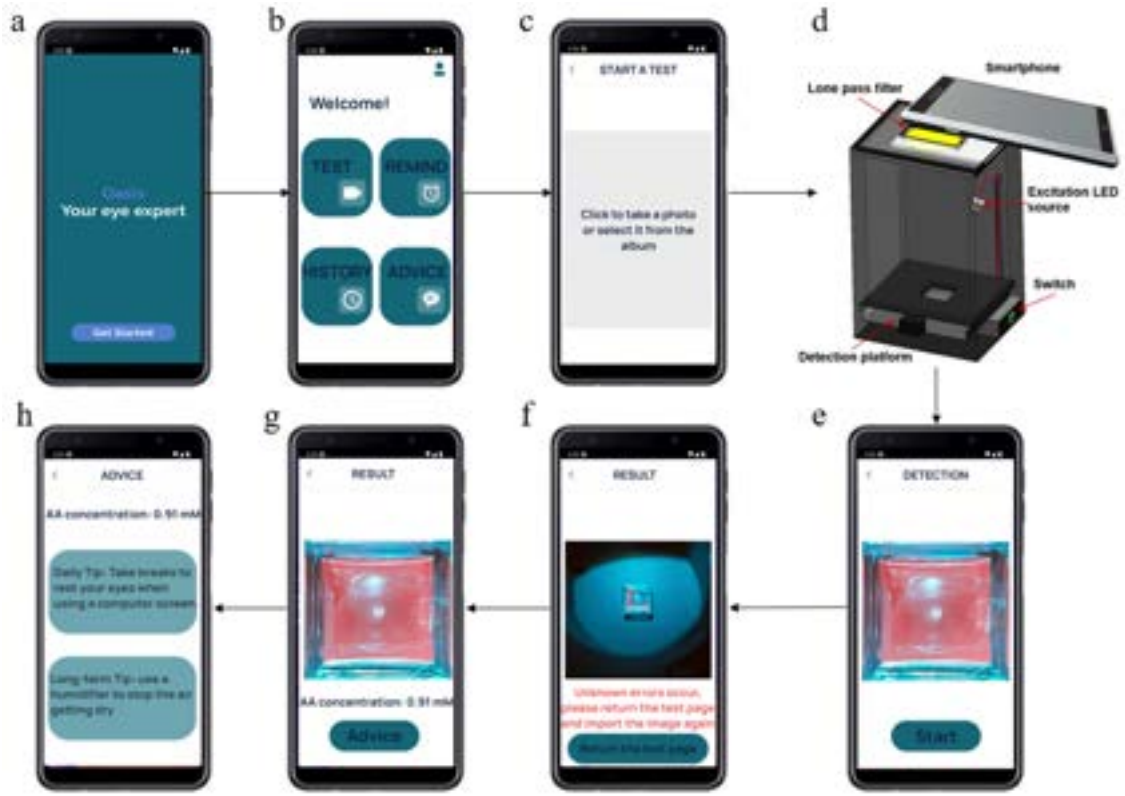


Figure 5: Demonstration of the user interface of the smartphone application (a) Onboarding (b) Homepage: navigation to other pages (c) Take the photo or upload from the album to start a test (d) Black readout device [26] (e) Import image and waiting for the detection (f) Unknown errors occur, return to the test page (g) Measure AA concentration (h) Appropriate suggestions based on the test result

fluid volume is usually around 6 µL (2.73 – 12.75 µL) [35] which indicates that Figure 6.a is more similar to the actual condition and this calibration curve is more suitable to act as the underlying principle for the smartphone application. However, for DED patients, the tear volume at the test is usually below the average and around 3 µL [36]. This might cause difficulties when trying to collect the tear sample. One possible solution could be using the porous polyester rods [37] to collect tear samples, which is a more rapid, user-friendly way for the collection of tear fluid. After that, the sample can be transferred to the contact lens sensors and continue the analysis.

3.3 Temperature analysis

Effects of temperature deviation in tear fluid on fluorescent intensity were investigated by carrying out six sets of experiments with samples temperatures ranging from 28 to 43 ^oC. Results indicated that fluorescent intensity was not directly correlated to the tear fluid temperature. Figure 6.e showed that with different sample temperatures, the calibration curves were nearly parallel with differences in the starting and endpoints due to the deviating environment of taking photos since the readout of fluorescent sensors can be affected by the intensity of the excitation light. This could be mitigated by maintaining and analysing the background lighting conditions during the smartphone readout. Figure 6.f showed that for each temperature, the fluorescent intensity for the same AA concentration was similar, indicating that the fluorescent effects were independent of the tear fluid temperature. The blue region highlighted the normal human body temperature, ranging from 36.4^{o}C to 37.1^{o}C [38], the greyscale value for each concentration was highly identical in this area, proving that temperature influence the fluorescence is negligible.

3.4 Storage time analysis

Experiments were carried out to study how long the sample can keep fresh storing at 5^{o}C and the effects of light on fluorescent performance. Samples were tested after 1 hour, 3 hours, 6 hours, 12 hours, 24 hours, 2 days, 3 days, 4 days, 5 days, 7 days, 10 days 15 days of storage with lighted and no light conditions in the fridge at 5^oC. From Figure 7, fluorescent intensity was lower in lighted conditions than in no-light conditions. This is because the chemical bonds in BSA-Au NCs are photo-sensitive and easily dissociated under lighted conditions. Au^+ undergoes a reduction reaction under light conditions from Au^+ to Au, then aggregates to form Au cluster, reducing its connection to BSA, leading to the reduction in the intensity [39]. Figure 7 showed that the greyscale value

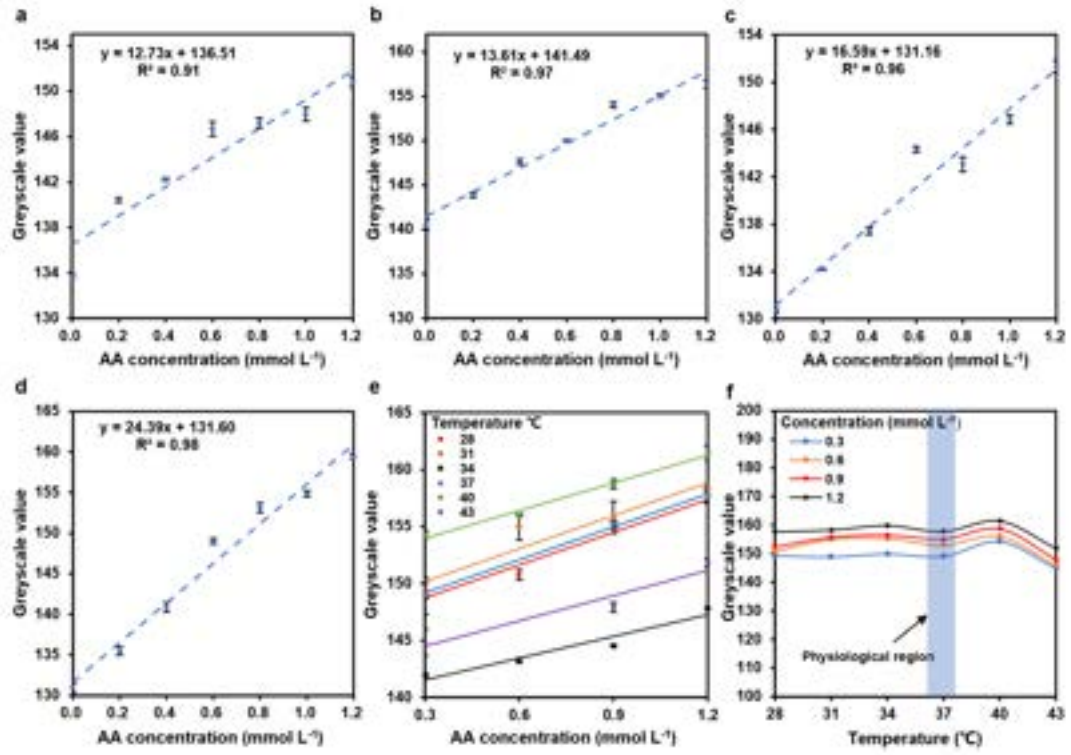


Figure 6: Optimisations on the volume of detection and characterisations of AA fluorescent sensor with ATF. Calibration curves on temperature dependency for tear sample volumes of 10,120,240,360 µL with AA concentration range from 0 - 1.2 mmol L^{-1}. (a) Calibration curve for sample volume of 10 µL, 25^oC. (b) Calibration curve for sample volume of 120 µL, 25^oC. (c) Calibration curve for sample volume of 240 µL, 25^oC. (d) Calibration curve for sample volume of 360 µL, 25^oC. (e) Calibration curves for sample temperature ranging from 28 to 43^oC. (f) Temperature dependency of greyscale value with varying AA concentration, temperature range from 28-43^oC

decreased within the first 120 hours, followed by a sharp rise. This is due to the degeneration of proteins in NCs, forming white cloud precipitates after a long period of storage. The precipitate can affect the image algorithm by increasing greyscale value (Figure 8). This is because the algorithm assigns white parts with greyscale value of 255; white precipitate reflects light, pulls up the average greyscale value of the identified region of interest. As a result, the recommended maximum storage time of samples is 120 hours at 5^oC in dark.

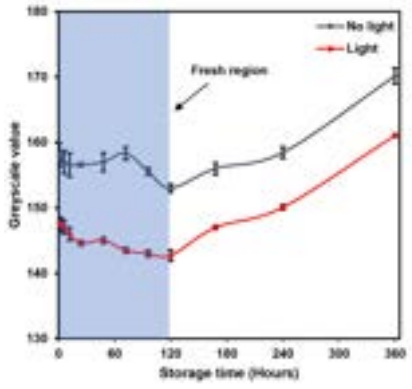


Figure 7: Change in greyscale values corresponding to different storage times. Blue region is fresh sample; white region is degraded sample.

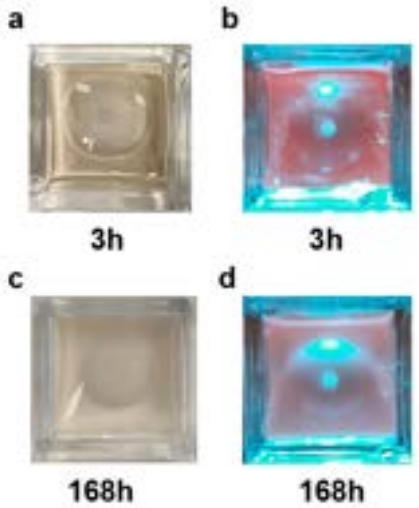


Figure 8: The comparison between a fresh sample and a non-fresh sample. (a) Sample stored for 3 hours under room light (b) Excited sample stored for 3 hours under dark environment (c) Sample stored for 168 hours (d) Excited sample stored for 168 hours

3.5 Operation time analysis

Users would start the AA detection by the smartphone application as long as they have captured the image. However, sometimes these users may be required to retake another picture due to different reasons. As a result, the sensitivity analysis of the fluorescence dependent on the operation time of 5 minutes, 10 minutes, 30 minutes, 1 hour,3 hours, 5 hours, 12 hours and 1 day should be carried out. According to Figure 9, the fluorescent intensity showed a slightly decreasing trend as the operation time proceeded. Images are recommended to capture within the blue region from 0 minutes to 5 minutes after AA is added into the quenched mixture of BSA-Au NCs and $KMnO_4$). Otherwise, the fluorescent intensity would be inaccurate when taking a photo after 10 minutes. After 300 minutes of the quenched mixture recovered by the additional AA, fluorescent intensities would be lower and unable to prove the disease diagnosis because of the instability of the quenched mixture of BSA-Au NCs and $KMnO_4$.

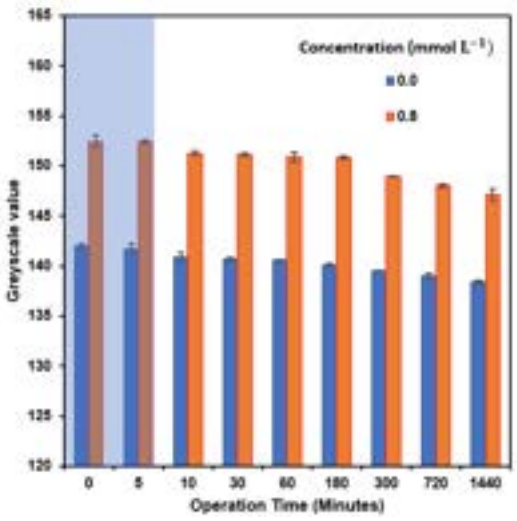


Figure 9: Different greyscale values corresponding to operation time. The blue region shows where samples within recommended operation time for image capture

4 Conclusion

In summary, the quantitative detection of AA was developed based on the change of fluorescent intensities after the quenching reaction between BSA-Au NCs and $KMnO_4$. The sensor obtained a stable performance under different temperature settings and was considered to be independent of temperature. As for the storage facilities of the sensor, the optimal settings for BSA-Au NCs is at 5^oC without light, and the detection was indicated to be accurate within 5 days. Moreover, the fluorescent effect of the sensor could last for 5 minutes during the operational period. Based on the obtained similar calibration equations of the sensor, the sample volume analysis showed that the fluorescent intensity was more correlated to AA concentrations rather than sample volume. The fluorescent detection for AA was examined in both Tris buffer solution (pH = 7.4) and ATF (pH =7.4), indicating that the fluorescent intensity was directly proportional to AA concentrations ranging from 0 to 1.2 mmol L^{-1}. In order to achieve point-of-care DED diagnosis, the developed system includes fluorescent AA sensors, 3D printed readout box, and a smartphone application, which demonstrates capturing images, quantitative analysis, and reporting DED severity stages. With the aid of this personalized system, some early-stage DED patients would be able to confirm diagnostic results and receive efficient treatments. Meanwhile, the government could reduce the expenditure on ophthalmic health care, and alleviate the burden of NHS.

5 Outlook

Considering the near future of tear sensor development on DED diagnosis, continuous multi-biomarker monitoring would be more valuable and accurate compared to one single biomarker detection at POC platform. Apart from AA, some indicators within tear fluid such as pH, Na^+, Mg^{2+}, Zn^{2+}, Ca^{2+} and K^+, were also examined and applied to improve the accuracy of DED diagnosis [17].

In this research, basic regression models and algorithms were used for data processing. AI models in this case could be an improvement. It can provide an accurate and robust prediction of concentrations of and report disease severity stages, subsequently decreasing the misjudgment possibilities of the smartphone diagnosis of ocular diseases. Machine learning algorithms can also be used to analyze user health status based on existing data. For example, supervised learning algorithms [40] could classify, predict, and detect abnormalities in raw sensing data. The use of machine learning can find hidden patterns, and correlations between data points and effectively process high-dimensional biosensing data acquired by sensors measuring multiplex biomarkers. As part of the implementation of AI in biosensing systems, a large amount of healthcare information will be required, such as medical history, streaming biosensing data, and medical imaging data. With the use of smartphone-based biosensing application, smartphones can upload the data onto the cloud server for processing by using embedded processors, after receiving all medical data collected by wearable sensors. Meanwhile, the cloud server data could optimise AI algorithms, leading to a better prediction of DED

with a shorter processing time.

Regarding the commercialisation of this personalised POC diagnosis of DED, the product would consist of a readout device-black box and contact lens sensors. The biosensing software can be downloaded from software shops such as Google play and Apple store. However, due to the use of contact lenses, the target user groups would initially focus on junior and senior generations. This is because some children and elderly people may experience difficulties using contact lenses. To face wider user generations, this problem can be solved by developing a supplementary tear fluid collection method. Introducing porous polyester rods to collect tear samples could be one of the options, as they can transfer the sample to the contact lens platform to make a detection.

6 Acknowledge

The authors would like to express their gratitude to the members of the Biosensors group for their valuable and continuous support. Particular thanks to Miss Yuqi Shi for her valuable feedback and great support throughout the entire project and Mr Yihan Zhang for his patient guidance on the coding.

References

[1] Dan Wen et al. “Ultrathin Pd nanowire as a highly active electrode material for sensitive and selective detection of ascorbic acid”. In: *Biosensors and Bioelectronics* 26.3 (2010), pp. 1056–1061. ISSN: 0956-5663. DOI: https://doi.org/10.1016/j.bios.2010.08.054.

[2] Sarah E. Bohndiek et al. “Hyperpolarized [1-13C]-Ascorbic and Dehydroascorbic Acid: Vitamin C as a Probe for Imaging Redox Status in Vivo”. In: *Journal of the American Chemical Society* 133.30 (2011). PMID: 21692446, pp. 11795–11801. DOI: 10.1021/ja2045925.

[3] Jie Feng et al. “Polyethyleneimine-templated copper nanoclusters via ascorbic acid reduction approach as ferric ion sensor”. In: *Analytica Chimica Acta* 854 (2015), pp. 153–160. ISSN: 0003-2670. DOI: https://doi.org/10.1016/j.aca.2014.11.024.

[4] Robert D. Hancock and Roberto Viola. “Improving the Nutritional Value of Crops through Enhancement of l-Ascorbic Acid (Vitamin C) Content: Rationale and Biotechnological Opportunities”. In: *Journal of Agricultural and Food Chemistry* 53.13 (2005). PMID: 15969504, pp. 5248–5257. DOI: 10.1021/jf0503863.

[5] Tingting Zhao et al. “Fluorescent color analysis of ascorbic acid by ratiometric fluorescent paper utilizing hybrid carbon dots-silica coated quantum dots”. In: *Dyes and Pigments* 186 (2021), p. 108995. ISSN: 0143-7208. DOI: https://doi.org/10.1016/j.dyepig.2020.108995.

[6] Masoud Rohani Moghadam et al. “Chemometric-assisted kinetic–spectrophotometric method for simultaneous determination of ascorbic acid, uric acid, and dopamine”. In: *Analytical Biochemistry* 410.2 (2011), pp. 289–295. ISSN: 0003-2697. DOI: https://doi.org/10.1016/j.ab.2010.11.007.

[7] Linda K. Massey, Michael Liebman, and Susan A. Kynast-Gales. “Ascorbate Increases Human Oxaluria and Kidney Stone Risk”. In: *The Journal of Nutrition* 135.7 (July 2005), pp. 1673–1677. ISSN: 0022-3166. DOI: 10.1093/jn/135.7.1673.

[8] Juanjuan Liu et al. ““Switch-On” Fluorescent Sensing of Ascorbic Acid in Food Samples Based on Carbon Quantum Dots–MnO2 Probe”. In: *Journal of Agricultural and Food Chemistry* 64.1 (2016). PMID: 26652202, pp. 371–380. DOI: 10.1021/acs.jafc.5b05726.

[9] Ngamjit Kasetsuwan et al. “Effect of Topical Ascorbic Acid on Free Radical Tissue Damage and Inflammatory Cell Influx in the Cornea After Excimer Laser Corneal Surgery”. In: *Archives of Ophthalmology* 117.5 (May 1999), pp. 649–652.

[10] Roswell R. Pfister and Christopher A. Paterson. “Ascorbic acid in the treatment of alkali burns of the eye.” In: *Ophthalmology* 87 10 (1980), pp. 1050–7.

[11] Eric B Papas. “The global prevalence of dry eye disease: A Bayesian view”. In: *Ophthalmic and Physiological Optics* 41.6 (2021), pp. 1254–1266. DOI: https://doi.org/10.1111/opo.12888.

[12] Parwez Hossain et al. “Patient-reported burden of dry eye disease in the UK: a cross-sectional web-based survey”. In: 11.3 (2021). ISSN: 2044-6055. DOI: 10.1136/bmjopen-2020-039209.

[13] “The management of dry eye”. In: 54.1 (2016). Ed. by, pp. 9–12. ISSN: 0012-6543. DOI: 10.1136/dtb.2016.1.0378.

[14] Gary N. Foulks. “The Correlation Between the Tear Film Lipid Layer and Dry Eye Disease”. In: *Survey of Ophthalmology* 52.4 (2007), pp. 369–374. ISSN: 0039-6257. DOI: https://doi.org/10.1016/j.survophthal.2007.04.009.

[15] Eiki Goto et al. “Impaired functional visual acuity of dry eye patients”. In: *American Journal of Ophthalmology* 133.2 (2002), pp. 181–186. ISSN: 0002-9394. DOI: https://doi.org/10.1016/S0002-9394(01)01365-4.

[16] Maurizio Rolando and Manfred Zierhut. “The Ocular Surface and Tear Film and Their Dysfunction in Dry Eye Disease”. In: *Survey of Ophthalmology* 45 (2001), S203–S210. ISSN: 0039-6257. DOI: https://doi.org/10.1016/S0039-6257(00)00203-4.

[17] Ali K. Yetisen et al. “Scleral Lens Sensor for Ocular Electrolyte Analysis”. In: *Advanced Materials* 32.6 (2020), p. 1906762. DOI: https://doi.org/10.1002/adma.201906762.

[18] Vatinee Y Bunya et al. “Variability of tear osmolarity in patients with dry eye”. In: *JAMA ophthalmology* 133.6 (2015), pp. 662–667.

[19] Motoko Kawashima and Kazuo Tsubota. “Tear lipid layer deficiency associated with incomplete blinking: a case report”. In: *BMC ophthalmology* 13.1 (2013), pp. 1–3.

[20] Kun Liu et al. “Online Electrochemical Monitoring of Dynamic Change of Hippocampal Ascorbate: Toward a Platform for In Vivo Evaluation of Antioxidant Neuroprotective Efficiency against Cerebral Ischemia Injury”. In: *Analytical Chemistry* 85.20 (2013). PMID: 24090233, pp. 9947–9954. DOI: 10.1021/ac402620c.

[21] S.P Arya, M Mahajan, and P Jain. “Non-spectrophotometric methods for the determination of Vitamin C”. In: *Analytica Chimica Acta* 417.1 (2000), pp. 1–14. ISSN: 0003-2670. DOI: https://doi.org/10.1016/S0003-2670(00)00909-0.

[22] Mustafa Özyürek et al. “Spectrophotometric determination of ascorbic acid by the modified CUPRAC method with extractive separation of flavonoids–La(III) complexes”. In: *Analytica Chimica Acta* 588.1 (2007), pp. 88–95. ISSN: 0003-2670. DOI: https://doi.org/10.1016/j.aca.2007.01.078.

[23] Petra Koblová et al. “Development and validation of a rapid HPLC method for the determination of ascorbic acid, phenylephrine, paracetamol and caffeine using a monolithic column”. In: *Analytical Methods* 4.6 (2012), pp. 1588–1591.

[24] Zhihua Wang, Xu Teng, and Chao Lu. “Carbonate interlayered hydrotalcites-enhanced peroxynitrous acid chemiluminescence for high selectivity sensing of ascorbic acid”. In: *Analyst* 137.8 (2012), pp. 1876–1881.

[25] Grahame J. Kelly and Erwin Latzko. “Prospect of a specific enzymic assay for ascorbic acid (vitamin C)”. In: *Journal of Agricultural and Food Chemistry* 28.6 (1980), pp. 1320–1321. DOI: 10.1021/jf60232a051.

[26] Yuqi Shi. “3D printed smartphone readout device”. In: (2022).

[27] Won-Suk Kim et al. “Ascorbic Acid Assays of Individual Neurons and Neuronal Tissues Using Capillary Electrophoresis with Laser-Induced Fluorescence Detection”. In: *Analytical Chemistry* 74.21 (2002). PMID: 12433096, pp. 5614–5620. DOI: 10.1021/ac025917q.

[28] Shan Huang et al. “A CdTe/CdS/ZnS core/shell/shell QDs-based “OFF–ON” fluorescent biosensor for sensitive and specific determination of L-ascorbic acid”. In: *RSC Advances* 4.87 (2014), pp. 46751–46761.

[29] Xianxiang Wang et al. “An ascorbic acid sensor based on protein-modified Au nanoclusters”. In: *Analyst* 138.1 (2013), pp. 229–233.

[30] Juanjuan Liu et al. ““Switch-On” Fluorescent Sensing of Ascorbic Acid in Food Samples Based on Carbon Quantum Dots–MnO2 Probe”. In: *Journal of Agricultural and Food Chemistry* 64.1 (2016). PMID: 26652202, pp. 371–380. DOI: 10.1021/acs.jafc.5b05726.

[31] Yuqi Shi et al. “Ophthalmic sensing technologies for ocular disease diagnostics”. In: *The Analyst* 146.21 (2021), pp. 6416–6444. ISSN: 0003-2654. DOI: 10.1039/d1an01244d.

[32] Yuqi Shi et al. “Fluorescence Sensing Technologies for Ophthalmic Diagnosis”. In: *ACS sensors* 7.6 (2022), pp. 1615–1633. ISSN: 2379-3694. DOI: 10.1021/acssensors.2c00313.

[33] Peter B. Luppa et al. “Point-of-care testing (POCT): Current techniques and future perspectives”. In: *TrAC Trends in Analytical Chemistry* 30.6 (2011), pp. 887–898. ISSN: 0165-9936. DOI: https://doi.org/10.1016/j.trac.2011.01.019.

[34] Suzanne Hagan, Eilidh Martin, and Amalia Enríquez-de-Salamanca. “Tear fluid biomarkers in ocular and systemic disease: potential use for predictive, preventive and personalised medicine”. In: *Epma Journal* 7.1 (2016), pp. 1–20.

[35] M. Esmaeelpour et al. “Tear film volume and protein analysis in full-term newborn infants”. In: *Cornea* 30 (2011), pp. 400–404.

[36] William D. Mathers and Thomas E. Daley. “Tear Flow and Evaporation in Patients with and without Dry Eye”. In: *Ophthalmology* 103.4 (1996), pp. 664–669. ISSN: 0161-6420. DOI: https://doi.org/10.1016/S0161-6420(96)30637-4.

[37] D T Jones, Dagoberto Monroy, and Stephen C. Pflugfelder. “A novel method of tear collection: comparison of glass capillary micropipettes with porous polyester rods.” In: *Cornea* 16 4 (1997), pp. 450–8.

[38] T. W. Hartgill, T. K. Bergersen, and J. Pirhonen. “Core body temperature and the thermoneutral zone: a longitudinal study of normal human pregnancy”. In: *Acta Physiologica* 201.4 (2011), pp. 467–474. DOI: https://doi.org/10.1111/j.1748-1716.2010.02228.x.

[39] Kairi Otto et al. “Thermal decomposition study of HAuCl4·3H2O and AgNO3 as precursors for plasmonic metal nanoparticles”. In: *Journal of Thermal Analysis and Calorimetry* 118 (2014), pp. 1065–1072.

[40] Yihan Zhang et al. “Wearable artificial intelligence biosensor networks”. In: *Biosensors and Bioelectronics* 219 (2023), p. 114825. ISSN: 0956-5663. DOI: https://doi.org/10.1016/j.bios.2022.114825.

Digital Twins to Address Flowsheeting Limitations

Emma Pajak and Cameron Aldren
Department of Chemical Engineering, Imperial College London, U.K.

Abstract As a rapidly growing field, the flowsheeting industry's fundamental importance to process design is illustrated by its lucrative nature. Flowsheeting software, as with any assumption-based engineering modelling, faces limitations. Digital twins offer potential advancements that could address the limitations of flowsheeting, such as poor modelling accuracy, limited customisation, accumulation of errors, and poor cost estimation. Whilst research has explored unit operation digital twins, there has not been an endeavour to apply them specifically to the limitations of flowsheeting. Therefore, this project aimed to explore the use of digital twins of unit operations to specifically address flowsheeting limitations. In line with achieving this aim, a pump, heat exchanger, and reactor were selected, coded in Python, and subsequently embedded in the open-source flowsheeting software, DWSIM. Data for the digital twins were either sourced from manufacturers or generated in ASPEN, before processing through methods such as neural networks or polynomial regression. The key findings included: the pump library demonstrating a more accurate cost estimation compared to traditional models; the grey box reactor digital twin addressing assumptions of idealised models, improving accuracy; and the heat exchanger's preliminary success in its application to multiple fluid cases, showing potential to reduce the data required by digital twins. It was concluded that with consideration of the limitations around data availability, paired with further engineering theory implementation, unit operation digital twins have the potential to offer improvements to flowsheeting. Looking at the applications of this potential, from a manufacturer's perspective, digital twins of their equipment could offer compatibility validation and real system performance predictions which would improve customer confidence and, in turn, equipment sales.

Keywords: *Digital Twin, Flowsheeting, Unit Operations, Grey Box, Neural Network*

Introduction

Flowsheeting, or process simulation, is a fundamental tool in process design used across the breadth of chemical engineering. It supports material and heat balance modelling, equipment design, sensitivity analysis, and cost analysis [1]. Flowsheeting software can provide first estimations of physical feasibility, equipment sizing, cost and commercial viability for a chemical process that would otherwise be overly cumbersome by hand. A taught module in most chemical engineering degrees, flowsheeting is a crucial skill [2].

As per the principle known as Moore's Law that states the capability of computers doubles every two years, intensive simulations are continually becoming more accessible [3]. This has facilitated process simulation becoming a highly active area of research. By its nature as an advancing field, and fundamental importance in commercial engineering projects, flowsheeting software development has become an increasingly financially lucrative area. The latest market research predicts that the digital manufacturing industry will be worth 120 billion USD by 2030 [4]. Popular proprietary software packages include ASPEN, which mainly focuses on a sequential modular process modelling approach, and gPROMS which, conversely, is an equation-oriented software [5] [6].

As with any engineering modelling based on assumptions, flowsheeting faces limitations – specifically, challenges with assumptions providing a poor representation of 'real system behaviour'. Furthermore, as flowsheeting often models multiple interconnected unit operations, a singular error can accumulate throughout the flowsheet, further damaging the accuracy of the design. Beyond this crucial downfall, unit operations can be rigid by design, rendering them non-customisable, and, therefore, unsuitable for bespoke processes. Flowsheeting cost estimation can also be a source of uncertainty in process design; costing packages are generally based upon correlations with a typical model CAPEX in the accuracy range of -50 to 100% [7]. Equipment design also holds inaccuracies as it treats sizing as a continuum; this can cause issues as manufacturers, unless specialists, offer a discontinuous set of equipment sizes. Whilst flowsheeting undoubtedly remains a crucial stage of process design, the accuracy of software recommendations and results hinges directly upon the specific assumptions employed - and implicitly adopted - by idealised models of unit operations. Inaccuracies with flowsheeting necessitate pilot-scale process development, which is expensive: to give an example, BP made a 19.5 million USD investment in a pilot-scale plastic recycling plant, in 2019 [8]. From an accessibility perspective, whilst open-source, community software exists, better-developed and integrated proprietary flowsheeting software licences are expensive and therefore less accessible for small-medium enterprises (SME) and start-up companies.

In summary, the flowsheeting field currently faces the following limitations: poor modelling accuracy, accumulation of errors in a flowsheet, limited customisation, and poor cost estimation. These limitations promote an opportunity to consider what improvements and concepts could be applied to further the current technology.

The term digital twin does not have a single discrete definition [9]. Rather, there are many definitions that

encompass similar key themes and ideas, each with slight nuances. Fundamentally, a digital twin is a "virtual copy of a physical asset, process, system or environment" [10] either created using real-world data, or, using real-time data from sensors on the physical version of the entity being simulated [11]. Although digital twins are an active area of research in the chemical industry, it is apparent that their specific application to the limitations of flowsheeting has not been fully explored.

Digital twins offer numerous potential advancements on traditional flowsheeting models; they use real process data as opposed to relying solely upon assumptions, in theory yielding a unit operation that models real system behaviour as opposed to solely approximating it. Further to model accuracy improvements, digital twins have the potential to address poor costing estimation, as exact process equipment data can be provided, introduce more flexibility and customisation of models, and, if based on an open-source platform, could offer great improvements to the process simulation capabilities of SMEs. In addition to bridging the gap on pre-existing flowsheeting limitations, the use of digital twins could further homogenise the flowsheeting stage of process design. With costing data sourced from manufacturers, as opposed to a single CAPEX figure, the output of a system of digital twins could essentially provide a 'shopping list' to the user, detailing the make, model, and cost of equipment they must procure.

Background

Many companies, including ASPEN, offer commercial-scale digital twin solutions to aid their clientele. These solutions, however, focus less on process design, and more on process operation applications, including operation scheduling, maintenance, and process control [12]. Such industrial attention has reaped substantial rewards for ASPEN's clientele; for example, the speciality chemicals manufacturer, Momentive, employed the use of an ASPEN digital twin, facilitating a 25% reduction in site inventory and a 40% decrease in supply lead time [5]. Therefore, with such tangible benefits offered by digital twin technology, attention has turned to employ their use in the earlier stages of a processing plant's lifecycle.

Within a subsection of ASPEN's 'Plant Digital Twin', is the capacity for the digital twin to be used as a process model. This approach aims to unify the process model with digital twins once the plant has been constructed – using machine learning techniques to 'calibrate' the model to align with real operation [5]. Research by Boschet et al. highlights a similar approach to linking digital twin production to the conceptual design phase of process development. This approach translates a flowsheet into a 'template' for a digital twin, with a view that it can be filled with sub-models of the equipment as and when the data is made available through testing [13]. This approach to digital twin development, using an existing chemical process to provide the requisite data to develop the digital twin, appears to have been adopted in further industrial instances. As of yet, there has been limited focus on developing digital twins for use during process design, to influence equipment selection and other significant financial decisions before construction.

Beyond digital twins of entire chemical processes, there have been efforts to produce unit operation digital twins that can be connected to produce a flowsheet model. Some flowsheeting software packages, such as DWSIM, allow users to develop a custom neural network operation, based on a dataset [14]. This would allow a user to develop a simple digital twin of a piece of equipment, should they have a relevant dataset available. Such a unit operation conforms to the 'black box' approach from machine learning, wherein complex neural networks are used to model a process. Although very accurate at modelling the nuances of the specific system, black boxes have no capacity for a user to interpret how the model translates the inputs onto the outputs [15]. As a result, such an operation can be rigid and does not allow for user insight to modify how the unit operation is calculated.

On the other hand, white box models, based entirely upon the physics of the process, have flexibility as the parameters of the model can be easily manipulated. However, these techniques become very complex, especially in 3D flow scenarios [16] and, as such, in the context of flowsheeting, require significant simplifications to make the flowsheet solvable, due to computational power limitations. Therefore, it is proposed that hybrid 'grey-box' models, using both data modelling methods, such as neural networks, and information based on the physics of the process, would allow the model to be solved in a reasonable time frame, whilst achieving a higher level of accuracy than an over-simplified white box model [17].

Therefore, this project intends to explore the research gap of using digital twins of unit operations to specifically address flowsheeting limitations. The following objectives have been constructed to support the achievement of the project aim:

- Identify three suitable unit operations, and the limitations of their traditional models, to produce digital twins that address these shortfalls.
- Implement digital twins in open-source process simulation software DWSIM using Python code.
- Conduct analysis to compare digital twins to traditional flowsheeting models to determine benefits and drawbacks.
- Explore and discuss the future of grey box digital twins within flowsheeting and highlight any obstacles to their production.

Methodology

In line with the project's aim, three unit operations were selected, created, and tested to best investigate the potential of digital twins in addressing the limitations of flowsheeting. A complete flowsheet comprises numerous different operations and therefore a well-balanced investigation of their digital twin implementations is required. A pump, heat exchanger, and reactor were chosen as, owing to their governing engineering relationships, they range in complexity.

The digital twins were coded in Python to then be embedded in a flowsheeting software. Accordant with the recognition that proprietary software is somewhat more developed than its open-source counterparts, and with the view of addressing the limitations of flowsheeting, the open-source software DWSIM (version 8) was selected [18]. DWSIM was specifically chosen as it has the infrastructure to easily integrate custom scripts directly into the software. Additionally, it is CAPE-OPEN compliant – CAPE-OPEN is a universal set of standards for process modelling software – meaning it supports common thermodynamic property packages. Also, CAPE-OPEN unit operations are more easily translated to other flowsheeting software [19].

Pump

To model exact pump behaviour, pump curves were sourced from Grundfos, a global pump manufacturer, as the datasets were readily available in an XLSX format for the entirety of their product range [20]. The datasets contained the following discrete data: head, power, efficiency and NPSH, all against flowrate. Polynomial regression models were applied to the datasets such that any desired pressure change within the given operating range could be achieved. An assumption was made that the pump would be operated with an inverter.

A python script defining the digital twin was implemented into the DWSIM user interface, which allows for a user-defined input stream and desired pressure change. The outlet pressure is calculated and compared to the pump curve to ensure that operation was feasible, i.e., the desired outlet pressure is within the operating range of the pump. Then, the outlet stream fluid enthalpy is calculated using **Equation 1**, where m is mass flowrate [kg s^{-1}], h is enthalpy [kJ kg^{-1}], P is power [kW], and η is efficiency [-]:

$$m_{out}h_{out} = m_{in}h_{in} + P\eta \qquad [1]$$

The outlet pressure and enthalpy sufficiently define the pure, single-phase stream as per Gibbs' phase rule. A layer of validation was implemented to ensure the pump digital twin was operating within the range of its physical counterpart. For example, checking the temperature to ensure the fluid is entirely in the liquid phase, and ensuring the difference between the inlet pressure (P_{in}) and NPSH, taken from the pump curve, is greater than the bubble pressure (P^*_{bubble}) for the liquid, i.e., the real pump will not cavitate (**Equation 2**).

$$P_{in} - NPSH > P^*_{bubble} \qquad [2]$$

Pump Library

With the pump successfully addressing the accuracy limitation, it was further developed to explore its applicability in tackling the additional flowsheeting limitations around process equipment selection and costing. For the pump digital twin, a single pump dataset is inputted to DWSIM, whereas the pump library can ingest multiple pump datasets. The workflow of the pump library is, given a user-defined inlet stream and desired pressure change, the code simulates the digital twin for each pump dataset provided, selecting the optimal pump for the given scenario, whilst still outputting the outlet stream data. The optimal pump is determined by a simple optimisation model, where the user can select to optimise by price or power – i.e., if optimising based on price, DWSIM will output the least expensive pump that is still suitable for the given instance. This approach also used the aforementioned layer of validation to reject any non-viable pumps before the optimisation selection step. To assess the improved accuracy offered by the pump library, its cost and energy usage estimations for a set of scenarios were subsequently compared to an ASPEN pump model.

Heat Exchanger

The heat exchanger digital twin aimed to accurately predict the outlet temperature of the two streams, given two fully defined inlet streams. As the complexity of the heat exchanger is greater than that of the pump, its behaviour is dependent upon multiple different variables; therefore, a neural network approach was adopted, as it was apparent that a neural network would best represent the complexities of real system behaviour and non-idealities. To gain further insight, and develop the heat exchanger beyond the black box approach, the neural network was informed by heat transfer theory to explore whether a digital twin built on a water-water (water cold and water hot stream) dataset could accurately predict heat exchanger behaviour for other fluids.

The dataset fed to the neural network comprised four inlet and two outlet variables: the streams' mass flowrate, m_{in}^{C}, m_{in}^{H}; the inlet temperature of the cool and hot streams, T_{in}^{C}, T_{in}^{H}, and the temperature of the outlet streams, T_{out}^{C}, T_{out}^{H}. Whilst it is plausible that manufacturers have testing rigs capable of generating this dataset, a synthetic version was used as, currently, these datasets are not readily available to consumers.

The synthetic water-water dataset was simulated using an ASPEN heat exchanger model by performing a four-variable sensitivity analysis on the inlet operating parameters to determine the hot and cold outlet stream temperatures. Furthermore, the identical model was used to simulate testing data for different fluid cases. It was ensured that the model operated solely in the liquid phase, to avoid further complexities during this initial investigation.

Instead of training the neural network on mass flowrate and temperature, it was trained on a pre-processed dataset, containing the variables $(mC_p)^H$, $(mC_p)^C$, T_{in}^H and T_{in}^C, capturing the thermal inertia of the fluid – enabling the digital twin to be tested for its applicability to other fluids. The neural network is then able to predict T_{out}^H and T_{out}^C for a given instance. The DWSIM implementation of the heat exchanger calls the pre-trained neural network and feeds the inlet temperature and heat capacity flowrate terms (the product of the mass flowrate and fluid specific heat capacity) for both streams and predicts the temperature of both outlet streams.

Testing was conducted to verify if the heat exchanger digital twin yielded accurate predictions of the temperature of the outlet streams. A single test set involved randomly selecting 100 different permeations of inlet conditions from the testing dataset, collecting the digital twin output predictions for each permeation, and calculating a single R^2 value to compare the predictions against the actual values. It was found that running a single test set multiple times caused the R^2 value to vary considerably, hence it was decided to run each test set 25 times to calculate the mean and standard deviation of the R^2 value to account for sampling bias. It was desired to understand the impact of the number of water-water training data points on the accuracy of prediction for each case. Hence, this testing procedure was repeated for each case using neural networks of varying training data points, n.

Reactor

To investigate the potential of a reactor digital twin, a simple, acid-catalysed esterification reaction of ethanol and acetic acid to form ethyl acetate was chosen. Due to the complex nature of a reactor, two different approaches were considered to produce a reactor digital twin.

The first approach was informed by reaction engineering to build upon the pre-existing, traditional reactor models by addressing their limitations by using real data in lieu of a poor ideal assumption. More specifically, this digital twin incorporated a real residence time distribution (RTD) to remove the assumption of perfect plug flow and, therefore, conformed to the more desirable grey box modelling approach. The second approach, like the heat exchanger, employed a neural network fit to predict the behaviour of the reactor. This methodology will be referred to as a black box approach as the inlet variables are fed to the neural network – which is essentially a black box that outputs the result of the reactor in the form of fully defined outlet streams.

Approach 1: Grey Box

An experimental RTD was sourced from a research paper [21]. As an RTD is specific to a flowrate, reactor geometry, and volume, the digital twin built is specific to the reactor from which the RTD was captured. **Equations 3 & 4**, sourced from H. S. Fogler, were applied such that conversion could be obtained from the digital twin [22].

$$E(t) = \frac{C(t)}{\int_0^\infty C(t)\,dt} \qquad [3]$$

$$\bar{x} = 1 - \int_0^\infty e^{-kt} E(t)\,dt \qquad [4]$$

C(t) is the RTD for a pulse tracer test, E(t) is the probability density function, k is the first order rate constant, t is time, and $\bar{x}$ is the mean conversion. To understand if the digital twin offered improvements to the ideal model, the testing approach involved varying the rate constant across a broad range, to highlight whether conversion predictions differed between the two.

Approach 2: Black Box

The black box reactor's neural network ingested data that informed it on the impact of reactor temperature, feed flowrate, feed ratio, and reactor volume on reactant conversion. For simplicity, the reaction was defined as isothermal. Whilst it was appreciated that this is a cumbersome dataset that would not be available from a manufacturer, this approach was followed to gain a greater understanding of the use of neural networks to model digital twins of unit operations with multiple dependent variables. Hence, similarly to the heat exchanger, sensitivity analysis was conducted in ASPEN to synthesise a large dataset both for training the neural network and testing its accuracy.

Kinetic data were not required for the neural network as it is a black box which has an inherent appreciation of the complex relationships between inlet and outlet variables. However, the synthetic model, in ASPEN, used simple Arrhenius kinetics to simulate the dataset [23]. Unlike the grey box, this approach did not require comparison to the ideal model; since the neural network is built using a 'real system' dataset, if its predictions are accurate, it is inherently, by definition, a true representation of the real system. Hence, testing was rather carried out to determine the accuracy of the neural network's fit to the dataset. Importantly, statistical analyses were performed to better understand the impact of the number of training data points, n, on the accuracy of predictions.

Results

In line with the third project objective, the relevant outcomes of the three-unit operation analyses are shown to establish their capabilities in serving as improvements on traditional models.

Pump

Building upon traditional flowsheeting software, the output of the pump library, as well as a complete stream table, includes a pump curve displaying where the user's pump is operating, as shown in **Figure 1**.

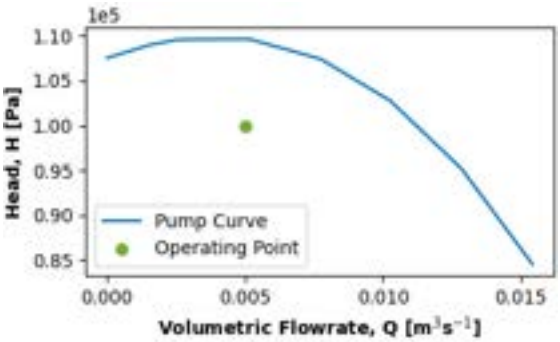


Figure 1 Pump curve of volumetric flowrate against head with operating point displayed from DWSIM pump library simulation. Data for this curve was sourced from Grundfos [20].

For the pump library, if an optimal pump is selected, a pop-up occurs, notifying the user of the model of the optimal pump, as well as its power or price requirement – depending on whether power or price optimisation was selected upon running the pump library. In the eventuality where an optimal pump could not be provided, as none of the pumps in the library had suitable operating ranges for the user's requirements, or the validation parameters (e.g., temperature and NPSH) were violated for all pumps, an error is passed to the command line.

Table 1 Table showing results of costing and power requirement comparison between ASPEN model and DWSIM digital twin.

Scenario		1	2	3
Pressure Change [bar]		10.0	5.0	1.0
Flowrate [kg s^{-1}]		5.0	1.0	5.0
Cost [£]	DWSIM	5710	5370	2090
	ASPEN	5810	3900	4320
Power [kW]	DWSIM	9.01	3.92	0.92
	ASPEN	8.46	1.70	0.99

The output of the pump library is dependent on the data supplied and, therefore, the performance of the pump library, relative to the optimisation functions, would improve monotonically as more pumps are included. For this analysis, 20 Grundfos pumps were provided to the digital twin. As this is a relatively small number of pumps, compared to the product range offered by Grundfos, operating points were carefully selected such that these results highlight the potential benefits of the pump library. Therefore, simulations were run for pressure increases from 1 to 10 [bar] and flowrates from 1 to 10 [kg s^{-1}], with key results summarised in **Table 1**.

For most simulations, the DWSIM pump cost was lower than ASPEN's, e.g., scenario 3 - £2090 compared to £4320, supporting the hypothesis posited by Walter et. al, that ASPEN costing accuracy lies within a -50 to 100% of the actual value [7]. In contrast, the ASPEN energy usage was generally lower than DWSIM, e.g., scenario 2 – 3.92 [kW] compared to 1.70 [kW]. A 130% difference in energy usage is surprising, as ASPEN does account for a pumping efficiency parameter, but the exact pump curve efficiency taken in the DWSIM module appears to provide a markedly different value to ASPEN's estimation. These inconsistencies justify the need for a digital twin as they can offer an accurate cost estimation.

Heat Exchanger

For the case of the heat exchanger, having verified the digital twin was successfully determining the outlet stream temperatures for the water-water fluid case, it was further tested on different fluid cases, **Table 2**.

Table 2 Table of heat exchanger digital twin fluid test cases.

	Case 1	Case 2	Case 3
Cool Side	Ethanol	Wtr-Eth*	Heptane
Hot Side	Ethanol	Water	Heptane

** Water-Ethanol mixture 1:1 molar ratio*

As the R^2 value of the testing predictions is the key parameter for this digital twin's performance, its mean and standard deviation are plotted in **Figure 2**.

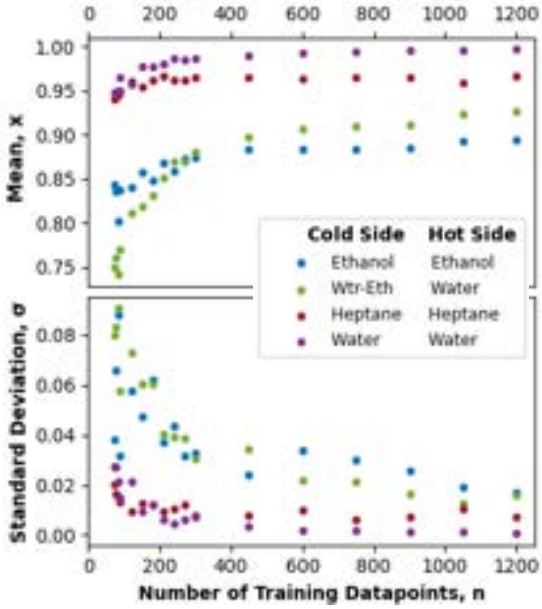


Figure 2 Graphs showing statistical analysis of the impact of the number of neural network training points, n, on the R^2 of heat exchanger predictions. Top graph: mean R^2 vs n. Bottom graph: standard deviation of R^2 vs n.

It is apparent from the water-water curves (purple data points in **Figure 2**) that the heat exchanger digital twin provides accurate predictions of the

outlet stream temperatures, even when trained with relatively few data points. For example, at n = 120 where, the mean R^2 value was 0.982, and the standard deviation was 0.006, meaning the neural network consistently gave accurate estimations of the testing data. The three remaining curves for cases 1-3 follow similar trends for both the mean and standard deviation: as the number of training data points increases, so does the mean of the R^2 values, and the standard deviation decreases. This analysis verifies that the digital twin, although built using water-water heat exchanger data, serves as a somewhat good predictor of heat exchanger behaviour for the other fluid cases. Thus, reducing the number of physical tests required from a manufacturer as one dataset can accurately simulate predictions for other fluids. Although promising, there is a need for further investigation into this approach as its potential, and limitations, need full and careful consideration. Such further investigation could involve the use of a real heat exchanger, rather than a synthetic dataset.

Reactor

Approach 1: Grey Box

To assess the potential benefits of the grey box approach over the traditional ideal model, a plot of conversion against the kinetic constant for both the digital twin (blue) and ideal plug flow reactor model (green) was developed, as shown in **Figure 3**.

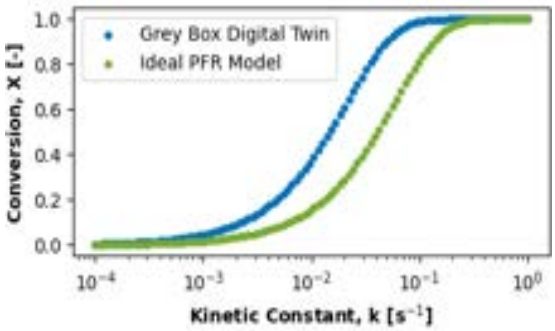


Figure 3 Graph of conversion against varying kinetic constant for an idealised reactor model (green) and the grey box reactor digital twin, incorporating a residence time distribution (blue) to better model real system behaviour.

Between a kinetic constant of 0.001 [s^{-1}] and 0.1 [s^{-1}], there is a marked difference in conversion, with the greatest difference at k = 0.032 [s^{-1}]. At which point, the ideal plug flow reactor (PFR) model predicts a conversion of 0.42, compared to 0.78 predicted by the digital twin. The difference in conversion verifies the need for a digital twin as an improvement on the idealised PFR model. This significant difference in conversion is present because of axial dispersion within the reactor, which yields an asymmetrical residence time distribution.

The substantial benefit provided using an RTD is its incorporation of the impact of axial dispersion. In their study, Obonukut and Bassey, [21], highlighted that, despite having a high Reynolds and Peclet number, their real reactor experienced substantial axial dispersion. It is in instances such as these, where the idealised model yields such an under prediction of conversion, that the fundamental insight offered by a digital twin can significantly improve the accuracy of predictions.

Approach 2: Black Box

As aforementioned, the 'black box' neural network inherently contains an appreciation for the exact nature of any non-idealities present in the system. Therefore, analysis was not required to determine the accuracy of the prediction as inherently it incorporates all non-idealities as based upon a real dataset. Instead, the KPI of this digital twin was not based on a comparison to an ideal model, but rather, the fit of the neural network to the parent dataset. Like the heat exchanger digital twin, the accuracy of the black box digital twin's prediction was measured through the R^2 value. **Figure 4** shows a graph of the standard deviation and mean of the R^2 value against the number of data points used to train the reactor neural network.

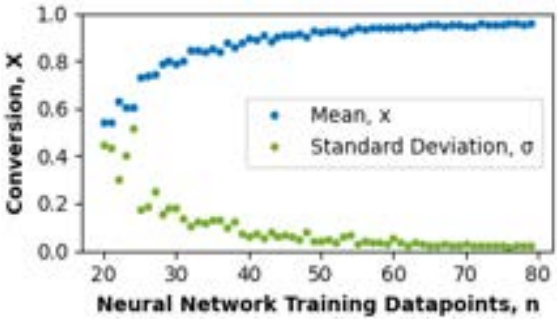


Figure 4 Graph of the relationship between the mean and standard deviation of the R^2 value for the black box reactor digital twin, against the number of neural network training data points, n, for reactor approach 2.

At 43 training data points, n, an R^2 of 0.9 is achieved, and, at 66 data points, an R^2 of 0.95 is achieved. Equally, once the training data points surpass 40, the standard deviation is lower than 0.065. This illustrates that the digital twin can both accurately and precisely predict the outlet conditions of the reactor effluent, given a set of initial conditions.

From a computational standpoint, it would be feasible to run a neural network trained on a number of data points in the hundreds, even thousands, to achieve an even stronger R^2 value. However, the limitations stand with the acquisition of such substantial amounts of data, specific to the reactor being modelled. Whilst these results show the black box reactor approach can be highly accurate in modelling the real behaviour of a reactor, the required data comes at a cost. It is unrealistic to expect manufacturers to be able to provide such a detailed dataset that essentially documents a full-variable sensitivity analysis of their equipment for all conceivable reactions.

Discussion

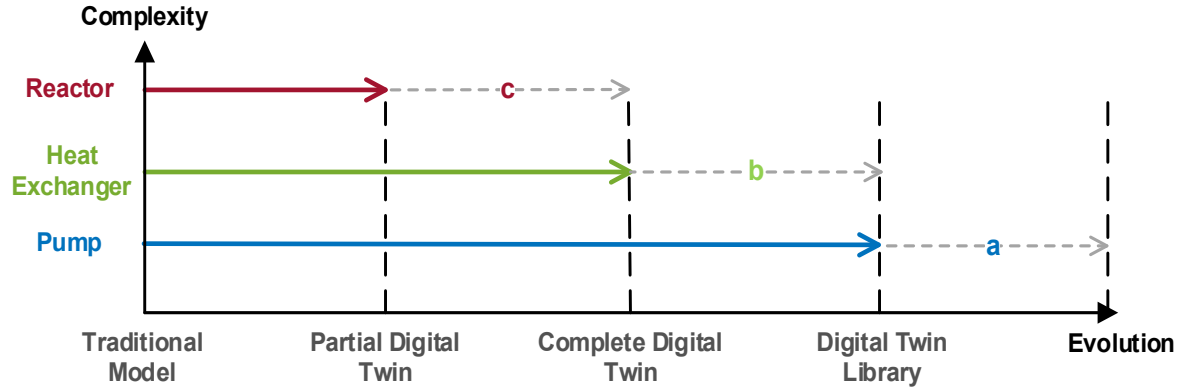


Figure 5 Schematic of digital twin 'roadmap', demonstrating the evolution of the unit operation digital twins developed within this project; labels a, b and c represent the next step of each digital twin. Whilst the figure plots the three discrete digital twins produced, as these twins are vehicles to represent varying complexity levels, all pieces of process equipment, with their corresponding complexity, exist on this evolutionary axis.

The varying complexity of the three-unit operation digital twins resulted in three digital twins at different stages of evolution. These different stages are plotted on an evolutionary axis, as shown in **Figure 5,** with the next steps for each digital twin highlighted by arrows, a, b, and c. In line with evaluating the potential of unit operation digital twins to address the limitations of traditional flowsheeting models, it is necessary to discuss their adherence to the grey box model definition, highlight any obstacles to their production, and discuss the potential uses of the finished products.

Pump Library

As the production of the pump library utilises both real data models and fluid mechanics theory, it adhered well to the definition of a grey box model. It showed success in addressing the flowsheeting limitation of inaccurate cost estimation and equipment selection. To build upon this success and take the next step in the development of the pump library digital twin, as shown in **Figure 5** arrow a, key challenges on data availability and collection need to be overcome.

Whilst pump curve graphs are readily available from manufacturers, currently, Grundfos is one of the few manufacturers where the raw data can be downloaded, as opposed to just being able to view graphs. The availability of the raw data in an appropriate format is fundamental to building and embedding the pump digital twin in a flowsheeting environment. There is potential for online tools to be employed to extract the data, however, this itself has its own limitations and is dependent on the graph format. Even with data being available in a convenient format, to fully realise the potential of the pump library, collecting this data needs to become a more automated process. As discussed in the results, the pump library's prediction monotonically improves with a larger pump library; hence, greater automation of the data collection process is important. Furthermore, whilst a primitive optimisation algorithm has been implemented in this project, further research into the multi-objective optimisation of pump choice would further facilitate a more flexible, user-customisable, selection.

Heat Exchanger

As the heat exchanger digital twin is built using a neural network informed by heat transfer theory, it too adheres to a grey box model definition. Within this category there are some digital twins, like the heat exchanger, that have a significant reliance on data modelling techniques, thus resulting in some black box-related restrictions, e.g., limitations in capacity for a user to understand the translation from input to output. The heat exchanger did show preliminary success through its potential to improve upon the poor accuracy of traditional flowsheeting models. Furthermore, it demonstrated potential in using a single heat exchanger case to predict the behaviour of different fluid cases. To further develop the heat exchanger digital twin, as shown in **Figure 5** by arrow b, the following challenges on data availability and application of heat transfer theory need to be overcome.

Whilst manufacturers do have suitable datasets of heat exchanger operation, it is something not currently readily available to consumers. To verify the findings from this project, further heat exchanger analysis must be conducted with real data. Additionally, in this early stage of investigation, an assumption of fully liquid operation was made – further considerations of how this digital twin can be applied to all heat exchanger instances, e.g., boilers and condensers, are required. Further validation of the potential to generalise the heat exchanger for use on other fluids by training the neural network on heat capacity flowrates is also recommended.

Reactor

The first approach to developing a reactor digital twin, as it applied real data to reaction engineering

theory, conformed well to the definition of a grey box model. In contrast, approach 2 was a paragon of a black box model as it was based solely on a neural network. Whilst approach 2 resulted in an accurate digital twin of a complex unit operation, it bears significant resemblance to the approach demonstrated by Boschet et al. [13], as this digital twin can only be developed once the equipment has been built and operational data collected. It is due to this, and the fact a manufacturer would not be able to provide such a dataset, that approach 2 (black box) will not be considered a viable approach for future reactor digital twins.

The grey box approach to producing a reactor digital twin showed significant success in addressing the flowsheeting limitation of poorly approximating real system behaviour. This was achieved through the removal of the perfect plug flow assumption by incorporating a real RTD. A natural next step in the advancement of this digital twin, as shown by arrow c, would be in addressing limitations present in the availability of the data required to produce the digital twin. RTDs are specific to a reactor's geometry and inlet flowrate and, therefore, for an unrestricted application to flowsheeting, multiple RTDs would be needed to build a digital twin of a single reactor. It is apparent that most reactor manufacturers have RTD data internally, but this data is not easily accessible to a consumer. A further consideration would be generalising this modelling for more complex kinetics and different types of chemical reactors.

Summary of Unit Operations

Whilst the three unit operations provide a demonstration of the potential of digital twins to address the limitations of flowsheeting, due to their differing complexities, each holds its own challenges in realising said potential. A shared theme between these digital twins is in the varying availability of consistently formatted, relevant, and complete data. Naturally, as systems of higher complexity are governed by an increasing number of variables, sourcing data that matches these requirements becomes increasingly difficult.

As per their definition, grey box models hinge upon two key pillars, models informed by real data, and incorporation of the process' physics. As such, is foreseen that with improvements in data availability and further focus on the implementation of engineering theory, the potential of digital twins of unit operations within flowsheeting could be further realised.

Flowsheeting Applications of Digital Twins

With the results validating the improvements made by unit operation digital twins, it is equally important to consider where their potential can be applied within flowsheeting and in a broader chemical engineering context. A key benefit provided by these digital twins is that there is a substantial incentive for manufacturers to produce digital twins of their process equipment to further incentivise customers. For instance, a digital twin of a physical piece of equipment could be distributed with every purchase, thus enabling customers to run operational tests virtually, saving time and resources. For this to be feasible, ensuring the digital twins are CAPE-OPEN compliant is important as this would facilitate customers importing them into a CAPE-OPEN compliant flowsheeting software of their choosing. Equally, before the acquisition of the equipment, digital twin application programming interfaced (APIs), similar to the pump library, could be implemented on a manufacturer's website, such that customers receive recommendations of the most suitable piece of equipment for their requirements.

Alternatively, from a process design perspective, the digital twins could be further developed and implemented by flowsheet software companies. This would incentivise manufacturers to provide datasets for their product range, to promote their process equipment to customers. As demonstrated with the pump library this would ensure that the best unit operation is put forward to customers, encouraging competitive pricing from the manufacturers.

Digital Twin Applications Beyond Flowsheeting

Although the application of digital twins in flowsheeting is an area of significant potential, it is important to recognise that the current research areas of operation scheduling, maintenance, and process control, are still highly relevant. Therefore, the prospect of integrating these flowsheeting digital twins with these operations digital twins could be an area of further exploration. A key consideration is that the flowsheeting digital twins, produced in this project, are steady state models, not dynamic models which are more suitable for process control applications.

As such, further exploration into the production of dynamic unit operation digital twins could facilitate the implementation of an operations digital twin. Furthermore, in the context of process control optimisation, a highly specific and accurate digital twin is required to model the behaviour of the system in response to disturbances with a sufficient degree of accuracy. Therefore, the dynamic digital twin would need further tuning, once construction and testing have been completed, to give a better emulation of the real system. This calibration of a dynamic model to real system behaviour adopts a similar approach to that highlighted by R. Beck in ASPEN's white paper on the future of digital twins [5].

Conclusion

In conclusion, this project aimed to investigate the potential of employing unit operation digital twins to address flowsheeting limitations. This overarching aim was supported by four key objectives: building digital twins that serve as an improvement on idealised models, implementing the digital twins into DWSIM, conducting analysis to determine their benefits and challenges, and, finally, exploring where the potential of the digital twins could be applied within the flowsheeting industry.

The methodology was underpinned by the selection of three suitable unit operations – a pump, heat exchanger and reactor. These unit operations offered varying levels of complexity, number of dependent variables, and data requirements, hence serving as appropriate focal points of the investigation.

After constructing the digital twins' Python code, the DWSIM implementations enabled analysis that determined their potential as improvements upon current models. The pump library showed strong potential as it saw up to a 50% improvement in cost estimation, as well as facilitating the selection of process equipment. In the instance of the reactor digital twin, in addressing the ideal model assumption of perfect plug flow through the application of a real RTD dataset, its output gave a closer approximation of the real system's behaviour, with up to a 36% improvement made in conversion estimation. The heat exchanger also showed similarly promising accuracy, with R^2 values in the range of 0.85 to 0.98 for different fluids, when trained on over 300 data points.

There are two fundamental aspects to a grey box digital twin: the use of models built upon real data, and an incorporation of engineering theory. As well as recognising the benefits of the digital twins, consideration was given to identify the challenges they currently face. This project observed that improvements in data availability alongside a continued focus on implementing the engineering theory would alleviate these challenges.

It has been concluded that, with consideration of the identified limitations, the potential of unit operation digital twins could serve as improvements to traditional flowsheeting models. As such, it is conceivable that there is sufficient incentive for equipment manufacturers and flowsheeting software providers to develop this technology further. Furthermore, there is future potential for these flowsheeting digital twins to be homogenised with pre-existing operations digital twins, facilitating the production of a singular digital twin to support flowsheeting, dynamic modelling, and control system optimisation.

Outlook

Having established digital twins have the potential to address flowsheeting limitations, this outlook aims to outline possible avenues for development upon this project. Although recommendations will be made for each unit operation, focus will be given to the heat exchanger and reactor digital twins because, as shown in **Figure 5**, they are not as developed on the evolution scale relative to the pump library.

Concerning the pump library, the next steps could consider how to integrate multiple manufacturers into one digital twin. This would involve considering how to incorporate pump curve datasets of different formats as, naturally, the format will differ by manufacturer.

The grey box reactor approach successfully removed the assumption of perfect plug flow; however, further traditional, idealised assumptions remain. More specifically, the assumption of isothermal operation is a severely limiting one; temperature has a significant impact on reaction kinetics, resulting in poor approximations of reactor behaviour. For example, Arrhenius kinetics observes an exponential relationship with respect to temperature. Building upon the current grey box model, an investigation on how temperature could be more accurately captured by the digital twin would offer a significant improvement of real system behaviour predictions. Additionally, pairing the RTD with a specific set of reaction kinetics would allow inference on how the digital twins could be applied realistically in flowsheeting, as the grey box model was applied across a wide range of kinetic constants to verify the difference between its conversion and that of an ideal PFR.

Furthermore, if digital twins are to be used to inform equipment design and costing, credibility on their accuracy is a necessity. Taking the heat exchanger as an example, verifying the potential demonstrated in this project using real heat exchanger rig data would be the first step in fostering this credibility. Moreover, a similar approach can be taken with the reactor digital twin as its results are based upon the synthetic ASPEN model dataset.

Due to the nature of flowsheeting encompassing a wide range of unit operations, it is pertinent to consider others beyond the three explored in this project and analyse any further complications and challenges they may bring.

References

[1] L. Evans, J. Boston, H. Britt, P. Gallier, P. Gupta, B. Joseph, V. Mahalec, E. Ng, W. Seider and H. Yagi, "ASPEN: An Advanced System for Process Engineering," *Computers & Chemical Engineering,* vol. Volume 3, no. Issues 1-4, pp. Pages 319-327, 1979.

[2] Z. N. Pintarič and Z. Kravanja, "Towards Outcomes-Based Education of Computer-Aided Chemical Engineering," *Computer Aided Chemical Engineering,* vol. 38, pp. 2367-2372, 2016.

[3] R. R. Schaller, "Moore's law: past, present and future," *IEEE Spectrum,* vol. 34, no. 10.1109/6.591665, pp. 52-59, 1997.

[4] Oliver Wyman, "Projected size of the global market for digital manufacturing in 2030, by industry segment (in billion U.S. dollars) [Graph]," Statista, 2016.

[5] R. Beck, "The Digital Twin and the Smart Enterprise," Aspen Technology, Massachusetts, 2019.

[6] Siemens PSE, "Siemens PSE," 2022. [Online]. Available: https://www.psenterprise.com/products/gproms. [Accessed 05/12/2022].

[7] O. Walter, A. Tremel, M. Prenzel, S. Becker and J. Schaefer, "Techno-economic analysis of hybrid energy storage concepts via flowsheet simulations, cost modeling and energy system design," *Energy Conversion and Management,* vol. 218, no. 112955, 2020.

[8] Edie Newsroom, "BP invests £19.5m in innovative pilot plant for 'unrecyclable' plastics," Faversham House Ltd, 2019.

[9] D. Jones, C. Snider, A. Nassehi, J. Yon and B. Hicks, "Characterising the Digital Twin: A systematic literature review," *CIRP Journal of Manufacturing Science and Technology,* Vols. 29, Part A, no. 1755-581, pp. 36-52, 2020.

[10] Unity, "What is a Digital Twin?," 2022. [Online]. Available: https://unity.com/solutions/digital-twin-definition. [Accessed 05/12/2022].

[11] AWS Amazon, "What Is Digital Twin Technology?," 2022. [Online]. Available: https://aws.amazon.com/what-is/digital-twin/. [Accessed 05/12/2022].

[12] A. M. Madni, C. C. Madni and S. Lucero, "Leveraging Digital Twin Technology in Model-Based Systems Engineering," *Systems,* vol. 7, no. 1, 2019.

[13] S. Boschet, C. Heinrich and R. Rosen, "Next Generation Digital Twin," in *TMCE 2018*, Las Palmas de Gran Canaria, 2018.

[14] D. Medeiros, "LinkedIn," 24 April 2018. [Online]. Available: https://www.linkedin.com/pulse/integrating-chemical-process-simulator-tensorflow-daniel-medeiros/. [Accessed 08/12/2022].

[15] J. Benitez, J. Castro and I. Requena, "Are artificial neural networks black boxes?," *IEEE Transactions on Neural Networks,* vol. 8, no. 5, pp. 1156-1164, 1997.

[16] Q. Meng, Y. Wang, X. Yan and Z. Li, "CFD assisted modeling for control system design: A case study," *Simulation Modelling Practice and Theory,* vol. 17, no. 4, pp. 730-742, 2009.

[17] P. Aivalioitis, K. Georgoulias, Z. Arkouli and S. Makris, "Methodology for enabling Digital Twin using advanced physics-based modelling in predictive maintenance," *Procedia CIRP,* vol. 81, pp. 417-422, 2019.

[18] D. Medeiros, "DWSIM Process Simulation, Modelling and Optimization Techical Manual," 2018. [Online]. Available: https://dwsim.inforside.com.br/docs/mobile/tech_manual.pdf. [Accessed 08/12/2022].

[19] D. Piñol, J. C. Rodriguez, M. Halloran, W. Drewitz, I. Richard Szczepanski, M. Pons, M. Woodman, P. Banks and J. v. Baten, "Thermodynamic and Physical Properties v1.0 CAPE-OPEN," *COLAN,* no. 1.08.008.DOC, pp. 1-87, 2011.

[20] Grundfos, "Grundfos," 2022. [Online]. Available: https://www.grundfos.com/uk/about-us. [Accessed 05/12/2022].

[21] M. Obonukut and P. Bassey, "Residence Time Distribution of A Tubular Reactor," *International Journal Of Scientific Research And Education,* vol. 4, no. 1, pp. 4767-4777, 2016.

[22] H. Fogler, Elements of Chemical Reaction Engineering 4th Edition, Philadelphia, PA: Prentice Hall, 2005.

[23] W. D. Seider, D. R. Lewin, J. Seader, S. Widago, R. Gani and K. M. Ng, Product and Process Design Principles - Synthesis, Analysis, and Evaluation (4th Edition), John Wiley & Sons, 2017.

Production of Carbon Nanotubes Through Electrolytic Reduction of Carbon Dioxide in a Molten Carbonate Salt

Marco Bruno Tome Freire, Daniel Binks
Department of Chemical Engineering, Imperial College London, U.K.

Abstract – Combating the release and encouraging the capture of CO_2 is humanity's main aim to reduce effects of global warming. In this study, a procedure published in literature for electrochemically reducing CO_2 in molten lithium carbonate electrolyte and produce carbon nanotubes (CNTs) was replicated. Lithium carbonate salt was the chosen electrolyte, iron and nickel were the chosen cathode and anode materials respectively. Electrolysis was carried out at 800 ℃ with a current density of 200 mA/cm^2, to achieve the following process: $Li_2O + CO_2 \rightarrow Li_2CO_3$ in the electrolyte, followed by electrochemical lithium carbonate splitting: $Li_2CO_3 \rightarrow C_{CNT} + O_2 + Li_2O$. The following methods of electrolyte and product analysis were investigated: simultaneous thermal analysis (STA), X-ray diffraction spectroscopy (XRD), Raman spectroscopy and X-ray fluorescence spectroscopy (XRF). STA was concluded to be a suitable technique for detecting changes to the electrolyte composition because of the electrolysis process, Raman enabled analysis of the structure of the carbonaceous deposit and XRF is best for determining compositional changes to the electrode materials. The relative novelty of this CNT production technique meant that it was challenging to produce carbon nanotubes and issues with electrode oxidation (Fe cathode to F_2O_3/Fe_3O_4 and Ni anode to NiO) were encountered. However, having set out the initial steps, future application of the written methodology would be easy to follow for further optimisation and upscaling.

I. INTRODUCTION

It is well understood that global warming is one of the main challenges that humans face on earth. Carbon dioxide (CO_2), the most bountiful anthropogenic greenhouse gas, has been increasingly emitted at an extremely rapid rate since the 1800s. Global CO_2 emission in billion metric tons has increased from 0.03 in 1800 to 34.81 in 2020[1]. Greenhouse gases (GHGs) like CO_2 absorb and emit thermal radiation from the sun, trapping this energy within the Earth's atmosphere and causing the planet to heat up. Current predictions estimate that temperatures worldwide will increase by 2 ℃[2].

To counteract this, there has been global initiative to limit GHG emissions and decrease the current concentration of CO_2 in the atmosphere. For the UK, this plan involves committing to a fully decarbonised power sector by 2035 and a ban on sales of all diesel/petrol powered cars by 2030[3]. To achieve the full decarbonisation of the nation's power grid there are multiple plans in place, including 50 GW of offshore wind capacity and a large investment into low carbon hydrogen production with the aim of achieving 5 GW capacity by 2030[3].

We know the implementation of renewable energy technology will not be enough to meet our current needs. Therefore, there is great importance on the technology of Carbon Capture and Storage (CCS) and Carbon Capture and Utilisation (CCU) to help renewables to reach net zero.

The difference between these two types of carbon dioxide reducing technologies is in the processing of captured CO_2. CCS takes the captured carbon dioxide and stores it in large, underground rock formations where it remains permanently, whereas CCU utilises the carbon dioxide in other processes.

If demand exists for a product made using CO_2, the case for CCU as the better option is strong. There are reduced costs in transportation and storage of the CO_2 and a greater potential to generate revenue as the captured gases can be sold to process plants. Industries that require carbon dioxide currently obtain it from collecting the waste products of burning fossil fuels. Considering fossil fuels will be part of our energy economy for years, this will prevent much CO_2 release.

Decarbonisation for a country such as the UK must be a priority, but carbon-based fuels will remain important to the global economy.

CCU cannot provide the emission mitigation rate of carbon capture and storage (CCS), but considering the UK, who's entire storage capacity for CO_2 is offshore, CCU could mitigate emissions from inland point sources.

Inland CO_2 emitting plants are prime targets for CCU activities, as they are unlikely to be part of CCS networks, which will be localised to coastal regions. Point sources generating of order 1 $MtCO_2e$ per year (24 of the UK's top 60 stationary GHG emitters emit 1 $MtCO_2e$ per year or less) could, in the short term, enable greater CO_2 processing by utilisation than by storage[29].

Up to 3.65 tCO_2 (per tonne of carbon nanotubes) potentially could be sequestered, with negligible CO_2 emissions[30]. Furthermore, the estimated energy requirement of carbon nanotube production is 40 times lower using the CCU method compared to the conventional process[31].

A current product that is expensive to produce but has many applications are carbon nanotubes (CNTs). CNTs are small, hollow tubes made exclusively of carbon atoms. Single-walled nanotubes have diameters in the range of 0.8 – 2.4 nm[4] and can stretch to a few micrometres in length. This arrangement of atoms produces strong covalent sp^2 bonds, which lead to the tubes possessing ultimate intrinsic tensile strength in the 100-200 GPa range[5]. They are as stiff as diamond and around 10x as strong as steel. Additionally, their thermal capacity is incredibly high, 20x larger than steel, making them resistant to thermal expansion. Chemically they are stable so are resistant to corrosion.

CNTs can be used in construction due to their desirable mechanical properties. There is also a growing interest in the medical research community for CNTs in efficient drug delivery and even as a treatment for aggressive cancerous tumours. The unique chemical properties of carbon combined with mechanical benefits of this specific structure make CNTs the ideal material for a range of medicinal techniques. Further applications can be seen on Figure 1. Clearly CNT production will be of greater importance in the future as the demand for them increases.

Figure 1: Usage and applications of CNTs[6]

CNTs are currently produced via three different methods: electric arc discharge, chemical vapour deposition, and laser ablation. These methods are energy intensive and expensive, prompting research into an alternative, more efficient production method. A potential alternative for production of CNTs is the reduction of CO_2 gas to solid carbon by molten carbonate salt electrolysis. Not only does this aid in the process of CCU, but a useful product with many applications is also made.

In molten lithium carbonate (the salt used in this experiment) a series of electrochemical processes occur in the conversion of CO_2 to CNTs. A voltage is applied in an electrolysis chamber where at a metal cathode CO_2 is split and CNTs are formed, and O_2 is produced at a metal anode[7]. Molten carbonates are necessary for CO_2 dissolution and therefore the continued creation of CNTs. First there is a reduction of the carbonate ion to form CNTs, oxygen and an oxide:

$$Li_2CO_3 \rightarrow C_{CNT} + O_2 + Li_2O$$

Then, CO_2 added to the electrolyte (through gas bubblers) dissolves and reacts exothermically with lithium oxide to recycle and reform lithium carbonate:

$$CO_2 + Li_2O \rightarrow Li_2CO_3$$

We have documented a method of CNT production at a lab-scale process through a custom salt heating cycle. The aim of this study was to provide a clear and complete method of CNT production and post-electrolysis analysis not presently found in research. This can be further optimised and upscaled in future developments.

II. BACKGROUND

It is important to first discuss the current methods of CNT production to gain an appreciation of how difficult and expensive these procedures are.

In electric arc discharge, a direct current is established between graphite electrodes kept in an inert atmosphere (helium or argon). High temperatures between the electrodes (3000-4000℃) causes carbon sublimation. Sublimated graphite is deposited at the negative electrode or the walls of the chamber where the process is carried out. These deposits contain CNTs[8]. The arc discharge method is a high energy process and requires high precision control which deters the scalability of the production of graphene and hence its possible applications[9].

In chemical vapour deposition (CVD), a substrate is prepared with a layer of metal catalyst particles (nickel, cobalt, iron). The substrate is then heated between 600-1200℃ and a mixture of gases are injected into the reactor. This includes a process gas (ammonia, nitrogen, or hydrogen) and a carbon-containing gas (ethylene, acetylene, methane). CNTs grow on the catalyst particles in the reactor and are collected after the reactor is cooled. The catalyst particles can remain at the bottom or top of the growing CNTs[10]. Some existing challenges are that post treatment of mass-produced CNTs is difficult. Furthermore, CVD-grown CNTs have poor crystallinity[11].

In the laser ablation technique, a high-power laser was used to vaporise carbon from a graphite target at high temperature while an inert gas (argon, helium) is bled into the chamber[12]. The basic principle of laser ablation is simple and easy to perform, but it is expensive and the production rate of CNTs could be improved[13].

There have been many research papers studying the production of CNTs via high temperature electrolysis in molten carbonate salts. However, the kinetics of the key electrochemical reactions and the relationship between the structures of the formed carbon nanotubes and the reaction conditions are still not well understood. Many electrode materials including monel, steel, galvanised steel and copper have been investigated. A wide variety of alkali and alkaline earth carbonate salts with different mixing ratios have been tested as the electrolyte. To find agreement on the optimal combination of electrode material, electrolyte composition and other reaction variables, further investigation is necessary. The parameters which can be used to tune the CNT structures also need to be further investigated.

We must also consider the ability of the CNT production system to be incorporated into existing gas and coal-powered energy plants. Lau et al[14]. demonstrate a thermodynamic model analysis for a molten Li_2CO_3 electrolysis system incorporated within a combined cycle (CC) natural gas power plant to produce both CNTs and oxygen. There are several major energy efficiency losses in the original CC plant. This plant system then generates electricity at higher

efficiency due to pure oxygen looped back to the gas turbine input from the CO_2 splitting. Enriched oxygen combustion allows for the combustion chamber to reach higher temperatures and combustion efficiencies, improving thermal energy efficiencies of the gas turbine as well as the steam turbine. Furthermore, the hot CO_2 product from the gas turbine is an excellent reactant for the molten electrolysis.

III. METHODOLOGY

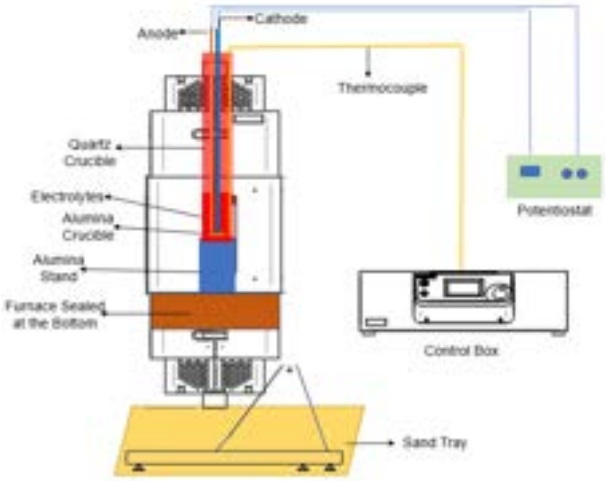


Figure 2: Diagram of furnace and peripheral equipment used in electrolysis

Figure 2 was adapted from a design by Zhixu Zhu and other group members in Dr. Anna Hankin's research group and displays the experimental setup.

A TS1 12/125/400 tube furnace along with a vertical stand and a control box was purchased from *Carbolite Gero*. TS1 indicates the furnace has a single heating zone and can be split into two halves for easy access to the crucible inside. 12 indicates the maximum furnace operating temperature is 1200°C. 125 indicates the maximum diameter of tubing that the furnace can accommodate in mm. 400 indicates the heating length of the tube furnace in mm.

The furnace was customised to have only one open end; the bottom of the furnace was left uncut. The upper aperture was designed with a diameter of 100 mm. The maximum wall thickness of the cylindrical crucible to be placed in the furnace was 3 mm, leaving a 2 mm gap between the furnace and the crucible.

A gantry was assembled with aluminium profiles supplied by *KJN Aluminium Profile*. The gantry was designed by Zhixu Zhu and Dr. Inyoung Jang and the adapted diagram can be seen in Figure 3. Clamps were set onto the gantry to hold the experimental electrodes.

40 g of the electrolyte, Li_2CO_3 (Alfa Aesar, 99.0% purity in powdered form), was added into an alumina crucible (15cm length, 10cm width, 6 cm depth). The heating process included heating the salt to 800°C at a slow rate of 4 °C/min, and the cooling rate was even slower at 0.5 °C/min to 650°C to prevent thermal shock. The reaction chamber needs to withstand high temperature. It was important to choose appropriate materials for safety concern.

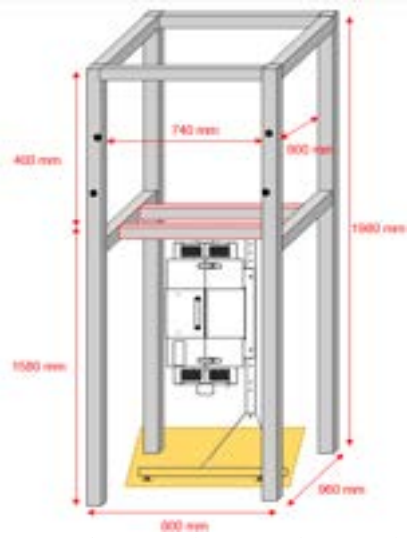


Figure 3: Diagram of gantry used to support furnace and electrolysis cables

A custom designed quartz crucible was a 600 mm tall cylinder with a 106 mm outer diameter and a 101 mm internal diameter, a diagram of which can be seen in Figure 4. The heating length of the furnace was only 400 mm, meaning that the quartz crucible was elevated by an alumina crucible (60 cm height with 12 cm diameter).

The rectangular crucible was placed inside the large quartz crucible for added safety.

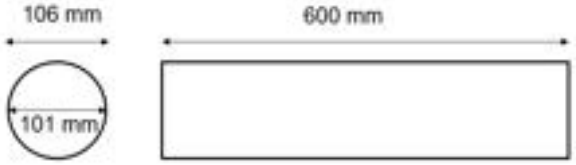


Figure 4: Quartz crucible dimensions

A specifically designed lid for two spiral disk wire electrodes was used, as can be seen in Figure 5. The lid has the function of holding the gas inlet and electrodes in position, therefore preventing the contact of the anode and cathode, leading to a short-circuit.

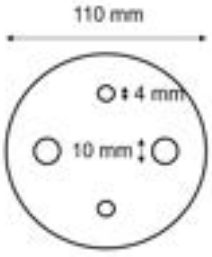


Figure 5: Quartz crucible lid dimensions

Cables were connected to the electrodes via silver wire, which is extremely heat resistant. Silver also has the best conductivity of any known metal[15]. For extra protection, the cables were insulated with alumina tubes. High purity alumina is usable in both oxidising and reducing atmospheres to 1925 °C[16]. The potentiostat

was placed outside of the fume cupboard to prevent any signal interference from the furnace.

5 m of nickel (99.95% purity) and iron (99.98% purity) wires of 1 mm diameter were purchased from *GoodFellow*. The metals were polished and acid washed to remove any impurity and oxides on the surface before putting to use as electrodes.

The two electrodes were spaced apart with a 5 cm piece of alumina tubing.

Electrolysis was performed at current densities of 200 mA/cm^2 at a temperature of 800 ℃**.** Electrolysis time was planned to be 2 hours, but electrolysis was not completed due to a large increase in resistance during operation. The potentiostat used was an *Autolab PGSTAT302N* produced by *Metrohm*.

After electrolysis, the products deposited at the cathode and anode were scraped off and were subjected to identification analysis.

Compositional analysis of the experimental electrolyte in its original and thermally decomposed form (Li_2CO_3 and Li_2O) was conducted using x-ray diffraction (XRD) on an Xpert Pro PANalytical XRD. The x-ray generator had a tension of 40 kV and a current of 20 mA. A full scan range from 5 to 90 degrees was used, with a step size of 0.0334 and time per step of 40 seconds. Following this, simultaneous thermal analysis was conducted using a Netzsch STA 449 F5 Jupiter. Simultaneous thermal analysis (STA) applies thermogravimetry and differential scanning calorimetry at the same time. Thermogravimetric analysis measures the mass of a sample over time as temperature changes. The molecular weights of the compounds were used to calculate the theoretical mass change of our sample. This was used to calculate the moles of lithium carbonate converted to lithium oxide in our electrolysis.

Raman spectroscopy using a *Senterra II* was conducted on CNT products and commercial CNTs. X-ray fluorescence (XRF) analysis, a non-destructive technique used to determine the elemental composition of a sample aided electrode product identification. A PANalytical Epsilion XRF Spectrometer was used.

IV. RESULTS AND DISCUSSION

Experimental Setup:

The furnace design was customised, leaving space for the crucible's expansion while keeping good extent of seal to prevent heat loss. The gantry was designed to hold the peripherals of the reaction setup, such as cable lines.

The electrodes chosen were iron and nickel wire because of their common appearance in literature. However, there are a multitude of electrode materials available for use in electrolysis. There aren't preferred electrode materials in literature and so an investigation must be made into their corrosion resistance, product yield, and product morphology.

Three materials were compared for crucible fabrication: nickel, quartz, and alumina (aluminium oxide) ceramic. Nickel can work as an anode whilst simultaneously being the reaction container. This can save space in the electrochemical cell, and electrode surface area can be maximised. Nickel also has the highest mechanical strength among the three materials. However, at temperatures greater than 700 $°C$, nickel will oxidise and will therefore cease to be conductive. Not only this, but its mechanical strength will also decrease. Quartz reactors were reported in literature to contain high temperature molten salts up to 1000 degrees for methane pyrolysis[17].

After running several experiments, it was discovered that molten Li_2CO_3 was etching into the quartz crucible. According to the supplier *Multi-Lab*, lithium, sodium and potassium salts can be used as flux for the quartz containers[18]. The quartz crucible also cracked due to the volume change of molten Li_2CO_3 as it cooled rapidly.

Alumina crucibles have been reported many times as electrochemical cells for electrolysis of molten carbonate salts[19].

To ensure CNT growth occurred at a constant rate with minimisation of defects, the current of electrolysis was fixed. The value of current density chosen was 200 mA/cm^2. Previous studies found that densities between 200-400 mA/cm^2 were sufficient for CNTs to form, with an increased current density favouring the generation of carbon products of lower particle size[20]. The potentiostat had a maximum current rating of 1 A (1000mA) and the set current was dependent on both the current density and surface area of the electrode

$$I = J \times A \quad (1)$$

$$where\ I = current\ [mA],$$

$$J = current\ density\ [mA \cdot cm^{-2}],$$

$$A = electrode\ surface\ area\ [cm^2]$$

It is preferrable to have a large surface area with many nucleation points for product to form. The electrode wire had a diameter of 0.1 cm and a length of 9 cm. The calculation did not include the area of the wire tip as this was assumed negligible. The fixed current calculated for the system was 0.56549 A.

$$A = (\pi * 0.1) * 9 = 2.83\ cm^2 \quad (2)$$

$$I = 200 \times 2.83 = 565.49\ mA \quad (3)$$

Analysis of Molten Salt:

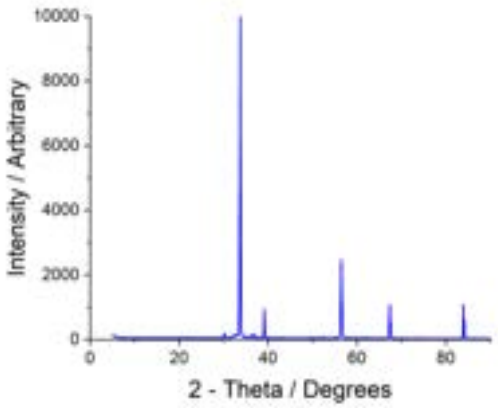


Figure 6: XRD measurement of pure lithium oxide

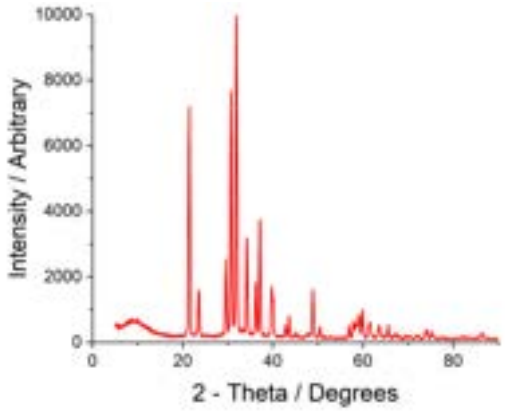


Figure 7: XRD measurement of pure lithium carbonate

Following this, a scan of a 50:50 mixture of the two compounds was conducted. Major peaks between angles of 20 and 40 degrees of the pure components are represented in the scan of the mixture.

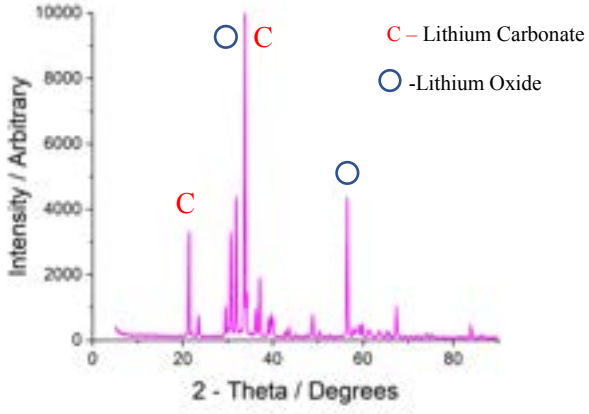


Figure 8: XRD measurement of a 50:50 mixture of lithium oxide and lithium carbonate

X-ray diffraction analysis was carried out to produce a calibration curve for the salt sample. By measuring a range of different known ratios of a Li_2CO_3: Li_2O mixture, calibration values can be obtained. There was no addition of extra CO_2 during electrolysis, and it was planned that the Li_2CO_3 would be slowly depleted as the experiment continued. This way we would be able to obtain the concentration ratio of the compounds in our final electrolyte and understand how well the electrolysis was performing. This is a technique that is not performed in literature for the assessment of this CNT production system. Usual methods involve the continual addition of CO_2 with comparison of CNT yield. However, XRD was abandoned as it was not possible to perform quantitative comparison between samples.

Figure 7 shows the XRD spectra for pure Li_2CO_3. There are clear peaks of high intensity at 2θ values of ~ 21 and 32, with smaller peaks in the range of 20-40. Outside of this range there is very little activity.

Figure 6 is the spectra for pure Li_2O and it has very little background activity compared to the Li_2CO_3 spectra, which makes the peaks easily identifiable. At $2\theta = 33^\circ$, a high intensity peak is present. The next largest occurs at $2\theta = 56^\circ$ and there are 3 other smaller peaks present.

Figure 8 is an analysis of a 50:50 mixture of both lithium carbonate and lithium oxide and the first two spectra can be used to confirm this. There are peaks at 21, 32 and 56, which match to the main peaks in the individual salt spectra. Lithium carbonate has high intensity peaks at 21 and 32, whilst lithium oxide has high intensity peaks at 32 and 56. The presence of all main peaks confirm that both salts are present.

XRD could be used later, however, as a technique for revealing the local and global features of CNTs' lattice and crystalline phase, domain sizes, and impurities, as written by Das, *et al*[21]. By characterising key features of CNTs, which can cause them to have a wide variety of properties, manufacturers will be able to select production methods which favour their desired CNT shape. Techniques such as scanning electron microscopy are popular for characterisation of CNTs, but they only unveil local features.

Example calculation for a 60:40 mixture of Li_2O:Li_2CO_3:

$$Li_2CO_3\ Molecular\ Weight: 73.891$$
$$Li_2O\ Molecular\ Weight: 29.88$$
$$CO_2\ Molecular\ Weight: 44.01$$
$$Li_2CO_3 \rightarrow Li_2O + CO_2$$
$$Calculating\ percentage\ mass\ of\ Li_2O\ formed:$$
$$0.6 \times \left(\frac{29.88}{73.891}\right) = 0.2426 \quad (4)$$
$$Theoretical\ remaining\ mass\ percentage:$$
$$0.2426 + 0.4 = 0.6426 = 64.26\% \quad (5)$$

We found that the final experimental remaining mass percentage for the corresponding mixture ratio was not equal to this theoretical value. This is because the sample could not be perfectly mixed, and the mixture was not homogeneous. Furthermore, any moisture in the sample could inflate the measured mass. These errors were minimised by stirring samples before analysis to prevent heterogeneity and keeping samples sealed in vials until time for analysis to prevent exposure to humid air.

Differential scanning calorimetry measured the difference in the amount of heat required to increase the temperature of a sample (our lithium carbonate/lithium oxide mixture), and a reference sample. This indicated points of enthalpy change, where exothermic or endothermic reactions occur throughout the measurement. Temperature-driven lithium carbonate decomposition to lithium oxide and carbon dioxide is expected to be an endothermic process and this is expected to occur through a decrease in mass of the sample.

Another purpose of STA analysis was to discover the maximum temperature that the electrolysis could be ran at. This was key to prevent thermal decomposition of Li_2CO_3 during electrolysis and preventing the escape of CO_2 without being used to create CNTs. Furthermore, allowing the mixture to

decompose thermally runs the risk of carbon monoxide being produced, which is toxic and a large hazard if released in the lab. There are multiple measures in place to mitigate this risk, such as carbon monoxide alarms, but ideally none would be produced.

Samples analysed were 20:80, 40:60, 50:50, 60:40, 80:20 mixtures of Li_2CO_3 and Li_2O, as well as pure Li_2CO_3. The first trial of simultaneous thermal analysis used a mixture of lithium carbonate and lithium oxide with a mass ratio of 60:40. A purge and protective gas of nitrogen was used to at rates of 60 ml/min and 20 ml/min, respectively, to prevent any reaction. The sample was heated to 1100 ℃ at a rate of 20 ℃/min and consequently held at this temperature for 10 minutes. Thermal decomposition temperature for Li_2CO_3 occurs at 1310℃, but this effect is reported to occur at lower temperatures, close to the melting point[22]. This is clearly the case in the STA measurements, where thermal decomposition consistently began at 850℃. Figure 9 shows the results of this first trial.

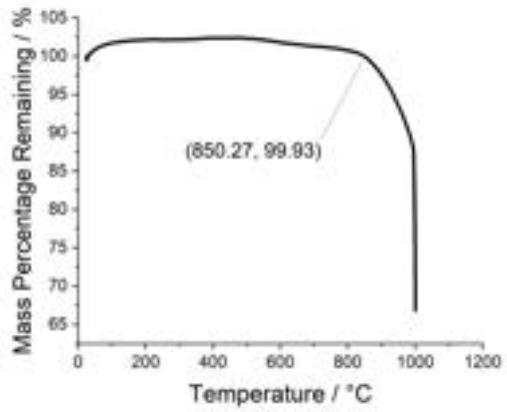


Figure 9: STA measurement of a 60:40 mixture of Li_2CO_3:Li_2O highlighting initial decomposition temperature

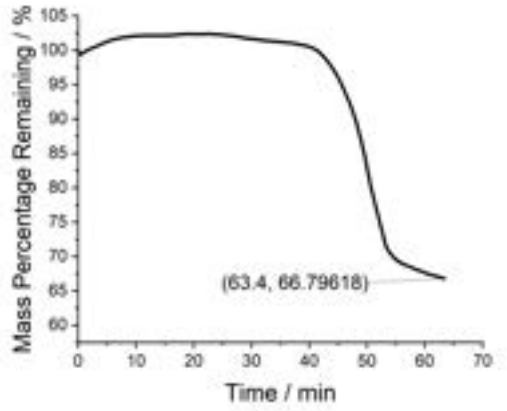


Figure 10: STA measurement of a 60:40 mixture of Li_2CO_3:Li_2O highlighting final remaining mass percentage

The results from the STA analysis of the 60:40 salt sample show a clear trend. The mass starts to decrease gradually at 850 ℃ then shows a much larger decrease at 1000 ℃. This decrease in mass correlates to the decomposition of lithium carbonate. As seen from literature, at temperatures close to its melting point, the lithium carbonate decomposes into lithium oxide and carbon dioxide[23], which was removed from the system as a gas. As the temperature increased further, the rate of decomposition did not change significantly, which suggests a higher temperature does not necessarily increase the rate of decomposition and a temperature lower than 1000℃ might be sufficient if held long enough.

The STA results indicate a temperature of below 850℃ is the ideal temperature to run the electrolysis at as the only thing causing decomposition will be the applied current. Further analysis of the remaining mass fraction shows that, from an initial mass of 36.88 mg, 14.00 mg was lost as carbon dioxide. Using the known initial ratio in calculations show that 106% of the available Li_2CO_3 mols decomposed. This suggests the initial mixture may not have been as uniformly mixed, leading to an alternative ratio to the expected one.

A further test was run on a sample with the same composition but with the end temperature set to 1000℃ and held for 15 minutes. This showed the same initial decrease at ~850 ℃ and then a greater rate of decrease at 1000 ℃. However, after holding for 15 minutes the rate of change of mass had still not reached a plateau, suggesting there was more CO_2 still to be released. The final recorded mass percentage was 66.80%, as can be seen in Figure 10. With the theoretical final mass percentage being 64.26%, this test had released all but 6.85% of the potential CO_2 (assuming a perfect mixture). To ensure complete thermal decomposition, the entire experimental results were conducted at 1100 ℃ with a holding time of 20 minutes.

STA analysis also allowed the thermodynamic description of the process to be found, as seen in Figure 11. A peak on the graph represents an endothermic reaction while a trough represents an exothermic reaction. In the differential scanning calorimetry (DSC) curves for Li_2CO_3: Li_2O, there are visible peaks at 420 ℃ and 690 ℃. The solid-solid phase transition of Li_2CO_3 is observed at 420 ℃ and the melting point is observed at 690 ℃.

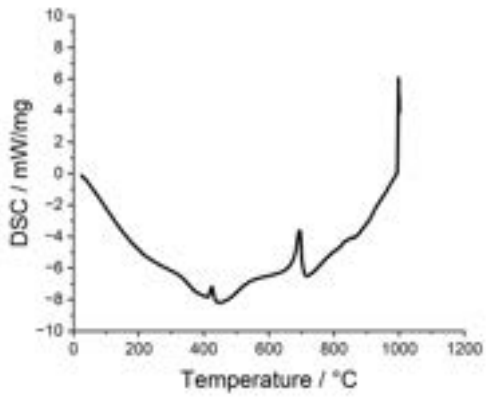


Figure 11: Differential scanning calorimetry curve of a 60:40 mixture of Li_2CO_3:Li_2O

Table 1: Comparison of calculated and experimental percentage mass loss

Li_2CO_3:Li_2O Ratio	Final Mass Percentage / %	Expected Mass Percentage / %
20:80	85.51	88.09
40:60	82.11	76.18
50:50	61.02	70.22
60:40	66.80	64.26
80:20	47.30	52.35
100:0	41.08	40.44

On average, percentage error between experimental and theoretical mass percentage was 6.5%. The largest error occurred in the 50:50 mixture, while the smallest occurred in the pure sample. The 50:50 measurement was repeated using a sample from the same mixture as the initial measurement, and the result had a 10.03% error. It was concluded that the mass of each compound in the mixture was incorrectly measured. Percentage error was attributed to poor mixing of the two salts in each sample. It was observed that lithium carbonate powder would clump and 'stick' together easily. This means the samples would not be uniformly distributed during preparation.

A problem encountered during STA was an unknown reaction that occurred during the heating of the sample. A green crystalline substance was produced that caused the sample crucible to be stuck to the STA weighing pan. This could also be attributed to percentage error. After removal using dilute hydrochloric acid, sample sizes were reduced from 30 mg to 8-10 mg, and a crucible lid was used. Furthermore, sample crucibles were cleaned with 37% hydrochloric acid after each use.

Concentrated hydrochloric acid is corrosive and can cause severe damage if it encounters skin, eyes or is ingested. Furthermore, at high concentrations it can be volatile and release HCl gas that can cause irritation to eyes, throat and lungs if inhaled. To minimise the risks involved, all personnel wore appropriate PPE (safety glasses, lab coat and gloves) whilst using the acid. The acid was only used inside a fume cupboard and the cleaning vessel was a glass beaker inside a large tray to prevent spillages.

All STA curves can be seen in the Appendix, from Figures 1 to 18.

Commercial CNT Analysis:

Raman spectroscopy using a *Senterra II* was conducted on CNT products and commercial CNTs as a method of identifying product purity. Raman spectroscopy measures the scattering of photons. There is a source of monochromatic light in the machine which interacts with the sample material, resulting in the energy of the laser photons being shifted up or down. This can help in the identification of molecules and the study of chemical bonding and intramolecular bonds. A magnification lens of 100x was used to obtain the highest Raman signal. The commercial MWCNTs (multi walled carbon nanotubes) were of varying sizes: 10-40 nm, 110-170 nm, and 9 – 13 nm. Each sample was analysed at 4 different points to confirm the results were accurate. All graphs show there was minimal variation between the points and the important peaks all appeared at the same values. For every calculation, an average value was used. There was one exception. This test appeared in the 110-170nm sample and shows a large spike in the 800-1000 cm^{-1} region. As this peak was not reproducible it is assumed to be an anomaly, likely caused by impurities in the sample

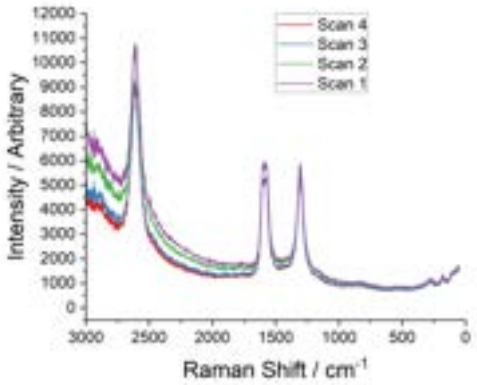


Figure 12: Raman measurement of CNTs sized 10-40 nm

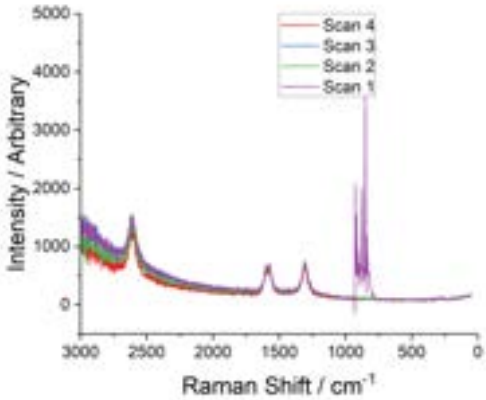


Figure 13: Raman measurements of CNTs sized 110-170 nm

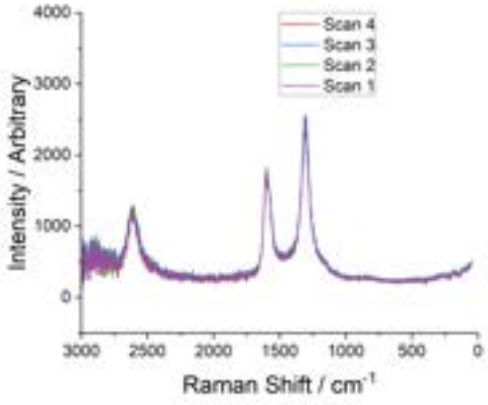


Figure 14: Raman measurement of CNTs sized 9 - 13 nm

The samples studied were all multi-walled carbon nanotubes (MWCNTs). These are larger than single-walled tubes and consist of multiple layers of increasing diameters stacked on top of each other. This gives the tube increased thermal and chemical stability. The properties of single walled and multi-walled CNTs vary and allow for them to have widely different uses in industry. Due to the much larger size and strength of MWCNTs, they are easier to produce and there is a much lower risk of damage or collapse during treatment processes. Therefore, in the initial stages of this investigation it is desirable to produce MWCNTs over single walled. Once the process has been tested and optimised, it will be possible to attempt to produce the single walled variety.

Raman data beyond Raman shift beyond 3000 cm^{-1} were removed for clarity because this is an area of noise. Full graphs including noise can be seen in Appendix Figures 21, 22, and 23.

Raman spectra usually consist of two main peaks that correspond to the disorder-induced mode (D band) and the high frequency E_{2g} first order mode (G band). An important parameter for assessing the quality of CNTs is the intensity ratio (I_D/I_G). This is used to evaluate graphitisation and is typically in the range of 0.5 - 1.8 for CNTs, with variations occurring due to differing CNT diameters. A lower[24] intensity ratio is favourable as this signifies a higher degree of graphitisation, which is the degree to which the carbon atoms form a close-packed hexagonal structure. A consistent carbon structure leads to predictable and consistent material properties. The D band can also be referred to as the main defect band[25] and is visible when there are defects present in the carbon aromatic structure. Therefore, a lower intensity is ideal for a stronger CNT as the graphene sheets that make up the CNT will have a higher degree of graphitisation. The intensity ratio for the 10-40 nm, 110-17 nm, and 9 – 13 nm CNTs were 0.98, 1.06, and 1.39 respectively. All these ratios are within the accepted range and consistent with values from previous studies.

Table 2: Frequencies and Intensity Ratio of all CNT samples

Diameter / nm	D band / cm^{-1}	G band / cm^{-1}	Intensity Ratio
9 - 13	1314	1599	1.39
10 - 40	1306.5	1599	0.98
110 - 170	1308	1582.5	1.06

However, they are considerably higher than literature values. Past studies have recorded ratios of 0.6 and 0.7. This is a positive finding as it shows CNTs with a lower degree of graphitisation and therefore weaker structure can still be sold commercially. This allows us to have greater flexibility when producing CNTs from electrolysis and potentially sacrifice product quality somewhat to reduce financial and energy costs.

There exists a weak correlation between the diameter and intensity ratio. As the diameter increases from 9 nm to 170 nm, the intensity ratio decreases from 1.39 to 1.06. This suggests that tubes with a larger diameter have fewer defects. During conventional production, CNTs must undergo acid treatment, and this can cause damage to tubes or even total collapse. A larger tube is composed of more graphene sheet layers (walls), which increases the chemical stability of the tube, therefore making them less susceptible to acid damage.

It is likely that after the molten electrolysis the electrodes will need to be acid washed to remove the solidified salts and any impurities. Therefore, it is desirable to have operating conditions that produce larger CNTs as they will show greater resistance to acid and produce a higher yield overall.

Product Analysis:

Electrolysis failure was caused by the formation of an oxide layer on both the iron and nickel electrodes. Post electrolysis, it was clear to see the level which the molten electrolyte reached in the crucible as it was marked by a brown layer formed during the experiment. Half of the electrodes' surface area were exposed to the air, and when the furnace reached temperatures of 800℃, the electrodes were oxidised. Figure 15 shows the result of oxidation. The sample collected from the iron cathode was composed of 98.076% iron oxide while the sample from the nickel anode was composed of 96.046% nickel oxide (full elemental breakdown seen in Appendix Figures 19 and 20; corresponding compositional graphs are shown in Appendix figures 24 and 25).

Figure 15: Image of oxidised electrodes post electrolysis

Raman spectroscopy was used to analyse the products found on the electrodes, specifically the iron electrode as this was where CNTs should form. The Raman spectra, shown in Figure 16, did not show bands at the same frequency as found in the commercial carbon nanotubes. Moreover, it did not correspond with any amorphous carbon forms. Therefore, carbon – and CNTs - were ruled out as the potential product in this very first experiment. Literature analysis of the bands obtained indicated it was a form of iron oxide. A Raman spectra of iron oxide is shown in Figure 17[26].

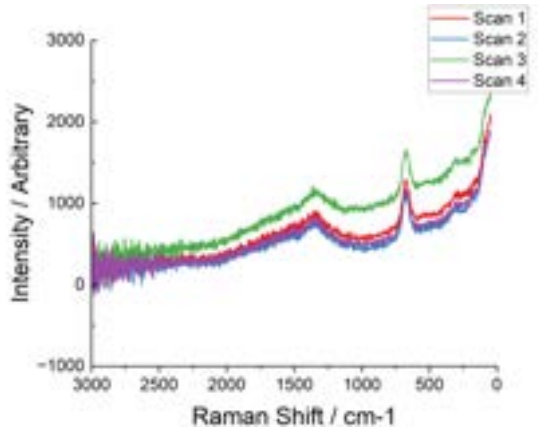


Figure 16: Raman measurements of iron electrode product

Spectrum A shows a very similar pattern to the one obtained from the electrolysis product, with the main band in an almost identical position. This suggested the product sample is most likely a form of iron oxide, specifically magnetite (Fe_3O_4). This can also explain the sudden increase in voltage observed during electrolysis as magnetite has a much higher resistance than pure iron and for a given current, will increase the voltage significantly.

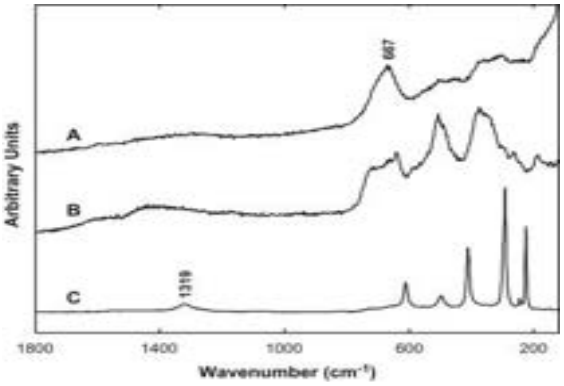


Figure 17: Raman of magnetite nanoparticles at laser powers: (A) 0.35 mW; (B) 4.93 mW; (C) 8.85 mW[26]

Future Research:

Introduction of CO_2 gas - CO_2 bubbler is a component which injects CO_2 gas into the molten salt. This is necessary because over time, Li_2CO_3 will be converted to Li_2O as the carbon nanotubes are produced. Once all Li_2CO_3 has been converted, electrolysis will not be able to continue. CO_2 is bubbled through the molten salt to continually replenish Li_2CO_3. Parkinson *et al*[27]. reported that quartz injectors were used for CH_4 pyrolysis in molten salts and can withstand high temperature up to 1000 °C. Safety is of utmost importance when dealing with high temperature electrolysis and flammable gases. Any CO_2 bubbled inside the reactor must be passed through filters to remove possible entrained O_2 and H_2O. Experiments must be conducted to investigate the role of surface area to volume ratios of bubbles generated by varying the bubble size using different flared tips[27].

Using a mixture of carbonates – it is well known that using a combination of different carbonates can drastically reduce melting temperature of the salt. This is beneficial because conducting electrolysis at a lower temperature saves energy and therefore money. However, there are many areas of investigation that could be chosen. This includes: electrode stability in the chosen combination of carbonate salts, oxygen evolution in different combinations of carbonate salts, thermal balance, and morphology of CNTs produced[28].

Investigate impact of reaction temperature on product – the temperature of the reaction affects the convection currents and physical properties of the salt once molten. This can impact the morphology of the carbon product. The structure of CNTs heavily influences the properties and overall strength and usefulness in industry so must be as close to perfect as possible. Once a successful method of production has been tested, the reaction temperature should be varied to study the effects on the CNT structure.

V. CONCLUSION

This study has explored a potential pathway for electrochemically reducing CO_2 in molten lithium carbonate electrolyte and produce CNTs. Various analytical techniques have been calibrated for post-electrolysis use: STA was used to produce a calibration curve of lithium carbonate and lithium oxide mixtures. This will be compared against crystallised molten mixtures produced post electrolysis to quantitatively assess the decomposition of the carbonate during electrolysis, and the theoretical yield of CNT. Commercially available CNTs of varying diameter and length were analysed using Raman spectroscopy to provide a reference spectrum to compare all electrolysis products against. This will determine whether a carbon product was formed and how close to commercial grade the quality is. XRF will support identification of any electrode product.

Only one full electrolysis experiment was carried out in this time frame and carbon nanotubes were not produced. Therefore, a working method cannot be confirmed but conclusions can still be made. Due to both electrodes oxidising, electrode shape is clearly an important factor and efforts must be made to ensure the electrode is fully submerged in the molten electrolyte throughout the running to prevent exposure to air.

VI. REFERENCES

1. Tiseo, I. Historical carbon dioxide emissions from global fossil fuel combustion and industrial processes from 1750 to 2020. https://www.statista.com/statistics/264699/worldwide-co2-emissions/ (2022).
2. Climate Action Tracker. Addressing global warming. https://climateactiontracker.org/global/temperatures/ (2022).
3. Climate Change Committee. Sixth Carbon Budget.

https://www.theccc.org.uk/publication/sixth-carbon-budget/ (2020).
4. Navas, H. *et al.* Unveiling the Evolutions of Nanotube Diameter Distribution during the Growth of Single-Walled Carbon Nanotubes. *ACS Nano* **11**, 3081–3088 (2017).
5. Takakura, A. *et al.* Strength of carbon nanotubes depends on their chemical structures. *Nat Commun* **10**, 3040 (2019).
6. Meijo Nano Carbon Co. What can we do with carbon nanotubes? https://meijo-nano.com/en/applications/use.html.
7. Licht, S. *et al.* Carbon Nanotubes Produced from Ambient Carbon Dioxide for Environmentally Sustainable Lithium-Ion and Sodium-Ion Battery Anodes. *ACS Cent Sci* **2**, 162–168 (2016).
8. *Carbon Nanotube Reinforced Composites*. (Elsevier, 2015). doi:10.1016/C2012-0-06123-6.
9. Agrawal, A. & Yi, G.-C. Sample pretreatment with graphene materials. in 21–47 (2020). doi:10.1016/bs.coac.2020.08.012.
10. Shah, K. A. & Tali, B. A. Synthesis of carbon nanotubes by catalytic chemical vapour deposition: A review on carbon sources, catalysts and substrates. *Mater Sci Semicond Process* **41**, 67–82 (2016).
11. Kumar, M. & Ando, Y. Chemical Vapor Deposition of Carbon Nanotubes: A Review on Growth Mechanism and Mass Production. *J Nanosci Nanotechnol* **10**, 3739–3758 (2010).
12. Guo, T., Nikolaev, P., Thess, A., Colbert, D. T. & Smalley, R. E. Catalytic growth of single-walled manotubes by laser vaporization. *Chem Phys Lett* **243**, 49–54 (1995).
13. Das, R., Shahnavaz, Z., Ali, Md. E., Islam, M. M. & Abd Hamid, S. B. Can We Optimize Arc Discharge and Laser Ablation for Well-Controlled Carbon Nanotube Synthesis? *Nanoscale Res Lett* **11**, 510 (2016).
14. Lau, J., Dey, G. & Licht, S. Thermodynamic assessment of CO2 to carbon nanofiber transformation for carbon sequestration in a combined cycle gas or a coal power plant. *Energy Convers Manag* **122**, 400–410 (2016).
15. mamalos. Our silver conductors. https://www.mamalos.com/our_silver_conductors.html.
16. Accuratus. Aluminum Oxide, Al2O3 Ceramic Properties. https://accuratus.com/alumox.html.
17. Becker, T., Richter, M. & Agar, D. W. Methane pyrolysis: Kinetic studies and mechanical removal of carbon deposits in reactors of different materials. *Int J Hydrogen Energy* (2022) doi:10.1016/j.ijhydene.2022.10.069.
18. Hansen, T. Lithium Carbonate. https://digitalfire.com/material/lithium+carbonate.
19. Wu, H. *et al.* One-pot synthesis of nanostructured carbon materials from carbon dioxide via electrolysis in molten carbonate salts. *Carbon N Y* **106**, 208–217 (2016).
20. Wu, H. *et al.* One-pot synthesis of nanostructured carbon materials from carbon dioxide via electrolysis in molten carbonate salts. *Carbon N Y* **106**, 208–217 (2016).
21. Das, R., Hamid, S., Ali, Md., Ramakrishna, S. & Yongzhi, W. Carbon Nanotubes Characterization by X-ray Powder Diffraction – A Review. *Curr Nanosci* **11**, 23–35 (2014).
22. Argandoña, G., Aresti, M., Blanco, J. M., Muel, E. & Esarte, J. Li2CO3 as Protection for a High-Temperature Thermoelectric Generator: Thermal Stability and Corrosion Analysis. *Applied Sciences* **11**, 7597 (2021).
23. Argandoña, G., Aresti, M., Blanco, J. M., Muel, E. & Esarte, J. Li2CO3 as Protection for a High-Temperature Thermoelectric Generator: Thermal Stability and Corrosion Analysis. *Applied Sciences* **11**, 7597 (2021).
24. Wang, J., Tu, J., Lei, H. & Zhu, H. The effect of graphitization degree of carbonaceous material on the electrochemical performance for aluminum-ion batteries. *RSC Adv* **9**, 38990–38997 (2019).
25. GTK. Science Blog: Studying the graphitization temperature and degree of graphite crystallinity with Raman spectroscopy. https://www.gtk.fi/en/current/science-blog-studying-the-graphitization-temperature-and-degree-of-graphite-crystallinity-with-raman-spectroscopy/.
26. Li, Y.-S., Church, J. S. & Woodhead, A. L. Infrared and Raman spectroscopic studies on iron oxide magnetic nano-particles and their surface modifications. *J Magn Magn Mater* **324**, 1543–1550 (2012).
27. Parkinson, B., Patzschke, C. F., Nikolis, D., Raman, S. & Hellgardt, K. Molten salt bubble columns for low-carbon hydrogen from CH4 pyrolysis: Mass transfer and carbon formation mechanisms. *Chemical Engineering Journal* **417**, 127407 (2021).
28. Wang, X., Liu, X., Licht, G., Wang, B. & Licht, S. Exploration of alkali cation variation on the synthesis of carbon nanotubes by electrolysis of CO2 in molten carbonates. *Journal of CO2 Utilization* **34**, 303–312 (2019).
29. Hankin, A. *et al. Assessing the economic and environmental value of carbon capture and utilisation in the UK*. https://www.imperial.ac.uk/molecular-science-engineering/publications-and-outputs/briefing-papers/assessing-the-economic-and-environmental-value-of-carbon-capture-and-utilisation-in-the-uk/ (2019).
30. Lau, J., Dey, G. & Licht, S. Thermodynamic assessment of CO2 to carbon nanofiber transformation for carbon sequestration in a combined cycle gas or a coal power plant. *Energy Convers Manag* **122**, 400–410 (2016).
31. Kushnir, D. & Sandén, B. A. Energy Requirements of Carbon Nanoparticle Production. *J Ind Ecol* **12**, 360–375 (2008).

A Framework For Conducting A Meta-Analysis And Implementing Machine Learning On Clinical Trial Data For Rheumatoid Arthritis

Isabelle Ho, Megan Caunce
Department of Chemical Engineering, Imperial College London

1. Abstract

Clinical trials are inherently inefficient, with 90% of treatments failing to make it through the three main stages. Since data-sharing of Individual Patient Data (IPD) by sponsors of clinical trials is not commonplace, the ability to build and learn from past outcomes for similar treatments is limited. By accessing IPD and implementing Machine Learning (ML), factors contributing to the progression of a disease can be better understood, aiding in the progress of drug development. This paper details the construction of a framework for analysis and prediction of efficacy of a treatment, with a dichotomous outcome. The traditional approach of meta-analysis (using summary data) was implemented alongside building of a classification model using ML techniques (using IPD). The framework is worked through with clinical trial data for Rheumatoid Arthritis treatments, sourced from the Yale Open Data Access Project (YODA), with an outcome measure of ACR 20. Results of the meta-analyses suggested Sirukumab and combined Golimumab and MTX therapies have a positive effect on participants of the included trials regarding the odds of them achieving ACR 20. For the implementation on two trials testing Sirukumab, the ML model provides a proof-of-concept result illustrating that the framework can be applied to any data set.

2. Introduction

Clinical trials are the primary way to understand the efficacy and toxicity of a new treatment on human participants. There is a high failure rate for treatments in clinical trials, with the four main reasons behind this being the lack of clinical efficacy, unmanageable toxicity, poor pharmacokinetic properties and lack of commercial interest[1] The average cost of developing a new drug has been estimated as $2-3billion. [2]

Rheumatoid Arthritis (RA) is the most common form of autoimmune arthritis. As a lifelong, progressive musculoskeletal disease, RA comes with added systemic complications due to the side effects of often widespread, severe inflammation. Early diagnosis and immediate treatment with an effective treatment plan is key to avoid long-term complications. Research shows that if an aggressive treatment plan is started during the first 3 months of the onset of symptoms, temporary or sustained remission is a realistic outcome for some patients. [3]

Since 2003 the average time from the onset of symptoms to diagnosis and the beginning of treatment has been approximately 9 months. This can be reduced by improving awareness of RA for both the public and healthcare professionals, as well as sufficient funding for the treatment and diagnosis of RA. However, the chance of remission for each patient is bounded by their response to the current first and second-line treatments.

DMARDs are the common initial treatment for RA, with the most popular being Methotrexate (MTX). The remission rates with MTX monotherapy fluctuate between 30% and 50% for patients with longer than 1 year disease duration and less than one year disease duration respectively. [4] Typical combination therapies that may also be prescribed immediately after diagnosis consist solely of conventional DMARDs. Biologic DMARDs (bDMARDs) were developed in the late 1990s, inhibiting either the proinflammatory cytokine Tumour Necrosis Factor alpha (TNF-α) or the Interleukin-6 (IL-6) cytokine. Studies show that combination therapy approaches employing both MTX and a bDMARD lead to better outcomes than MTX alone.[5] The development of new drugs had been difficult of late, because phase III trials have commonly compared the trial treatment with anti-TNF drugs – prioritising advancing current therapeutic strategies rather than bringing new drugs to market. [6]

The socioeconomic cost of RA is vast, costing the NHS £560m yearly (2010) and the wider economy £1.8billion – due to 75% of people being diagnosed are of working age and 1/3 of these people are estimated to have stopped working completely within a year of being diagnosed. As of 2018, Adalimumab was the single medicine on which hospitals in the UK spent the most, at a cost of £400m a year. [7] bDMARDs are considerably more costly than conventional DMARDs.

Since the cause of RA is unknown, and the prognosis guarded, the development and deployment of therapeutics is based solely in the understanding of the pathogenesis of the disease.[8]

The aim of this project is to evaluate the efficacy of common RA treatments, Golimumab and Sirukumab, through a meta-analysis of selected trials from the Yale University Open Data Access (YODA) Project using the efficacy measure ACR 20. Then to build a framework using Machine Learning (ML) to predict the efficacy of Sirukumab based on clinical and demographic factors. This has the potential for expansion to consider other bDMARDs and combination treatments for RA to foster a more data-led understanding of the efficacy of RA treatments.

Meta analysis has been used for decades to pool results of several studies – reducing bias and increasing statistical power compared to individual trial outcomes. Key requirements for success are a carefully considered analysis of study heterogeneity and robust inclusion

Factor	Description
Exposure of Interest	Participants must be experiencing RA during the study duration.
Participants	Restrict to only studies with an all-adult population (18+). This is the case for all of the studies for RA treatment in the YODA database.
Reported Outcomes	All trials must report at least one ACR20 measurement. The meta-analysis included trials with ACR20 measurements taken in the window between Week 14 and Week 16.
Study design	Trials included contained the same dosage of the treatment and a well-defined placebo group that received only placebo treatment until the ACR20 measurement has been taken.

Table 1: Inclusion criteria

criteria, to ensure the power of the conclusions from the pooled result, using either a fixed or random-effects model.

Big data and ML is undoubtedly a force for change within the pharmaceutical industry and medicine, with McKinsey forecasting up to $100 billion in yearly value, in part due to enhanced decision making. [9]Compared to the prevalence of meta-analysis in studies on efficacy for treatments of RA, ML is scarcely explored. Other examples of the use of ML in medicine use include predictions of interactions between drugs, analysis of medical images and patient monitoring.[10] Formulation of a classification model for systematically predicting the efficacy of anti-cancer drugs using proteomics and phosphoproteomics data has been completed with high accuracy.[11] Such models are valuable especially for progressive conditions, allowing the most therapeutically and economically effective treatment be delivered first to allow the best prognosis possible for the patient.

Shared information from clinical trials is typically limited to summary data at most. Of late there has been a strong movement for sharing individual patient data (IPD) for research purposes, as well as allowing researchers to work more efficiently by building on previous findings. [12]

3. Methodology

3.1.1 Step 1 – Data collection

In this paper, the Yale University Open Data Access (YODA) [13] project is used to access summary data of various trials researching drugs treating RA. Information on each trial can be gathered to make informed decisions on trials to use in the analysis. Summary level data is generally shared online but to access individual level patient data (IPD) via projects such as YODA or through individual sponsors must be made.

An inclusion criterion is important for selecting the trials for analysis. For meta-analysis, efforts should be made to ensure the included trials are similar enough in factors such as reported outcomes, study design and exposure of interest so that the pooled effect can offer insight into the field of study. For machine learning, the inclusion criteria should ensure that the trials analysed are of the same drug to understand the efficacy for certain patient individuals. The minimum inclusion criteria used is outlined in Table 1.

3.1.2 Step 2 – Meta-Analysis

The summary measure carried forward through the meta-analysis section is the odds-ratio, as unlike the risk ratio it can be used in case-control studies, so has a wider scope.

To conduct a meta-analysis using the odds ratio as a summary measure, the only quantitative data needed is the sizes of the placebo and treatment group as well as the number of participants from each exposure group that experienced the outcome of measure, for example ACR 20. For each exposure-placebo group, the Table 2 was populated using summary data gathered at the time of measurement of ACR 20:

	Exposed	Non-Exposed
Achieved ACR 20	a	b
Did not achieve ACR 20	c	d

Table 2: Outcome table for each trial

Using the information in Table 2, a sample estimate of the odds ratio can be calculated as: [14]

$$Sample\ estimate\ of\ the\ odds\ ratio = \frac{a \times d}{b \times c} \quad (1)$$

3.1.2.1 Confidence Interval for the Individual Studies

The confidence interval (CI) describes the uncertainty associated with the estimation of the odds ratio for each individual study and for the summary, estimates following the fixed and random-effect meta-analysis methods. For the individual studies, it is mostly dependent on the size of the study – with smaller studies having larger CI's due to their inherently less precise estimates of effect size. The CI's associated with the summary odds ratios are dependent on the number of studies considered and their individual precision of estimation.

The upper and lower limits for the 95% CI (UL and LL respectively) for the individual trials are calculated using equation 2 and 3 respectively [15].

$$UL = \exp\left(ln(OR) + 1.96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right) \quad (2)$$

$$LL = \exp\left(ln(OR) - 1.96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right) \quad (3)$$

If the calculated confidence interval of the odds ratio for a trial contains only values greater than 1, this suggests that there are greater odds of experiencing the outcome

measure for the treatment group compared to the placebo group.

3.1.2.2 How Fixed and Random effect models differ

Both fixed and random effect models were considered in the meta-analysis as it can be efficiently executed using the PythonMeta package. The most appropriate model can be decided using the I^2 statistic, a measure of the inconsistency across studies (the percentage of variation across studies that is due to heterogeneity). [16]

$$df = k - 1 \quad (4)$$

$$Q = \sum_{i=1}^{k} W_i(Y_i - M)^2 = \sum_{i=1}^{k} \frac{(Y_i - M)^2}{V_i} \quad (5)$$

$$I^2 = \frac{Q - df}{Q} \times 100\% \quad (6)$$

Where Y_i: effect size estimates, M: mean of the effect size estimates, k: number of studies included in the meta-analysis.

If I^2 is less than 50% then a fixed-effects model may be more appropriate as it indicates only small inconsistency between study results, and the studies are considered homogeneous. [17] The fixed-effect model assumes that the only variance in the odds ratio between trials is due to the within-studies estimation error, whereas the random effects model assumes normally distributed odds ratios and aims to estimate the mean of a distribution of effects. Compared to the fixed effect model, the random-effect model is more likely to assign smaller studies higher weights.

The fixed and random effect models answer different questions. By obtaining a summary OR and CI for the fixed-effect model, the question at hand is 'What's the best (single) estimate of the effect?' whereas with the random-effects model the question is 'what is the best estimate of the average effect?' due to assumptions surrounding between-study heterogeneity. [17]

3.1.2.3 Fixed Effect Model

The Maentel-Haenszel (MH) method for calculation of the odds ratio is used for estimation of the standard errors of the effect estimates under the fixed effect model. This is a good alternative to the commonly used inverse variance method which tends to perform poorly in the case low event rates or small study sizes. [17] The MH method's approach to weighting is different depends on the effect measure being used, which is not the case for the inverse variance method.

The pooled odds ratio estimate by the MH method is generated using the following formula:

$$\widehat{OR}_{MH} = \frac{\sum_{i=1}^{k}\left(\frac{a_i d_i}{n_i}\right)}{\sum_{i=1}^{k}\left(\frac{b_i c_i}{n_i}\right)} \quad (7)$$

Where

$$n_i = a_i + b_i + c_i + d_i \quad (8)$$

And the corresponding 95% CI is found using the Robins, Breslow and Greenland variance formula. [18]

DerSimonian and Laird method for Tau

The random-effects model requires the calculation of τ^2, the between-study variance. This originates from the heterogeneity in effect size alongside the within-study estimation error.

An estimation of τ^2 was completed using the DerSimonian and Laird (DL) method , chosen as it is well documented compared to its counterparts[19]. The DL method overestimates τ^2 on average and can incur substantial bias when the number of studies is small. The summary odds ratio and 95% confidence interval can be calculated for the random-effects model as outlined in the paper 'A simple confidence interval for meta-analysis'. [20]

3.1.3 Step 3 – Machine Learning

3.1.3.1 Gathering clinical trial data

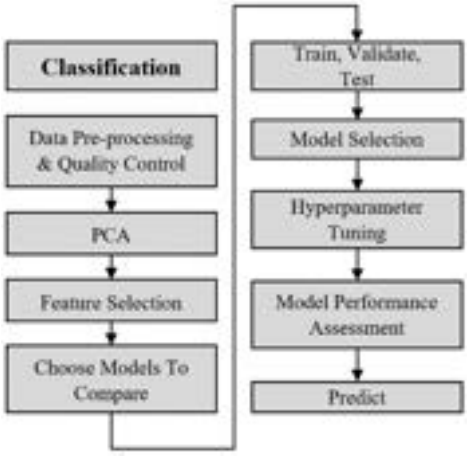


Figure 1: A flowchart of steps of applying ML to analyse clinical trial data

The IPD should be gathered for each included trial. However, in the absence of IPD, a pseudo-population can be generated from summary data to represent the patient demographic and outcome data. The following demographic input variables are populated: trial (drug of choice), dosage (mg), frequency (of drug administration), previous use of anti-TNF therapy, age, and gender. The output measure used in this paper is whether the patient achieved an ACR20 response at week 16. It is up to the user's discretion to change this output accordingly.

For continuous variables such as age, a random normal distribution can be generated, with the standard deviation and mean as the ones listed in the summary statistics for that variable. To generate the categorical variables such as gender, use a random uniform distribution to generate a number between 0 and 1 in the same proportions as the variable's summary statistic.

In this project, country of origin is represented as individual features with binary inputs. It is felt that by assigning numbers 1-n to a 'n' countries may cause the model to interpret some countries more important than others based on the order of the list.

Model Name	Description
kNN	Pattern recognition algorithm. A datapoint from the testing set is classified by looking at the classification of the training set and finding the 'k' closest relatives. Inherently performs feature selection – well suited to large datasets with high dimensionality.
Linear SVM	Supervised learning algorithm, parametric model. Uses the training data to find an optimal hyperplane which can be used to classify the test set data points.
Logistic Regression	Finds a logistic curve using maximum likelihood, as the y value is a binary outcome.
Kernel SVM	This model is not parametric. Works similarly to linear SVM however now fits data that is not linearly separable. More expensive to train than the Linear SVM, ad as it is more complex it is easier to overfit
XGBoost	Implements the gradient boosting decision tree algorithm – new models are created to predict the errors of previous models and correct the models. All are added together to make a final prediction.
CatBoost	Similar to XGBoost, although only builds symmetric trees, uses ordered boosting and supports more feature types other than just numerical and categorical– saving time on pre-processing
Random Forests	An ensemble technique - extension of a decision tree algorithm. Constructs many decision trees with the training set, then for your test data to the closest tree on the data scale.
Decision Trees	Supervised learning algorithm. Separates data points into two similar categories at a time, building a flowchart where the categories become more finally similar as you move through the layers
Naïve Bayes	Calculates the probability of the datapoint being in each class using Bayes Theorem.

Table 3: Brief explanation of 9 common machine learning algorithms [36]–[41]

3.1.3.2 Pre-processing

Scaling

Scaling data is a standard procedure in machine learning. A standard scalar is used which scales the data to unit variance and centres the data around 0. Many algorithms benefit from the process of scaling, for example, gradient based algorithms such as XGBoost or distance-based algorithms such as KNN. [21]

The input features to describe a patient were fed into a dataset labelled as 'X' and the output of choice, in this case ACR20, fed into the dataset labelled as 'y'. Both datasets are then split into test and train datasets using an 80:20 split (train:test).

Feature selection

Feature selection is used to make ML models more accurate. It can help decrease over-fitting of data as it helps to decrease the chances of making decisions based on noise.[22] As feature selection removes redundant data, it helps decrease training time which is important for models when using large training datasets. As this is a framework, the scope to optimise is an important consideration so that the model can be applied to a range of datasets with varying size.

There are a few methods that are commonly used for feature selection. Lasso regression was chosen in this project due to its speed, effectiveness, and widespread use across ML projects. Lasso (Least Absolute Shrinkage and Selection Operator) regression, also known as L1 regularisation is a technique used to tune the model by adding a penalty to the error function.[23] The penalty used is the sum of the absolute value of the feature coefficients. L1 regularization introduces sparsity in the dataset, and it can use to perform feature selection by eliminating the features that are not important. This is done by shrinking the coefficients of each feature, with those having a coefficient of 0 being redundant features that can be removed from the datasets containing them.

PCA analysis

Principal component analysis (PCA) is an unsupervised, non-parametric statistical technique for dimensionality reduction in ML and was invented by Pearson in 1901. [24] The technique involves mapping feature information onto a new dimension transforming them to principal components (PC). Clustering is a type of classification and one way to perform this analysis is by applying PCA. It is technique that transforms data with many dimensions to a fewer number of dimensions whilst retaining as much information as possible. PCA uses orthogonal linear transformation to project the data onto a new coordinate system so that the greatest variance by some scalar projection of the data comes to lie on the first coordinate or first principal component. The second greatest variance is the second coordinate and so on. [24]

The graph produced can be visually analysed to spot the different clusters, representing the classes. PCA is used to confirm the number of classes in our dataset to allow confidence in the progression to the next step.

PCA can also be used as an outlier detection method. Datapoints not grouped close to any specific clusters can be identified and investigation can be performed to understand why, on an individual patient level, these patients are outliers. If such patients are deemed outliers, they may be removed from the dataset.

3.1.3.3 Model selection

Nine common ML algorithms considered for the model are briefly explained in Table 3. Despite the possibility that more complex models, such as deep neural networks may be more accurate, they are more complex and harder to explain– so less accessible for our audience. Such models were not considered to be used in this framework. This trade-off between accessibility and accuracy is justified to ensure uptake of ML in clinical trial data analysis. ML is still in the adoption phase for

clinical trial analysis, it is important for the ML models chosen to relatively easy to understand.

Scikit-learn, a Python library, is the main package used in this paper and to pick the most suitable estimator, their documentation suggests that for classification models of less than 100,000 samples, to try use a Linear SVC model. If this does not work or is deemed unsuitable, as long as the data is numerical, to try using K Neighbours Classifier and again, if this does not seem suitable to use SVC or ensembles like random forest classifiers. [25]

Each algorithm is trained using the training dataset and the test set is uses this trained algorithm and its effectiveness can be measured by generating an accuracy score. Accuracy is calculated using equation 9 [21]

$$Accuracy = \frac{TN + TP}{TN + FN + TP + FP} \quad (9)$$

Where TP: true positive, TN: true negative, FP: false positive, FN: false negative prediction.

A k-fold cross validation of with 10 folds (k=10), generates scores to validate the accuracy of the training set. The data is split into 'k' different sample datasets. The model iterates a process of training and testing on the data in each sample dataset generating an accuracy score for each iteration. An average accuracy score is calculated and outputted, as well as its standard deviation.

3.1.3.4 Model development

Once an algorithm is chosen, the next step is to tune the hyperparameters. In this paper, three algorithms are chosen to tune and to have the results of the tuning compared, with the view to picking the model with the highest test-set accuracy score. It is at the user's discretion how many algorithms to choose to compare. Three are chosen to show a wide view of possible algorithms and how they can be compared, and how their accuracy scores can change by tuning the hyperparameters.

A list of all hyperparameters can be found in the documentation for each algorithm issued by the developers of the package it is contained within, so that they can be included in the model for tuning. A random search is used initially to explore a wide range of hyperparameters, and it implements a fit and score method. [26] Following this, a more specific and exhaustive grid search can be applied to find the optimal values of the hyperparameters. This method of using random search followed by grid search is useful, as grid search applies the model for every combination of parameters and can be inefficient to start with, whereas random search can be used before it to narrow down the search and improve efficiency. The optimal parameters can then be applied to the model and a new set of accuracy scores can be found. This should be repeated for each model.

3.1.3.5 Model evaluation

Another way of evaluating a model other than the accuracy scores as presented above, is analysis of receiver operating characteristic, ROC, curves. They are probability curves which visualise the trade-off between the true positive rate and false positive rate for a predictive model using different probability thresholds. The area under the curve, AUC, represents the degree or measure of separability. The best-case scenario, when AUC = 1, is when the model chooses the randomly positive instance higher than a randomly chosen negative one. [27]

3.1.3.6 Prediction

Once a model has been chosen and optimised, the reader is able to then predict the efficacy of a drug for a certain patient based on their individual characteristics.

4. Results & Discussions

4.1 Meta–analysis

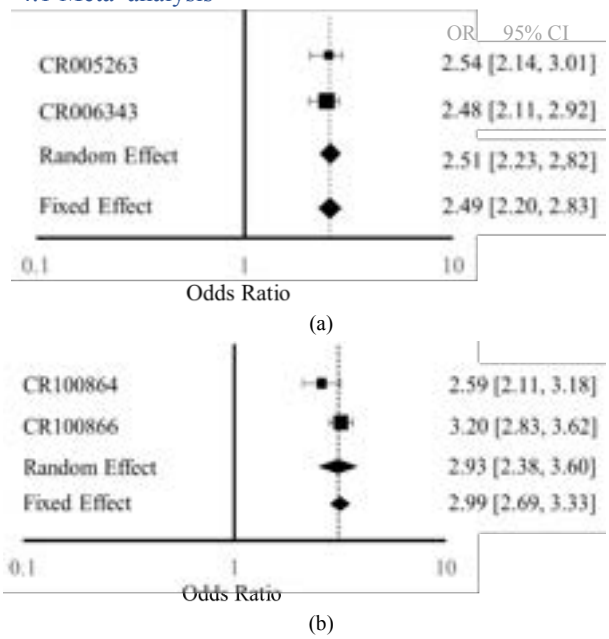


Figure 2: A forest plot representing, (a)Intervention with Golimumab (+MTX) 50mg q2w (b) Intervention with Sirikumab 100mg q4wm with ACR20 measured at week 16.

From the 20 studies available related to RA available on the YODA database, 16 were removed as per the inclusion criteria outlined in Figure 2. The remaining trials contained 1991 participants and were split into two subsets for meta-analysis as they tested different treatments, Golimumab (Study ID Numbers CR005263 and CR005263) and Sirukumab (Study ID Numbers CR100864 and CR100866). [28]–[31]All trials provided dichotomous data for the instances of achievement of ACR 20 in both the placebo and treatment groups, listed as percentages.

For a combination treatment of Golimumab (50mg dose every 2 weeks) and MTX, the results of the meta-analysis method suggest that compared with placebo, the treatment method has a positive effect on the odds of achieving ACR 20 in both trials. Overall, under the fixed-effect model, the OR was 2.4 with a 95% confidence

interval of [2.2, 2.83]. There was minimal heterogeneity between the trials implying that the use of the fixed-effect model is more appropriate, in this case. Since only 2 trials were included, the difference in the combined odds ratios achieved by the two methods is small.

Usage of a monotherapy of Sirukumab (100mg every 4 weeks) exhibits that treatment had a positive effect compared to placebo, in both trials. Small differences between the odds ratio across two trials is likely due to within study error. Under the random-effect model, the OR was 2.94 and there was a 95% confidence interval around this of [2.38, 3.6]. Similarly to the Golimumab and MTX combined therapy, the combined effect of Sirukumab compared to placebo is positive with regards to ACR 20. The random-effect model was employed here due to the calculated I^2 score of 67.51% (>50%). This moderately high I^2 statistic brings uncertainty as to whether the studies in this meta-analysis can be considered from the same population. Subgroup analysis could be employed to explore this further. As expected, the confidence interval for the combined effect size calculated using the random-effect model (3.6) is wider than that of the one calculated using the fixed-effect model (3.33), as there is a higher degree of heterogeneity between trials. There is little difference between the combined odds ratios for both methods, with the random and fixed effect method resulting in 2.99 and 2.93 respectively.

In both meta-analyses, the random-effect model placed more weight on smaller trials. The combined effect sizes for both methods in each meta-analysis are very similar. Therefore, higher weights on small studies have had little effect on the combined effect size estimate.[31] In calculating the combined effects for Figure 2a, the ratio of weights for CR005263 to CR006343 was 0.37 for the fixed-effect model compared to 0.90 for the random-effect model. Here the fixed-effect ratio more closely aligns with the ratio of trial size, 0.32, as expected.

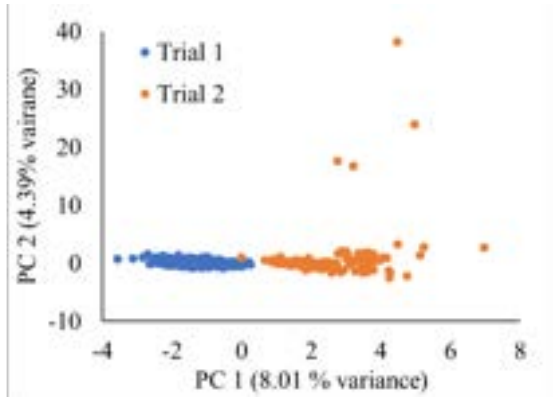
A similar story played out for the calculation of the combined effects in Figure 2b, with the ratio of weights for CR100864 to CR100866 of 0.52 for the fixed-effect model compared to 0.73 for the random-effect model. Again, the fixed-effect model aligns with the ratio of trial size (0.53) more closely, and the difference between the ratio of random-effect weights is closer to the size ratio for this meta-analysis as the size ratio is larger. The meta-analyses suggest that both therapies have a positive overall effect on the number of participants achieving ACR 20. This is in agreement with earlier meta-analyses conducted by Aaltonen et al.[32]

Overall, the efficacy of the treatment methods is clear, however this global analysis does not consider why such therapies may be more effective on one participant compared to another. A personalised level analysis using ML is valuable for understanding differences in efficacy.

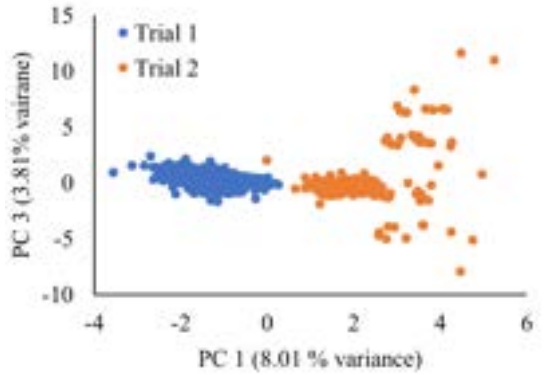
4.2 Machine learning

In this paper, the trials CR005263 (trial 1) and CR006343 (trial 2) were chosen using the inclusion criteria and are used for analysis using machine learning.

4.2.1 PCA



(a)



(b)

Figure 4: Two PCA plots to show the clustering of datapoints. (a) PC2 vs PC1, (b) PC3 vs PC1

In this classification application, there are 2 classes: patients from trial 1 and patients from trial 2. Generating a PCA plot is helpful to confirm that the class information is valid and to identify if the data contains more than the expected number of classes. In Figure 4, 2 classes are observed (with some outliers). If it was clear there were more clusters than expected, this would require further investigation. In this case, what is observed matches the prediction (2 classes present), and the next step can be taken.

It is worth noting that with IPD it is worth investigating outliers and potentially eliminating them from the dataset to avoid overfitting of the model later on in the process. However, since a pseudo-population was analysed, it was not deemed necessary to eliminate outliers in the interest of retaining all data, to ensure the methods remain general for use with different datasets.

Total variance is lower than expected at 16% which is less than the expected total variance for PCA analysis which is usually between 70-80% [33]as this is a randomly generated pseudo population these results are not expected to be comparable to IPD from a clinical trial.

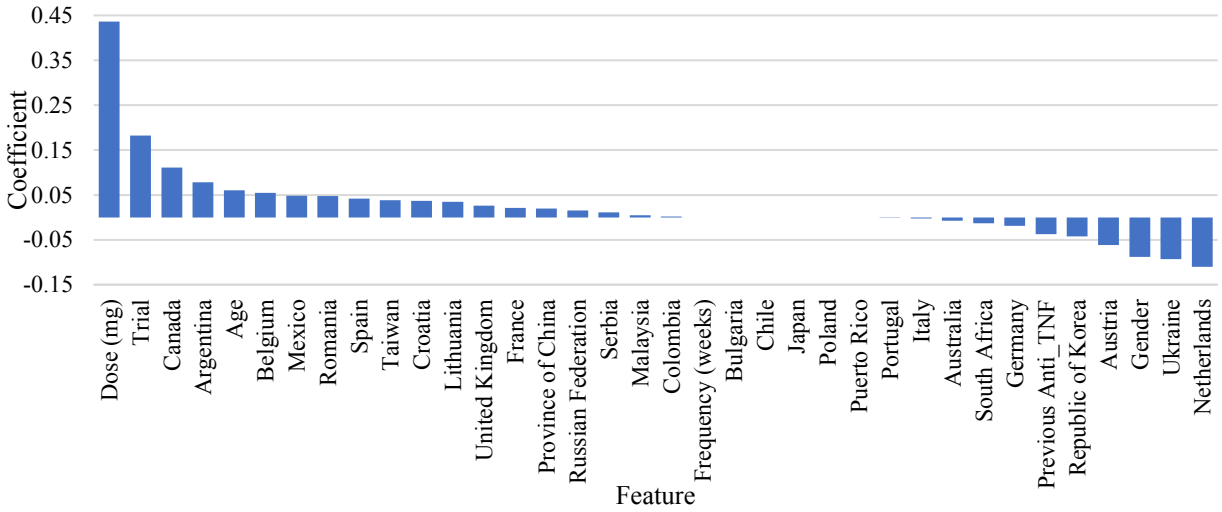
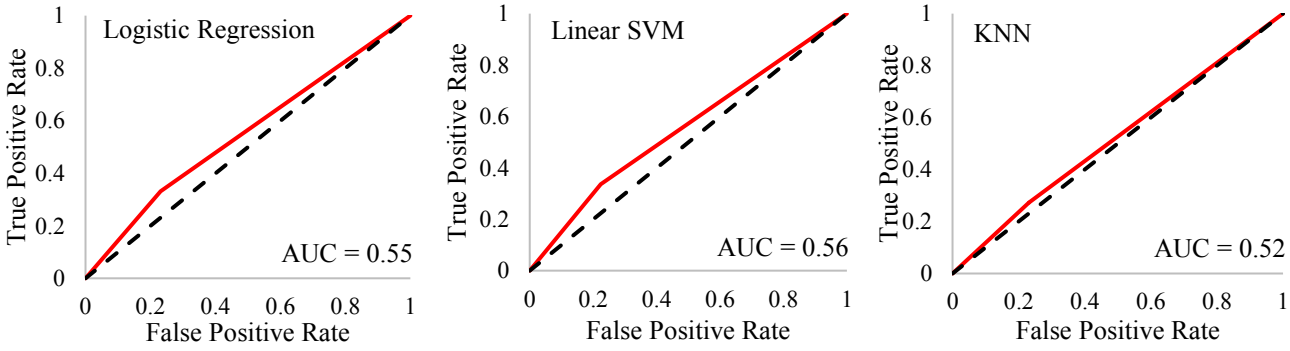


Figure 3: Feature selection graph showing the features importance based on L1 or Lasso regularisation. Features on the x axis sorted from left to right, highest coefficient to lowest.

Model Name	Training set accuracy	Cross-validation accuracy mean	Cross-validation accuracy standard deviation	Test set accuracy
kNN	71.88	56.77	2.25	58.63
Linear SVM	60.40	58.44	2.83	58.24
Logistic Regression	61.29	59.03	3.42	58.04
Kernel SVM	63.98	58.00	3.05	57.25
XGBoost	66.98	59.33	2.28	57.25
CatBoost	74.09	59.57	2.38	56.67
Random Forests	100.0	55.40	4.56	54.31
Decision Trees	100.0	52.95	2.13	54.12
Naïve Bayes	42.15	41.41	1.07	43.73

Table 4: Accuracy scores using different algorithms based on the pseudo-population dataset. The table is sorted from highest to lowest test accuracy



(a) (b) (c)

Figure 5: ROC graphs for logistic regression (a), Linear SVM (b), KNN (c) with AUC scores shown

4.2.2 Feature selection

After performing L1 regularisation on our data, we found that the following features had a coefficient of 0: frequency of dosage and the countries Bulgaria, Chile, Japan, Poland, Puerto Rico as seen in Figure 3. These features were then removed from the 'X' dataset and the dataset was split again into test and train groups. Reasons why only certain countries are redundant are not explicit, as in generating the pseudo-population, the generated ACR20 binary outcome is independent of the generated country. This discounts the possibility of environmental factors affecting the presentation of RA symptoms. If IPD was applied to the model, we might expect more obvious correlation between countries with colder climates, for example, and the outcome measure ACR 20. However, country data is not necessarily representative, as many countries have a vast difference in climate depending on location within the country, and the duration of time a participant has been living there can vary. It would have been valuable to have had ethnicity

data also, to grasp the effects of nature and nurture on the presentation of RA. Country data was included in this project as summary data is limited and so all information was used.

Dosage (mg) had the greatest magnitude of coefficient. This is likely because both trials have 3 groups of dosage groups: 0 (placebo), 50 and 100mg. Therefore, there was variance in the dataset that led the model to be the most sensitive to dosage. Frequency of dosage is a redundant feature as both trials used had the same frequency of drug administration of 2 weeks. In this case, as all frequency data points have the same value, it will not influence the model and this result confirmed this prediction. If this process was applied to trials with varying frequency of dosage, this may not be the case. Thus, it is worth keeping this feature in the dataset as the trials having the same frequency of dose administration may not always be the case.

4.2.3 Model evaluation

As can be seen in Table 4, by using the default hyperparameters for each algorithm, kNN had the highest test accuracy score and Naïve Bayes had the lowest.

The training set accuracy score was 100 for random forest and decision trees classifiers, indicating the models had overfitted the training data. To avoid this, one can look to increase the number of samples or decrease the number of features in the dataset. All test set and cross validation scores were low (below 60), suggesting all models underfitted the testing data. Scores under 60 are generally considered poor.[34] Low scores can be anticipated in this application since the pseudo-population dataset, being randomly generated, did not inherently contain any patterns for the model to find. As this is a proof-of-concept project, this example with a pseudo population is not representative of how well the models would perform with real data.

Each cross validated score in Table 4 shows underfitted scores. The standard deviation was small for each cross validation (less than 5), showing that each model performed quite consistently.

kNN, linear SVM and logistic regression were chosen to move onto the next step of tuning the hyperparameters to optimise the model and to increase the accuracy scores. A random search is employed followed by a grid search to attain the optimal parameters. Here it was found that tuning had no effect on the randomly generated pseudo population dataset, and no higher accuracy score was achieved for any model we attempted to tune.

It is important to evaluate the type of data the algorithm will be working with and the context of the application of the model. For example, some ML algorithms are complex such as deep neural networks but algorithms such as logistic regression or KNN are less so. For this framework, ease of understanding is important to ensure accessibility and ease of use for all. If there are technical barriers in providing research teams these tools, it will be expected that it will not be taken up as readily.

4.2.4 Predictions

This ML framework is designed to allow predictions of the efficacy of a drug based on certain patient characteristics. Whilst it is not applied in this proof-of-concept project it is the aim that the reader will be able to apply the model to their datasets of choice.

Predictions could be made to analyse the importance of features within clinical trials and a more specific selection criterion for patients can be made to avoid unnecessary time spent on feature selection. Additionally, by introducing a non-subjective tool into the allocation of treatment, it could help to reduce bias and help find the most suitable medication for a patient quicker. This improvement of efficiency on a larger scale and would lead to a reduced cost in the overall healthcare system. The next step after this project would be to implement the ML framework with IPD from YODA.

4.3 Limitations

At the beginning of this project, a research proposal was submitted to YODA detailing our project title, aims and methods and the sponsors of the trials, GSK and Jansson Pharmaceutics were also contacted. Due to time constraints, the IPD access was not granted in the timeframe of this project from YODA. Restrictions outlined in the sponsors' internal policy meant that they were not able to disclose patient level data.

This presented a unique challenge and allowed exploration of generating pseudo-populations for trials using summary data which is readily available on the YODA website. Randomly generating demographic and outcome data results in lower accuracy scores than what one could expect when IPD is fed into the framework. An advantage of the lack of access to IPD means the resulting framework is more general, and robust for different study designs.

ACR 20 has a low threshold of response, arguably an ACR50 or ACR70 score would speak more about the drugs efficacy and is what should be aimed for to significantly improve quality of life for patients. However, there is a trade off as the size of the groups achieving ACR50 and ACR70 are likely to be considerably smaller than those chieving ACR 20, meaning a larger percentage error in measures such as the odds ratio. ACR20 being a dichotomous measure is arguably too high-level as it does not incorporate information about the relative improvement of different symptoms for the patient in the trial. It would be better to use the measure 'bdACRhybrid', as proposed by the American College of Rheumatology in 2009.[35] This measure is much more sensitive to change, incorporates a requirement for joint count improvement and preserves the ACR20/50/70 system so can be analysed alongside trials which just use ACR20, for example.

Due to the lack of comparable studies within the YODA database for RA, the number of trials included in each meta-analysis is smaller than what is typical in other systematic reviews. The strength of the assumptions involved in meta-analyses can come into question with a

small number of trials, especially the strength of decisions because of the I^2 statistic. Ideally there would be a higher degree of homogeneity between trials for the same treatments to increase the ability to pool their results; this is an argument for a higher degree of data sharing in the pharmaceutical industry which would offer a greater scope for discovering new insight using clinical trial data.

5. Conclusion

In conclusion, two different methods were used to analyse multiple clinical trial data. The traditional method of meta-analysis is implemented to analyse global level data on the RA drugs Golimumab and Sirukumab to understand treatment efficacy. In using this technique, it was found from the trials studied that there are greater odds of achievement of ACR20 in the exposed vs the placebo groups.

The technique of machine learning was then used to analyse patient level data. Whilst the ML aspect of this project is not results based, a proof-of-concept result is provided based on clinical trial data for the RA drug Sirukumab and have shown that this framework can be applied to any data set. The concept of applying machine learning for analysing clinical trials is a growing field of work and it is hoped that this continues forward to help develop the efficiency of the process of clinical trials and the general healthcare industry in terms of time and capital investment. As discovered during this project, accessing patient level data can be a significant limitation in the implementation of ML in analysing clinical trials. This project has shown the potential of ML and how it can be incorporated into clinical trial analysis, as it improves the understanding of treatment effect heterogeneity by studying the relationships between demographic factors and the outcome ACR 20.

6. References

[1] R. K. Harrison, "Phase II and phase III failures: 2013-2015," *Nat Rev Drug Discov*, vol. 15, no. 12, pp. 817–818, Nov. 2016, doi: 10.1038/NRD.2016.184.

[2] J. A. DiMasi, H. G. Grabowski, and R. W. Hansen, "Innovation in the pharmaceutical industry: New estimates of R&D costs," *J Health Econ*, vol. 47, pp. 20–33, May 2016, doi: 10.1016/J.JHEALECO.2016.01.012.

[3] J. Listing *et al.*, "Clinical and functional remission: even though biologics are superior to conventional DMARDs overall success rates remain low – results from RABBIT, the German biologics register," *Arthritis Res Ther*, vol. 8, no. 3, p. R66, Apr. 2006, doi: 10.1186/AR1933.

[4] K. Chatzidionysiou and P. P. Sfikakis, "Low rates of remission with methotrexate monotherapy in rheumatoid arthritis: review of randomised controlled trials could point towards a paradigm shift," *RMD Open*, vol. 5, no. 2, p. e000993, Jul. 2019, doi: 10.1136/RMDOPEN-2019-000993.

[5] J. Listing *et al.*, "Clinical and functional remission: even though biologics are superior to conventional DMARDs overall success rates remain low – results from RABBIT, the German biologics register," *Arthritis Res Ther*, vol. 8, no. 3, p. R66, Apr. 2006, doi: 10.1186/AR1933.

[6] Y. Tanaka, "Recent progress in treatments of rheumatoid arthritis: an overview of developments in biologics and small molecules, and remaining unmet needs," *Rheumatology*, vol. 60, no. Supplement_6, pp. vi12–vi20, Dec. 2021, doi: 10.1093/RHEUMATOLOGY/KEAB609.

[7] "NHS England » NHS set to save £150 million by switching to new versions of most costly drug." https://www.england.nhs.uk/2018/10/nhs-set-to-save-150-million-by-switching-to-new-versions-of-most-costly-drug/ (accessed Dec. 14, 2022).

[8] I. B. McInnes and G. Schett, "The Pathogenesis of Rheumatoid Arthritis," *https://doi.org/10.1056/NEJMra1004965*, vol. 365, no. 23, pp. 2205–2219, Dec. 2011, doi: 10.1056/NEJMRA1004965.

[9] "How big data can revolutionize pharmaceutical R&D | McKinsey." https://www.mckinsey.com/industries/life-sciences/our-insights/how-big-data-can-revolutionize-pharmaceutical-r-and-d (accessed Dec. 11, 2022).

[10] A. Rghioui, J. Lloret, S. Sendra, and A. Oumnad, "A Smart Architecture for Diabetic Patient Monitoring Using Machine Learning Algorithms," *Healthcare 2020, Vol. 8, Page 348*, vol. 8, no. 3, p. 348, Sep. 2020, doi: 10.3390/HEALTHCARE8030348.

[11] H. Gerdes *et al.*, "Drug ranking using machine learning systematically predicts the efficacy of anti-cancer drugs," *Nature Communications 2021 12:1*, vol. 12, no. 1, pp. 1–15, Mar. 2021, doi: 10.1038/s41467-021-22170-8.

[12] K. el Emam, S. Rodgers, and B. Malin, "Anonymising and sharing individual patient data," *BMJ*, vol. 350, Mar. 2015, doi: 10.1136/BMJ.H1139.

[13] "The YODA Project." https://yoda.yale.edu/ (accessed Dec. 14, 2022).

[14] "Odds Ratio Meta-analysis (Mantel-Haenszel and Exact) - StatsDirect." https://www.statsdirect.co.uk/help/meta_analysis/mh.htm (accessed Dec. 14, 2022).

[15] S. Tenny and M. R. Hoffman, "Odds Ratio," *Encyclopedia of Genetics, Genomics, Proteomics and Informatics*, pp. 1388–1388, May 2022, doi: 10.1007/978-1-4020-6754-9_11771.

[16] M. Borenstein, L. v. Hedges, J. P. T. Higgins, and H. R. Rothstein, "A basic introduction to fixed-effect and random-effects models for meta-analysis," *Res Synth Methods*, vol. 1, no. 2, pp. 97–111, Apr. 2010, doi: 10.1002/JRSM.12.

[17] "Cochrane Handbook for Systematic Reviews of Interventions." https://handbook-5-1.cochrane.org/.htm (accessed Dec. 15, 2022).
[18] J. Robins, N. Breslow, and S. Greenland, "Estimators of the Mantel-Haenszel Variance Consistent in Both Sparse Data and Large-Strata Limiting Models," *Biometrics*, vol. 42, no. 2, p. 311, Jun. 1986, doi: 10.2307/2531052.
[19] D. Jackson, J. Bowden, and R. Baker, "How does the DerSimonian and Laird procedure for random effects meta-analysis compare with its more efficient but harder to compute counterparts?," *J Stat Plan Inference*, vol. 140, no. 4, pp. 961–970, Apr. 2010, doi: 10.1016/J.JSPI.2009.09.017.
[20] K. Sidik and J. N. Jonkman, "A simple confidence interval for meta-analysis," *Stat Med*, vol. 21, no. 21, pp. 3153–3159, Nov. 2002, doi: 10.1002/SIM.1262.
[21] Andriy Burkov, *The Hundred-Page Machine Learning Book*. 2019.
[22] "Feature Selection to Improve Accuracy and Decrease Training Time - MachineLearningMastery.com." https://machinelearningmastery.com/feature-selection-to-improve-accuracy-and-decrease-training-time/ (accessed Dec. 14, 2022).
[23] "1.13. Feature selection — scikit-learn 1.2.0 documentation." https://scikit-learn.org/stable/modules/feature_selection.html (accessed Dec. 14, 2022).
[24] "View of A Review of Principal Component Analysis Algorithm for Dimensionality Reduction." https://publisher.uthm.edu.my/ojs/index.php/jscdm/article/view/8032/4199 (accessed Dec. 15, 2022).
[25] "Choosing the right estimator — scikit-learn 1.2.0 documentation." https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html (accessed Dec. 11, 2022).
[26] "sklearn.model_selection.RandomizedSearchCV — scikit-learn 1.2.0 documentation." https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html (accessed Dec. 11, 2022).
[27] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982, doi: 10.1148/RADIOLOGY.143.1.7063747.
[28] D. Aletaha *et al.*, "Efficacy and safety of sirukumab in patients with active rheumatoid arthritis refractory to anti-TNF therapy (SIRROUND-T): a randomised, double-blind, placebo-controlled, parallel-group, multinational, phase 3 study," *The Lancet*, vol. 389, no. 10075, pp. 1206–1217, Mar. 2017, doi: 10.1016/S0140-6736(17)30401-4.
[29] T. Takeuchi *et al.*, "Sirukumab for rheumatoid arthritis: The phase III SIRROUND-D study," *Ann Rheum Dis*, vol. 76, no. 12, pp. 2001–2008, Dec. 2017, doi: 10.1136/ANNRHEUMDIS-2017-211328.
[30] J. H. Leu, O. J. Adedokun, C. Gargano, E. C. Hsia, Z. Xu, and G. Shankar, "Immunogenicity of golimumab and its clinical relevance in patients with rheumatoid arthritis, psoriatic arthritis and ankylosing spondylitis," *Rheumatology (United Kingdom)*, vol. 58, no. 3, pp. 441–446, Mar. 2019, doi: 10.1093/RHEUMATOLOGY/KEY309.
[31] J. Kay *et al.*, "Golimumab in patients with active rheumatoid arthritis despite treatment with methotrexate: A randomized, double-blind, placebo-controlled, dose-ranging study," *Arthritis Rheum*, vol. 58, no. 4, pp. 964–975, Apr. 2008, doi: 10.1002/ART.23383.
[32] K. J. Aaltonen, L. M. Virkki, A. Malmivaara, Y. T. Konttinen, D. C. Nordström, and M. Blom, "Systematic review and meta-analysis of the efficacy and safety of existing TNF blocking agents in treatment of rheumatoid arthritis," *PLoS One*, vol. 7, no. 1, Jan. 2012, doi: 10.1371/JOURNAL.PONE.0030275.
[33] "Principal Components (PCA) and Exploratory Factor Analysis (EFA) with SPSS." https://stats.oarc.ucla.edu/spss/seminars/efa-spss/ (accessed Dec. 15, 2022).
[34] "What is a good accuracy score? Simply explained." https://stephenallwright.com/good-accuracy-score/ (accessed Dec. 13, 2022).
[35] reginap, "A Proposed Revision to the ACR20: The Hybrid Measure of American College of Rheumatology Response AMERICAN COLLEGE OF RHEUMATOLOGY COMMITTEE TO REEVALUATE IMPROVEMENT CRITERIA," 2007, doi: 10.1002/art.22552.
[36] H. Bhavsar and M. H. Panchal, "A Review on Support Vector Machine for Data Classification," *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, vol. 1, no. 10, pp. 2278–1323, 2012.
[37] S. Zhang, M. Zong, X. Zhu, D. Cheng, and X. Li, "Learning k for kNN classifi-cation," *ACM Trans. Intell. Syst. Technol*, vol. 8, no. 43, 2017, doi: 10.1145/2990508.
[38] "1.9. Naive Bayes — scikit-learn 1.2.0 documentation." https://scikit-learn.org/stable/modules/naive_bayes.html (accessed Dec. 14, 2022).
[39] S. B. Kotsiantis, "Decision trees: a recent overview," *Artif Intell Rev*, vol. 39, pp. 261–283, 2013, doi: 10.1007/s10462-011-9272-4.
[40] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, "A comparison of random forest variable selection methods for classification prediction modeling," *Expert Syst Appl*, vol. 134, pp. 93–101, Nov. 2019, doi: 10.1016/J.ESWA.2019.05.028.
[41] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: Unbiased boosting with categorical features," *Adv Neural Inf Process Syst*, vol. 2018-December, pp. 6638–6648, 2018, Accessed: Dec. 15, 2022. [Online]. Available: https://mljar.com/machine-learning/catboost-vs-xgboost/

Synthesis of Coconut Fatty Acid DMAE Esters for Betaine Biosurfactants

Huan Yang Tan and Rishabh Prasad
Imperial College London, Department of Chemical Engineering, U.K.

Abstract

Speciality chemicals for use in personal care products and cosmetic formulations, can be derived from vegetable oils and have the potential to replace traditional petrochemicals. Cocoyl ester betaine (CEB) is an amphoteric surfactant with use in cosmetic products that can be produced in a two-step process. This study investigated the first step, which is a transesterification of either coconut oil triglycerides or fatty acid methyl esters (FAME) with 2-dimethylaminoethanol (DMAE). This produces a coconut fatty acid DMAE ester (CFADE) intermediate desirable at high purity for subsequent reaction. The product was characterised using ^{1}H-NMR confirming formation of CFADE. Transesterification using oil and FAME substrates were compared using a microwave reactor which afforded product samples with only 51.1 and 32.5 % CFADE content for oil and FAME respectively at stoichiometric reactant ratios. Following the estimation of equilibrium constants, the severe limitations were revealed which led to investigations into shifting the position of equilibrium by varying the excess of DMAE. This method offered improvements but required molar excesses of over 5 (for FAME) and 10 (for oil) to approach 100% purity. Methanol, the by-product of transesterification of FAME and DMAE was continually removed using vacuum distillation. Over 96% purity was obtained at a 2:1 DMAE to FAME molar ratio operating at 50 mbar and 60^{o}C.

1. Introduction

Mineral oil derived from petroleum has been an important raw material in the chemical industry since the 1940s. However due to finite petrochemical sources and an ever increasing demand for environmentally friendly processes, natural fats and vegetable oils have garnered attention as a feedstock for the production of oleochemicals utilised in products such as pharmaceuticals, cosmetics and detergents. Compared to their petroleum counterparts, vegetable oils have economic and ecological advantages. They display low toxicity, are biodegradable and can be obtained from renewable resources. The use of vegetable oils in the chemical industry also opens up the pathway to new synthetic routes not previously accessible (Baumann et al., 1988).

Another common use of vegetable oils is for the manufacturing of biodiesel by a process called transesterification, converting triglycerides into fatty acid methyl esters (FAME) as seen in Figure 3. FAME additionally serves a role as an intermediate for conversion into other products. Using FAME in synthesis pathways is sometimes preferrable to vegetable oils for reasons such as them having lower viscosity and fewer impurities (Belousov et al., 2021).

In the beauty and personal care industry, specialty chemicals derived from vegetable oils can have roles in cosmetic formulations such as thickeners, emollients and surfactants. The fatty acid profiles of the parent vegetable oils determine their physiochemical characteristics, available reaction pathways and the properties of the final product. For example, ricinoleic acid found in castor oil has hydroxyl functionality, which also enhances its emollient properties. Coconut oil in particular is an excellent feedstock due to its high lauric acid content (C-12:0), giving it good foaming properties and making it suitable to make surfactants from (Boateng et al.,2016).

Surfactants are molecules that lower the surface energy between interfaces such as those between oil and water which are prevalent in many cosmetic formulations. They have roles such as emulsifiers or foaming agents (Yea et al., 2021). Surfactants are made up of a hydrophilic and hydrophobic group and can be categorised according to the chemical nature of its polar part. These categories consist of anionic, cationic, amphoteric and non-ionic surfactants (Hayes & Smith, 2019).

A large majority of the surfactants used in industry are petroleum derived further motivating research in the field of biosurfactants, by which the hydrophobic portion can be substituted with bio-based carbon (Van Bogaert et al., 2007)

2. Background

Betaines are an important class of chemicals employed as amphoteric surfactants in many personal care products such as shampoos and liquid soaps. These molecules are zwitterionic and are characterised by their quaternary ammonium (NH_4^+) and carboxylate (COO^-) group (Clendennen & Boaz, 2019). Alkyl betaines (Figure 1.A) and especially alkyl amido betaines are two amphoterics used widely commercially. The hydrophobic tail groups of alkyl betaines are typically produced from an alkyl

dimethyl amine feedstock derived from petroleum sourced alpha olefins. The alkyl dimethyl amine can also be synthesised from fatty alcohols, either petro- or oleo-derived (Shah et al., 2016).

On the other hand, alkyl amido betaines obtain their hydrophobe from oleochemical fatty acids. Alkyl amido betaines are produced industrially by reacting fatty acids or esters of fatty acids with a linker molecule. In the case of cocoamidopropyl betaine (CAPB) (Figure 1.C), coconut oil fatty acids or fatty acid esters first undergo amidation with the linker molecule 3-dimethylaminopropylamine (DMAPA) (Figure 1.B). The amide intermediate is subsequently reacted with monochloroacetic acid (MCA) to form the final product. CAPB has become an essential surfactant since the 1960s (Floyd & Jurczyk, 2008) with 3,328 personal care and cosmetic products containing it as of 2022, according to EWG's Skin Deep database. Previous studies report that the amidation step requires high temperatures in the ranges of 120-160 °C (Herrwerth, 2008) and 150-180 °C (Clendennen & Boaz, 2019). Alkyl amido betaines possess an amide bond which can be broken down and biodegraded by natural enzymes, making it superior in terms of environmentally friendliness to alkyl betaines (García, Campos & Ribosa, 2007).

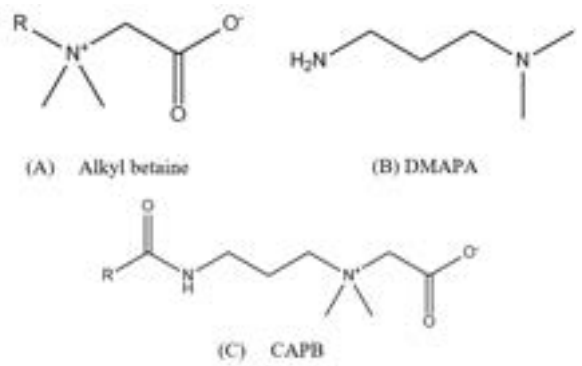


Figure 1: Molecular structures of Alkyl betaine, DMAPA and CAPB(alkyl amido betaine), where R is a fatty acid chain with lipid numbers between (8:0) and (14:0)

This study focuses on the synthesis pathways for cocoyl ethyl betaine (CEB) (Figure 2) which falls under a class of betaine called ester betaines and are characterised by their ester bond. CEB is also derived from coconut oil but while the hydrophobic tail groups for CEB and CAPB originate from the same feedstock, the key difference between the two is the linker molecule used. A study tested subjects sensitive to CAPB, with 1% DMAPA, resulting in all subjects displaying allergic reactions (Foti et al., 2003). For this reason, a CAPB-like surfactant using an alternative linker is desirable for the portion of the population allergic to CAPB. DMAE which contains hydroxyl functionality is a good option as the tertiary amino group is sterically hindered, reducing nucleophilicity towards the carbonyl group of the fatty acid ester as the reaction is a transesterification. The 2-carbon alkyl chain between the amino and hydroxyl group, further increases nucleophilicity of the hydroxyl group (Nagumalli, Jacob & Gamboa, 2020). CEB also has the biodegradable properties of CAPB, as the ester bond readily undergoes hydrolysis. The ammonium group withdraws electron density, making the ester carbonyl prone to attack from nucleophiles (Hellberg, Bergström & Holmberg, 2000).

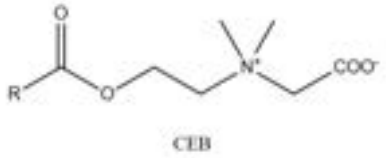


Figure 2: CEB molecular structure

CEB has very similar applications and properties to CAPB but have been reported significantly less extensively in literature and could not be found in any commercial products listed on EWG's Skin Deep's database. CEB has previously been made in a two-step process using ethyl esters of coconut oil fatty acids as the substrate. The fatty acid ethyl esters were first reacted with DMAE in an enzyme catalysed transesterification process. The second step is a reaction of the ester intermediate with MCA, and is analogous to the CAPB process (Burk et al., 2016).

An economical industrial process for CEB that has milder transesterification conditions than the amidation in the production of CAPB is desirable, as it would have lower costs and environmental implications. Additionally, bypassing the initial conversion into fatty acid esters and performing DMAE transesterification with coconut oil to form high purity ester intermediate would reduce the number of required processing steps, resulting in the same advantages.

This study carries out an investigation into the transesterification of FAME and coconut oil to form coconut fatty acid DMAE esters (CFADE).

3. Methodology

3.1 Chemicals

Sodium methoxide powder (95%, CH_3NaO) and 2-dimethylaminoethanol (>99.5%, $C_4H_{11}NO$) were purchased from Sigma-Aldrich. Methanol (CH_3OH), sodium chloride (NaCl), and sodium sulphate anhydrous (Na_2SO_4) purchased from VWR Chemicals were all in analytical reagent grade. The coconut oil was purchased from Sevenhills Wholefoods.

3.2. Catalyst selection

The catalyst chosen for FAME production and the transesterification of oil or FAME with DMAE was CH_3NaO. This is because it forms the required methoxide anion (CH_3O^-) while preventing saponification (which is highly undesirable), as opposed to other alternatives.

3.*3 FAME production*

FAME was produced by transesterification using a 6:1 methanol to coconut oil molar ratio with sodium methoxide (1 wt% oil) catalyst. About 300g of coconut oil was added to a 500 ml dual-necked flask and heated to 60°C using a

water bath in a reflux setup. The magnetic stirrer was set to 600rpm to ensure the homogenisation of the reaction mixture. Once the desired temperature was achieved, 90g of methanol containing 3g of dissolved catalyst was added to the flask and reacted for 1 hour. Upon completion, the content in the flask was transferred to a separatory funnel and left undisturbed until two distinct liquid phases could be observed. The bottom glycerol layer was disposed of, leaving the upper product layer of, which was washed with 10% brine solution to remove dissolved impurities (water, methanol, glycerol, and catalyst). This process was repeated as necessary until the aqueous discharge reached a neutral pH. Lastly, anhydrous sodium sulphate was added to remove water traces from the finished product. The liquid was then separated from the solid and stored in a bottle to be kept in a freezer. This process produced approximately 300g of 98.5% FAME.

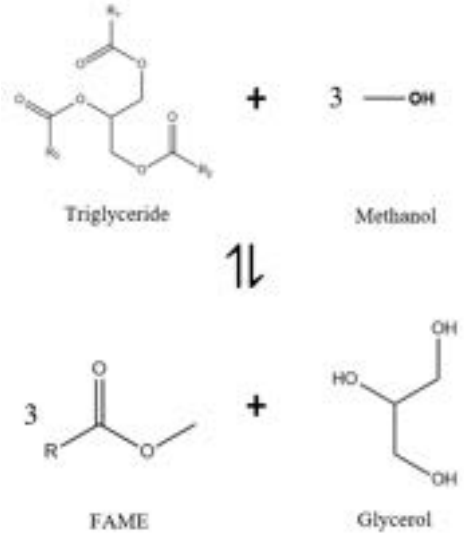


Figure 3: Reaction scheme for transesterification of oil to FAME, where R represents a fatty acid chain

3.4 Transesterification using microwave

Two sets of experiments were conducted with each substrate, oil and FAME. For the first set, a molar ratio of 3:1 for DMAE (2.67 g) to coconut oil (6.36 g) and a 1:1 molar ratio for DMAE (2.57 g) to FAME (6.16 g) was used. These ratios correspond to reaction stoichiometry shown in Figures 4 and 5. The reaction mixture was added to a 20ml borosilicate glass vial with sodium methoxide (1 wt% oil/FAME) and a magnetic stirrer. The reaction mixture was then made homogenous by vigorously shaking. The reaction temperature was held at 120°C, controlled by a thermometer probe inserted through the capillary tube in the PTFE cap. The temperature was held for different durations (1, 2, 5, 10, 20 and 30 minutes) and stirred at 600 rpm. At the end of the reaction, the vial was quickly cooled with compressed air to 55°C. The heating and cooling time was roughly 3 minutes combined.

For the second set of experiments, different DMAE to Oil molar ratios between 4:1 – 10:1 were used. And DMAE to FAME molar ratios in the range of 2:1 – 5:1 were investigated. This experimental procedures was identical to the previous one, apart from a fixed holding time of 2 minutes.

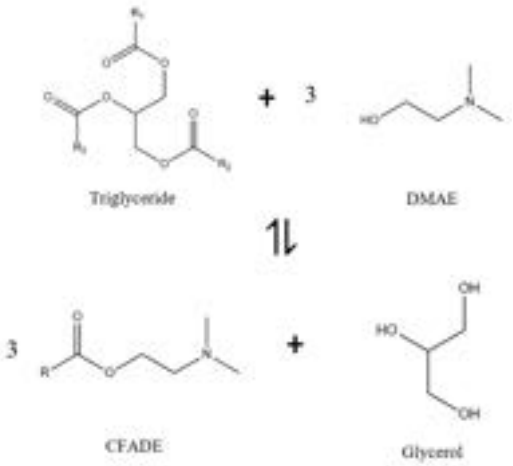


Figure 4: Reaction scheme for transesterification of oil to CFADE

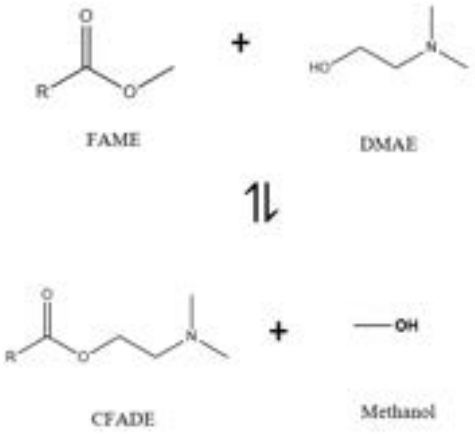


Figure 5: Reaction scheme for transesterification of FAME to CFADE

3.5 Transesterification using rotary evaporator

Vacuum distillation was used to carry the experiments at reduced pressure. The experiments were performed using a range of pressures from 300 to 30 mbar while varying the molar ratios of DMAE to FAME from 1:1 to 4:1. Likewise, sodium methoxide (1 wt% FAME) and stirrer were added to a 100ml two-necked round bottom flask alongside the reactants (FAME and DMAE) before starting the experiments. The flask was set to rotate at 100rpm on an axis in a 60°C water bath. All the reactions were run for 20 minutes.

3.6 Sample work-up

After each reaction, the crude product was either afforded as a gel, semi-gel or liquid. For the first two cases, the gel was first broken up by vigorous stirring. Next the crude product was mixed with roughly 5-10 ml of 10% brine solution and then transferred to a centrifuge tube. It was then centrifuged for 10 minutes at 9000 rpm until two visible phases could be observed. The top layer contained

CFADE with unreacted oil or FAME, while the bottom aqueous phase contained the removed CH_3NaO catalyst, DMAE and methanol or glycerol depending on the reaction type. The top layer was carefully separated using a transfer pipette to another tube. Anhydrous Na_2SO_4 was used to dry the washed product. Once dried, the CFADE product sample (clear liquid) was transferred to a 20ml glass vial and frozen for further usage.

3.7 Product characterisation using ^{1}H-NMR spectroscopy

^{1}H-NMR spectra data were obtained using a Joel JNM-ECZ400S/L1 FT NMR spectrometer running at 400 MHz. Each NMR sample was prepared by adding 35µL of product and 820µL of deuterated chloroform ($CDCL_3$) to an Eppendorf tube and vigorously shaking it before transferring to an NMR tube. A metric for CFADE content in a sample was derived by comparing the obtained peak integral ratio to the expected ratio for 100% CFADE content (Equation 1).

$$CFADE\ Content\ (\%) = \frac{3A_2}{2A_1} \times 100\% \qquad (1)$$

Where A is the integration of the corresponding signal shown in Table 1.

4. Results and Discussion

4.1 Analysis of ^{1}H-NMR spectra

Figures 6 and 7 show various ^{1}H-NMR spectra used to characterise the formation of CFADE. Signals present in ^{1}H-NMR spectra obtained using the method outlined in section 3.7 were assigned by analysing expected molecular structures and utilising online resources such as nmrdb.org. The proton assignments are tabulated in Table 1.

Without an internal standard, CFADE content was calculated semi-quantitatively using Equation 1. This method neglected potential side-products and impurities, assuming the sample to only contain CFADE and unreacted fatty acid feedstock .

Figure 6 shows the ^{1}H-NMR sample prepared from a reaction with oil and DMAE. The integration of δ 0.86 ppm which corresponds to the terminal -CH3 present in both triglycerides and CFADE is normalised to 3. The peak integration of 1.52 at δ 2.55 ppm which corresponds to -CH_2N- in CFADE, is then compared to the expected ratio of 2, yielding a CFADE content of 76% in the sample. Selecting the region for the integral at δ 2.55 ppm generally resulted in errors of ±0.05, manifesting as ± 2.5% CFADE content.

Table 1 ^{1}H-NMR spectra peak assignment

Signal	Chemical shift, δ (ppm)	Proton Assignment
1	0.86	Terminal -CH_3
2	2.55	-CH_2N-
3	4.15	-CH_2O-
4	3.65	-OCH_3

Comparison of Figures 6 and 7 highlight an important difference between the oil and FAME transesterifications. Unexpected peaks are observed in Figure 6 for oil reaction at around δ 3.59 ppm which may be attributed to the formation of intermediates and side products such as partial glycerides. This behaviour is observed in traditional transesterification of triglycerides with methanol when the reaction has not gone to completion due to its stepwise nature (Noureddini & Zhu, 1997) . Partial glycerides act as surfactants which can stabilize emulsions and complicate separation of the product(Kishore, 2017).

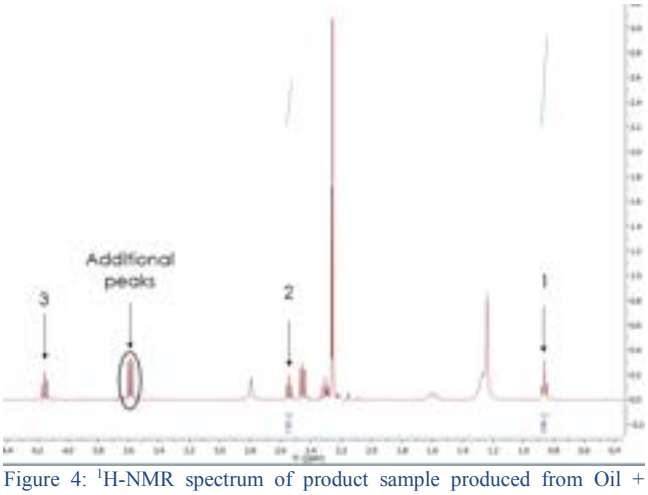


Figure 4: ^{1}H-NMR spectrum of product sample produced from Oil + DMAE reaction using method in section 3.4 using an 8:1 DMAE to oil ratio. Numbers correspond to the signals in Table 1

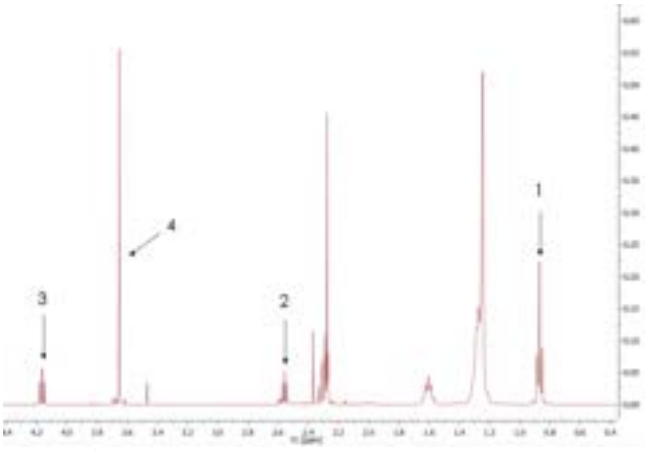


Figure 5: ^{1}H-NMR spectrum of product sample produced from FAME+ DMAE reaction using method in section 3.4 using a 1:1 DMAE to FAME ratio. Numbers correspond to the signals in Table 1

4.2 Equilibrium Limitations of Microwave Reactions

The results of the DMAE transesterifications with oil and FAME at stoichiometric ratios (3:1 and 1:1) following the methodology outlined in section 3.4 can be seen in Figure 8. It is observed that the CFADE content of samples reach a fixed value and do not change significantly with time for both feedstocks. This indicates the existence of a chemical equilibrium, which is reached by both reactions within 1 minute. The average CFADE content is 51.1% for the oil substrate and 32.5% for the FAME substrate.

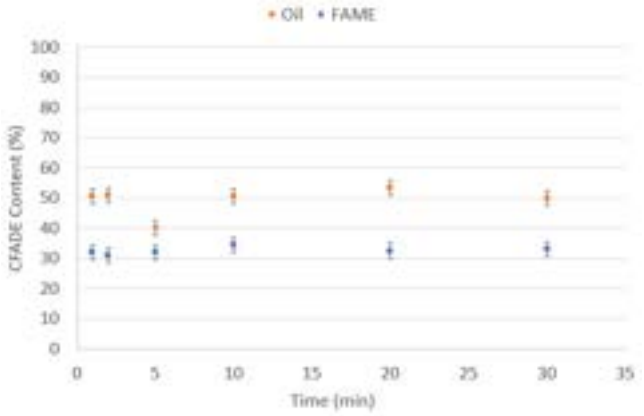


Figure 8: Oil or FAME reacted with DMAE in stoichiometric ratios at 120℃ using microwave-assisted heating

A method was developed to provide an estimate for the average constants of equilibrium (K) for both reactions. This was carried out by relating the metric for CFADE content (Equation 1) a,nd the ratio between components in the product samples, to the concentrations in the reaction mixtures.

In the case of the reactions with FAME, the value of K could be calculated using the relative ratio of peak integrations in the ^{1}H-NMR spectra. As per Table 1, the signal at δ 2.55 ppm corresponds to (-CH_2N-) protons which are unique to CFADE while that at δ 3.65 ppm corresponds to (-OCH_3) in FAME . Normalisation of each signal to the number of protons can be used to compute a concentration ratio. Equations 2 and 3 were used to carry out these calculations.

$$K(FAME) = \frac{[CFADE][CH_3OH]}{[FAME][DMAE]} = \frac{x^2}{(a-x)(b-x)} \quad (2)$$

$$\frac{CFADE\ signal}{FAME\ signal} = \frac{\frac{A_2}{2}}{\frac{A_4}{3}} = \frac{[CFADE]}{[FAME]} = \frac{x}{(a-x)} \quad (3)$$

Where a and b correspond to the initial moles of FAME and DMAE, and x is the moles of CFADE formed at equilibrium. A_1 and A_4 correspond to the peak integrals as listed on Table 1.

In the case for the reactions with oil, signals corresponding to glycerol backbone protons in triglycerides were unclean and could not be integrated accurately. This meant that the relative signal of triglycerides in the sample could not be obtained. Instead, CFADE content was related to the consumption of oil, where total consumption of oil would produce a sample containing 100% CFADE. This is however assumes that any potential side reactions do not consume the oil. Equations 4 and 5 were used to carry out these calculations.

$$K(Oil) = \frac{[CFADE]^3[Glycerol]}{[Oil][DMAE]^3} = \frac{(3x)^3(x)}{(c-x)(b-3x)^3} \quad (4)$$

$$CFADE\ content\ in\ sample = \frac{x}{c} \quad (5)$$

Oil is assumed to only consist of coconut oil triglycerides, c refers to its initial moles, b refers to the initial moles of DMAE and x refers to moles of oil consumed.

The standard Gibbs free energy change (ΔGº) can also be determined for both reactions using Equation 6.

$$\Delta G° = -RT \ln(K) \quad (6)$$

Where R=8.314 J mol^{-1} K^{-1} and T=393 K

The sample that produced figure 6 was produced from a reaction with DMAE and oil in the molar ratio 8:1. As discussed in section 4.1, it had a 76% CFADE content. Using Equation 5, a value for x of 0.76 is obtained. Finally using Equation 4, a value for K(Oil) of 0.2 is obtained. K values for FAME and Oil reactions were obtained and averaged to yield the results in Table 2.

The obtained K values for both reactions do not favour the formation of CFADE, meaning that there are quite significant equilibrium limitations. The ΔGº for the reactions with oil substrate is almost 3 times greater than that of the FAME substrate. This can be justified by referring to Figures 4 and 5, and realising that in both reactions the bonds broken and formed are quite similar. In the case of oil and DMAE, 3 times as many bonds are broken and formed.

There is an additional caveat of this analysis. The relative concentrations of CFADE samples can only be applied to the concentrations of the entire reaction mixture if all reactants and products remain in the reaction phase until equilibrium is reached. The appearance of a gel phase (discussed in section 4.4) in the reaction mixtures is inherently associated with a change in phase behaviour and therefore only reactions resulting in a homogenous liquid phase were considered for these calculations. For both substrates after allowing the liquid reaction mixtures to settle, there was no apparent phase separation and therefore all components were assumed to be mutually soluble.

Table 2 Equilibrium constants (K) and standard Gibbs free energy change (ΔGº) for each reaction at 120 ºC

Reaction	K	ΔGº (KJ/mol)
Oil + DMAE	0.230	4.798
FAME + DMAE	0.593	1.707

4.3 *Overcoming equilibrium limitations*

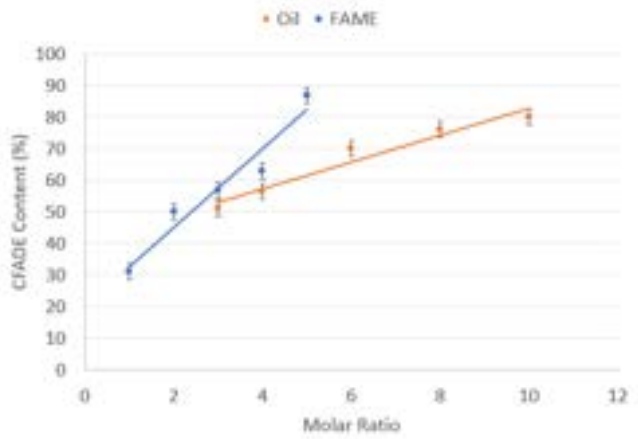


Figure 9: Oil or FAME reacted with excess DMAE at 120°C using microwave-assisted heating

The previous section highlights the problem of equilibrium limitations preventing the production of high purity CFADE, which is required for CEB synthesis. According to the *K* expressions in Equations 2 and 4, an increase in reactant concentration will shift the equilibrium towards the product side. As per the method described in section 3.4, microwave transesterifications were carried out with both coconut oil and FAME substrates with a varying excess of DMAE. The results of these experiments can be seen in Figure 9. As expected, CFADE content in product samples increases steadily with increasing mole ratio. The purity of CFADE increases at a faster rate w.r.t molar ratio for the reactions using FAME. They eventually overtook the oil substrate reactions at a molar ratio of 3:1 and reach CFADE content of 87% at a ratio of 5:1, while the reactions with oil reach 80% at a ratio of 10:1 and also seem to plateau.

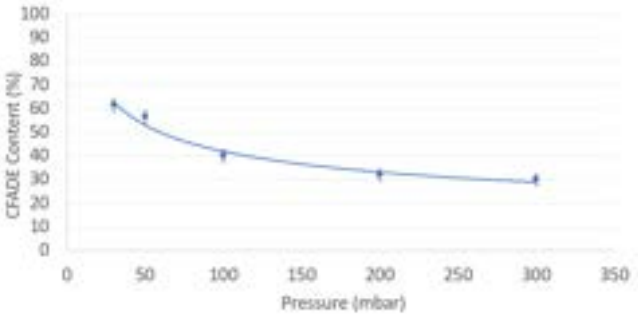


Figure 10: FAME reacted with DMAE at stoichiometric ratio for 20 minutes using rotary evaporator heated by a 60°C water bath under vacuum.

The reactions using FAME as a substrate offered an alternative method to overcome the equilibrium limitations. Due to the boiling point differences between methanol (64°C) and DMAE (133°C), continuous removal of the methanol would also promote a shift in equilibrium and was carried out using a rotary evaporator as described in section 3.5.

Aspen Plus V11 was used to approximate methanol and DMAE as a binary mixture using the NRTL method. The purpose of this was to choose initial operating pressures resulting in a methanol enriched vapour phase. Figure 10 shows that increasing the vacuum increases sample CFADE content up to 61.5% at 30mbar for a 1:1 molar ratio at 60°C.

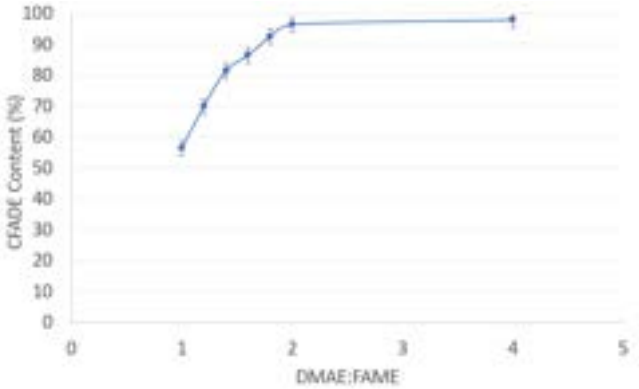


Figure 11: FAME reacted with excess DMAE for 20 minutes using rotary evaporator heated by a 60°C water bath at 50 mbar

The final set of experiments combined an excess of DMAE and rotary evaporation. The operating pressure was chosen to be slightly higher at 50 mbar, as at maximum vacuum operation, the fishy odour of DMAE was detected in the condensate. This is due to the vaporisation of DMAE along with methanol, which must be avoided to maximise DMAE concentration in the reaction phase. As seen in Figure 11, this method was highly effective, resulting in a trend that approaches a purity of 100%. A maximum CFADE purity of 98% was obtained at a DMAE to FAME ratio of 4:1, however this is a marginal increase from a 2:1 ratio that has 96.5% CFADE purity.

For this reason it is suggested that a 2:1 ratio at 50mbar and 60°C would be the optimal conditions to synthesise high purity CFADE which can then undergo reaction with SCA to form CEB.

4.4 *Gel phase behaviour of the crude product*

As mentioned before, the crude product was either afforded as a gel, semi-gel or liquid, depending on the molar ratio used, which is also described in Figure 12. A possible theory for the gel formation is due to either DMAE or CFADE acting as a gelator and forming oleogels in the product mixture. Factors that could cause this are the intermolecular forces, such as hydrogen bonding and van der Waals interactions between the gelator and fatty acids chains (Li et al., 2022). Additionally, these forces are affected by the fatty acid structure, that is, the number of unsaturated carbons and the carbon chain-length. Furthermore, gel formation is a major concern as it will change the viscosity of the crude product, which may give rise to complications in industrial processes.

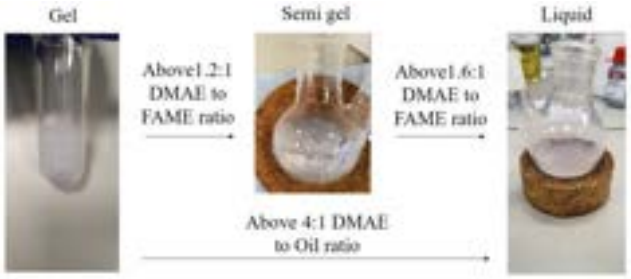


Figure 12: Gel behaviour of crude product mixture

5. Conclusions

By using ^{1}H-NMR as a characterisation technique, product formation from both substrates was confirmed by identifying the corresponding peaks of (-CH_2N-) at δ 2.55 ppm and (-CH_2O-) at δ 4.15 ppm, which are found in CFADE. Transesterification of both coconut oil and FAME feedstocks are heavily equilibrium limited as showcased by their low values of equilibrium constant, 0.230 for oil and 0.593 for FAME. FAME has advantages as a substrate as it does not form partial glyceride intermediates, and has potential for vacuum distillation. 96.5-98% CFADE was produced in a rotary evaporator at 2:1 or 4:1 DMAE to FAME ratio at 50 mbar and 60 °C. It would be of interest to use this high purity product in the second step of CEB production, consisting of a reaction with SCA.

The ^{1}H-NMR characterization techniques employed were sufficient as an initial study, however an internal standard would be desired to make the method more quantitative. This would allow for more robust calculations of thermodynamic quantities. The positive values of ΔG^o suggest that investigation into temperature effects to assess spontaneity at different conditions may be of relevance.

Further study of the VLE phase behaviour and identification of azeotropes between methanol and DMAE would allow for further optimisation of vacuum distillation for high purity CFADE.

6. Outlook

DMAE is still primarily produced from ethylene oxide, a petroleum derivate. A recent paper proposed a method to synthesis DMAE from glycolaldehyde using the reductive amination reaction (Favaree et al.,2021). The method could be investigated for future work to produce reactant-grade DMAE. If successful, it will improve the 'greeness' of the biosurfactant.

7. References

Baumann, H., Bühler, M., Fochem, H., Hirsinger, F., Zoebelein, H. & Falbe, J. (1988) Natural Fats and Oils-Renewable Raw Materials for the Chemical Industry. *Angewandte Chemie (International Ed.)*. 27 (1), 41-62. 10.1002/anie.198800411.

Belousov, A. S., Esipovich, A. L., Kanakov, E. A. & Otopkova, K. V. (2021) Recent advances in sustainable production and catalytic transformations of fatty acid methyl esters. *Sustainable Energy & Fuels*. 5 (18), 4512-4545. 10.1039/d1se00830g.

Bialek, A., Bialek, M., Jelinska, M. & Tokarz, A. (2016) Fatty acid profile of new promising unconventional plant oils for cosmetic use. *International Journal of Cosmetic Science*. 38 (4), 382-388. 10.1111/ics.12301.

Boateng, L., Ansong, R., Owusu, W.B., Steiner-Asiedu M. (2016) Coconut oil and palm oil's role in nutrition, health and national development: A review. Ghana Med J.;50(3):189-196. PMID: 27752194; PMCID: PMC5044790.

Burk, C. H., Clendennen, S. K., Boaz, N. W. (2016). Betaine esters and process for making and using. US 9487805

Chaudhary, P., Kumar, B., Kumar, S. & Gupta, V. K. (2015) Transesterification of Castor Oil with Methanol – Kinetic Modelling. *Chemical Product and Process Modeling*. 10 (2), 71-80. 10.1515/cppm-2014-0032.

Clendennen, S. K. & Boaz, N. W. (2019) Chapter 14 - Betaine Amphoteric Surfactants—Synthesis, Properties, and Applications. In: Hayes, D. G., Solaiman, D. K. Y. & Ashby, R. D. (eds.). *Biobased Surfactants (Second Edition)*. , AOCS Press. pp. 447-469.

Erhan, S. Z. (2005) *Industrial Uses of Vegetable Oil*. AOCS Publishing.

Faveere, W. H., Van Praet, S., Vermeeren, B., Dumoleijn, K. N. R., Moonen, K., Taarning, E., B. F. Sels, B., F. (2021) *Angew. Chem. Int. Ed. 60*, 12204.

Faveere, W. H., Van Praet, S., Vermeeren, B., Dumoleijn, K. N. R., Moonen, K., Taarning, E. & Sels, B. F. (2021) Toward Replacing Ethylene Oxide in a Sustainable World: Glycolaldehyde as a Bio-Based C2 Platform Molecule. *Angewandte Chemie International Edition*. 60 (22), 12204-12223. 10.1002/anie.202009811.

Floyd, D. J. & Jurczyk, M. (2008) Amphoteric Surfactants: Synthesis and Production, In: Zoller, U. (ed.) *Handbook of Detergents, Part F* surfactant science series 142, United States, CRC Press, pp. 231-232.

Foti, C., Bonamonte, D., Mascolo, G., Corcelli, A., Lobasso, S., Rigano, L. & Angelini, G. (2003) The role of 3-dimethylaminopropylamine and amidoamine in contact allergy to cocamidopropylbetaine. *Contact Dermatitis*. 48 (4), 194-198. 10.1034/j.1600-0536.2003.00078.x.

García, M. T., Campos, E. & Ribosa, I. (2007) Biodegradability and ecotoxicity of amine oxide based surfactants. *Chemosphere (Oxford)*. 69 (10), 1574-1578. 10.1016/j.chemosphere.2007.05.089.

Hayes, D. G. & Smith, G. A. (2019) Chapter 1 - Biobased Surfactants: Overview and Industrial State of the Art. In: Hayes, D. G., Solaiman, D. K. Y. & Ashby, R. D. (eds.). *Biobased Surfactants (Second Edition)*. , AOCS Press. pp. 3-38.

Herrwerth, S., Leidreiter, H., Wenk, H. H., Farwick, M., Ulrich-Brehm, I. & Grüning, B. (2008) Highly Concentrated Cocamidopropyl Betaine – The Latest Developments for Improved Sustainability and Enhanced Skin Care. *Tenside Surfactants Detergents*. 45 (6), 304-308. 10.3139/113.100387.

Hellberg, P.-E., Bergström, K. Holmberg, K. (2000), Cleavable surfactants. J Surfact Deterg, 3: 81-91.

Kelleppan, V. T., King, J. P., Butler, C. S. G., Williams, A. P., Tuck, K. L. & Tabor, R. F. (2021) Heads or tails? The synthesis, self-assembly, properties and uses of betaine and betaine-like surfactants. *Advances in Colloid and Interface Science*. 297 102528. 10.1016/j.cis.2021.102528.

Kishore, K. (2017) *Partial Glycerides -An Important Nonionic Surfactant for Industrial Applications: An Overview Nonionic surfactants View project Lead-free pervoskite materials for Capacitor application View project.*

Krist, S. (2020) Introduction. In: Anonymous *Vegetable Fats and Oils*. Cham, Springer International Publishing. pp. 1-26.

Levison, M. I. (2008) Surfactant Production: Present Realities and Future Perspectives. *Handbook of Detergents, Part F* surfactant science series 142, United States, CRC Press, pp. 13-14.

Li, Q., Zhang, J., Zhang, G., & Xu, B. (2022). l-Lysine-Based Gelators for the Formation of Oleogels in Four Vegetable Oils. *Molecules (Basel, Switzerland)*, *27*(4), 1369. https://doi.org/10.3390/molecules27041369

Nagumalli, S. K., Jacob, C. C. & Gamboa da Costa, G. (2020) A rapid and highly sensitive UPLC-ESI-MS/MS method for the analysis of the fatty acid profile of edible vegetable oils. *Journal of Chromatography. B, Analytical Technologies in the Biomedical and Life Sciences.* 1161 122415. 10.1016/j.jchromb.2020.122415.

Noureddini, H., Zhu, D. (1997) Kinetics of transesterification of soybean oil. *J Amer Oil Chem Soc* **74**, 1457–1463.

Panda, A., Kumar, A., Mishra, S. & Mohapatra, S. S. (2020) Soapnut: A replacement of synthetic surfactant for cosmetic and biomedical applications. *Sustainable Chemistry and Pharmacy*. 17 100297. 10.1016/j.scp.2020.100297.

R. C, S., Kipkemboi, P. K. & Rop, K. (2020) Synthesis, Characterization, and Evaluation of Solution Properties of Sesame Fatty Methyl Ester Sulfonate Surfactant. *ACS Omega*. 5 (44), 28643-28655. 10.1021/acsomega.0c03698.

Saxena, N., Pal, N., Ojha, K., Dey, S. & Mandal, A. (2018) Synthesis, characterization, physical and thermodynamic properties of a novel anionic surfactant derived from Sapindus laurifolius. *RSC Advances*. 8 (43), 24485-24499. 10.1039/c8ra03888k.

Shah, J., Arslan, E., Cirucci, J., O'Brien, J., & Moss, D. (2016). Comparison of Oleo- *vs* Petro-Sourcing of Fatty Alcohols via Cradle-to-Gate Life Cycle Assessment. *Journal of surfactants and detergents*, *19*(6), 1333–1351. https://doi.org/10.1007/s11743-016-1867-y

Van Bogaert, I. N. A., Saerens, K., De Muynck, C., Develter, D., Soetaert, W. & Vandamme, E. J. (2007) Microbial production and application of sophorolipids. *Applied Microbiology and Biotechnology*. 76 (1), 23-34. 10.1007/s00253-007-0988-7.

Yea, D., Lee, Y., Park, K. & Lim, J. (2021) Synthesis of eco-friendly fatty acid based zwitterionic biosurfactants from coconut oil sources and characterization of their interfacial properties. *Journal of Industrial and Engineering Chemistry (Seoul, Korea)*. 97 287-298. 10.1016/j.jiec.2021.02.012.

Yusuff, A. S., Porwal, J., Bhonsle, A. K., Rawat, N. & Atray, N. (2021) Valorization of used cooking oil as a source of anionic surfactant fatty acid methyl ester sulfonate: process optimization and characterization studies. *Biomass Conversion and Biorefinery*. 1 10.1007/s13399-021-01663-y.

Zhou, H., Lu, H. & Liang, B. (2006) Solubility of Multicomponent Systems in the Biodiesel Production by Transesterification of Jatropha c urcas L. Oil with Methanol. *Journal of Chemical and Engineering Data*. 51 (3), 1130-1135. 10.1021/je0600294.

Wrong Way Behaviour of Packed Bed Reactors: Investigating Fischer-Tropsch and the Contact Process

Ritik Attwal and Harkaran Dhaliwal
Department of Chemical Engineering, Imperial College London, U.K.

Abstract Wrong way behaviour is a curious phenomenon which is best described as the transient temperature increase in a reactor caused by a decrease in the feed temperature or an increase in the feed velocity. It has been modelled with various complexities but all models in literature assume constant velocity and first order kinetics; models in literature fail to develop system parameters which are representative of industrial processes. This work develops a two-phase model with variable velocity, generalised to n^{th} order kinetics and applied to two industrially relevant processes – Fischer-Tropsch and the Contact Process. The model was implemented in gPROMS. Base system parameters were developed for the aforementioned processes to investigate the presence of wrong way behaviour and it was not observed for either process at industrial conditions. A sensitivity analysis was performed, and it was determined that for Fischer-Tropsch the likelihood of wrong way behaviour increases with larger temperature rises across the reactor. For the Contact Process, the suppression of wrong way behaviour is attributed to highly effective dispersion. Future scope is wide and could include accounting for radial effects, pressure drop, rigorous modelling of kinetics and temperature-pressure dependence of thermophysical properties.

Keywords: *wrong way behaviour, n^{th} order kinetics, variable velocity, Fischer-Tropsch, Contact Process, gPROMS*

1. Introduction

Wrong way behaviour (WWB) is a curious phenomenon which is best described as the transient temperature increase in a reactor caused by a decrease in the feed temperature. Intuitively, one would expect the temperature profile for a reactor to decrease with a decrease in feed temperature for all points along the reactor. Instead, as a result of a reduction in feed temperature, the reaction rate in the upstream section of the reactor decreases and causes a higher concentration of reactants to reach still-hot catalyst in the downstream section. This leads to a faster reaction rate in the downstream section. Consequently, additional heat is generated downstream leading to a new higher temperature peak compared to the initial steady-state profile. In accordance with the properties of the system, a cooler temperature wave reaches the downstream section and forces the system to settle to a new, cooler steady state - marking the end of the transient behaviour. It should be noted an increase in the (superficial) feed gas velocity can also produce WWB and follows the same reasoning as previously explained. It is of industrial relevance to investigate this behaviour as systems usually operate relatively close to reaction runaway temperatures due to economic pressures, hence the transient temperature peak may be significant enough to enable runaway. The induced transient temperature rise may also damage the catalyst or the reactor. Furthermore, fluctuations in feed velocity/temperature are not unusual on a plant due to a multitude of reasons ranging from control system sensitivities, start-up procedures or human input, thus this investigation is necessary.

The concept of WWB was first predicted by Crider and Foss (1966) [1], who observed a transient drop in temperature following a step increase in feed temperature whilst modelling packed bed tubular reactors. It was observed by various other researchers over time; Sharma and Hughes (1979) [2] observed this behaviour whilst exploring fixed bed reactor systems for the oxidation of carbon monoxide. As a result of these observations, further investigation into the effect was undertaken; Table 1 presents selected studies in chronological order and displays the type of model used, observations made and limitations of the studies, thereby providing an overview of the developments in the modelling/understanding of WWB.

Table 1: *A summary of observations and limitations of wrong way behaviour modelling that can be found in literature.*

Study & Model	Key Observations	Limitations
Mehta et. Al (1981) [3] One-Dimensional-Single Phase-Pseudo-Homogeneous	The analysis indicated that high transient temperatures develop only in sufficiently long reactors, thus, the impact of the wrong way behaviour is not encountered in short reactors.	Dispersion effects of heat and mass were neglected. Interfacial effects between gas and solid catalyst were neglected. Predicts erroneously that temperature discontinuities may exist in the bed. Assumption of constant velocity.
Pinjala et. Al (1988) [4] One-Dimensional-Single Phase-Pseudo-Homogeneous- That accounts for the axial dispersion of heat and species	Study showed dispersion decreases the magnitude of wrong way whilst also prolonging the shift to a new steady state.	Interfacial effects between gas and solid catalyst were neglected. Radial effects were neglected. Assumption of constant velocity.

Chen and Luss (1989) [5] One-Dimensional-Two Phase-Heterogeneous	Confirmed the predictions of Pinjala et al. (1988) [4]; in reactors with a high conversion rate, temperature excursions are negligible due to a decrease in feed temperature, but the excursions are more prevalent in reactors with intermediate conversion rates. Also observed the similarity between the two-phase model and pseudo-homogenous model when all feed temperatures have a unique steady state.	Assumption of constant velocity. Radial effects were neglected. Two-phase model could not predict the backward migration of the temperature wave.
Ganesan and Khaitan (2020) [6] One-Dimensional-Two Phase-Heterogeneous	Considered variable velocity and demonstrated that velocity contraction due to reaction dominates velocity expansion due to temperature. Therefore, models with constant velocity are more likely to exhibit wrong way behaviour.	Assumption of constant feed velocity for changes in feed temperature exaggerated effects of variable velocity. More robust methods could be used to obtain system parameters. Investigation was limited to a single industrially relevant process.

Variable velocity is not considered in older studies; Ganesan and Khaitan (2020) [6] emphasise the importance of taking it into consideration. There is a problem with assuming constant velocity since temperature changes and gas contractions/expansions would affect gas velocity competingly. Due to the density-temperature relationship, temperature changes impact gas velocity and this effect is prevalent in exothermic reactions which are characterised by an increase in temperature along the reactor; because of this the gas expands resulting in an increase in velocity. Gas contraction/expansion occurs as a result of the change in the number of moles from reactants to products as reaction progresses.

Considering the importance of modelling variable velocity when investigating WWB, this study builds upon the work of Ganesan and Khaitan (2020) [6] and improves on it by presenting a model which is generalised to nth order reactions and validates whether system parameters are truly representative of the process under investigation. Furthermore, this study investigates WWB in the following cases: Fischer-Tropsch (FT) and the Contact Process (CP). These two cases were specifically chosen, firstly, due to their industrial relevance but more significantly due to their comparatively differing nature as highlighted in Table 2. Furthermore, the criterion of intermediate conversion for WWB as suggested by Chen and Luss (1989) [5] is satisfied by these reactions. It is important to note that a single step of the Contact Process is modelled in this study - the oxidation of sulphur dioxide. Additionally, modelling FT in this study serves as a validation of the study conducted by Ganesan and Khaitan (2020) [6].

Table 2: *A summary of key features of Fischer-Tropsch and the oxidation of sulphur dioxide in the Contact Process which are relevant to this study.*

Feature	Fischer-Tropsch	Contact Process
Reaction	$nCO + (2n+1)H_2 \rightarrow C_nH_{2n+2} + nH_2O$	$2SO_2 + O_2 \leftrightarrow 2SO_3$
Mode of Operation	Wall-cooled	Adiabatic
Degree of Gas Contraction	Significant	Negligible
Type of Reaction	Gas-to-Liquid (GTL)	Gas Phase
Reaction Order	1st w.r.t H_2 [7]	0.5th w.r.t SO_2 [8]
Type of Reaction	Exothermic	Exothermic

The objectives of this study for each case (FT and CP) are to:

1. Develop a set of base system parameters which represent an industrial reactor
2. Investigate if WWB is present using the set of base system parameters
3. If WWB is not present, investigate when it could occur through sensitivity analysis of system parameters

2. Methods

2.1. Modelling Approach

In order to investigate the presence of WWB in each of the case studies (FT and CP), a system of mass and energy balances were derived. These balances represent a bed of radius r and length L where feed enters with a velocity $u_{f,0}$ and temperature T_f. The bed is composed of void space (through which the reaction fluid flows in the axial z direction) and spherical catalyst particles with a radius r_p as shown in Figure 1. The bed voidage ε, therefore, is defined as the fraction of volume in the bed unoccupied by catalyst.

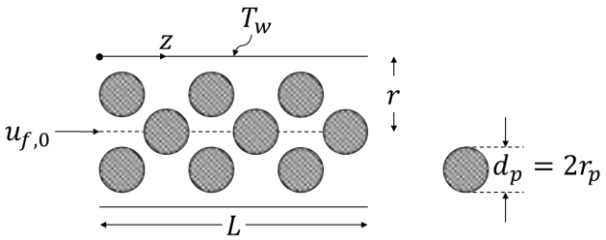


Figure 1: *Generalised representation for the scope of the model.*

In the case of FT, the system previously described represents a single tube within a multi-tubular reactor where a coolant (likely steam) flows around the tubes. In the case of CP, the system previously described represents a single (the first) stage/pass within a multi-pass reactor where the reaction occurs adiabatically (cooling occurs between stages). This is depicted in Figure 2 where the systems under investigation in this study are highlighted in green and shaded sections represent catalyst packing.

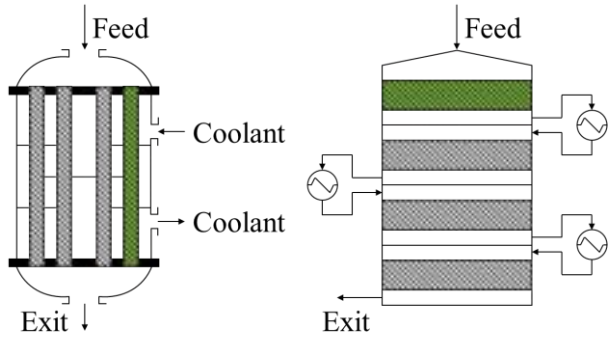


Figure 2: *Simplified diagram for a wall-cooled Fischer-Tropsch reactor (left). Simplified diagram for a multi-stage sulphur dioxide oxidation reactor in the Contact Process (right). Shaded sections represent catalyst packing.*

A series of step changes in reactor feed temperature T_f and velocity $u_{f,0}$ were conducted to test for the presence of WWB for each of the case studies. The step change ranges were chosen to best reflect likely deviations one could encounter during normal operation.

2.2. Governing equations

Presented here, in dimensionless form for an n[th] order reaction, are the governing mass and energy balances for the gas phase and solid catalyst; definitions for the corresponding boundary conditions, scaling equations and dimensionless groups follow. At bulk conditions, the dimensionless total concentration, mole fraction and temperature are denoted by C_T, y and θ respectively.

Balances and definitions:

$$\frac{\partial(u \cdot C_T)}{\partial z} = \alpha M C_T (y - y_s) \tag{1}$$

$$\frac{1}{Le}\frac{\partial(y \cdot C_T)}{\partial t} = -\frac{\partial(u \cdot y \cdot C_T)}{\partial z} - MC_T(y - y_s) + \frac{1}{Pe_M}\frac{\partial}{\partial z}\left(C_T \frac{\partial y}{\partial z}\right) \tag{2}$$

$$\frac{C_T}{Le}\frac{\partial \theta}{\partial t} = -uC_T\frac{\partial \theta}{\partial z} + \frac{1}{Pe_H}\frac{\partial^2 \theta}{\partial z^2} - H(\theta - \theta_s) - U(\theta - \theta_w) \tag{3}$$

$$MC_T(y - y_s) = \eta_s (y_s \cdot C_T)^n exp\left(\frac{1}{\theta_r} - \frac{1}{\theta}\right) \tag{4}$$

$$\left(1 - \frac{1}{Le}\right)\frac{\partial \theta_s}{\partial t} = \beta \eta_s (y_s \cdot C_T)^n exp\left(\frac{1}{\theta_r} - \frac{1}{\theta}\right) + H(\theta - \theta_s) \tag{5}$$

$$\eta_s = \frac{3}{\phi_s^2}\left(\frac{\phi_s}{tanh(\phi_s)} - 1\right) \tag{6}$$

$$\phi_s^2 = \phi_0^2\, exp\left(\frac{1}{\theta_r} - \frac{1}{\theta}\right)(y_s \cdot C_T)^{n-1} \tag{7}$$

$$X = 1 - y \tag{8}$$

Boundary conditions:

At $z = 0$:

$$u_f = uC_T$$

$$u_f(1 - y) = -\frac{C_T}{Pe_M}\frac{\partial y}{\partial z} \tag{9}$$

$$u_f(\theta_f - \theta) = -\frac{1}{Pe_H}\frac{\partial \theta}{\partial z}$$

At $z = 1$:

$$\frac{\partial \theta}{\partial z} = 0$$

$$\frac{\partial y}{\partial z} = 0 \tag{10}$$

Scaling equations:

$$C_T = \frac{C_T'}{C_{T,f,0}'} = \frac{P/RT}{P/RT_{f,0}} = \frac{\theta_f}{\theta}$$

$$\theta = \frac{RT}{E};\ \theta_r = \frac{RT_r}{E}$$

$$y = \frac{y_A'}{y_{A,f}'};\ u = \frac{u'}{u_{f,0}'} \tag{11}$$

$$z = \frac{z'}{L};\ t = \frac{1}{Le}\frac{u_{f,0}' t'}{\varepsilon L}$$

Dimensionless parameters:

$$M = \frac{3Lk_{mc}}{u_{f,0}' r_p}(1 - \varepsilon)$$

$$\alpha = -y_{A,f}'\left(\frac{a + b - c}{a}\right)$$

$$U = \frac{2Lh_w}{u_{f,0}' r C_{T,f,0}' c_{p,g}}$$

$$H = \frac{3Lh_f}{u_{f,0}' r_p C_{T,f,0}' c_{p,g}}(1 - \varepsilon) \tag{12}$$

$$\beta = \frac{R}{E}\frac{(-\Delta H) y_{A,f}'}{c_{p,g}}$$

$$Le = 1 + \frac{\rho_c c_{p,c}}{\varepsilon C_{T,f,0}' c_{p,g}}(1 - \varepsilon)$$

$$Pe_M = \frac{Lu'_{f,0}}{D_a}$$

$$Pe_H = \frac{Lu'_{f,0}C'_{T,f,0}c_{p,g}}{\lambda_a}$$

$$\phi_0^2 = \frac{u'_{f,0}(n+1)r_p^2}{2D_eL(1-\varepsilon)}$$

$$\frac{k(\theta_r)}{k_0} = exp\left(-\frac{1}{\theta_r}\right) = \frac{u'_{f,0}}{Lk_0\left(y'_{A,f}\cdot C'_{T,f,0}\right)^{n-1}}$$

Key assumptions behind the model are detailed here:

1. Radial effects are negligible
2. Rate of reaction follows n[th] order kinetics with respect to a single reactant
3. Liquid volume in FT is insignificant compared against gas volume so it can be ignored
4. Pressure drop through the bed is negligible
5. Total concentration is independent of composition and follows the ideal gas law
6. Concentration difference between bulk fluid and particle surface can be ignored
7. $c_{p,c}$, $c_{p,g}$, D_a, h_f, h_w, k_{mc}, λ_a, ρ_c and viscosity are constant in θ, P, z and t

To preserve the general applicability of the model developed in this study, simplified kinetics are considered for both cases (FT and CP) where the rate expression depends exclusively on one reactant. A further simplification is made in treating the reversible oxidation of sulphur dioxide in CP as irreversible. A simple calculation for the equilibrium constant (and conversion) proves that, for the conditions considered in this study, there is an overwhelmingly strong bias towards the formation of sulphur trioxide. Care was taken to ensure that the parameters used to calculate dimensionless parameters were consistent for a given type of catalyst. In the case of FT, a cobalt-based catalyst is considered as this enables the use of simple kinetics for a low-temperature FT process [7]. In the case of CP, a vanadium-based catalyst is considered as it is most used in industry.

2.3. Modelling strategies

The system of equations defined above was solved numerically in gPROMS using a backwards finite difference solver as a discretised and distributed system. To ensure representative initial guesses were used, an initialisation procedure was employed which solves an easy-to-solve version of the model at steady state.

A systematic approach was required to ensure that the feed temperature was not so high as to trigger thermal runaway - a convenient mathematical characteristic of thermal runaway was exploited to achieve this. An indication for the onset of runaway is where the first and second derivative of gas temperature (with respect to reactor position) exceed zero – where the profile displays a point of inflection pointing upwards. This was monitored in gPROMS and, for a given case, the feed temperature that triggers runaway was determined. Models were run to investigate WWB with a feed temperature 10 K below that which triggers runaway.

An example runaway profile is shown for FT in Figure 3. The profile shows a characteristic inflection point indicating runaway and displays a drop in temperature after the conversion reaches 100% (and the reactant has been completely consumed).

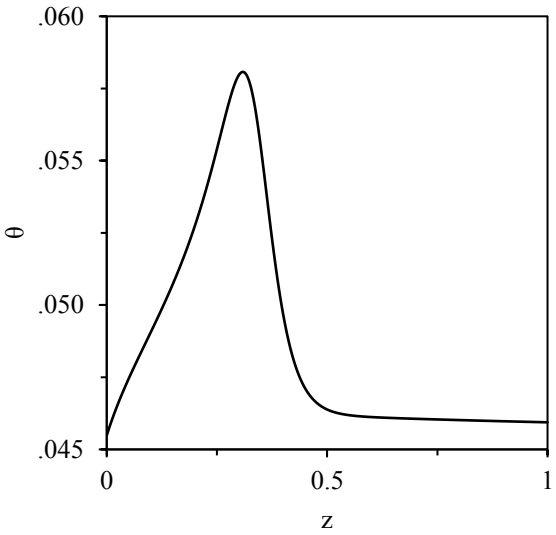


Figure 3: *Thermal runaway profile for Fischer-Tropsch which displays a characteristic point of inflection in θ with respect to z.*

Table 3 details the values for the base (dimensionless) system parameters. The dimensional quantities used to calculate the dimensionless parameters can be found in the Supplementary Information.

Table 3: *A summary of dimensionless base system parameters.*

Parameter	FT	CP
Le	179	745
M	15	33
$\propto$	-0.311	-0.055
Pe_M	1582	48
Pe_H	1597	3
H	133	26
U	31	0
ϕ_0	5.590	2.008
β	0.125	0.025
θ_r	0.047	0.079

3. Results and discussion

3.1. Model validation

To ensure model system parameters are representative of their respective case studies, modelled conversion and temperature rise across the bed are compared against literature in Table 4 and Table 5, respectively. It should be

noted that conversion comparisons are for single pass conversions and in the case of CP the conversion quoted is for a single (the first) stage of the multi-stage reactor only as this is what the model is valid across.

Table 4: *Conversion and reactor temperature rise predicted by model.*

Case	Conversion [%]	Temperature Rise [K]
FT	25	14
CP	45	90

Table 5: *Conversion and reactor temperature rise from literature.*

Case	Conversion [%]	Temperature Rise [K]
FT	18 – 89 [7]	12 [9]
CP	50 [10] – 70 [13]	130 [10] – 180 [11]

It is evident that the system parameters for FT capture the conversion and temperature rise one would expect from a reactor in service in industry. The system parameters for CP yield a conversion and temperature rise which are in fair agreement with literature.

It is important to note that the model presented in this study assumes constant diffusivity, thermal conductivity and total concentration; the temperature dependence of these three parameters was neglected. To ensure using this simplified approach was justified, a complex model which included temperature dependence was investigated using gPROMS. No major discrepancies were noticed between the complex and simplified models in the surface: rate, temperature and mole fraction with values agreeing within a maximum range of ±0.5%.

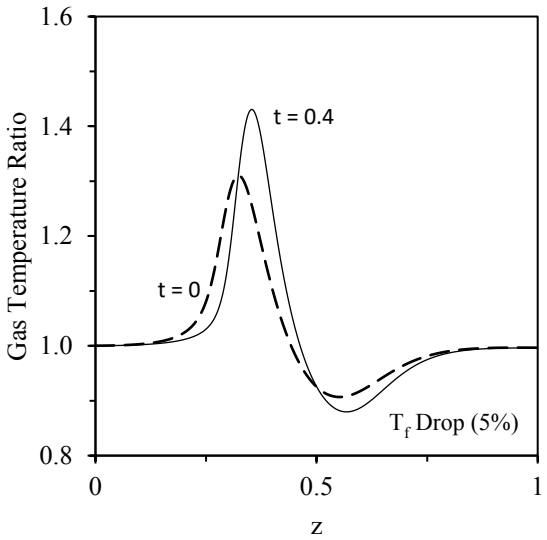


Figure 4: *Gas temperature ratio for Fischer-Tropsch between a model that assumes constant velocity to a model that uses variable velocity.*

Figure 4 demonstrates the importance of accounting for variable velocity. It displays a FT process displaying WWB where it undergoes a transient increase in bulk gas temperature following a reduction in feed temperature. This is evidenced in the relative difference between the two curves. The ratio of gas temperature considering constant velocity to gas temperature considering variable velocity is plotted against z (dimensionless position along the reactor). A value of greater than 1 on the y axis means the gas temperature is overpredicted when considering constant velocity. This is exactly what is observed on Figure 4 where a significant (maximum 40%) overprediction can be seen. Therefore, it is apparent that using a constant velocity model exaggerates the magnitude of WWB observed.

3.2. Fischer-Tropsch

3.2.1. Base system parameters analysis

Figure 5 shows the trend between θ (bulk gas temperature) and z for a T_f (feed temperature) drop of 1%. The trend shows the progression of transient behaviour over time where the broken curves show initial ($t = 0$) and final ($t = 0.16$) steady states. Unbroken curves in between the steady state profiles are for intermediate time intervals. It is important to note that t denotes dimensionless time and for FT, 1 dimensionless time unit is approximately equal to 1600 seconds. It is evident from Figure 5 that for a step change of 1%, there is no WWB. This is because the temperature at any given position along the reactor, for all times, is below that of the $t = 0$ trend. The behaviour exhibited is what one would intuitively expect. Larger T_f drops also do not result in WWB and instead progressively lead to the reaction becoming quenched and the temperature profile flattening out. No WWB was observed for FT for any magnitude of T_f drop using system parameters representative of an industrial reactor.

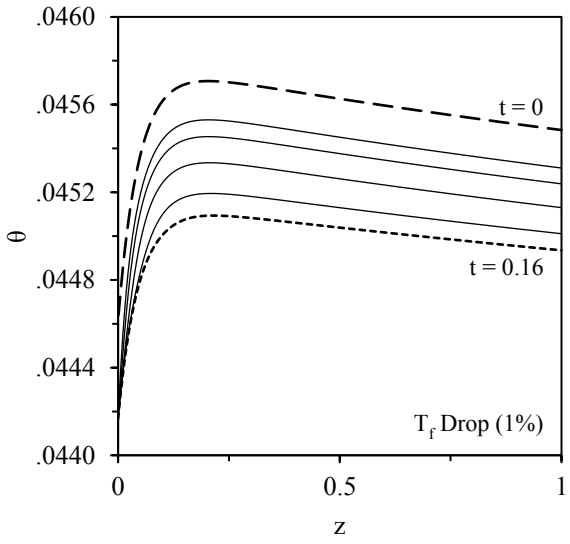


Figure 5: *Response of dimensionless bulk gas temperature to a feed temperature drop of 1% for Fischer-Tropsch.*

In the case of increasing u_f (feed gas velocity) for FT, apparent WWB was observed. Figure 6 displays the magnitude of the effect by displaying the difference between the maximum transient temperature and maximum initial steady state temperature over non dimensionless time.

This difference can be thought of as a proxy for the presence/tendency of WWB where positive values are indicative for the presence of WWB. The effect is not a cause for any significant concern as the largest occurrence was only 0.13K for a 100% increase in u_f (which is an extremely severe increase in u_f).

This is corroborated by Figure 13 (found in the Supplementary Information) which displays θ_s (the catalyst surface temperature) as a ratio of the θ; as this is always close to unity this shows the catalyst is also not significantly heated. Hence there is practically no risk of runaway or catalyst damage.

It is important to note that although Figure 6 shows a characteristic of WWB (as there is an increase in peak temperature following an increase in u_f) it is not strictly WWB. This is because the temperature increase is not transient, and this is evidenced in Figure 6. There still is a positive temperature difference even after the system has settled at the new steady state.

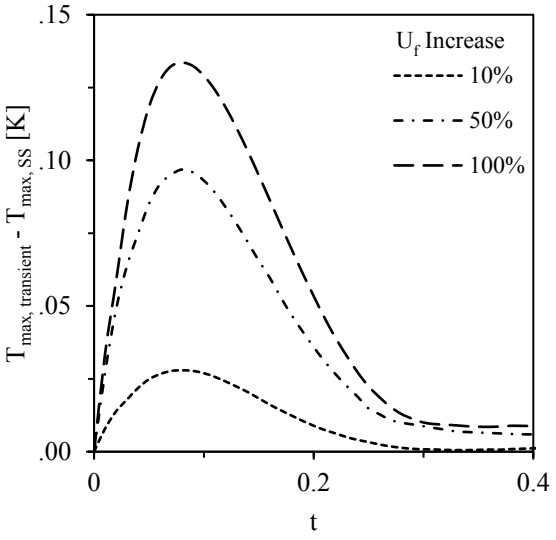


Figure 6: *Wrong way behaviour tendency as a function of dimensionless time for various feed velocity increases for Fischer-Tropsch.*

More curiously, the apparent WWB (albeit at a small scale) is still observed at significantly larger increases in u_f as shown in Figure 6. Intuitively, one would expect the temperature to fall and converge in accordance with the rate of reaction; instead, the behaviour is more prevalent. This is explained by Figure 7 which shows the relationship between an increase in u_f and X (the conversion). It is shown that even after large increases in u_f, the decrease in X is comparatively insignificant. For instance, a 50% increase in u_f leads to only a 7% decrease in X whereas this can be achieved by a contrastingly low 5% drop in T_f. One reason for this lack of dependency, is that the FT process has a product distribution, therefore instead of the decreasing residence time sharply impacting conversion a compromise is made with selectivity. This means that instead of the desired long-chain hydrocarbons being produced, shorter chain hydrocarbons are produced with increasing velocity. This would also explain why Figure 7 plateaus at around $X = 0.1$; as at this point the residence time is so low, that it is probable only methane and other short-chain hydrocarbons are being produced.

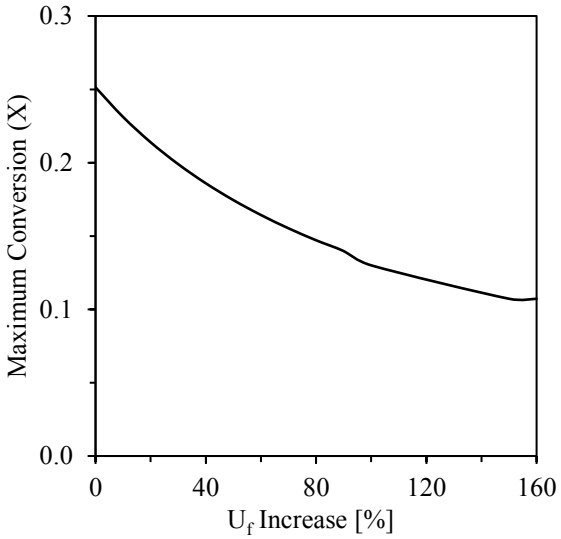


Figure 7: *Maximum conversion as a function of feed velocity for Fischer-Tropsch.*

As a result of this, it can be concluded that changes in feed velocity do not have a significant impact on the FT system. This is most prevalent for feed velocity change ranges which are representative of typical operating variations. Consequently, the apparent WWB can be explained by the fact that a higher velocity feed has a marginally higher spatial average temperature. This would lead to a higher bulk temperature and would explain why the magnitude of the apparent WWB seen was so small, why the change is not transient, and why it increased with increasing velocity.

Evidently using system parameters which are representative of an industrial reactor yields no WWB in the case of FT. This can be attributed to the effect of wall cooling. Although bulk concentration of reactants did transiently increase (following moderate temperature drops which were representative of typical operating variations), this increase in bulk concentration of reactants was identified by a transient increase in the ratio of bulk concentration to catalyst surface concentration (as seen in Figure 14 in the Supplementary Information). Therefore, there was potential for WWB to occur in the reactor. However, the hotspot of the reactor is unable to effectively utilise this increase in concentration to locally increase the rate of reaction and hence induce a transient temperature increase (and thus facilitate WWB). This is attributed to the low 14K peak temperature rise in the reactor. Hence, overall due to the coolant effectively supressing large deviations in temperature no WWB is exhibited in the base case of FT. It is probable a process with a higher temperature rise would be more susceptible to exhibit WWB. Furthermore, in the case of velocity increases WWB is even less likely to occur compared to

temperature drops for FT. This is because the bulk concentration of reactants insignificantly increased following moderate velocity increases which were representative of typical operating variations. The reason for this insignificant change was due to the lack of dependency of reaction rate on velocity increase. Contrastingly, temperature changes have a strong effect on the reaction rate. This is attributed to the Arrhenius relationship between temperature and reaction rate, hence why bulk concentration increases were more prevalent for temperature drops. This dependency is even more profound due to the use of first order kinetics. Therefore, both the low peak temperature rise and negligible increase in the bulk concentration effectively suppress the potential of WWB to occur following velocity increases for FT.

3.2.2. Sensitivity analysis

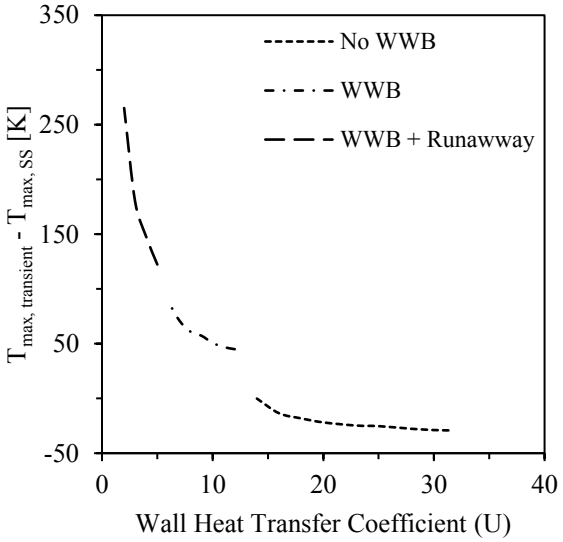


Figure 8: *Wrong way behaviour tendency as a function of dimensionless wall heat transfer coefficient for Fischer-Tropsch.*

The postulation stipulated in the previous section is supported by Figure 8. The figure is split into three distinct regions: No WWB, WWB and WWB with runaway. It is evident that there is a strong inverse relationship between U (the dimensionless wall heat transfer coefficient) and the magnitude/presence of WWB. This can be explained by the reasoning that a lower U value is accompanied by a larger temperature rise in the reactor. This is because as U is decreased the cooling system is unable to cool the heat generated by the reaction to the same degree, leading to a higher temperature peak. At values of U around 13, which corresponds to a temperature rise of around 89K, the hotspot of the reactor was able to effectively utilise the temperature rise to locally increase the rate of reaction. Hence, a transient temperature increase is induced (and thus WWB was observed). As U is decreased even further the temperature rise increases comparably, eventually this temperature rise is significant enough to increase the local rate of reaction at the hotspot rapidly. This leads to a considerable amount of additional heat being released, which quickly increases the rate of reaction even further. This interdependent cycle leads to the system entering a runaway state, with temperature only decreasing once the reaction undergoes completion.

It is important to note that a 5% drop in T_f was employed to investigate this relationship. This was specifically employed to simulate a plausible action an operator would take to compensate for a large drop in U. During the analysis of base FT conditions, it was found that a 5% feed temperature drop can effectively quench the reaction, thus justifying its use as an appropriate response to a drop in U. Furthermore, greater sudden adjustments to operating conditions may not be feasible.

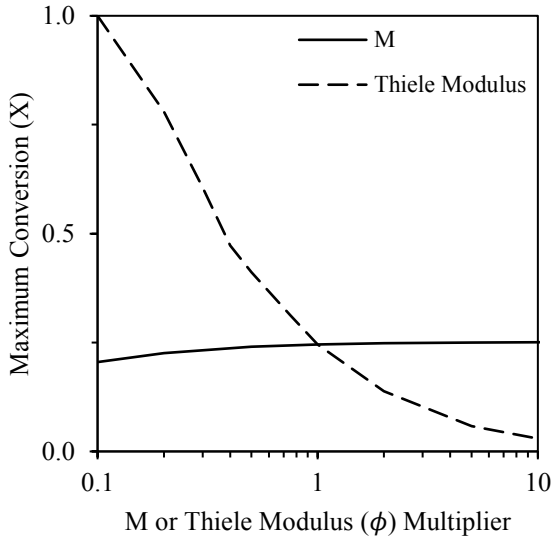


Figure 9: *Sensitivity of maximum conversion with respect to interfacial mass transfer coefficient and the Thiele modulus for Fischer-Tropsch.*

Figure 9 shows the effect of changing M (dimensionless interfacial mass transfer coefficient) and ϕ (Thiele modulus) on the maximum conversion. The changes are made independently with all other parameters equal to the base system parameters. The effect of changing a parameter by a factor equal to a multiplier (while keeping all other parameter values equal to their base parameter values) gives one an idea of the effect of a parameter (in isolation) on WWB. The aim of this graph is to determine if M (film diffusion) or ϕ (pore diffusion) is mass transfer limiting. Maximum conversion is being used as a proxy for mass transport effectiveness as mass transfer limitations hinder conversion. A large value of M represents small film diffusion limitations whereas a small value of ϕ represents small pore diffusion limitations (as the isothermal effectiveness factor tends to unity for smaller values of ϕ). Figure 9 shows maximum conversion decreases strongly with increasing ϕ and remains practically constant for changes in M. This suggests that, for FT, pore diffusion is mass transfer limiting. This is consistent with literature [7] and intuitively makes sense. This is because liquid (paraffinic) hydrocarbon product is present in catalyst pores which would pose a high degree of mass transfer resistance. When compared to a gas, it is not unreasonable to expect a liquid to linger in catalyst

pores for longer and occupy a given pore volume for a greater amount of time. This would drive a lower time-averaged catalyst porosity and in turn hinder pore diffusion.

3.3. Contact Process

3.3.1. Base system parameters analysis

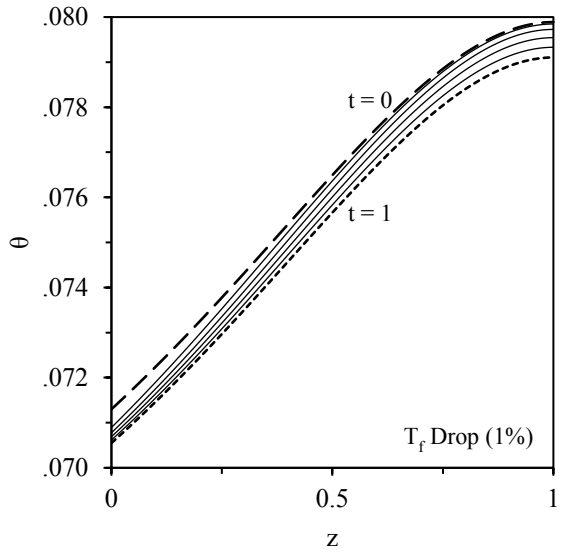


Figure 10: *Response of dimensionless bulk gas temperature to a feed temperature drop of 1% for the Contact Process.*

It is evident from Figure 10 that there is no WWB for a T_f drop of 1% for CP. As with FT, increasing magnitudes of drops in T_f result in progressive quenching of the reaction and no change in T_f yielded WWB for CP using system parameters representative of an industrial reactor. For reference, 1 dimensionless time unit for CP is approximately equal to 800 seconds.

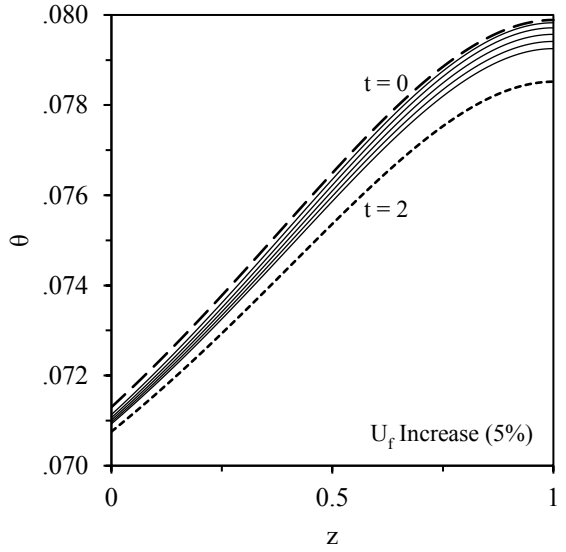


Figure 11: *Response of dimensionless bulk gas temperature to a feed velocity increase of 5% for the Contact Process.*

Figure 11 shows no WWB for a U_f increase of 5%. As with drops in T_f, no WWB is observed for CP for any magnitude of U_f increase and as these step changes increase, the reaction is progressively quenched.

Using a set of base system parameters which are representative of an industrial reactor yields no WWB in the case of CP. This could be attributed to a high degree of dispersion in the reactor. Dispersion manifests itself (in the model) through the Peclet numbers Pe_M and Pe_H – which in the case of CP are small. Small Peclet numbers translate to a high degree of dispersion. With increasing dispersion, a reactor can be thought to tend towards the behaviour of a continuously stirred tank reactor where the contents of the reactor experience a high degree of mixing and rapidly reach a state of homogeneity. What this means in the context of WWB is that, should a concentration imbalance form due to a temperature drop, it would be rapidly dissipated; a significant excess of reactant would not find its way to a hotspot. If a higher concentration of reactants manages to reach still-hot catalyst, the excess heat generated would be rapidly dissipated. Cumulatively, this would suppress the expression of WWB.

A curious behaviour can be observed with the trends above where the gas temperature does not abruptly change to the same value of a new feed temperature. Instead, the gas temperature tends toward the new feed temperature with time. This suggests that there is a resistance to change in the gas temperature at the very start of the reactor. This could also be attributed to dispersion.

Physically, the CP reactor is very short - it has an aspect ratio of less than 0.06. The lack of WWB is consistent with the studies conducted by Mehta et. Al (1981) [3] and Pinjala et. Al (1988) [4]. It is not unreasonable to consider the short reactor as effective in dispersion axially. The same cannot be said about dispersion radially as the diameter is very large for CP.

3.3.2. Sensitivity analysis

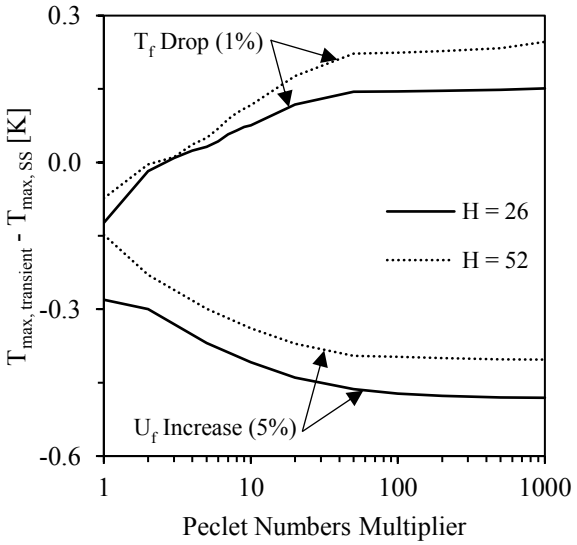


Figure 12: *Effect of dispersion and dimensionless interfacial heat transfer coefficient on the tendency for wrong way behaviour in the Contact Process.*

Figure 12 shows the influence of changing Peclet numbers on the tendency of WWB. The base Peclet number for mass and heat are both (simultaneously) multiplied by the same multipliers and all other parameters are left equal to their base parameter values. Figure 12 shows an increased tendency for WWB with increasing Peclet numbers when the feed temperature is dropped. The opposite is true for feed velocity increases. Physically, what this means is that with decreasing dispersion, the tendency for WWB increases for feed temperature drops and decreases for feed velocity increases. This supports the hypothesis put forward previously as to why no WWB is observed upon a feed temperature drop. A physical reasoning for the decreased tendency of WWB with a feed velocity increase could be the fact that turbulence increases with velocity. An increased turbulence would facilitate mixing, therefore, has an opposing effect to the decreased dispersion that comes with higher Peclet numbers.

The solid line trends (which have a value for H equal to the base value) eventually plateau with increasing Peclet numbers which suggests that some other factor becomes limiting in the expression of WWB when dispersion is sufficiently poor in mitigating a concentration imbalance and dissipating heat. This other limiting factor is shown to be H in Figure 12. The broken line trends show that, with an increase in the value of H by a factor of 2, the tendency toward WWB increases (shown by an increase in how positive the value of the y axis becomes versus the solid line trend). The physical effect of increasing H is better heat transfer between catalyst and bulk fluid. Therefore, the trend suggests that when dispersion is poor (for large Peclet numbers) an increased heat transfer between catalyst and bulk fluid is required to increase the fluid temperature overall and show WWB. This conclusion comes with the assumption that the maximum steady state temperature is constant between the solid and broken line trends – which was what was observed.

An analysis for M and ϕ was conducted for CP in the same fashion as FT. CP exhibited sensitivity toward both parameters but more so towards ϕ (but it is not severely pore diffusion limited like FT). An analysis of the tendency for WWB with respect to H (dimensionless interfacial heat transfer coefficient) was also conducted and found that increasing values of H tend towards WWB but the tendency plateaus and never quite achieves WWB. The physical effect of increasing H is better heat transfer between catalyst and bulk fluid, so the trend above suggests this is no longer limiting with large enough H. It is likely that the reaction rate becomes limiting at this point and the maximum amount of heat that is being generated at the catalyst surface is transferred to the fluid but this is not enough to result in WWB. Further details can be found in the Supplementary Information in Figure 15 and Figure 16, respectively.

The degree of WWB observed for CP is practically insignificant as it corresponds to a temperature difference (between transient and steady state) in the order of less than 0.3K. Nonetheless, these trends warrant discussion.

4. Conclusions

This study presents a dynamic two-phase, one dimensional reactor model generalised for nth order kinetics which accounts for variable velocity. This was successfully implemented in gPROMS for two industrially relevant processes/cases which differ greatly in nature – the wall-cooled Fischer-Tropsch process and the adiabatically operated Contact Process. A set of base system parameters are presented which successfully represent an industrial reactor for each case. The presence of WWB is investigated in this study for each case and it is evident that there is no WWB for both cases using models which are based on industrial parameters. This is an assuring finding as this study implies that the problems that can result from WWB are likely not present for these industrially significant processes. A sensitivity analysis was also successfully performed to determine what sort of deviations from industrially representative parameters could result in WWB. The magnitude of deviations in system base parameters which cause WWB, are so large that they are physically unlikely and lead to an insignificant degree of WWB.

For FT (which is a wall-cooled reactor), this study shows that the likelihood of WWB increases with larger temperature rises across the reactor which can be caused by a poor wall heat transfer coefficient/insufficient cooling. Also, the impact of feed velocity increases is weak in causing WWB or generally affecting the gas temperature. For CP, this study attributes the suppression of WWB to highly effective dispersion in the reactor which manifests itself in the small Peclet numbers.

Whilst the model provided physically explainable results, one should be aware of the limitations/assumptions behind the model. There is scope for lifting these assumptions to improve how well these models represent their respective systems. For example, in the case of CP the assumption of negligible radial effects is likely unsuitable as the reactor diameter is large so this could warrant developing a two-dimensional model. Other complexities that can be introduced include: more rigorous modelling of kinetics (rather than simple n^{th} order with respect to a single reactant), including reactor pressure drop and accounting for temperature-pressure dependance of thermophysical properties. Nonetheless, the model could be easily applied as-is to other processes to investigate WWB.

Nomenclature

ΔH	heat of reaction [J mol^{-1}]
a, b, c	stoichiometric coefficients
C_T	dimensionless total concentration
C	dimensionless bulk concentration, $C = C_T \cdot y$
C_s	dimensionless surface concentration
C_T'	total concentration [mol m_{bed}^{-3}]
$c_{p,g}$	specific gas/fluid heat capacity [J k^{-1} mol^{-1}]
$c_{p,c}$	specific solid catalyst heat capacity [J k^{-1} kg^{-1}]
D_a	axial mass dispersion coefficient [m_{bed}^2 s^{-1}]
D_e	effective diffusivity [$m_{surface}^2$ s^{-1}]
E	activation Energy [J mol^{-1}]
H	dimensionless interfacial heat transfer coefficient
h_f	interfacial heat transfer coefficient [W $m_{surface}^{-2}$ K^{-1}]
h_w	wall heat transfer coefficient [W m_{wall}^{-2} K^{-1}]
$k(\theta)$	rate constant at temperature θ [mol$^{(1-n)}$ $m_{bed}^{-3(n-1)}$ s^{-1}]
k_0	pre-exponential factor [mol$^{(1-n)}$ $m_{bed}^{-3(n-1)}$ s^{-1}]
k_{mc}	interfacial mass transfer coefficient [m_{bed}^3 $m_{surface}^{-2}$ s^{-1}]
L	reactor tube length [m]
Le	Lewis number
M	dimensionless interfacial mass transfer coefficient
n	reaction order
P	reactor operating pressure [bar]
Pe_H	Peclet number for heat
Pe_M	Peclet number for mass
R	ideal gas constant [J mol^{-1} K^{-1}]
r	reactor tube radius [m]
r_p	particle radius [m]
T	temperature [K]
t	dimensionless time
t'	time [s]
U	dimensionless wall heat transfer coefficient
u	dimensionless velocity
u'	velocity [m s^{-1}]
$u'_{f,0}$	reference feed velocity [m s^{-1}]
X	conversion
y	dimensionless mole fraction
y'_A	reactant mole fraction [$mol_{reactant}$ mol_{total}^{-1}]
z	dimensionless axial position
z'	axial position in reactor [m]

Greek letters

α	expansion factor; negative for contraction
β	dimensionless adiabatic temperature rise
ε	bed voidage [m_{void}^3 m_{bed}^{-3}]
η_s	effectiveness factor
θ	dimensionless temperature
λ_a	axial heat dispersion coefficient [W_{bed}^{-1} K]
ρ_c	solid catalyst density [kg m^{-3}]
ϕ_0	Thiele modulus at $\theta = \theta_r$
ϕ_s	Thiele modulus at $\theta = \theta_s$

Subscripts

c	solid catalyst
f	feed
g	gas
r	reference
s	surface
SS	steady state
w	wall

References

[1] Crider, J.E. and Foss, A.S. (1966), Computational studies of transients in packed tubular chemical reactors. *AIChE J.*, 12: 514-522. DOI: https://doi.org/10.1002/aic.690120322

[2] Sharma, C. S., & Hughes, R. (1979). The behaviour of an adiabatic fixed bed reactor for the oxidation of carbon monoxide I: General parametric studies. *Chemical Engineering Science*, 34(5), 613-624. DOI: https://doi.org/10.1016/0009-2509(79)85106-4

[3] Mehta, P.S., Sams, W.N. and Luss, D. (1981), Wrong-way behavior of packed-bed reactors: 1. The pseudo-homogeneous model. *AIChE J.*, 27: 234-246. DOI: https://doi.org/10.1002/aic.690270210

[4] Pinjala, V., Chen, Y.C. and Luss, D. (1988), Wrong-way behavior of packed-bed reactors: II. Impact of thermal dispersion. *AIChE J.*, 34: 1663-1672. DOI: https://doi.org/10.1002/aic.690341010

[5] Chen, Y.C. and Luss, D. (1989), Wrong-way behavior of packed-bed reactors: Influence of interphase transport. *AIChE J.*, 35: 1148-1156. DOI: https://doi.org/10.1002/aic.690350710

[6] Ganesan, R. and Khaitan, V. (2021) Wrong Way Behaviour of Packed Bed Reactors: Influence of Variable Velocity for a Gas-to-Liquid Reactor, *Imperial College Chemical Engineering Research*, 3: 130-139.

[7] Post, M.F.M., Van't Hoog, A.C., Minderhoud, J.K. and Sie, S.T. (1989), Diffusion limitations in fischer-tropsch catalysts. *AIChE J.*, 35: 1107-1114. DOI: https://doi.org/10.1002/aic.690350706

[8] Dunn, J. P., Koppula, P. R., G. Stenger, H. & Wachs, I. E. (1998) Oxidation of sulfur dioxide to sulfur trioxide over supported vanadia catalysts. *Applied Catalysis B: Environmental.* 19 (2), 103-117. DOI: https://doi.org/10.1016/S0926-3373(98)00060-5

[9] Hooshyar, N., Vervloet, D., Kapteijn, F., Hamersma, P. J., Mudde, R. F. & van Ommen, J. R. (2012) Intensifying the Fischer–Tropsch Synthesis by reactor structuring – A model study. *Chemical Engineering Journal*. DOI: https://doi.org/10.1016/j.cej.2012.07.105

[10] Gosiewski, K. (1993) Dynamic modelling of industrial so2 oxidation reactors part I. model of 'hot' and 'cold 'start-ups of the plant. *Chemical Engineering and Processing: Process Intensification*. 32 (2), 111-129. DOI: https://doi.org/10.1016/0255-2701(93)85021-7

[11] Kiss, A. A., Bildea, C. S. & Grievink, J. (2010) Dynamic modeling and process optimization of an industrial sulfuric acid plant. *Chemical Engineering Journal*. 158 (2), 241-249. DOI: https://doi.org/10.1016/j.cej.2010.01.023

[12] Pratt, J. W. (2012) A Fischer-Tropsch synthesis reactor model framework for liquid biofuels production. DOI: https://doi.org/10.2172/1055628.

[13] Fogler, H. S. (2006) Elements of chemical reaction engineering 4th ed., 2. printing, Pearson international. ed. edition. Upper Saddle River, NJ, Pearson Education International

[14] Rastegar, S. O. & Gu, T. (2017) Empirical correlations for axial dispersion coefficient and Peclet number in fixed-bed columns. *Journal of Chromatography A.* 1490 133-137. DOI: https://doi.org/10.1016/j.chroma.2017.02.026

[15] Li, C. & Finlayson, B. A. (1977) Heat transfer in packed beds—a reevaluation. *Chemical Engineering Science.* 32 (9), 1055-1066. DOI: https://doi.org/10.1016/0009-2509(77)80143-7

[16] Jorge, L.M.M., Jorge, R.M.M. & Giudici, R. Experimental and numerical investigation of dynamic heat transfer parameters in packed bed. *Heat Mass Transfer* **46**, 1355–1365 (2010). DOI: https://doi.org/10.1007/s00231-010-0659-6

[17] Odunsi, A. O., O'Donovan, T. S. & Reay, D. A. (2016) Temperature stabilisation in Fischer–Tropsch reactors using phase change material (PCM). *Applied Thermal Engineering.* 93 1377-1393. DOI: https://doi.org/10.1016/j.applthermaleng.2015.08.084Get

[18] NIST Standard Reference Database Number 69 NIST Chemistry WebBook. DOI: https://doi.org/10.18434/T4D303

[19] Loewert, M., Hoffmann, J., Piermartini, P., Selinsek, M., Dittmeyer, R. and Pfeifer, P. (2019), Microstructured Fischer-Tropsch Reactor Scale-up and Opportunities for Decentralized Application. *Chem. Eng. Technol.,* 42: 2202-2214. DOI: https://doi.org/10.1002/ceat.201900136

[20] Li, Z. & Gariboldi, E. (2021) Review on the temperature-dependent thermophysical properties of liquid paraffins and composite phase change materials with metallic porous structures. *Materials Today Energy.* 20 100642. DOI: https://doi.org/10.1016/j.mtener.2021.100642

[21] Levenspiel, O. (1999) *Chemical reaction engineering.* 3. ed. edition. New York

[22] Livbjerg, H. & Villadsen, J. (1972) Kinetics and effectiveness factor for SO2 oxidation on an industrial vanadium catalyst. *Chemical Engineering Science.* 27 (1), 21-38. DOI: https://doi.org/10.1016/0009-2509(72)80138-6

[23] Hong, R., Li, X., Li, H. & Yuan, W. (1997) Modeling and simulation of SO2 oxidation in a fixed-bed reactor with periodic flow reversal. *Catalysis Today.* 38 (1), 47-58. DOI: https://doi.org/10.1016/S0920-5861(97)00038-2

Manipulating the electrode-electrolyte interface for improved electrocatalytic performance for Oxygen Reduction at the cathode of Anion Exchange Membrane Fuel Cells

Bora Kuzuoglu and Fayyad Uddin
Department of Chemical Engineering, Imperial College London, U.K.

Abstract

Anion Exchange membrane fuel cells (AEMFCs) have recently received increasing attention, as they contain the ability to produce respectable power densities, overcoming the issues related with using Pt-based catalysts found in Proton Exchange membrane fuel cells (PEMFCs). While extensive research has focussed on the development of such catalysts, there is limited information available on the role of ionomers in fuel cells, in particular for AEMFCs. The role of ionomers in the catalyst layer is essential, as can significantly improve the electrocatalytic performance and stability of the catalyst itself. Currently, the ionic polymer Nafion®, synthesised via the copolymerisation of tetrafluoroethylene, shows the highest performance as this agent. The limitations associated with Nafion® include high synthetic costs, which involve the use of environmentally harmful per-fluorinated compounds, whilst also being humidity and temperature sensitive, limiting its application process. Considerable attention has been drawn to vinyl imidazolium (IM) based ionic liquids as ionomers. Although these ionomers have shown to poison Pt/C catalysts in acidic medium, there has been steady process of the study of these compounds and their use in basic conditions; and in general, they show good performance. This study covers the synthesis and characterisation of alternative ionomers based on poly(ionic liquid)s and compares their electrochemical performance to the commercially available Nafion® ionomers in alkaline conditions.

1 Introduction

The well-established and growing interest in fuel cells makes it one of the most prominent technologies on the course to meet global sustainability goals. Research on fuel cells is primarily focused on proton-exchange membrane fuel cells (PEMFC) and the anion-exchange membrane fuel cells (AEMFC), with an emphasis on PEMFC due to the availability of highly conductive and durable proton exchange membranes (PEM) namely Nafion® (Favero, Stephens, & Titirici, 2020). Both fuel cells are limited by the oxygen reduction reaction (ORR) at the cathode due to sluggish kinetics. A large amount of expensive Pt-based catalyst is used to overcome this fault. Although research has been successful in reducing Pt loading, its increasing price is still an issue from a commercial standpoint (Holdcroft, 2014).

Efforts on substituting Pt with an inexpensive alternative has shown promise with developments of transition metals/nitrogen supported on carbon. Recently, the emergence of highly conductive anion exchange membranes (AEM) has boosted the research of AEMFCs. In such conditions, the ORR proceeds much faster and thus other, cheaper catalysts can rival the activity of Pt catalysts (Gao, et al., 2017). In particular, iron/nitrogen supported on carbon (FeNC) has shown encouraging performances in alkali environments (Favero, Stephens, & Titirici, 2020).

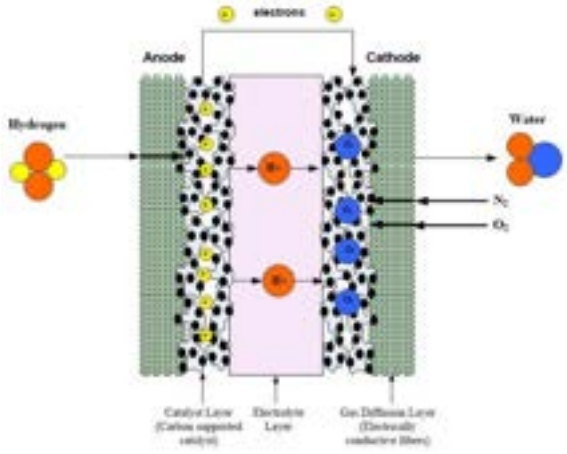


Figure 1. Schematic of the PEMFC with the catalyst layer (Spiegel, 2019)

Nafion® is the current state of the art polymer electrolyte membrane, but it has issues for large-scale development owing to its high synthetic cost, dependence on fluorine-based synthesis as well as temperature and humidity sensitivity (Holdcroft, 2014). The ionomer and catalyst form the catalyst layer (CL), which is the region where fuel and oxidant are converted to products. The ionomer plays a great role in determining the overall performance of the fuel cell, by acting as a binder to give mechanical integrity to the catalyst layer, through providing proton or hydroxide conductivity, by regulating water management and finally by facilitating oxygen transport. **Figure 1** shows a schematic of a proton exchange membrane fuel cell, featuring the proton exchange membrane and the two catalyst layers, deposited on a gas diffusion layer and composed of catalyst and ionomer

Owing the limitations of Nafion®, research has been focused on the development of alternative proton exchange membrane and on the development of new anion exchange membranes. However, little research has

been devoted to the study of such polymers as ionomer and their role in the catalyst layer.

Recent studies have shown that ionic liquids (IL) or poly(ionic liquid)s (PILs) can provide high oxygen concentration and transport (Wang, et al., 2019). ILs are molten salts with melting points below 373K. Notable characteristics of ILs include low vapour pressure, high thermal/electrochemical stability, and relatively low toxicity. An attractive aspect of ILs is its versatility, allowing desired properties to be emulated, for example providing high oxygen solubility. PILs are formed upon polymerisation of ILs which adds polymer-like qualities in addition to those associated with the IL. As a result, PILs are very suitable to be studied as ionomer in PEMFC and AEMFC. In addition, PILs offer even further versatility since the morphology of the chain can be tuned, allowing to independently optimize their ionic conductivity, mechanical stability, oxygen transport and hydrophobicity.

In this manuscript we will present the synthesis of imidazolium-based block copolymers and co-PILs Their structure and thermal integrity is studied with NMR, TGA and elemental analysis. Finally, their performance and cathodic catalyst layer ionomer is tested in rotating disk electrode (RDE) and gas diffusion electrode (GDE) configuration, to draw structure-property relationships that will guide the further development of such ionomers.

2 Background

A major limitation of Nafion® includes performances in high current density applications, since the presence of the ionomer intensifies the local oxygen starvation, increasing the oxygen transport resistance (Li, Intikhab, Malkani, Xu, & Snyder, 2020). Additionally, the specific adsorption of Nafion®'s hydrophilic sulfonate functional group causes the blockage of active sites and hinders the performance of the ORR. Research has been made to mitigate these effects through the use of IL additives to Nafion®-containing ink, resulting in enhanced ORR kinetic activity (Li, et al., 2020).

Zhang et al. have reportedly improved the electrocatalytic performance of ORR in a RDE setup by introducing PILs as replacement ionomers for Nafion®, specifically using copolymers containing protonated imidazolium (Zhang, Yang, Zhang, & Fang, 2019). Styrene was also present in the copolymers. Styrene provides hydrophobicity, increases the structural integrity and chemical stability of the polymer (Lu, et al., 2013). Imidazolium provides high anionic conductivity through its conjugated structure, which allows its cationic charge to delocalize across the structure. A high proportion of imidazolium is required to maintain a high ionic conductivity and facilitate oxygen transport. There is a limit to how much imidazolium improves performance, as having too much causes the hydrophobicity of the CL to decrease. Hydrophobicity of the CL is crucial to prevent the flooding of the cathode, blocking oxygen diffusion (Wang, et al., 2018). The performance of the copolymers have been comparable to Nafion® (Zhang, Yang, Zhang, & Fang, 2019).

Previous research by Lu et al. covers various synthesis methods for imidazolium-based anion exchange membranes and compares various ORR performance parameters. By varying the polymer matrix material, a range of polymers were obtained and tested showing how the performance of imidazolium-based polymers is comparable to Nafion®, with Pt/C electrochemical surface area (in an AEMFC RDE setup) of a Nafion® reported at 79.4 ± 2.4 $m^2g_{Pt}^{-1}$ and that of imidazolium-based PIL at 68.8 ± 1.5 $m^2g_{Pt}^{-1}$ (Lu, et al., 2013). The analysis also revealed how the AEM performance is highly dependent on the ionomer and catalyst ratio (denoted as I:C) (Lu, et al., 2013).

The use of styrene/imidazolium copolymer to modify the catalyst layer inspired the synthesis of similar copolymers. **Figure 2** shows the structure of the copolymer before and after protonation with TFSI.

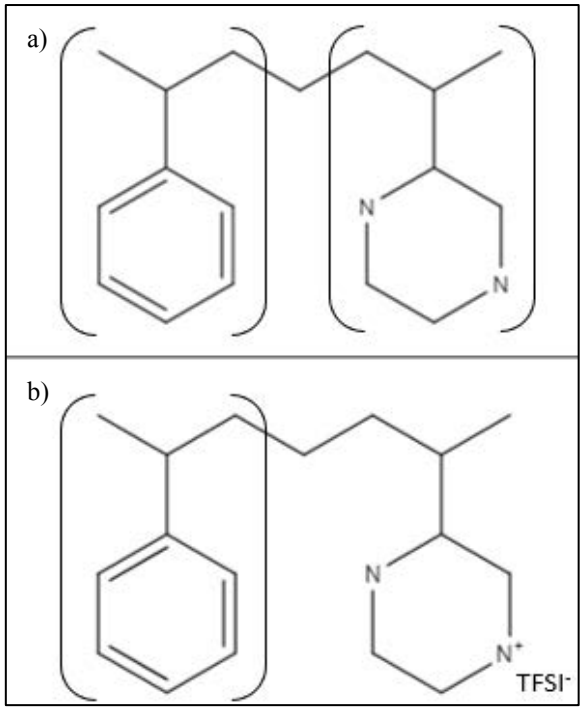


Figure 2. Structure of the PILs a) before ion exchange in TFSI b) after ion exchange

The aim of this research is to further examine the performance of imidazolium and styrene PIL's potential to be used as an alternative to Nafion®. The high oxygen affinity, good conductive properties, mechanical stability and hydrophobicity of the copolymer should help overcome the oxygen uptake resistance and result in enhanced ORR performance at the cathode. The performance will be measured using various electrochemical tests where various compositions will be explored to optimise the CL.

3 Methods

3.1 Polymer Synthesis

The polymers and PILs were synthesised with the following materials and procedures.

3.1.1 Materials

Toluene, diethyl ether, Azobisisobutyronitrile (AIBN), 1-Vinylimidazole, and styrene were purchased from Sigma-Aldrich, of which the 1-Vinylimidazole and styrene were purified with an alumina filter column and the rest were used as received.

3.1.2 Synthesis of Poly(1-Vinylimidazole-co-styrene) (Poly(St-co-Vim))

The desired amounts of 1-Vinylimidazole (VIm), styrene (St), and AIBN (see **Table 1**) were dissolved in anhydrous toluene under an argon atmosphere in a Schlenk flask. Three freeze-pump-thaw cycles were performed to remove oxygen present in the reaction vessel. The mixture was then placed in an oil bath heated to 60°C and left for approximately 12 hours. However, for the second synthesis, the oil bath was heated to 80°C. After the polymerisation reaction, the solution was dropped into cold diethyl ether solvent, which resulted in the production of a 'sticky' orange precipitate (Poly(St-co-VIm)). The precipitate was washed with hexane and methanol, using a Soxhlet set up and dried under vacuum at 80°C for 24 hours.

Molar Ratio of Vim/Sty	VIm (mL)	St (mL)	Toluene (mL)	AIBN (mg)
7.5:1	30	5	100	9
4:1	9	4.4	19	16

Table 1. Reactants for Free-Radical Polymerization reaction.

3.1.3 Protonation/Ion Exchange

The Poly(St-co-VIm) was synthesised using acid HTFSI and dissolved in ethanol. The mixture was then stirred at room temperature for 12 hours using a magnetic stirrer. To extract the ion-exchanged polymer, the mixture was dripped into de-ionized water, which produced a pale white precipitate, Poly(St-co-VImH+). The precipitate was collected via vacuum filtration, and then left to dry under a vacuum at 40°C for 24 hours.

3.2 Ink preparation

For ink preparation, the ratio of ionomer (mg) to catalyst (mg) will be denoted by I/C, where I/C = 1 corresponds to 1mg of ionomer per 1mg of catalyst (FePC). The catalyst inks for ORR catalysts were prepared using the following procedures. For Nafion®, FeNC catalyst (8mg) was mixed with de-ionised water (1mL), iso-propanol (1mL) and 5 wt% Nafion® of varying quantities from 40µL (I/C=0.25), 80µL (I/C=0.5), and 160µL (I/C=1). For Polymers and PILs, FeNC catalyst (8mg) was mixed with ethanol (2mL), and ionomer of varying quantities from 8mg (I/C=1) and 16mg (I/C=2). The catalyst inks were ultrasonicated in an ice bath for 20 minutes to form a homogenous solution. Once sonicated, 12µL of the ink was slowly and carefully distributed onto the surface of the glassy carbon working electrode and subsequentially left to dry for 20 minutes.

3.3 Rotating Disc Electrode

Electrochemical measurements were made using a Rotating Disk Electrode (RDE) method, alongside software NOVA 2.1 to record the data produced. As seen in **Figure 3**, the RDE setup consists of a three-electrode system; a catalyst-coated glassy carbon working electrode (catalyst loading 0.24 mg/cm^2), a graphene counter electrode, and an Ag/AgCl reference electrode, all in an electrolyte solution of 0.1M KOH to create a predominantly alkaline environment, , under which, metal-free catalysts offer good ORR performance. The performance of the inks was characterized using Linear Sweep Voltammetry (LSV, at 1600rpm) in the presence of oxygen (to remove any mass transport limitations), the electrochemical Impedance at room temperature in the presence of oxygen, and the Cyclic Voltammetry (CV) in the presence of Nitrogen.

Figure 3. Rotating Disk Electrode (RDE) configuration.

3.4 Gas Diffusion Electrode

Further electrochemical measurements were made using a Gaskatel Gas Diffusion Electrode (GDE) cell (see **Figure 4**) alongside NOVA 2.1 to record the data produced. The GDE setup consists of a three-electrode system; a carbon paper on which the catalyst ink has been sprayed on as the working electrode, a graphene counter electrode, and an Ag/AgCl reference electrode, with both the counter and reference electrodes submersed in an electrolyte solution of 1M KOH. The performance of the inks was characterized using Linear Sweep Voltammetry (LSV) in the presence of oxygen, the electrochemical Impedance at room temperature in the presence of oxygen, and the Cyclic Voltammetry (CV) in the presence of nitrogen. For each ionomer being tested, three repeats were conducted for each of the measurements which were then used to calculate and plot the average and error.

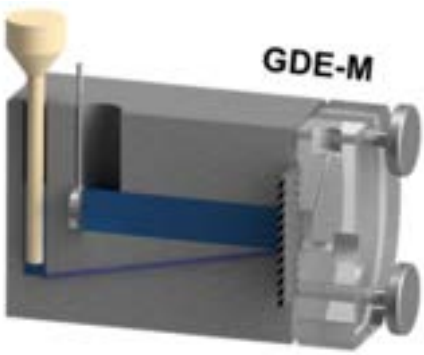


Figure 4. Commercial GDE cell from Gaskatel adapted at TY Darmstadt (GDE-M).

3.5 Elemental Analysis

A 400 MHz NMR device was used to obtain H-NMR and F-NMR spectra for the Poly(St-co-VIm) samples. Deuterated Chloroform, purchased from Sigma Aldrich, was used to dissolve the corresponding samples. Elemental analysis techniques were also conducted to analyse and confirm the monomeric ratio of the co-polymer PS-Vim and to determine the degree of protonation after the acidic substitution stage. Moreover, gel permeation chromatography was also utilised to determine the molecular number average (MN).

3.6 Thermal Analysis

A thermogravimetric analysis (TGA) was performed on a PIL-2 sample over a temperature range of 31°C - 500°C using a Netzsch TG 209F1 Libra®. The same analysis was extended to a sample of PIL-3 using a Netzsch STA 449F5 Jupiter® over 24°C - 600°C. The upper temperature limit of thermal stability was determined by the TGA and used to run a differential scanning calorimetry (DSC) analysis on PerkinElmer DSC 8000® across a thermally stable range to avoid damaging the equipment. The glass transition temperature (T_g) extracted from the DSC would be used to characterise the polymers together with the thermal stability analysis from the TGA.

4 Results and Discussion

4.1 Polymer Characterization

An elemental analysis had previously been performed following the synthesis of the polymers and a ratio of the two ionomers styrene and vinyl imidazole was obtained with data is summarised in **Table 2** (further data available in **Section 8.1** and **Section 8.2**). TGA revealed the upper temperature limit of thermal stability to be 350°C for PIL2 and 370°C for PIL3. To stay within the stable region, the DSC was operated in a temperature range between -100°C and 330°C. The DSC analysis did not reveal a glass transition temperature for both PIL2 and PIL3 when the samples were subject to being heated at 10°C/minute. The heating rate was then changed to 50°C/minute for the same temperature range, but this adjustment did not change the results.

Polymer	Styrene:Vinyl imidazole	Relative molecular weight	TGA (°C)
P1	2.2:1	Low	-
P2	2.3:1	Medium	-
P3	1:5	High	-
PIL1	2.2:1	Low	-
PIL2	2.3:1	Medium	350
PIL3	1:5	High	370

Table 2. Summary of obtained polymer properties

4.2 RDE Electrocatalytic Results

The RDE was used to test the electrocatalytic activity of the inks derived from the synthesised copolymers and commercial Nafion® via Linear Sweep Voltammetry (LSV), Impedance, and Cyclic Voltammetry (CV).

4.2.1 Linear Sweep Voltammetry (LSV)

LSV were performed in oxygen-saturated 0.1M KOH at a scan rate of 10 mV/s, to observe the kinetic and mass transport behaviour of the catalyst towards the ORR. As seen in **Figure 5**, the LSV plot can be separated into 3 regions: the kinetic region which is influenced by the kinetics and activity of the catalyst; the diffusion region which is influenced by oxygen diffusion; and mixed kinetic-and-diffusion region.

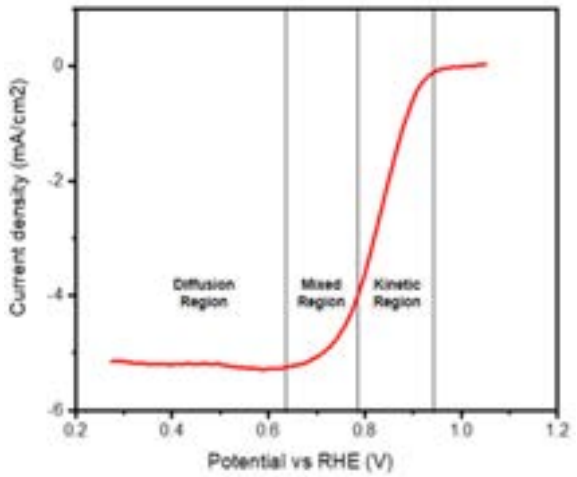


Figure 5. A typical LSV plot including the three observed regions (kinetic, mixed, and diffusion regions).

Varying the content of Nafion® based ionomer influences the performance of the proton membrane fuel cells. To perform a fair comparison between the commercially available Nafion® and the synthesised ionomers, the performance of Nafion® was tested at I/C ratios of 0.25, 0.5, and 1 in alkaline conditions using LSV. As seen in **Figure 6**, Nafion® with an I/C ratio = 1 had the greatest electrocatalytic performance with the fastest kinetics and hence served as a benchmark to compare the other ionomers against. Unexpectedly, I/C=0.25 yield better results than I/C=0.5

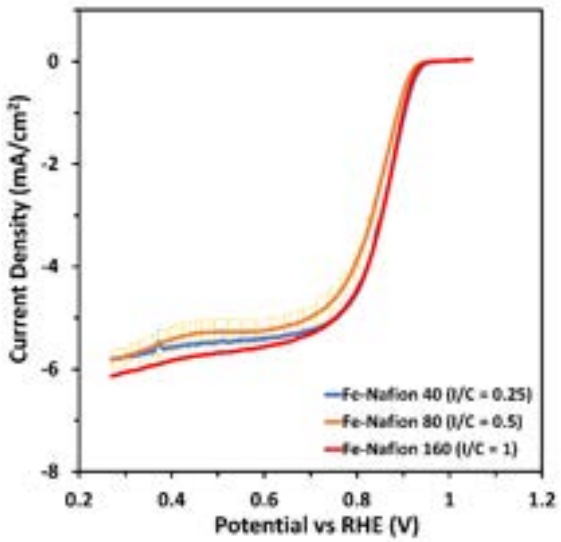


Figure 6. Linear Sweep Voltammetry plot (at 1600 rpm) of Nafion® at I/C = 0.25, 0.5, and 1.

The electrocatalytic performance of the different polymers at I/C = 1 (see **Figure 7**) and I/C ratio = 2 (see **Figure 8**) were obtained via LSV and compared to Nafion® at I/C ratio = 1. All polymers and PILs at either I/C ratio = 1 or I/C ratio = 2 exhibited worse kinetic properties and increased resistance to oxygen transport compared to Nafion®. These limitations were possibly due to adsorption on the catalysts' active sites or decreased accessibility to the active sites on the catalyst surface. Furthermore, during the loading of the ink onto the electrode, poor stability of the catalyst layer was observed which required multiple attempts and changes in composition in order to increase the mechanical stability.

PIL2 and PIL3, although worse in electrocatalytic activity compared to Nafion®, showed the greatest performance of the synthesised ionomers with both I/C ratios of 1 and 2 displaying identical electrocatalytic performance (as seen in **Figures 7** and **8**). Due to the better performance, PIL2 and PIL3 were further tested via Gas Diffusion Electrode (GDE) as explained in **Section 3.4**.

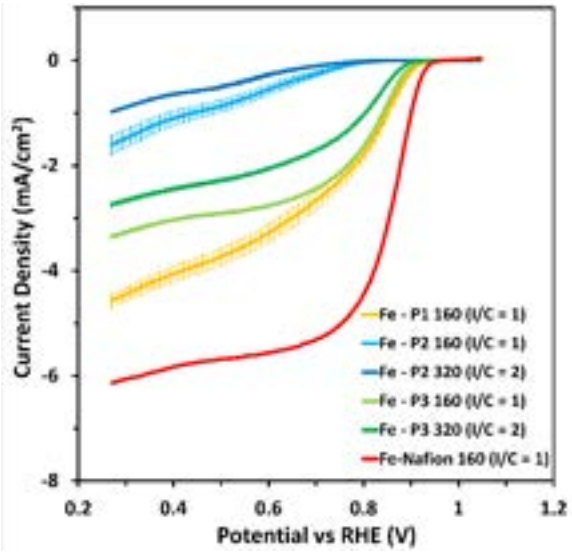


Figure 7. Linear Sweep Voltammetry plot (at 1600 rpm) of Nafion (I/C = 1) and P1, P2, and P3 (I/C ratio = 1 and 2).

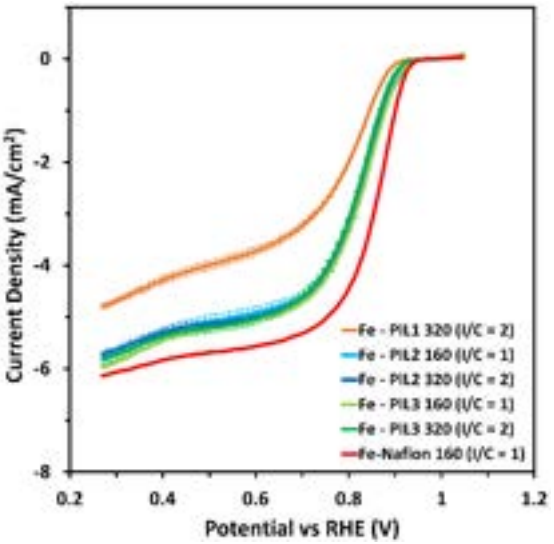


Figure 8. Linear Sweep Voltammetry plot (at 1600 rpm) of Nafion (I/C = 1) and PIL1, PIL2, and PIL3 (I/C ratio = 1 and 2).

4.2.2 Impedance

The RDE was rotated at 1600 rpm while sinusoidal perturbations in the frequency range $10^{-1} - 10^{5} Hz$ were applied to the system to collect impedance data with 10 frequencies tested per decade at an amplitude of $0.01\ V_{RMS}$. The results were examined on a Nyquist plot fitted to the equivalent circuit shown in Figure 9, composed of the Ohmic drop (Rs), polarization resistance (Rp), double layer capacitance (Cdl), resistance to the transport of oxygen (Ro) and an additional capacitance originated from the storage of oxygen. The results are displayed on a Nyquist plot as displayed on **Figure 9**. The electrical circuit that was fitted onto the data is shown as a schematic on **Figure 10**. The values for the fittings can be found summarised in **Section 8.4** under Supplementary Information.

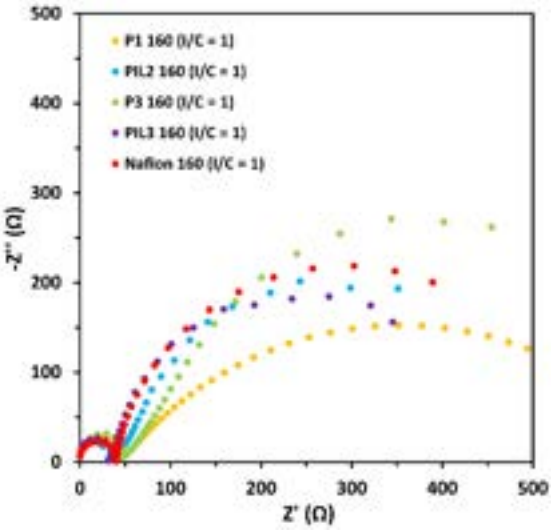


Figure 9. Electrochemical impedance spectroscopy Nyquist plot for various PILs with I/C ratio of 1.

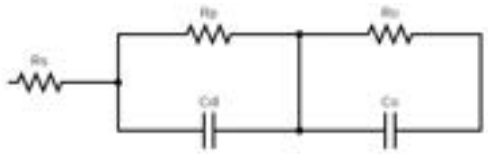


Figure 10. Schematic model of two resistances, which represent film polarization resistance and oxygen transport. Capacitance is denoted by C and the offset potential resistance denoted by Rs.

The polarization resistance only experienced some fluctuation but remained relatively constant, which is expected since the active site is not perturbed by the ionomer. As seen in **Figure 9**, PIL2 and PIL3 polymers have good oxygen transport since the resistance is considerably lower than the commercial benchmark Nafion®, confirming the predicted improvement of oxygen transport. It is also important to note that the ion exchanged polymers have excellent oxygen transport in comparison to their original state. This is likely due to the fluorinated anions, which provide excellent oxygen solubility. The only exception is P1/PIL1; P1 showed excellent oxygen transport, likely due to the short chain length, however, the low molecular weight also results in poor mechanical stability and PIL1 did not yield a stable enough layer to perform impedance measurements. A notable flaw in the impedance results is a negative offset impedance, whereby the Z' value start off at a negative value (i.e. Rs < 0). This peculiarity has no physical meaning as there cannot be a negative impedance, so the fittings of the Nyquist plot were conducted by taking Rs = 0.

4.2.3 Cyclic Voltammetry (CV)

CV is employed to observe and analyse the reduction and oxidation process of a molecular species. It is carried out in the presence of nitrogen so that oxygen reduction doesn't occur. The typical rectangular shape that occurs is due to the double layer capacitance and the redox peaks observed are considered to be due to electron transfer to the metal centre or the adsorption / desorption of oxygenated intermediates. The high potential peak in FePC is believed to originate from OH desorption and its position is correlated with the binding energy of the OH intermediate. The position of this peak is not expected to change with ionomer unless the ionomer has a strong interaction with the active sites.

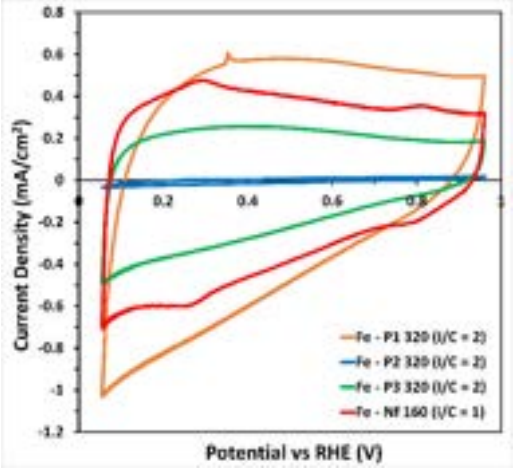


Figure 11. Cyclic Voltammetry plot of Nafion (I/C = 1) and polymers 1,2, and 3 (I/C = 2).

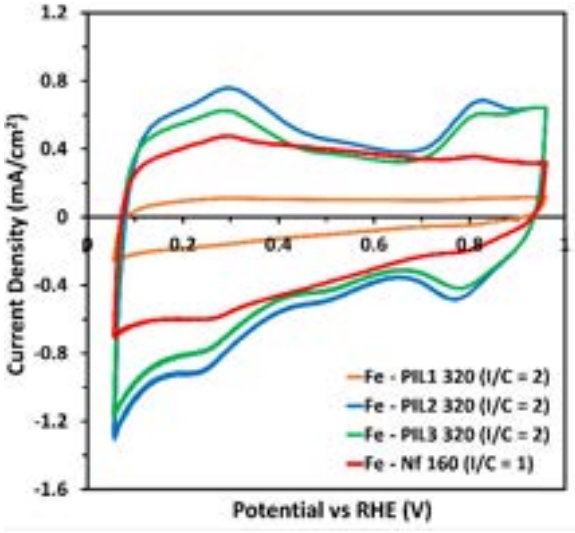


Figure 12. Cyclic Voltammetry plot of Nafion (I/C = 1) and PILs 1, 2, and 3 (I/C = 2).

In CV, the size of the cyclic voltammogram relates to the electrical double layer; a measurement of the surface area and the polarizability of the catalyst surface. A bigger width relates to a more accessible catalyst surface and a better polarizability of the ionomer. Changes in the size of the cyclic voltammogram can relate to different surface area accessibility, polarizability, or conductivity of the catalyst.

As seen in **Figure 11**, the polymers (I/C = 2) exhibit a smaller capacitance compared to Nafion® (I/C = 1). P2, P3, and PIL1 (see **Figure 12**) have a smaller width compared to Nafion®. All the polymers, and PIL1 do not exhibit the CV peaks characteristic of FePC. This can indicate poor conductivity, poisoning of the active site or lower accessibility to the catalyst active sites, which can also explain the low ORR performance reported with these polymers.

In contrast, PIL2 and PIL3 (I/C = 2) displayed a similar double layer capacitance and cyclic voltammogram peaks to Nafion® (see **Figure 12**). This suggests that there is good conductivity and better accessibility to the catalyst active sites. The position of the high potential peak does not change which indicates that the PILs do not poison or strongly interact with the active sites.

4.3 GDE Electrocatalytic Results

The RDE is a fast and simple technique, used to screen catalysts and ionomer. However, the experimental set up of RDE is very different from a real fuel cell, and results obtained in RDE are not always translatable to real devices. Therefore, it was decided to test the best performing synthesised ionomers in a Gaskatel GDE configuration, which is a better representation of performance in fuel cells. As seen in **Section 4.2**, PIL2 and PIL3 displayed the best in performance and hence were chosen for the GDE experiments of Linear Sweep Voltammetry (LSV), in the presence of oxygen, the electrochemical Impedance at room temperature in the presence of oxygen, and the Cyclic Voltammetry (CV) in the presence of nitrogen.

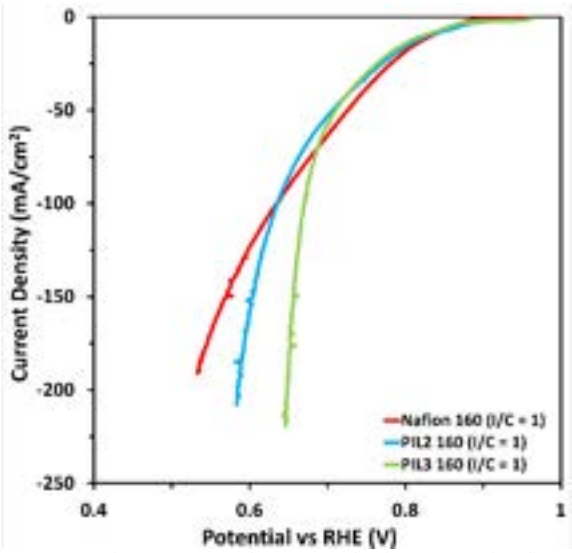


Figure 13. Linear Sweep Voltammetry plot from GDE of Nafion®, PIL2, and PIL3 at I/C = 1.

The electrocatalytic performance of PIL2 and PIL3 at I/C =1 (see **Figure 13**) was obtained via LSV and compared to against Nafion® at I/C = 1. The overpotential is the difference between the applied potential and the thermodynamic potential for ORR (1.23V). At low overpotential, Nafion® outperforms the PIL2 and PIL3, but at high overpotential, the PILs exhibit better activity. This suggests that the catalyst activity is slightly higher with Nafion® used as the ionomer whilst oxygen transport is better with PILs used as the ionomer.

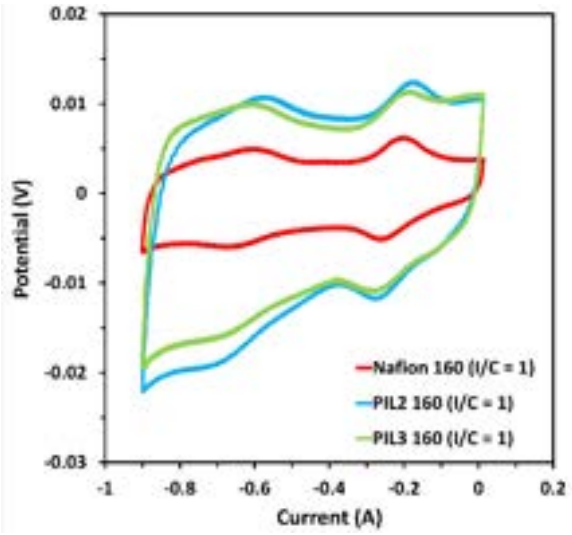


Figure 14. Cyclic Voltammetry plot from GDE of Nafion®, PIL2, and PIL3 at I/C = 1.

The GDE's CV performance of PIL2, and PIL3 of I/C = 1 were compared to the performance of Nafion® of I/C = 1 as seen in **Figure 14**. In both cases of PIL2 and PIL3, the ionomers display a larger double layer capacitance which indicates they either expose more surface area or that the ionomers have greater polarizability compared to Nafion®. Furthermore, the cyclic voltammogram peaks of the ionomers are also greater than that of Nafion®, suggesting that there is better conductivity and accessibility to the catalyst's active sites. From the GDE CV results, the actual performance of the synthesised ionomers is better than that of Nafion®.

5 Conclusions and Outlook

The polymers and poly(ionic liquids) were successfully synthesised, their structure was confirmed by NMR and the Styrene:Imidazolium ratio was determined by elemental analysis. Synthesis 1, 2, 3 were found to produce polymers with varying proportions of the monomers (styrene, imidazole) and molecular weight. TGA confirmed that the synthesised polymers PIL2 and PIL3 were stable up to 350°C and 370°C respectively.

The electrocatalytic testing from the RDE yielded informative results, showing that the polymers before ion exchange feature low conductivity and ORR performance. In contrast, the poly(ionic liquid)s showed better performance, with the exception of PIL1 which, due to its low molecular weight, led to a bad binding performance and low mechanical stability of the catalyst layer. PIL2 and PIL3 showed RDE performance comparable to Nafion® suggesting that they are very similar in conductivity and accessibility to the catalyst active sites. The impedance results suggest that the synthesised ionomers have less resistance in oxygen transport than Nafion®.

GDE was also used to confirm the performance of the best performing synthesised ionomers (PIL2 and PIL3) in a set-up more representative of a real fuel cell device. Despite showing similar results in RDE, PIL3 outperformed PIL2 in GDE, indicating that the higher proportion of imidazole monomers offered better oxygen transport, which is again further confirmed by the results obtained from Impedance.

For future research, the chain length and composition of polymers can be varied to study their effects on the electrocatalytic activity of the synthesised ionomers. Ex-situ characterization of the ionic conductivity, oxygen transport and water uptake of the polymers could guide the further development of structure-property relationships. Additionally, stability measurements combined with post-mortem analysis could shine light on the stability and degradation mechanism of the catalyst layer in presence of different ionomers. Finally, further research can allow the introduction of OH conductive monomers in order to improve the ionomer's ionic conductivity and performance.

6 Acknowledgements

The authors would like to extend their gratitude to Silvia Favero (Imperial College London) for their continuous support and guidance throughout this research project.

7 References

Favero, S., Stephens, I. E., & Titirici, M. M. (2020). Engineering the Electrochemical Interface of Oxygen Reduction Electrocatalysts with Ionic Liquids: A Review. *Advanced Energy and Sustainability Research*.

Gao, X., Yu, H., Jia, J., Hao, J., Xie, F., Chi, J., . . . Shao, Z. (2017). High Performance Anion Exchange Ionomer For Anion Exchange Membrane Fuel Cells. *RSC Advances*, 19153-19161.

Holdcroft, S. (2014). Fuel Cell Catalyst Layers: A Polymer Science Perspective. *Chemistry of Materials*, 381-393.

Li, Y., Intikhab, S., Malkani, A., Xu, B., & Snyder, J. (2020). Ionic Liquid Additives for the Mitigation of Nafion Specific Adsorption on Platinum. *ACS Catalysis*, 7491-7698.

Li, Y., Van Cleve, T., Sun, R., Gawas, R., Wang, G., Tang, M., . . . Neyerlin, K. C. (2020). Modifying the Electrocatalyst–Ionomer Interface via Sulfonated Poly(ionic liquid) Block Copolymers to Enable High-Performance Polymer Electrolyte Fuel Cells. *ACS Energy Letters*, 1726-1731.

Lu, W., Shao, Z.-G., Zhang, G., Zhao, Y., Li, J., & Yi, B. (2013). Preparation and characterization of imidazolium-functionalized poly (ether sulfone) as anion exchange membrane and ionomer for fuel cell application. *International Journal of Hydrogen Energy*, 9285-9296.

Spiegel, C. (2019, July 5). *Modelling the Catalyst Layers*. Retrieved from FuelCellStore: https://www.fuelcellstore.com/blog-section/modeling-the-catalyst-layers

Wang, M., Zhan, H., Thirunavukkarasu, G., Salam, I., Varcoe, J., Mardle, P., . . . Du, S. (2019). Ionic Liquid-Modified Microporous ZnCoNC-Based Electrocatalysis for Polymewr Electrolyte Fuel Cells. *ACS Energy Letters*, 2104.

Wang, S., Li, X., Wan, Z., Chen, Y., Tan, J., & Pan, M. (2018). Effect of hydrophobic additive on oxygen transport in catalyst layer of proton exchange membrane fuel cells. *Journal of Power Sources*, 338-343.

Zhang, F., Yang, M., Zhang, S., & Fang, P. (2019). Protic Imidazolium Polymer as Ion Conductor for Improved Oxygen Evolution Performance. *Polymers (Basel)*, 1268-1279.

8 Supplementary Information

8.1 NMR Results

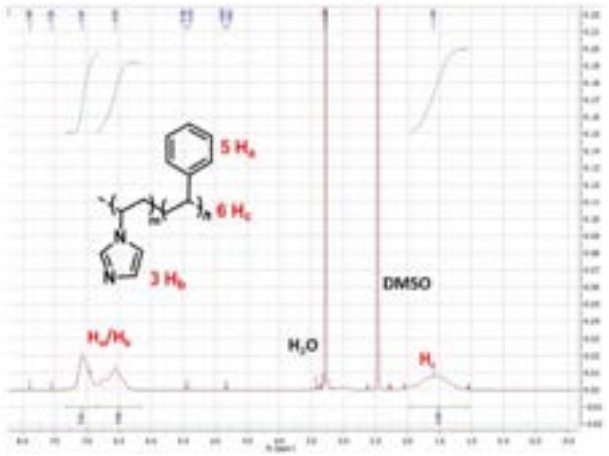


Figure 15. 400 MHz H-NMR results for the first P1

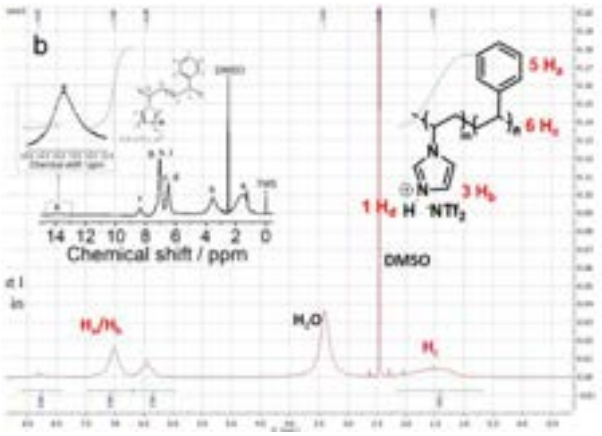


Figure 16. 400 MHz H-NMR results for the first PIL1

8.2 Elemental Analysis

Polymer	N Area	C Area	H Area	S Area	N %	C %	H %	S %	O2%	Styrene:Vynil-Imidazole
P1	5 290	37 017	9 802	0	4.2	47.4	47.5	0.0	0.9	**2.20**
P2	5 430	39 176	10 901	6	4.0	46.9	47.9	0.0	1.2	**2.30**
P3	13 588	28 009	8 952	14	11.6	38.5	45.6	0.0	4.3	**0.21**

Table 3. Raw Data from Elemental Analysis

Assessing the competitiveness of heat pump technologies in the UK domestic heating system

Kishan Amin and Aidan Sulway
Department of Chemical Engineering, Imperial College London, U.K.

Abstract

Heat pump technologies are becoming an increasingly popular low-carbon solution to replace natural gas-fired boilers in efforts decarbonise the domestic heating industry. The considered heat pumps are categorised into either electricity- or thermally-driven technologies. The former consists of air-, and ground-source heat pumps, whilst the latter includes thermally-driven absorption, and novel integrated-organic Rankine cycle heat pumps. In this study, the competitiveness of these technologies is assessed in the context of the UK domestic heating sector. Here, a whole-system UK heating model is used to evaluate the competitiveness of the investigated technologies over a range of market-informed hydrogen and electricity prices. Additionally, governmental net-zero policies and strategies are considered to investigate how the competitiveness of each technology varies as the domestic heating sector progresses along its decarbonisation pathway. It is found that high-performance air-source heat pumps are the most competitive technology for most resource price points. However, given the projected future electricity market volatility and plans for a functioning hydrogen supply infrastructure in the UK by 2035, it is expected that thermally-driven heat pumps will become increasingly favourable on approach to 2050. For price scenarios in which thermally-driven technologies are most competitive, it is expected that thermally-driven absorption heat pumps will serve high and medium heat demand households, whilst integrated-organic Rankine cycle heat pumps serve low demand households. Considering future hydrogen supply limitations, and the increasing electrification of other sectors in the UK, it is likely that a net-zero domestic heating sector will see a mix of thermally-driven and high-performance air-source heat pumps installed across UK households subject to resource availability.

1. Introduction

The average surface temperature of Earth has risen approximately 1 °C since the pre-industrial era (1880-1900) [1]. To address this issue, 196 parties from across the globe came together to form the Paris Agreement in 2015, which set a goal to limit global warming to well below 2 °C compared with pre-industrial levels [2]. To achieve this goal, countries aim to reduce greenhouse gas emissions as soon as possible to achieve net-zero emissions by 2050. In 2019 the UK committed to achieving net-zero across all sectors by 2050 and so the Carbon Budgets were established, providing a roadmap of carbon emission targets and key recommendations to meet net-zero ambitions [3]. In 2017 the domestic heating sector, made up of domestic space heating and hot water demands, accounted for 17 % of UK carbon emissions [4]. Currently, gas boilers supply 78 % of the UK's domestic heating demand [5]. Therefore, to decarbonise, the UK domestic heating sector must transition to low-carbon heating technologies such as electricity- or thermally-driven heat pumps [6].

The Sixth carbon budget [7] and the 2021 UK hydrogen strategy [8] indicate that hydrogen production must be scaled up significantly to achieve a net-zero future, this could also have implications to decarbonise the domestic heating sector. Hydrogen supply can be categorised into three colours: grey, blue, or green. The colour is dependent on the relative scale of carbon emissions arising from hydrogen production processes. Grey hydrogen is hydrogen produced via steam methane reforming (SMR) or autothermal reforming (ATR) using natural gas feedstocks, with all process emissions being released into the atmosphere [9][10]. Blue hydrogen follows the same production methods as grey hydrogen, but almost all of carbon emissions are captured and sequestered underground. Green hydrogen is produced via the electrolysis of water using renewable energy sources, such as solar or wind, to power the production process. Large scale green hydrogen production is a relatively new process. This, in conjunction with its dependence on renewable energy sources, causes green hydrogen to have limited supply and high prices in today's energy landscape [11]. However, with further developments in production technologies, increasing investment into green hydrogen production facilities, and decreasing costs in renewable energy, the price of green hydrogen is expected to gradually decrease over the next 20 years [11].

The International Energy Association identifies that by 2045, 50 % of the global heating demand must be met by heat pumps to reach global net-zero ambitions. To achieve this, it is estimated that heat pumps will represent 75 % of low-carbon heat technology sales in the UK by 2030 [6]. Additionally, the UK could be placing a ban on the installation of domestic fossil fuel boilers in new-build homes from as early as 2025 [5]. Heat pumps use power to transfer heat from a cold source to a hot sink [12], with their performance increasing with decreasing temperature difference between the cold source and hot sink. Their importance in low-carbon heating comes from being able to use low-carbon intensive power sources to drive the heat pump [13]. In line with their increasing importance, the research in this paper focuses solely on the role of heat-pump technologies in decarbonising the UK domestic heating sector.

There are two main categories of heat pumps: electricity- and thermally-driven heat pumps. Electric heat pumps are generally more efficient than electric boilers since they deliver more heat energy than electrical energy consumed [14]. The electric heat pump market is dominated by two types: air-source heat pumps (ASHPs), and ground-source heat pumps (GSHPs) [15]. The former

technology extracts heat from ambient air whilst the latter utilises subsurface ground temperatures. For this research, ASHPs are split into two design categories: high- (HP) and low- (LP) performance, with both types having large (L) or small (S) sizes [16]. The performance of an electric heat pump is measured using its coefficient of performance (COP), defined as the ratio between the useful heat output, and the electricity input [17].

Thermally-driven heat pumps could be significant on the pathway to net-zero heating for two key reasons. Firstly, green hydrogen has no scope-1 emissions as a fuel source used for heating. Scope-1 emissions are defined as the direct greenhouse gas emissions that are associated with resource usage [18]. Secondly, completely electrifying domestic heating would incur additional costs and could put strain on the national grid as peak electricity demand would significantly increase as other sectors seek electrification as a solution to decarbonisation [19]. Therefore using hydrogen-fuelled thermally-driven heat pumps can help reduce the strain on the national grid during times of peak energy demand. The fuel-to-heat (FHR) ratio is a performance parameter used for hydrogen heat pumps describing the ratio of useful heat output, to the heat provided by the fuel [20]. Similar to the COP for air-source heat pumps, the FHR is dependent on the ambient air temperature.

In the current thermally-driven heat pump landscape, there are two promising technologies: thermally-driven hydrogen absorption heat pumps (THHPs), and novel integrated-organic Rankine cycle heat pumps (ORCHPs). Gas-absorption heat pumps operate in a similar manner to electricity-driven heat pumps, however the system utilizes heat to drive the compressor. Initially, these heat pumps gained significant interest for refrigeration applications. However, work carried out by Scoccia *et al.* [21] identifies the promising application of such technologies in residential heating, where in some cases they may outperform electricity-driven heat pumps. Since the heat pump is gas-fired, low-carbon intensive hydrogen can be used as a fuel source. ORCHPs, which were investigated by Song *et al*. [20], are a novel heat pump technology. The work completed by Song *et al.* indicates, through thermal and economic assessment, that ORCHPs have significant potential to compete with existing domestic heating technologies. The operation of this technology involves an organic Rankine cycle, driven by fuel combustion, in series with a vapour-compression air-source heat pump cycle, with a heat transferred via an appropriate working fluid.

In a study by Petrović and Karlsson [22] the competitiveness of residential heat pumps is assessed in the context of the Danish energy system. They found that, for the Danish energy system, 24-28 % of the total heat demand after 2035 will be met by heat pumps and will be responsible for 66-70 % of heat from individual heating sources. However, it is worth noting that the only residential heat pumps assessed within this study were ASHPs and GSHPs with no consideration of thermally driven technology.

In a study by Wang and He [23], a national-scale techno-economic analysis of electric-driven heat pumps for decarbonising heat in Great Britain was performed. The study indicates that such technologies, under suitable policy and subsidy schemes, are cost-competitive against traditional gas boilers. However, similar to the work completed by Petrović and Karlsson [22], thermally-driven heat pumps were not considered in their analysis. Work completed by Olympios *et al.* [16] provided whole-system comparisons of electricity-driven heat pumps, and thermally-driven THHPs, boilers, and district heating in the UK. Findings from this study indicate that electricity-driven heat pumps are the least-cost decarbonisation pathway for the domestic heating sector. However, under certain price conditions, thermally-driven technologies are economically favourable.

In this research, the competitiveness of the heat pump technologies is assessed in the context of the UK domestic heating system. To perform this analysis, a modified version of the Energyscope model [24] is used to optimise the UK heating system with a focus on heat pump technologies. Further, heat pump techno-economic performance models developed in the CEP technology library [25] are used to integrate the investigated heat pump technologies into the whole-system heating model. To the best of the authors knowledge, novel thermally-driven ORCHPs are yet to be assessed in a whole-system context, thus forming the first novelty of the research.

The energy crisis in 2022 has caused unforeseeable disruptions in electricity and hydrogen markets. Therefore, the second novelty of this work builds on the work completed by Olympios *et al.* whilst considering current and projected future resource prices due to the energy crisis. Additionally, this work uses current governmental net-zero policies and strategies, to investigate how the competitiveness of each investigated technology varies as the UK domestic heating sector progresses along its decarbonisation pathway.

2. Methods

2.1 UK domestic heating system model

To optimise the UK domestic heating sector, a modified version of the Energyscope model [24] is used. It has the same conceptual formulation as the original model developed for the Swiss energy system in the year 2035. The model is a greenfield model that optimises the energy system considering a single snapshot year. The objective is the minimisation of the total annual cost of the energy system, defined as the sum of the annualized investment and operation and maintenance (O&M) cost of technologies, and fuel costs.

The heating model was initially simplified by the CEP Lab group. To simplify the original model, all technologies and infrastructures not related to heating are removed. Additionally, all input data is updated to be consistent with UK conditions in the year 2019. As such, the simplified model optimizes the investment and operating strategy of the considered technologies (Table 1) to meet the

decentralized heat demand (space heating and domestic hot water) over an entire year. Thus, the model aims to 'choose' the economically optimum technology to deliver heat to different households. Since the model represents the UK heating sector, it is important to accurately specify the number and size distribution of UK households (as different sizes have different levels of heating demand). As such, a clustering algorithm is used on the Cambridge housing model, [26] which identifies three representative types of UK household based on their heating demands (Table 2).

Table 1: Key technology-level inputs to the UK domestic heating system model. The numerical cost values and demand-level capabilities for each technology are derived from the CEP lab technology library [25], and the work from Song et. Al (ORCHP) [20]. All technologies are assumed to share the same maintenance cost, which is £162/yr (also derived from the CEP lab technology library)

Technology	Specific investment cost (£/kW)	Installation cost (£)	Satisfy household demands?		
			High	Medium	Low
HP ASHP (L)	548	3720	✓	✓	
HP ASHP (S)	677	3720		✓	✓
LP ASHP (L)	340	3720	✓	✓	
LP ASHP (S)	402	3720		✓	✓
ORCHP	489	3720	✓	✓	✓
GSHP	974	10300	✓	✓	✓
THHP	434	3720	✓	✓	

Table 2: Specification of household heating demand-level distribution over all UK households described by the Cambridge housing model

Household demand	Annual heat demand	Number of households (millions)
High	23200	2.20
Medium	13600	7.91
Low	7680	11.5

Another key input to the model is the UK ambient air and subsurface temperature profile. The ambient air temperature governs both household heating demands and the temperature-dependant performance of (above ground) heat pump technologies, whereas the subsurface temperature allows for evaluation of GSHP performance. Using annual weather data for London (2019), a KMedoids clustering algorithm is applied to identify twelve 'typical' weather days and one 'peak' weather day as inputs to the model.

In addition to the objective function, a series of practical feasibility constraints, such as the fact that heat demand must be satisfied at every hour, are included in the model. When considering the household-level installation of heat pumps, a thermal storage vessel must also be installed. As such, the model is constrained to allow a maximum installation of one heat pump technology plus thermal storage for each household demand type. For this research, the amount of thermal storage installed in each household is not assessed, as this will always be present in conjunction with heat pumps regardless of which heat pump technology is chosen. However, to accurately represent the total heating system cost, the costs associated with thermal storage installation in each household are considered. The model is set up to consider the scope-1 carbon emissions arising from using the selected heating technologies. As such, the carbon intensities of resources are nominally specified. In addition to the evaluation of system carbon emissions, it is also possible to constrain the total emissions to a nominally set limit. A key assumption underlying the model is that there exists a fully functional electricity and hydrogen supply infrastructure such that there is no constraint on the consumption of each resource. Further, the model assumes that the specific investment cost of the investigated technologies does not change over time. In reality, this is not the case since technology costs are naturally expected to slowly diminish over time. However, since all investigated technologies are of the same heat-pump nature, it can be assumed that their costs relative to each other will remain the same over time as all heat-pump technologies would improve at the same rate.

A major component of model development is the specification of the costs and performances of the investigated technologies. The CEP Lab group has previously developed a library of techno-economic performance models for novel and mature technologies such as ASHPs, THHPs, and GSHPs [25]. Each technology model assumes a domestic hot water and space heating demand of 55 ºC [16] and quantitatively describes relationships between performance (COP or FHR) and ambient air/ground temperature, which can be used with weather data to evaluate the required resource consumption to meet a given heat demand. These technology models are implemented into the UK heating model with key model inputs shown in Table 1.

2.2 Techno-economic performance of the ORCHP

As previously discussed, the integrated ORC heat pump is a novel technology which was conceptualised by Song *et al.* in 2021 [20]. The theoretical models developed by the research group are used to evaluate the specific investment cost (SIC), and temperature-dependant FHR profile of the technology. Further, it is found that the HP-ORC system orientation yields superior techno-economic performance compared to the ORC-HP orientation, thus only the HP-ORC is implemented in the UK heating model. The relative novelty of the technology requires some additional assumptions: i) ORCHP has the same installation and maintenance cost, and lifetime as an ASHP, and ii) ORCHP can operate at both high and low heating demands.

This represents possible large and small variants like ASHPs. The ORCHP is not treated like the THHP since the latter requires a large absorption cycle which constrains its minimum size.

2.3 Resource prices and carbon emissions

This section details the market research conducted to establish current and future resource prices, carbon intensities, and carbon emission targets. This allows simulations to be reflective of the economic and environmental climate leading up to 2050. A summary of the results from the market research can be seen in Tables 3, 4, 5 and 6. In 2021, the UK government committed to a completely decarbonised national grid by 2035 [27] and a report from the Northern Gas Networks titled the "H21 Project" [28] predicts hydrogen to be commercially available by the year 2035. As such, the years 2035 and 2050 are significant milestones on the pathway to net-zero and are consequently, in addition to 2022, the key years considered in this analysis. Considering the resource prices within the "H21 project", a wholesale-to-retail scale-up factor of 2 is applied. This is due to distribution and operation costs associated with the hydrogen network being estimated to contribute 25% each to the total cost paid by domestic consumers, with the wholesale cost contributing the remaining 50% [28].

2.3.1 Electricity prices

The conflict in Ukraine and worsening geopolitical climate has caused problems to the natural gas supply chain. This massively inflated the cost of electricity in the UK and highlighted the National Grid's reliance on natural gas, and the need to decarbonise. As the national grid increasingly shifts from natural gas to renewable energy, the cost of electricity is expected to recover to pre-2021 values over the next 28 years [29][30]. As a result, there are large variations in the predicted cost of electricity, and research carried out after the start of the energy crisis is prioritised. The 2022 cost of electricity is given as the 0.34 £/kWh energy price cap set by the UK government in September 2022 [31]. The BEIS predicts "as the (energy) system moves closer to 2050, wholesale prices will become increasingly volatile" [29]. The data in this report shows predicted deviations from the 2020 electricity price in 2035 and 2050, and their corresponding likelihood. For this analysis, 5% of the lower and upper tail of percentage point distribution graph are ignored as the probability of these occurring is deemed negligible. Further, predictions from additional literature provide much lower estimates than those seen past the 95th percentile, thus validating the assumption made [32]. Given the 5th and 50th percentile for 2035 and 2050 electricity price are close or equal to a £0 deviation, the lower bound is given as the 2020 price in both cases. An article written by The Cornwall Insight state the average pre-2021 wholesale electricity price is 50 £/MWh [30] which is supported by the Trading Economics' electricity price statistics [33]. Following a scale-up to consumer retail prices, this becomes 0.1 £/kWh, the lower bound for both 2035 and 2050. The 95th percentile in 2035 and 2050 shows a 95 £/MWh and 120 £/MWh increase over the 2020 price, respectively. Adding this to the 50 £/MWh price and applying the same conversions as previous, the upper bounds for 2035 and 2050 are 0.29 £/kWh and 0.34 £/kWh, respectively.

Table 3: Summary of current and projected electricity prices.

Retail price scenario (£/kWh)	Year		
	2022	2035	2050
Nominal	0.34	N/A	N/A
Low	N/A	0.10	0.10
High	N/A	0.29	0.34

2.3.2 Green Hydrogen prices

Market research carried out by pwc determines the wholesale price of green hydrogen in the UK should reach 1.75-2.00 $/kg by 2050 [11]. A hydrogen lower heating value (LHV) of 33.33 kWh/kg [16] and a USD/GBP conversion rate of 0.86 is used for converting to £/kWh. Using the central value of 1.875 $/kg it is calculated that the wholesale price of green hydrogen in 2050 would be 0.0485 £/kWh, with a retail price of 0.097 £/kWh. The low estimate of 1.75 $/kg gives a retail price of 0.09 £/kWh, only 0.007 £/kWh lower than the central estimate. Because of this, only the central value is taken, and low and high scenarios were not constructed as the error margin is deemed negligible compared with other resource price ranges. It is also determined that by 2035 the cost of hydrogen will likely reach 2.00-2.25 $/kg, the central value is taken and applying the same methodology gives a consumer retail cost of 0.11 £/kWh. Baldino et al. [34] estimate that green hydrogen could be produced for 0.126 £/kWh and thus the retail price would be 0.252 £/kWh. From the investigated literature, this is the highest estimated value of green hydrogen and so it is taken as the fuel's upper bound. Finally, the current 2022 price of green hydrogen is selected to be the expected cost of green hydrogen via electrolysis. A study by Gèrard et al. [35] predict the cost of green hydrogen produced via electrolysis to be 0.09 £/kWh, with a retail price to consumers being 0.18 £/kWh. Green hydrogen production costs are generally expected to decrease over time. This is a result of decreasing renewable energy costs, lessons learned from early hydrogen projects and technological advances in electrolyser and renewable technology [11].

Table 4: Summary of current and projected green hydrogen prices.

Retail price scenario (£/kWh)	Year		
	2022	2035	2050
Nominal	0.180	0.110	0.097
Error margin	N/A	±0.007	

2.3.3 Blue Hydrogen prices

As a result of the energy crisis, 2022 is assumed to have the highest cost of blue hydrogen as the price of natural gas is at an all-time high. A study from the Rocky Mountain Institute [36] expects blue hydrogen to cost 4.60 $/kg post-Ukraine war, corresponding to a wholesale cost of 0.119

£/kWh and a retail cost to consumers of 0.237 £/kWh. BloombergNEF's modelling on global hydrogen production estimate the cost of green hydrogen will not out-compete blue hydrogen until 2030 [36][37]. A report by the Energy Networks Association also supports this claim [38], they predict the price of blue hydrogen will slowly increase between 2030 and 2050 as green hydrogen becomes more competitive. A lower bound of 40 £/MWh and a 2050 value of 55 £/MWh are taken, corresponding to a retail price of 0.08 £/kWh and 0.11 £/kWh respectively. Northern Gas Networks' "H21 Report" [28] estimates that blue hydrogen must be sold at a price of 0.101 £/kWh to account for costs of establishing and operating the network, as well as producing hydrogen. They also predict a commercially ready hydrogen network by 2035. The nominal values for the blue hydrogen price are obtained from references dated before the start of the energy crisis. As such, a 35% uplift is applied to consider what could be more realistic future prices [39].

Table 5: Summary of current and projected blue hydrogen prices.

Retail price scenario (£/kWh)	**Year**		
	2022	2035	2050
Nominal	0.237	0.101	0.110
High	N/A	0.140	0.150

2.3.4 CO2 Emissions Target

From The Office for National Statistics' quarterly publication [40], the level of CO2 emissions in 2020 is given as 71000 kt CO2eq. and this is assumed to be the highest output over the next 28 years. The Sixth Carbon Budget published by The Climate Change Committee [7] gives a UK CO2 2035 emissions target 50% lower than the 2020 value to be on track for net-zero by 2050. This gives a 2035 carbon target of 36000 ktCO2eq./yr and is used for constructing scenarios during this year. Finally, a 2050 target of 0 kt CO2eq. was used, also discussed by the Climate Change Committee [41].

Table 6: Summary of CO2 emissions targets for each year considered.

Year	2022	2035	2050
Carbon emissions target (ktCO2eq./yr)	71000	36000	0

2.3.5 Carbon intensity of resources

Data from the National Grid [42] gives an average carbon intensity of 0.183 kgCO2eq./kWh from the start of 2022. Given the UK's commitment to a decarbonised national grid by 2035 [27], the carbon intensity of electricity is taken as 0 kgCO2eq./kWh for the years 2035 and 2050. Blue hydrogen supply has a low, non-zero carbon intensity. The Pembina Institute analysed the carbon intensity of blue hydrogen [43] and states the Shell Quest facility currently produces blue hydrogen with a carbon intensity of 54 kgCO2eq./GJ. Conversion into kWh gives a final blue hydrogen carbon intensity of 0.194 kgCO2eq./kW. However, it is to be noted that the current Quest facility has only 43% capture of greenhouse gases. Future projects are estimated to have 90-95% capture which would result in much lower carbon intensities [43].

3. Results

3.1 Performance analysis

Using technology models from the CEP Lab technology library [25] and Song *et al.* [20], heat pump performance is plotted against temperature (Figure 1).

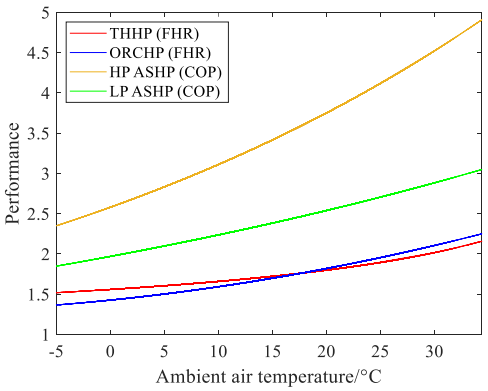


Figure 1: Comparison of heat pump performance as a function of ambient air temperature.

As shown, both high- and low-performance ASHPs have significantly better performance compared to the thermally-driven technologies which is due to low absorption cycle efficiency in the THHP, and poor heat-to-electricity conversion in the ORCHP. Considering the thermally-driven technologies, the THHP yields better performance at temperatures below 18 ºC. Thus for an average UK temperature of 12.8 ºC [44], it is expected that the THHP would be the optimal hydrogen technology on a performance basis. GSHP performance is not considered here due its dependence on ground temperatures, however considering the relatively high costs of GSHPs (Table 1) and their relatively large space requirements, it is expected that in most cases the technology would be suboptimal in comparison to other heat pumps.

3.2 Resource price sensitivity analysis

To assess the competitiveness of heat pump technologies for different demand-level households in response to variations in the specific price of resources (£/kWh), simulations over a range of green hydrogen and electricity prices are run using the UK domestic heating system model (Figure 2). For this analysis, blue hydrogen is held at its maximum price to represent the increasing popular opinion of green hydrogen becoming the most competitive commercially supplied type of hydrogen after 2030 [37] [38]. As such, if a thermally-driven technology is optimal, it would utilise green hydrogen as a fuel source. The nominal resource price ranges in this analysis are taken as the respective resource price upper and lower bounds from market research (Section 2.3). Further, the 2022 CO2 emissions target of 71000 ktCO2eq./yr (Section 2.3.4) is applied to the model such that all investigated technologies are feasible from a carbon emissions perspective.

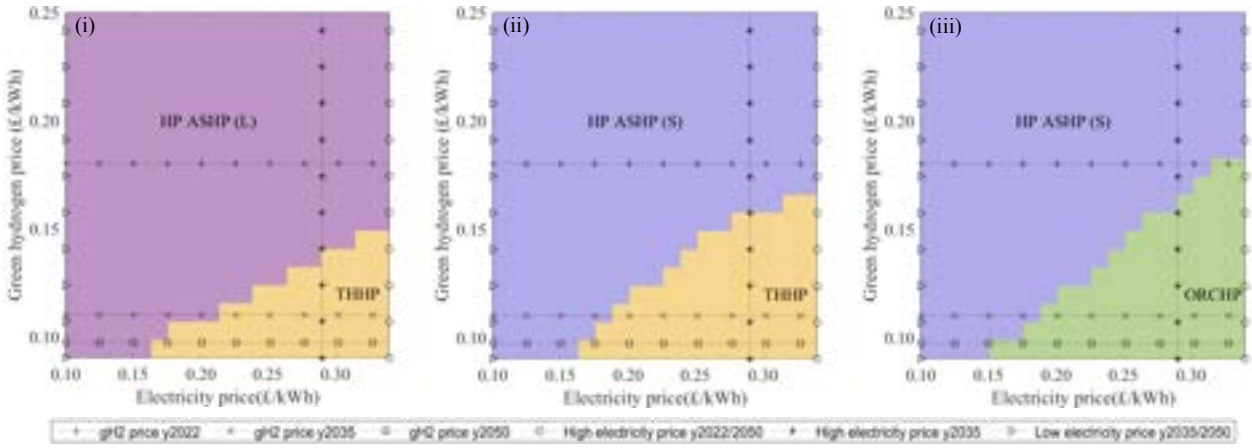


Figure 2: Assessment of optimal heat pump technologies over a range of retail green hydrogen and electricity prices for: (i) high, (ii) medium, and (iii) low demand households. Different colours represent which technology minimises the total UK domestic heating system cost for given resource prices. The horizontal and vertical lines denote current and projected green hydrogen and electricity prices for the years 2022, 2035, and 2050. The considered technologies include small (S) and large (L), high- and low-performance air-source heat pumps (HP ASHP/LP ASHP), thermally-driven absorption heat pumps (THHP), ground-source heat pumps (GSHP), and thermally-driven integrated organic Rankine cycle heat pumps (ORCHP).

3.2.1 High and medium demand households

Over the nominal price ranges, it is indicated that either HP ASHPs or THHPs are optimal to provide heat to high and medium demand households, with high demand households requiring the larger and more expensive HP ASHP (L) variant. Considering the 2022, and low electricity price-case resource prices for 2035 and 2050, both types of households favour the installation of HP ASHPs. Conversely, THHPs are favoured for the high electricity price-case resource prices in 2035 and 2050. HP ASHPs have the highest specific investment costs (Table 1) and performance (Figure 1) out of the investigated technologies (excluding GSHPs), with the latter property implying that HP ASHPs consume the least resources when supplying heat to meet a given demand. As such, considering the degree of overlap of the hydrogen and electricity price ranges, the relatively low resource costs from using HP ASHPs outweighs their high investment costs thus allowing them to be the optimal heat pump technology for most resource price points. However, when electricity and green hydrogen prices are respectively high and low, THHPs are favoured. This behaviour can be attributed to the resource prices causing the total resource cost from using THHPs to be lower than if HP ASHPs are used. Additionally, the lower investment costs of THHPs relative to the HP ASHP (Table 1) further contributes to their optimality in this case. Naturally, the selection of THHPs which utilise green hydrogen as a fuel source cause zero scope-1 carbon emissions from the UK heating system.

Neither ORCHPs or LP ASHPs are optimal for high and medium demand households. Although LP ASHPs have better COP performance than thermally-driven technologies, and lower investment costs compared to their high-performance counterparts, the lower resource consumptions of HP ASHPs are sufficient to offset the savings in investment costs achieved from using LP ASHPs. As previously discussed, ORCHPs have higher investment costs (Table 1) and would perform worse in the UK climate compared to THHPs (Section 3.1) thus making them the suboptimal hydrogen technology in this case

3.2.2 Low demand households

Simulations for low demand households yield similar results to that of medium demand households, where small HP ASHPs are optimal for most of the resource price points. However, in this case, when electricity and green hydrogen prices are respectively high and low, the ORCHP is favoured instead of the THHP. This is due to the nominal minimum size of the THHP (10kW) exceeding the heat demand for low households. As such, since the ORCHP has marginally worse techno-economic performance compared to the THHP but can satisfy low heating demands, the model would naturally select the ORCHP in this case. Considering the resource price-points for 2022, 2035, and 2050, the behaviour for low demand households is similar to that of high and medium demand households. However, in the case of low demand households, the HP ASHP is not optimal for 2022 resource prices. This behaviour, along with visual inspection of Figure 2, indicates that for high electricity prices, the green hydrogen price at which the HP ASHP is no longer optimal increases as household demand decreases. This is an expected trend since lower household demands imply lower resource consumption and smaller technology installed capacities, which allows the thermally-driven technologies to become more competitive against the better performing, higher costing HP ASHPs.

3.2.3 Blue hydrogen price variation

Despite the uncertainty in future blue hydrogen supply, a similar analysis is performed for the fuel whilst holding green hydrogen at its maximum price. It is found that the results for each household demand level follow the same behaviour and reasoning to that of the previously discussed green hydrogen price analysis, however in this case the thermally-driven technologies utilise blue hydrogen as a

fuel source due to the nominally set high price of green hydrogen.

3.3 Yearly snapshot analysis

Using results from market research, yearly snapshots are constructed for 2022, 2035, and 2050. As previously discussed in Section 2.3, resource prices in 2035 and 2050 have associated uncertainties, thus alternative scenarios for each snapshot year are also considered (Table 7). The relevant inputs for each scenario are then applied to the UK domestic heating model. Ultimately, this allows investigation of the optimal heat pump technologies and total heating system cost on a scenario basis as the domestic heating sector progresses along its decarbonisation pathway (Figure 3). In this section, the reasoning behind the nominal resource prices and carbon emissions limits used in each scenario can be found in Section 2.3.

Table 7: Market research-informed resource price scenarios considered in the yearly snapshot analysis. Corresponding numerical values can be found in Section 2.3.

Scenario	Blue H_2 price	Green H_2 price	Electricity price
	2035	2035	2035/2050
(a)	Nominal	Nominal	High
(b)	Nominal	Nominal	Low
(c)	High	Nominal	High
(d)	High	Nominal	Low

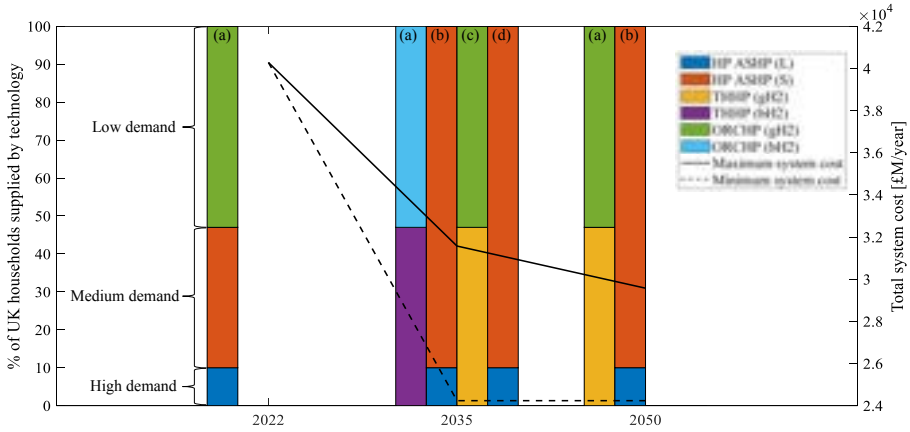


Figure 3: Assessment of the optimal proportion of UK households heat is supplied to by each investigated heat pump technology on a year-by-year scenario snapshot basis considering the years 2022, 2035, 2050. The sections of each bar correspond to the heating technology which minimises the total UK domestic heating system cost for each proportion of UK households (corresponding to high, medium, or low demand households) in each scenario. The lines reflect the variation in maximum, and minimum total system cost across the three years considering the total system cost arising from scenarios (a)-(d). The conditions in each scenario are described in Section 3.3 and Table 7, with numerical values found in Section 2.3

The 2022 scenario represents current-day fuel prices and carbon intensities, with a total system carbon emissions limit of 71000 $ktCO_{2eq.}$/yr. As shown, the green hydrogen-fuelled thermally-driven ORCHP is selected to supply heat to a large proportion of households in the UK (corresponding to low demand households). This is due to the low heat demand of these households, and the high price of electricity in 2022, causing the ORCHP to be economically favourable over the HP ASHP. Further, since the blue hydrogen price exceeds green hydrogen prices in 2022, it is not expected that any technologies would utilise blue hydrogen as a fuel. Although the electricity price is high relative to hydrogen prices in 2022, the remainder of the households have their heat supplied by either large or small HP ASHPs which represent high and medium demand households respectively. As discussed in Section 3.2.1, the alternative technology here would be the THHP. However, the superior performance of the HP ASHP allows it to be favourable in this case, regardless of the relatively lower hydrogen prices and THHP investment cost. Although these results indicate that ORCHPs are optimal to supply heat to low demand households, the application of thermally-driven heat pump technologies in the year 2022 is infeasible due to the current lack of a functional domestic hydrogen supply infrastructure in the UK. Aside from hydrogen supply issues, ORCHPs cannot be used in 2022 due to their technological immaturity. As such, it is expected that small HP ASHPs would be used in lieu of ORCHPs to supply heat to low demand households.

The transition from 2022 to 2035 sees a general decrease in resource prices, and a reduction in the carbon emissions limit to 36000 $ktCO_{2eq.}$/yr. As discussed in Section 2.3.3, it is assumed that there will be a functional domestic hydrogen supply infrastructure in the UK from 2035 onwards, thus allowing thermally-driven heat pump technologies to be feasible. It is found that HP ASHPs are the optimal heating technologies for scenarios (b) and (d), where the smaller size HP ASHP supplies heat to most UK households (corresponding to medium and low demands) and the larger size serves the remaining high demand households. In these scenarios, it is assumed that the electricity price will diminish back to 2020 levels. Since this electricity price is lower than the 2035 hydrogen prices, it is favourable to supply heat using HP ASHPs due to their low resource costs and high conversion efficiencies. These factors offset the high investment costs for HP ASHPs, allowing them to out-compete the thermally-driven technologies. Scenarios (a) and (c) implement a mixture of THHPs and ORCHPs as the optimal heat pump technologies. Here, ORCHPs serve the

highest proportions of households (low demands) whilst THHPs serve the remaining medium and high demand households. In both scenarios, the electricity price takes its maximum value on account of its projected market volatility. Scenario (a) assumes 2035 blue hydrogen prices to be consistent with predictions made before the energy crisis in 2022. As such, scenario (a) likely provides an underestimate of the blue hydrogen price thus causing a lower blue hydrogen price compared to green hydrogen in 2035. Hence, the thermally-driven technologies are fuelled by blue hydrogen in scenario (a). Scenario (c) considers a more realistic estimate of future blue hydrogen prices whereby a 35% uplift has been applied to the blue hydrogen price in scenario (a). Consequently, green hydrogen becomes cheaper than blue hydrogen, and so the thermally-driven technologies are fuelled by green hydrogen.

The 2050 snapshot looks at a fully decarbonised UK domestic heating sector. The projected green hydrogen price for this year assumes sufficient technological advances such that green hydrogen has the lowest specific resource price. Since the supply of blue hydrogen has a non-zero (although low) carbon intensity, the zero scope-1 carbon emissions constraint implies that thermally-driven technologies cannot be fuelled by blue hydrogen in 2050. Therefore, scenarios (c) and (d) are redundant in 2050 since they respectively yield the same results as scenarios (a) and (c). ASHPs are a feasible technology in 2050 due to the assumption of a decarbonised electricity supply from 2035 onwards (Section 2.3.5). In the year 2050, electricity prices are expected to be more volatile compared to 2035, causing the high electricity price to take a maximum value of 0.34 £/kWh. As such, scenario (a) models an electricity price which is over three times higher than the expected green hydrogen price in 2050. In this scenario, green hydrogen-fuelled thermally-driven heat pump technologies are favoured, and the behaviour mimics that of scenario (c) in 2035. As indicated by scenario (b) in 2050, HP ASHPs are the optimal heating technologies if electricity prices diminish back to 2020 levels by 2050. For this case, the electricity price is approximately equal to the green hydrogen price. Although the resource prices are similar, HP ASHPs are the most competitive technology due to their superior performances causing significantly lower resource consumption costs than the thermally-driven technologies. Further, the size distribution of HP ASHPs follows that of scenarios (b) and (d) in the year 2035.

The total system cost generally decreases along the decarbonisation pathway to 2050. The sharp decrease observed from year 2022 to 2035 represents the rapid decrease in all resource prices as they recover following the record-high prices witnessed in 2022. The wide range in system cost in 2035 and 2050 results from scenarios (b) and (d) investigating a nominal electricity price like those seen pre-2021 causing a low minimum total system cost. The minimum system cost also remains the same across these 2035 and 2050 as the minimum price of electricity is predicted to stay the same. The maximum system cost in both cases is attributed to the cases where thermally-driven technologies are favoured over electricity-driven technologies. This behaviour arises due to the poor performance of the thermally-driven technologies relative to the HP ASHP thus implying a higher resource consumption cost when they are chosen. Lastly, a decrease between 2035 and 2050 results from the decreasing price of green hydrogen between these years.

4. Conclusions

In this study, the competitiveness of novel and mature heat pump technologies was assessed in the context of the UK domestic heating sector across varying levels of household heating demands. Firstly, a thorough market analysis of current and projected future retail prices and carbon intensities of hydrogen and electricity was conducted. Then, along with techno-economic performance models for the investigated technologies, a modified version of the Energyscope model was used to identify the economically optimal heat pump technologies to supply heat to households over a range of resource prices. Lastly, governmental net-zero policies and strategies were considered to optimise the domestic heating sector as it decarbonises for the key years of 2022, 2035, and 2050.

It was found that the competitiveness of the investigated technologies was dependant on resource prices, and household heating demands. Owing to their superior performance, HP ASHPs were optimal over most of the investigated resource price ranges, with high demand households requiring a larger size compared to medium and low demand households. However, when electricity prices were high and hydrogen prices were low, a switch was made towards thermally-driven heat pumps. These consisted of THHPs serving high and medium demand households whilst ORCHPs serve low demand households. The thermally-driven heat pumps were found to have similar temperature-dependant performance behaviour, with the THHP outperforming the ORCHP in a UK climate. A further observation was that thermally-driven heat pump technologies became increasingly favourable as heating demand decreases due to lower demands closing the performance gap between the thermally-driven and HP ASHPs.

Although there were conditions under which thermally-driven technologies were optimal, their practical application is reliant on a functional domestic hydrogen supply infrastructure in the UK. As such, in line with the H21 report [28], it is likely that the use of thermally-driven heat pumps will only be feasible from 2035 onwards, causing HP ASHPs to be the only competitive heat pump technology in the year 2022.

Modelling of 2035 conditions indicated that either HP ASHPs, or a combination of THHPs and ORCHPs would be the optimal technologies to supply heat to households. Again, this was dependant on the relative cost of resources, where the thermally-driven technologies would naturally be fuelled by the cheapest type of hydrogen. The current energy crisis has caused a great deal of uncertainty in future

blue hydrogen prices. Therefore, if conditions do not improve, it is likely that the blue hydrogen price in 2035 will be greater than predictions made pre-energy crisis. As such, if electricity markets also remain disrupted due to the energy crisis, green hydrogen-fuelled thermally-driven heat pumps would be the most competitive technology in 2035. Given the novelty of green hydrogen as a fuel, and the need for sustainable hydrogen in decarbonising the wider energy sector, it is unlikely that there would be sufficient green hydrogen supply to meet all household heating demands in 2035. Thus, a more realistic scenario would be that a mixture of air-source and thermally-driven heat pump technologies would be installed across UK households subject to resource availability.

Simulations of a fully decarbonised domestic heating sector in 2050 yielded the same technology selection scenarios as 2035. However, the competitiveness of air-source heat pumps is reliant on two actions occurring by 2050: i) electricity markets stabilise, and ii) the UK achieves its 100% decarbonised electricity supply goal. If either of the conditions are not met, then thermally-driven technologies would dominate the domestic heat pump market. Given the potential case where electricity supply is decarbonised, and green hydrogen price predictions are accurate, HP ASHPs would be the optimal net-zero heating technology, so long as electricity prices remain below 0.15 £/kWh (low demand households), or 0.16 £/kWh (high and medium demand households). Considering the estimated 2050 electricity market volatility by BEIS [29] electricity prices will remain below these values 69 % and 70 % of the time, respectively. Thus, it is likely that HP ASHPs will be the economically favourable heat pump technology in 2050. However, for the remaining ~30 % of the time, electricity prices are predicted to substantially increase, which would lead to an unavoidably high total domestic heating system cost if HP ASHPs were utilised in all UK households. As such, the argument still exists for the use of thermally-driven heat pump technologies since they can help minimise the total UK domestic heating system cost by being installed in preparation for periods of inflated electricity prices.

In general, HP ASHPs were found to be the most promising current-day domestic heat pump technologies due to their technological maturity, superior performance, and reliance on a well-established electricity infrastructure. However, when the UK domestic heating sector was modelled to progress along its decarbonisation pathway to net-zero, THHPs and ORCHPs were predicted to become increasingly popular. The transition to these thermally-driven technologies would arise due to the ongoing advancements in green hydrogen production technology, electricity market volatility, and the need to reduce peak electricity demand from the UK national grid as other sectors seek electrification as a solution to decarbonisation.

5. Outlook

Although this work takes a comprehensive view of the UK domestic heating sector, energy markets, net-zero policies and strategies, and techno-economic performance of heat pump technologies, some areas of future research are recommended to build on this study.

From a technological perspective, further work should be carried out to develop and optimise the design of the ORCHP. Currently, its capacity constraints are an assumption and there may be a case where ORCHPs cannot satisfy low heating demands due to a practical size constraint. Additionally, optimising factors such as working fluids can lead to better economic performance of the ORCHP thus opening pathways for ORCHPs to serve medium and high demand households.

An extension to the model would be to individually consider different regions of the UK. Currently the model utilises weather data from London to model heating demands. However in colder areas, such as Scotland, heating demands would be higher. This, in conjunction with the fact that the performance of most heat pumps is dependent on ambient air temperature, would affect the optimal technology selection to some extent on a region-by-region basis. As previously discussed, the competitiveness of thermally-driven technologies is dependent on domestic hydrogen supply in the future. Therefore, if more robust information regarding future hydrogen supply becomes available, the model can be constrained to limit the consumption of hydrogen. A further development to this would be to also constrain hydrogen supply on a UK regional basis, since locations closer to hydrogen ‘hubs’ would most likely have access to a greater hydrogen supply compared to areas such as London. Additionally, electricity supply may also be constrained to model scenarios where the aim is to minimise peak electricity demand from the national grid as other sectors seek electrification as a solution to decarbonisation. This would result in a mix of heat-pump technologies supplying heat to each household demand-level (high/medium/low) as opposed to one technology per household demand-level thus allowing for a more thorough heat pump evaluation to be made whilst considering the resource requirements of other sectors in the UK.

6. References

[1] Lindsey, R. and Dahlman, L. (2022) Climate change: Global temperature, NOAA Climate.gov. NOAA Climate. Available at: https://www.climate.gov/news-features/understanding-climate/climate-change-global-temperature.

[2] The Paris Agreement | UNFCCC (2018) Unfccc.int. Available at: https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement.

[3] Carbon budgets (2016) GOV.UK. Department for Business, Energy & Industrial Strategy. Available at: https://www.gov.uk/guidance/carbon-budgets.

[4] Net Zero - Technical Report (2019) Climate Change Committee. Committee on Climate Change. Available at: https://www.theccc.org.uk/publication/net-zero-technical-report/

[5] UK Gas Boiler Ban - Complete Guide | EDF (2020) EDF. Available at: https://www.edfenergy.com/heating/advice/uk-boiler-ban.

[6] Net Zero by 2050: A Roadmap for the Global Energy Sector (2021) IEA. International Energy Agency. Available at: https://www.iea.org/events/net-zero-by-2050-a-roadmap-for-the-global-energy-system.

[7] Sixth Carbon Budget (2020) Climate Change Committee. Available at: https://www.theccc.org.uk/publication/sixth-carbon-budget/
[8] UK hydrogen strategy (2021) GOV.UK. Department for Business, Energy & Industrial Strategy. Available at: https://www.gov.uk/government/publications/uk-hydrogen-strategy
[9] The hydrogen colour spectrum (2021) National Grid Group. Available at: https://www.nationalgrid.com/stories/energy-explained/hydrogen-colour-spectrum
[10] Blue hydrogen | Low-carbon energy | Shell Catalysts & Technologies (2021) Shell Global. Available at: https://www.shell.com/business-customers/catalysts-technologies/licensed-technologies/refinery-technology/shell-blue-hydrogen-process.html
[11] Green hydrogen economy - predicted development of tomorrow (2021) PwC. Available at: https://www.pwc.com/gx/en/industries/energy-utilities-resources/future-energy/green-hydrogen-cost.html
[12] Heat Pump Systems (2015) Energy.gov. Department of Energy. Available at: https://www.energy.gov/energysaver/heat-pump-systems
[13] Installing heat pumps in the race to net zero (2022) RPS. Available at: https://www.rpsgroup.com/sectors/energy-consultants/renewable-energy/renewable-energy-markets/heat-pumps/
[14] Air Source Heat Pumps Explained | Low Carbon Heating (2020) EDF. Available at: https://www.edfenergy.com/heating/advice/air-source-heat-pump-guide
[15] Sajip, J. (2017) Types of Electric Heat Pumps and Their Advantages, MEP Engineering & Design Consulting Firm. New York engineers. Available at: https://www.ny-engineers.com/blog/types-of-electric-heat-pumps-and-their-advantages
[16] Olympios, A.V. et al. (2022) “Delivering net-zero carbon heat: Technoeconomic and whole-system comparisons of domestic electricity- and hydrogen-driven technologies in the UK,” Energy Conversion and Management, 262, p. 115649. Available at: https://doi.org/10.1016/j.enconman.2022.115649.
[17] Olympios, A.V. et al. (2020) “On the value of combined heat and power (CHP) systems and heat pumps in centralised and distributed heating systems: Lessons from multi-fidelity modelling approaches,” Applied Energy, 274, p. 115261. Available at: https://doi.org/10.1016/j.apenergy.2020.115261.
[18] Scope 1 and Scope 2 Inventory Guidance (2020) EPA. Environmental Protection Agency. Available at: https://www.epa.gov/climateleadership/scope-1-and-scope-2-inventory-guidance
[19] Olympios, A. et al. (2020) “Optimal design of low-temperature heat-pumping technologies and implications to the whole-energy system,” The 33rd International Conference on Efficiency, Cost, Optimization, Simulation and Environmental Impact of Energy Systems [Preprint].
[20] Song, J. et al. (2021) “Integrated Organic Rankine Cycle (ORC) and heat pump (HP) systems for domestic heating,” 34th International Conference on Efficiency, Cost, Optimization, Simulation and Environmental Impact of Energy Systems (ECOS 2021) [Preprint]. Available at: https://doi.org/10.52202/062738-0143.
[21] Scoccia, R. et al. (2018) “Absorption and compression heat pump systems for space heating and DHW in European buildings: Energy, environmental and economic analysis,” Journal of Building Engineering, 16, pp. 94–105. Available at: https://doi.org/https://doi.org/10.1016/j.jobe.2017.12.006.
[22] Petrović, S.N. and Karlsson, K.B. (2016) “Residential heat pumps in the future Danish Energy System,” Energy, 114, pp. 787–797. Available at: https://doi.org/10.1016/j.energy.2016.08.007.
[23] Wang, Y. and He, W. (2021) “Temporospatial techno-economic analysis of heat pumps for decarbonising heating in Great Britain,” Energy and Buildings, 250, pp. 111–198. Available at: https://doi.org/10.1016/j.enbuild.2021.111198.
[24] Moret, S. (2017) “Strategic energy planning under uncertainty.” Available at: https://doi.org/10.5075/epfl-thesis-7961.
[25] Olympios, A.V. et al. (2021) “Library of price and performance data of domestic and commercial technologies for low-carbon energy systems.” Available at: https://doi.org/ 10.5281/zenodo.5758943.
[26] Cambridge Housing Model (BEIS) (2016) Cambridge Energy. Available at: https://cambridgeenergy.org.uk/project/cambridge-housing-model-decc/.
[27] Plans unveiled to decarbonise UK Power System by 2035 (2021) GOV.UK. Department for Business, Energy & Industrial Strategy. Available at: https://www.gov.uk/government/news/plans-unveiled-to-decarbonise-uk-power-system-by-2035.
[28] Sadler, D., Anderson, H.S. and Sperrink, M. (2018) H21 North of England, H21. Available at: https://h21.green/projects/h21-north-of-england/.
[29] Review of Electricity Market Arrangements (2022) GOV.UK. Department for Business, Energy & Industrial Strategy. Available at: https://www.gov.uk/government/consultations/review-of-electricity-market-arrangements.
[30] Edwards, T. (2022) Energy prices to remain significantly above average up to 2030 and beyond, Cornwall Insight. Available at: https://www.cornwall-insight.com/press/energy-prices-to-remain-significantly-above-average-up-to-2030-and-beyond/.
[31] Energy Price Guarantee (2022) GOV.UK. Available at: https://www.gov.uk/government/publications/energy-bills-support/energy-bills-support-factsheet-8-september-2022.
[32] Schmitt, A. (2022) EU Energy Outlook 2050: How will the european electricity market develop over the next 30 years?, Energy BrainBlog. Available at: https://blog.energybrainpool.com/en/eu-energy-outlook-2050-how-will-the-european-electricity-market-develop-over-the-next-30-years/.
[33] United Kingdom Electricity Price - 2022 Data - 2013-2021 Historical - 2023 Forecast (2013) Trading Economics. Available at: https://tradingeconomics.com/united-kingdom/electricity-price.
[34] Baldino, C., O’Malley, J., Searle, S., Zhou, Y. and Christensen, A., 2020. Hydrogen for heating? Decarbonization options for households in the United Kingdom in 2050. International Council of Clean Transportation.
[35] H.J., Gérard, F., van Nuffel, L., Smit, T., Yearwood, J., Černý, O., Michalski, J. and Altmann, M., 2020. Opportunities for Hydrogen Energy Technologies Considering the National Energy & Climate Plans.
[36] Parkes, R. (2022) 'Green hydrogen imports will be cheaper than locally produced H2 in Europe from 2024': study, Recharge. Available at: https://www.rechargenews.com/energy-transition/green-hydrogen-imports-will-be-cheaper-than-locally-produced-h2-in-europe-from-2024-study/2-1-1218185.
[37] Green hydrogen imports will be cheaper than locally produced H2 in Europe from 2024 (2022) Globuc. Available at: https://globuc.com/news/green-hydrogen-imports-will-be-cheaper-than-locally-produced-h2-in-europe-from-2024-study/.
[38] (2020) Gas Goes Green Hydrogen Cost to Customer Report - November 2020. rep. Energy Networks Association. Available at: https://www.energynetworks.org/assets/images/Project%20Altair_H2%20Cost%20to%20the%20Customer_Nov%20update%20v4_final.pdf.
[39] Chestney, N. (2022) High gas prices spur green hydrogen investment -report, Reuters. Thomson Reuters. Available at: https://www.reuters.com/business/energy/high-gas-prices-spur-green-hydrogen-investment-report-2022-10-19/.
[40] Watkins, A. (2022) Climate change insights, natural and rural environments, UK, Office for National Statistics. Available at: https://www.ons.gov.uk/economy/environmentalaccounts/articles/climatechangeinsightsuk/november2022.
[41] Development of trajectories for residential heat decarbonisation to inform the Sixth Carbon Budget (Element Energy) (2020) Climate Change Committee. Available at: https://www.theccc.org.uk/publication/development-of-trajectories-for-residential-heat-decarbonisation-to-inform-the-sixth-carbon-budget-element-energy/.
[42] ESO Data Portal: Historic Generation Mix & Carbon Intensity (2020) nationalgridESO. Available at: https://data.nationalgrideso.com/carbon-intensity1/historic-generation-mix.
[43] Gorski, J., Wu, K.T. and Jutt, T. (2021) Carbon intensity of blue hydrogen production, Pembina Institute. Available at: https://www.pembina.org/pub/carbon-intensity-blue-hydrogen-production.
[44] Hampstead (Greater London) UK climate averages (2020) Met Office. Available at: https://www.metoffice.gov.uk/research/climate/maps-and-data/uk-climate-averages/gcpv7fnqu.

CFD Modelling of a Concentrated Photovoltaic-Thermal System with a Spectral Beam Splitter

Siew Wen Ng and Yong Yan Chai
Department of Chemical Engineering, Imperial College London, U.K.

Abstract

A numerical model of a spectral beam splitting-concentrated photovoltaic thermal (SBS-CPVT) system is developed and used to model such a collector. The spectral beam splitter (SBS) is specified to be a double-layer dichroic glass filter which transmits wavelengths 500-1100 nm to silicon-based photovoltaic (PV) cells and reflects the remaining spectrum to a thermal absorber. A 2D modelling approach in COMSOL Multiphysics was used to obtain the transmitted fraction and the optimum configuration for the SBS. A transmitted fraction of 0.52 was obtained with a convex SBS selected. The 2D results were then translated to a Multiphysics 3D model in order to evaluate system performance. Results show that incorporating an SBS into a CPVT system improved PV cell efficiency by 7.5%. The SBS-CPVT system developed also reported an optical and thermal efficiency (waste heat recovered at the PV cells only) of 90% and 33.7% respectively. The average cell temperature of the SBS-CPVT system also decreased by 14 °C to 41 °C compared to a non-SBS system with the outlet temperature of the heat transfer fluid reaching 38 °C. A parametric study on the effect of solar irradiance (I_0), ambient temperature (T_{amb}), windspeed (v), mass flow rate ($\dot{m}$) and inlet temperature (T_{in}) of the heat transfer fluid on the system performance was also conducted, revealing that a set of parameter values may be selected according to system performance requirements. The model developed can also be used to evaluate various SBS material, system design and absorber design.

1. Introduction

The UK expects to reach net zero carbon emissions by 2050, with solar energy being one of the most cost-effective renewables to achieve this target[1].

Solar energy can be harvested using solar photovoltaic (PV) cells which are typically p-n junction semiconductors. When photons in sunlight strike the surface of the cell, electron-hole pairs are generated within the depletion layer. This creates an electric field which results in separation of the holes and electrons into the p-n junction respectively. This separation continues until the difference in electrons and holes across the p-n junction becomes so great, giving rise to a potential difference across the junction. This drives electrons to holes in the p-junction through a circuit upon connecting a load to the cell. Cells are either single-junction or multi-junction. Multi-junction cells (GaAs-based) generally outperform single-junction cells, giving a record efficiency of 47.1%[2], compared to single-junction cells with a maximum theoretical efficiency of 33%[3]. This is due to reduced thermalisation and recombination losses in multi-junction cells[4]. However, they will likely be more costly due to their requirement of expensive materials[5]. The remaining solar energy that was not utilised will be wasted as heat.

Solar energy can also be extracted by converting it directly into heat. Such systems are referred to as solar thermal systems. Other than solar thermal systems, hybrid PV-thermal (PVT) systems have also been explored in the past. A PVT system is a combination of solar PV and solar thermal systems which can generate electrical and thermal energy simultaneously. PVT systems of different configurations have been developed in the past to utilise this waste heat as solar thermal energy, such as the flat-plate PVT and concentrated PVT (CPVT). Flat-plate PVTs usually consist of glass covers, fluid mediums, PV cells, absorbers, and a layer of insulation[6]. One of the challenges with PVT systems is that the thermal energy generated is of a low grade (< 60 °C), which can be improved by using a CPVT system. CPVT systems typically consist of parabolic concentrators, thermal absorbers, PV cells and a heat transfer fluid (HTF). Rays reflected from the parabolic concentrator will converge onto a focal point on the PV cells, and result in a high electrical and thermal output from a small PV cell area.

The PV cells in such systems are commonly Si-based, for which the usable solar spectrum, i.e., the bandgap, is about 500-1100 nm[7]. Only photons within the bandgap will be utilised for electricity conversion by the PV cells, while the excess spectrum incident on the PV cells will be dissipated as waste heat. This causes cell temperature to rise, especially under concentrated solar irradiation. Consequently, conversion efficiency can drop by as much as 0.08% K^{-1}, as observed by Radziemska[8]. To overcome this challenge, spectral beam splitting has been considered for hybrid (PVT) systems. A spectral beam splitter (SBS) will split the incident spectrum by transmitting the usable spectrum for electricity generation to the PV cells while directing the remaining spectrum to a thermal absorber, which also allows for more efficient use of the incident solar irradiation. Since then, various SBSs have been developed, including interference filters and liquid absorptive filters[9]. Interference filters, such as dichroic filters for CPVT systems are commonly thin multi-layer materials with non-absorptive high or low refractivity[9]. The incident spectrum splits at the layer boundaries of the SBS where a specified bandwidth is reflected and the remaining transmitted.

Liquid absorptive filters have been increasingly popular in recent years compared to dichroic filters, particularly nanofluids, since they can act as spectral splitters, carrier fluids for the thermal energy generated and coolants for the PV cells. While nanofluids offer multiple functions, more research needs to be done to enhance its stability under concentrated irradiance and high temperature[10]. In contrast, multi-layer dichroic

filters are relatively more stable and do not degrade as easily[11].

Hence, this paper aims to develop a 3D coupled optical-thermal CFD model of a CPVT system which utilises a dichroic double-layer SBS and evaluate the model's performance under varying operational conditions. The summary of this paper is as follows:

- A brief overview on the development of SBS-CPVT systems has been provided in Section 2 to highlight the significance of developing a model for such systems.
- The numerical modelling methodology is presented in Section 3.
- The model developed and a sensitivity analysis on the system performance is discussed in Section 4.
- The conclusion of this study is presented in Section 5.
- Future outlooks for this study are suggested in Section 6.

2. Background

The development of a 3D model that can simulate the solar radiation propagation process, heat transfer process and fluid flow in the solar collection system is crucial in the effort to further develop CPVT systems, optimum SBSs and so on. Literatures related to SBS incorporated solar energy collection system had been reviewed.

The waveband allocation for ray splitting by the SBS had been evaluated extensively to determine the optimum ray splitting properties that would achieve the desired system performance. The effect of the SBS cut-off wavelength on a system's electrical, thermal, and overall performances have been studied by Kandil et al.[12]. In their study, GaAs type PV cell and a dichroic mirror SBS has been used. Rays of a low wavelength are directed onto the PV cell while rays of a high wavelength are directed to a thermal absorber. It was reported that the electrical efficiency of the PV cell increases when more energy is directed to the PV cell. However, the rate of increase reduces when a larger portion of radiation is directed onto the PV cell due to thermal effects of the cell, resulting in a lower external quantum efficiency (EQE). Zhu, Li, and Yu[13] also reported a similar observation. The effect of the SBS splitting waveband on different types of solar cells were also studied by Gao et al.[14]. For GaAs type cells, which has short spectrum range (410-890 nm), it was found that when the SBS splitting waveband matches the solar cell's active spectrum range, a high EQE can be achieved.

Various types of SBS material have also been explored. Djafar et al[15] investigated a cold mirror and hot mirror type SBS. A cold mirror SBS is a dichroic filter that transmits infrared rays and reflects visible light, while a hot mirror SBS transmits and reflects rays in the opposite way. A multi-layer film-based SBS was designed by Wang et al[16] for a CPVT system. Besides that, a combined fluid and solid based SBS was proposed by Han et al[17]. Volumetric absorption was the ray splitting mechanism employed by the combined fluid-solid based SBS. The fluid-based filter material is an infrared light absorber while the solid-based filter material is a visible light absorber which absorbs short wavelength light not utilised by the Si-based PV cells. Improvements on the uniformity of flux on PV cells was observed when the combined fluid-solid based SBS was used.

The thermal energy utilisation method and its techno-economic viability had also been investigated. Wang et al.[18] reported a 1500 tonne CO_2 decarbonisation potential per year for the processing of milk products by using the thermal output of a CPVT system to produce steam along with electricity generation using the PV cells. The system was found to be economical as long as the investment cost of the SBS is lower than 85% of the concentrator cost. Peacock et al.[19] explored the integration of an organic Rankine cycle (ORC) into an SBS-CPVT system. The thermal energy harvested from that SBS-CPVT system was used to provide hot water and heating. The technoeconomic viability of this system was then examined at various locations. The study concluded that this setup is only economically feasible on a larger scale.

Several developed CPVT systems with SBSs were also reviewed. Wang et al.'s[20] system mainly consisted of flat mirrors arranged into a concentrator, secondary parabolic reflector, thermal receiver, and monocrystalline Si-cells covered by a 13-layer thin-film optical filter (from bottom to top). The average transmittance of the optical filter was 0.721 within the 250-2500 nm wavelength region. A numerical model of this system reported a PV efficiency of 30.5% and overall system efficiency 26.6% respectively. Widyolar et al.[21] simulated an ideal interference filter for a conventional SBS-CPVT system configuration with c-Si cells. The paper reported a PV cell efficiency of 23% with an SBS bandpass of 504-1126 nm and transmittance of 0.9. Interference filters were also shown to have superior transmission, reflection and spectral control compared to back-reflecting cell systems. Huaxu et al.[22] experimented on an SBS-CPVT system utilising Fresnel lens as the concentrator. The optical splitting film was an SiO_2/TiO_2 interference thin-film filter which reflects wavelengths of 400-1100 nm to the Si-cells, while transmitting wavelengths of 1100-2500 nm to a thermal absorber. Results demonstrated a reduction in cell temperature by 11 K. These studies all observed an improvement in system performance for an SBS-CPVT system compared to the same system without an SBS.

Most literature focused on PVT system with fluid-based or multi-layer film-based SBSs for which the optical properties of the SBS (i.e., transmittance and reflectance) can be altered relatively easily. Whereas research on optimum geometrical configurations of interference filter type SBSs, such as dichroic filters, is scarce. In conjunction to previous studies, this paper seeks to broaden existing research on SBS-CPVT systems by performing optical analysis on several SBS configurations and develop a 3D coupled optical-thermal model of the complete system.

3. Methodology

3.1 System description

The system examined in this study is a concentrated photovoltaic-thermal (CPVT) system that consists of a parabolic concentrator, a thermal absorber, a spectral beam splitter (SBS), and PV cells mounted onto the outer wall of a triangular duct containing water as the heat transfer fluid (HTF). The geometrical configuration of the system is shown in Figure 3.1.

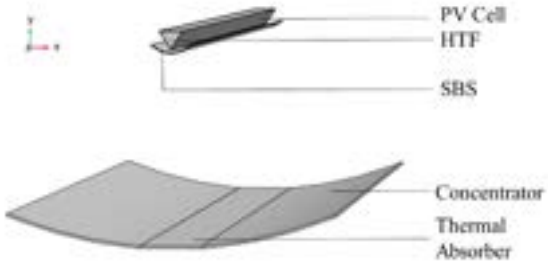


Figure 3.1: CPVT system configuration

The concentrator is a reflector with a reflectivity of 0.90, where radiation from the sun is reflected. The irradiance from the sun was set at a constant of 1000 W m^{-2} and incident perpendicularly to the reflection surface. The reflected rays were then concentrated towards the SBS. The SBS splits the radiation into transmitted and reflected rays. The transmitted rays were received by the PV cells while the reflected rays were absorbed by the thermal absorber.

3.1.1 System geometry

The distance between the concentrator, SBS and the triangular duct were determined according to the focal lengths of the concentrator and the SBS. The focal points of both the SBS and the parabolic concentrator fell at the centre of the triangular duct. The focal lengths of the concentrator (f_c) and SBS (f_s) were set at 420 mm and 25.2 mm respectively. The aperture of the concentrator (a_c) and the SBS (a_s) were specified at 670 mm and 84 mm respectively. The geometry of the system is shown in Figure 3.2. The thickness of all surfaces was set to 1 mm. The length of the CPVT system was chosen to be 3 m.

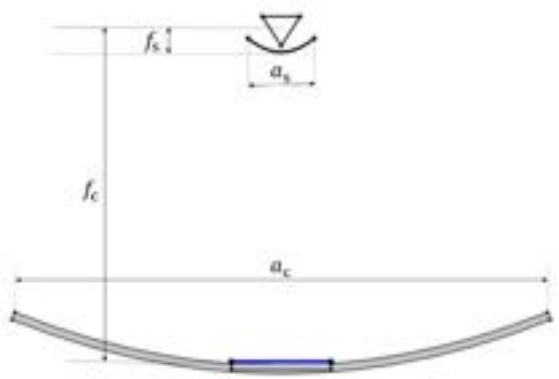


Figure 3.2: System geometry

The geometrical concentration ratio (GCR), optical concentration ratio (OCR) and optical efficiency (η_{opt}) of the system were calculated. As the concentration ratios and optical efficiency are an indicator of concentrator performance only, the SBS was excluded from the calculations. The GCR is the ratio of the effective concentrator area (A_c) to the total PV cell surface area (A_{cell}), as described by Eq. 3.1[23]:

$$GCR = \frac{A_c}{A_{cell}} \qquad \text{Eq. 3.1}$$

The OCR is defined as the ratio of the total incident flux on the PV cells (I_{cell}) to the total source flux, i.e., solar irradiation (I_0), as described by Eq. 3.2.

$$OCR = \frac{I_{cell}}{I_0} \qquad \text{Eq. 3.2}$$

The optical efficiency is defined as the OCR to GCR ratio, or the total power at the PV cells to total source power ratio, as described in Eq. 3.3.

$$\eta_{opt} = \frac{OCR}{GCR} = \frac{I_{cell}A_{cell}}{I_0 A_c} \qquad \text{Eq. 3.3}$$

The GCR, OCR and optical efficiency of the system in this study were found to be 6.2, 5.6 and 90% respectively.

3.2 Computational modelling

COMSOL Multiphysics is used to analyse geometrical optics and evaluate the heat transfer and fluid flow in the CPVT system. Propagation of concentrated rays passing through the SBS was simulated using a 2D optical model. The splitting proportion (transmitted fraction) of rays obtained through the 2D geometrical optics simulation was then imported to the 3D model. Heat transfer across the solid and fluid domains as well as fluid dynamics were evaluated in the 3D model through Multiphysics coupling. Splitting of rays at the SBS was excluded in the 3D model due to computational constraint. Using the 3D model developed a priori, a sensitivity analysis on the system performance was also conducted.

A finite element analysis method is employed by COMSOL in which the geometry domain is discretised into elements which form a mesh for computation. The stability of the results and computational cost are affected by the defined mesh size whereby a finer mesh size corresponds to a more stable result. However, a higher computational cost is associated with finer mesh sizes. The maximum mesh size that is compatible with the geometry and produces a stable result was determined. This mesh was then used in order to minimise computation costs.

3.2.1 2D model

A 2D model was developed to study ray propagation in the system. A non-sequential ray-tracing approach was taken by COMSOL geometrical optics to solve for the ray propagation pathway of the geometry. The direction of ray propagation is constant until the path intersects with a boundary that separates two mediums with different refractive indices. Rays at a boundary can be either refracted, reflected, or absorbed.

For refracted rays, the propagation of rays from a boundary is determined by the refractive index of the ray propagation medium. The direction of the refracted ray is extrapolated out from a boundary according to Snell's law[24], which is described by Eq. 3.4:

$$n_1 \sin\theta_i = n_2 \sin\theta_t \qquad \text{Eq. 3.4}$$

where n is the refractive index, θ_i is the angle of incidence and θ_t is the angle of transmittance. The medium which the rays propagate is indicated by subscripts 1 and 2. The speed at which the rays propagate through a medium (c) is determined using Eq. 3.5[24].

$$c = \frac{Speed\ of\ light\ in\ vacuum}{Refractive\ index} \quad \text{Eq. 3.5}$$

For reflected rays, the law of reflection which states that the angle of incidence is equal to the angle of reflectance is obeyed. The geometrical optics was evaluated with the assumption that the wavelength of the radiation is insignificant relative to the dimensions of the system. The negligence of the diffraction of rays is allowed following this assumption.

Light rays from the sun were simulated as polychromatic light with wavelengths in the range of 5 nm to 4500 nm. The concentrator was simulated as an illuminated boundary in the 2D model. Boundary conditions were specified to model the filter property of the SBS. Rays with wavelengths within the bandgap of 500-1100 nm was set to pass through the boundary while the remaining rays were reflected. The refractive index was defined as 1 for the SBS material. The refractive index value is dependent on the SBS material used. The reflection coefficient of the SBS was specified as 0.77[25], which accounts for the radiation absorbed by the SBS material. The properties of the SBS in this model are outlined in Table 3.1.

Table 3.1: Properties of the SBS in the 2D model

Property	Value
Transmitted wavelength	500-1100 nm
Refractive index	1
Reflection coefficient	0.77

Material discontinuities were specified at the outer surface of the SBS to exclude the release of reflected rays at these boundaries, which will contribute to stray light rays that produce noise in the result.

3.2.2 3D model

A 3D model was developed to simulate heat transfer and fluid flow in the CPVT collection system. The absorption and conversion of the optic rays into heat energy at the PV cells were represented in the 3D model. Heat fluxes within the geometry were determined by considering heat transfer in the solid and fluid domain.

The incident rays were set as monochromatic, with a wavelength of 660 nm in the 3D model to minimise computational cost. The number of incident rays was limited to 1000. More accurate results can be obtained by using higher number of rays at the expense of a higher computational cost. A balance between result accuracy and computational time was taken into account. Aluminium and copper were specified as the concentrator and triangular duct wall material respectively.

Rays absorbed by the PV cell were converted into heat and electricity. The total radiation reaching the PV cell (S) is given by Eq. 3.6, where r_{conc} is the reflectivity of the concentrator and τ_{SBS} is the transmitted fraction of the SBS. The value of τ_{SBS} was obtained through geometrical optics simulation across the SBS using the 2D model.

$$S = (I_0 A_c) r_{conc} \tau_{SBS} \quad \text{Eq. 3.6}$$

The radiation reaching the PV cell is converted into electrical and heat energy. The heat energy generated at the PV cell ($\dot{q}$) is given by Eq. 3.7, where η_{PV} is the electrical efficiency and α_{PV} is the absorptivity of the PV cell. The heat energy generated is collected by the HTF in the duct.

$$\dot{q} = S\alpha_{PV}(1 - \eta_{PV}) \quad \text{Eq. 3.7}$$

The electrical efficiency of the PV cell (η_{PV}) is dependent on PV cell temperature. This dependency is described by Eq. 3.8[26], where β is the temperature coefficient and T_{cell} is the average PV cell temperature in °C. The base electrical efficiency (η_0) is the electrical efficiency at a reference state where the cell temperature is at 25 °C and I_0 is 1000 W m^{-2}.

$$\eta_{PV} = \eta_0(1 - \beta(T_{cell} - 25°C)) \quad \text{Eq. 3.8}$$

In the solid domain, thermal energy is transferred from the PV cell to the HTF via conduction. Heat transfer in the solid domain is governed by Eq. 3.9, where k_s is the thermal conductivity of the solid, T_s is the temperature of the solid domain and $\dot{q}_s$ is the heat generated in the solid domain.

$$k_s \nabla^2(T_s) + \dot{q}_s = 0 \quad \text{Eq. 3.9}$$

The thermal boundary condition at the PV cell is described by Eq. 3.10, where $\boldsymbol{n}$ is the direction vector and q_{out} is the outward heat loss to the surroundings. These boundary conditions correspond to the convective heat flux to the surrounding and the surface-to-ambient radiation at the PV cell.

$$-\boldsymbol{n} \cdot q_{out} = q_0 \quad \text{Eq. 3.10}$$

For convection, q_0 is:

$$q_{conv} = h(T_{amb} - T) \quad \text{Eq. 3.11}$$

For surface-to-ambient radiation, q_0 is:

$$q_{rad} = \varepsilon\sigma({T_{sky}}^4 - T^4) \quad \text{Eq. 3.12}$$

where h is the convective heat transfer coefficient, T_{amb} is the ambient temperature, T is the cell temperature, T_{sky} is the sky temperature calculated using the correlation specified in Eq. 3.14[27], ε is the cell surface emissivity and σ is the Stefan-Boltzmann constant. The convective heat transfer coefficient is given by the correlation specified in Eq. 3.13, where v is the surrounding windspeed in m s^{-1} [28].

$$h = 2.8 + 3v \quad \text{Eq. 3.13}$$

$$T_{sky} = 0.0552{T_{amb}}^{1.5} \quad \text{Eq. 3.14}$$

The thermal boundary conditions assigned to the fluid domain are given by Eq. 3.15 for inflow and Eq. 3.17 for outflow. The inflow condition corresponds to the heat energy of the HTF at inlet. The outflow condition is specified to account for convection-dominated heat transfer at the outlet.

$$-\boldsymbol{n} \cdot q_{inlet} = \rho_f \Delta h \boldsymbol{v_f} \cdot \boldsymbol{n} \quad \text{Eq. 3.15}$$

$$\Delta h = \int_{T_{inlet}}^{T} c_p \, dT \quad \text{Eq. 3.16}$$

The inflow heat is given by q_{inlet}. ρ_f is the fluid density, Δh is the specific enthalpy change calculated from Eq. 3.16, $\boldsymbol{v_f}$ is the fluid velocity, c_p is the specific heat capacity of the fluid, T is the immediate fluid temperature and T_{inlet} is the fluid temperature at the inlet.

$$-\boldsymbol{n} \cdot q_{\text{outlet}} = 0 \qquad \text{Eq. 3.17}$$

For the fluid domain, the flow regime was determined using the Reynolds number (Re). Eq. 3.18 was used to calculate Re, where u is the fluid flow speed, d_h is the hydraulic diameter and μ is the dynamic viscosity. The hydraulic diameter is calculated from Eq. 3.19 where A is the cross-sectional area of the duct and p is the "wetted" perimeter of the duct.

$$Re = \frac{\rho_f u d_h}{\mu} \qquad \text{Eq. 3.18}$$

$$d_h = \frac{4A}{p} \qquad \text{Eq. 3.19}$$

The critical Re for laminar flow in a triangular duct is 1100 [29].

Navier-Stokes equations were used to model the fluid flow. Conservation of mass, momentum, and energy, described by Eq. 3.20, 3.21 and 3.22 respectively were satisfied.

$$\rho_f \nabla \cdot \boldsymbol{v_f} = 0 \qquad \text{Eq. 3.20}$$

$$\rho_f \boldsymbol{v_f} \nabla \cdot \boldsymbol{v_f} = -\nabla P + \nabla \cdot \boldsymbol{\tau} + \rho_f \boldsymbol{g} \qquad \text{Eq. 3.21}$$

$$\rho_f c_p \boldsymbol{v_f} \nabla \cdot T_f = \nabla \cdot (k_f \nabla T_f) \qquad \text{Eq. 3.22}$$

where P is the fluid pressure, $\boldsymbol{\tau}$ is the viscous stress tensor, $\boldsymbol{g}$ is gravitational acceleration, T_f is the fluid temperature and k_f is the conductivity of the fluid.

The fluid flow was assumed to be incompressible, in which the fluid density is expected to be constant with respect to time and space. The fluid was specified as a Newtonian fluid in which its dynamic viscosity is dependent on its thermodynamic state. This assumption is valid as the fluid in this system was chosen to be water, which obeys the properties of a Newtonian fluid under normal conditions.

For the fluid domain, a no-slip boundary condition was specified at the duct wall. The inflow boundary condition was given by the mass flow rate of the HTF, flowing in the normal direction to the boundary. The pressure at the outflow boundary was specified at 0 Pa.

3.3 System performance

Using the 3D model, the effects of the solar irradiance, ambient temperature, surrounding windspeed, mass flow rate of the HTF and inlet temperature of the HTF on the system performance were examined. The outlet temperature of the HTF and average PV cell temperature were obtained from COMSOL. System performance with and without an SBS was also evaluated by setting the transmitted fraction of the SBS to 1 (for non-SBS system) or 0.52 (for an SBS incorporated system).

Here, the performance indicators of the system are the PV cell efficiency, determined by Eq. 3.8, and the thermal efficiency of the HTF, determined by Eq. 3.23,

$$\eta_{th} = \frac{\dot{m} c_p \Delta T}{I_0 A_c} \qquad \text{Eq. 3.23}$$

where $\dot{m}$ is the mass flow rate of the HTF and ΔT is the temperature difference between the inlet and outlet of the HTF.

4. Results & discussion

4.1 2D modelling

The 2D model for the SBS-CPVT system was used for geometrical optics simulation to select the optimum SBS configuration which minimises optical losses, i.e., ray divergence.

A convex and concave SBS were considered with a flat plate SBS as a control. As the SBS material was assigned a unity refractive index, the transmitted rays through the upper SBS layer did not refract from their initial trajectories for all SBS configurations. For the reflected spectrum, the angle of reflection equals the angle of incidence. Thus, the reflected ray trajectory is dependent on the curvature of the bottom SBS layer.

The concave SBS (Figure 4.1) resulted in the greatest optical loss with no rays received at the thermal absorber. The flat plate SBS (Figure 4.2) demonstrated a focal point far from the thermal absorber, with some rays received by the thermal absorber. The convex SBS (Figure 4.3) was chosen as all reflected rays at the layer boundary fall onto the thermal absorber without any observable ray divergence.

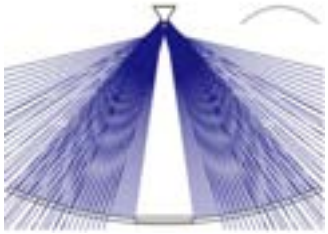


Figure 4.1: Ray trajectories of a concave SBS system

Figure 4.2: Ray trajectories of a flat plate SBS system

Figure 4.3: Ray trajectories of a convex SBS system

The transmitted fraction (τ_{SBS}) of the spectrum through the convex SBS is defined as the ratio of flux transmitted to the PV cells to the flux reaching the SBS boundary. The transmitted fraction for the convex SBS in this study was calculated to be 0.52. This value is

similar to Zhang et al.'s[25] value of 0.558 and Wingert et al.'s[30] value of 0.560. Dichroic mirrors were used as the SBS under a AM1.5 D solar irradiation by Wingert et al.[30].

4.2 3D modelling

4.2.1 SBS-CPVT system performance

The transmitted fraction of 0.52 was then used in the 3D model to account for the SBS component, under the assumption that the component's ray splitting properties remain constant throughout this study. This significantly reduced computational cost as the SBS component and optical simulation can be excluded from the 3D model.

Heat transfer and fluid flow were then coupled into the 3D model to study the system performance. The 3D model was evaluated under the base case condition specified in Table 4.1.

Table 4.1: Base case condition of the system

Parameter	Symbol	Value	Unit
Solar irradiance	I_0	1000	$W\ m^{-2}$
Ambient temperature	T_{amb}	298	K
Windspeed	v	1	$m\ s^{-1}$
HTF mass flow rate	$\dot{m}$	0.01	$kg\ s^{-1}$
HTF inlet temperature	T_{in}	298	K

Under the base conditions, the system achieved an outlet HTF temperature (T_{out}) of 311 K (38 °C), average cell temperature (T_{cell}) of 313 K (41 °C), PV cell efficiency (η_{PV}) of 19 % and thermal efficiency (η_{th}) of 34 %. It should be noted that the thermal efficiency for the SBS-CPVT system in this study only accounts for the waste heat recovered (WHR) by the HTF at the PV cells.

4.2.2 Sensitivity analysis

A parametric study on the effect of the operational parameters on the SBS-CPVT system performance was also conducted. The parameters in question were the base case parameters listed in Table 4.1. One parameter was varied at a time while the rest were kept constant at the base case value. The key performance indicators are the thermal and PV cell efficiency. The average and maximum cell temperature fell below the upper operating temperature limit of 620 K (347 °C) [30] for the silicon PV cells throughout this study.

4.2.2.1 Effect of parameters on the PV cell efficiency

Simulation results demonstrated that the PV cell efficiency decreased with an increase in the solar irradiance (Figure 4.4), ambient temperature (Figure 4.5) and HTF inlet temperature (Figure 4.8).

The transmitted flux through the SBS to the PV cells increases with an increase with solar irradiation. This increases the rate of thermalisation in the PV cells. Thus, the cell temperature rises, causing a drop in PV cell efficiency. Similarly, PV cell efficiency is lower at higher ambient temperature (Figure 4.5) as the cell temperature increases with an increase in the surrounding temperature. The PV cell efficiency decreases as the HTF inlet temperature increases (Figure 4.8). As the HTF inlet temperature increases, the temperature difference between the cell and the HTF decreases. Consequently, this leads to a decrease in the rate of heat transfer from the cells to the HTF, resulting in a higher cell temperature and lower PV cell efficiency.

Alternatively, PV cell efficiency increased with an increase in the HTF mass flow rate (Figure 4.7) and windspeed (Figure 4.6).

As the mass flow rate increases, the larger fluid flow allows more heat to be recovered at the HTF. Hence, more heat is conducted away from the PV cell, causing cell temperature to drop. This improved the PV cell efficiency. At higher windspeed, convection heat losses from the cell to the environment is enhanced. This is because the heat transfer coefficient increases linearly with windspeed (Eq. 3.13). Thus, cell temperature is lower, and PV cell efficiency is enhanced at a higher windspeed.

4.2.2.2 Effects of the parameters on the thermal efficiency

The thermal efficiency (WHR at PV cells) increased at higher solar irradiance (Figure 4.4), HTF mass flow rate (Figure 4.7) and ambient temperature (Figure 4.5).

As mentioned in Section 4.2.2.1, increasing the solar irradiance and ambient temperature leads to an increase in cell temperature. Thus, the thermal output at the HTF increases, which improves the thermal efficiency. Likewise, using a higher mass flow rate enhances the total heat capacity of the HTF, which increases the thermal output at the HTF.

In contrast, the thermal efficiency decreased with an increase in the HTF inlet temperature and windspeed. This is due to a decrease in rate of heat conduction to the HTF and higher rate of convection heat losses to the surroundings respectively. This leads to a lower thermal output at the HTF and consequently, a lower thermal efficiency.

It is also worth mentioning that the PV cell efficiency and thermal efficiency are inversely related with respect to a change in solar irradiance, ambient temperature and windspeed. Thus, improving the PV cell efficiency would compromise the thermal efficiency and vice versa. On the contrary, the thermal and PV cell efficiency are parallelly related with respect to the mass flow rate and inlet temperature of the HTF. As results suggest, increasing the mass flow rate and decreasing the inlet temperature of the HTF can potentially amplify both efficiencies.

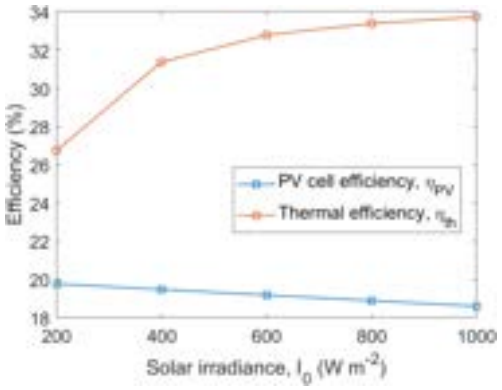


Figure 4.4: Effect of I_0 on the efficiencies at constant $T_{\mathrm{amb}}, v, \dot{m}, T_{\mathrm{in}}$

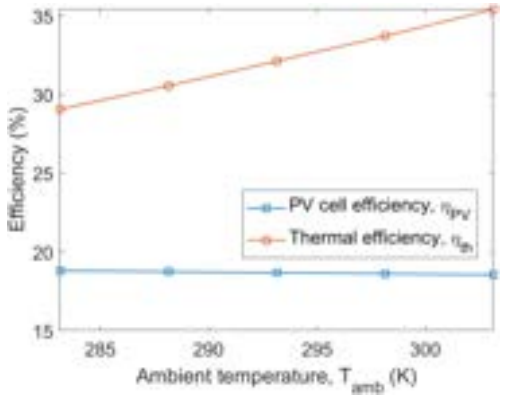


Figure 4.5: Effect of T_{amb} on the efficiencies at constant $I_0, v, \dot{m}, T_{\mathrm{in}}$

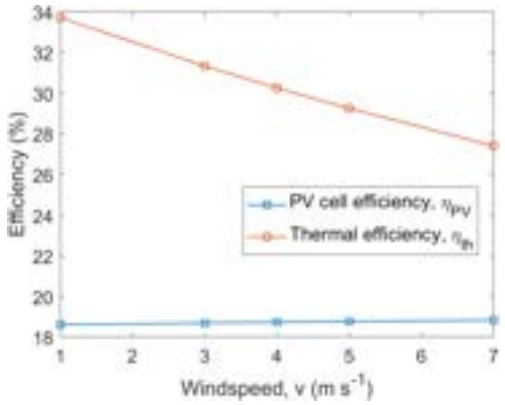


Figure 4.6: Effect of v on the efficiencies at constant $I_0, T_{\mathrm{amb}}, \dot{m}, T_{\mathrm{in}}$

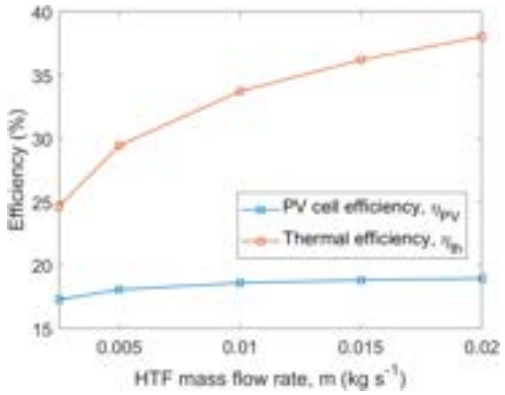


Figure 4.7: Effect of $\dot{m}$ on efficiencies at constant $I_0, T_{\mathrm{amb}}, v, T_{\mathrm{in}}$

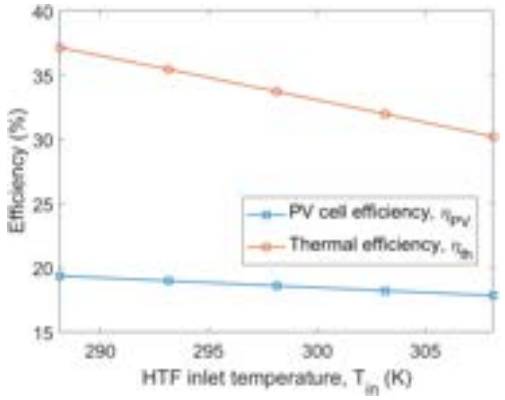


Figure 4.8: Effect of T_{in} on the efficiencies at constant $I_0, T_{\mathrm{amb}}, \dot{m}, v$

4.2.3 Flux & temperature distribution

As discussed in Section 4.2.2, the PV cell and thermal efficiencies depend heavily on the cell temperature, which also depends on the flux incident on the PV cells. Thus, the flux and temperature distribution of the PV cells are explored in more detail here.

Initially, a geometrical optical simulation was run on the 3D SBS-CPVT system. Light rays reflected from the parabolic concentrator is concentrated onto focal points at the PV cells, as shown in Figure 4.9. Thus, the solar flux distribution on the PV cells is non-uniform in the y-direction.

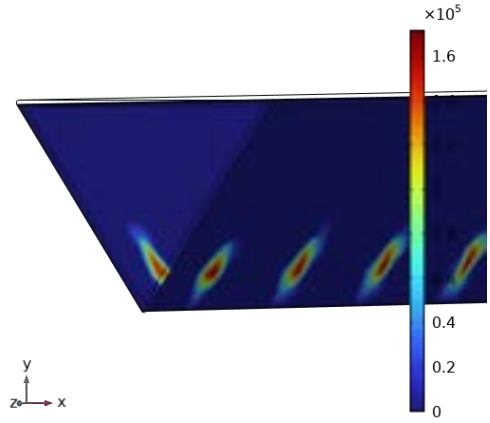


Figure 4.9: Flux distribution on the PV cells under the base case conditions (geometrical optics simulation only)

However, there was computational constraint upon coupling heat transfer, fluid flow and geometrical optics in the 3D model. Hence, an average flux which was uniform across the PV cell surface was obtained from a pure optical simulation on the 3D model. This average flux was used instead of the local flux as the incident flux on the cell surface for the coupled heat transfer and fluid flow simulations. This resulted in a uniform temperature distribution along the y-direction of the cell surface, as shown in Figure 4.10.

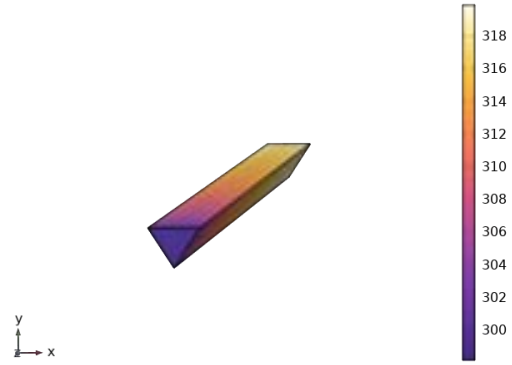


Figure 4.10: Cell surface temperature distribution under base case conditions (by using an average incident flux)

This method was employed as initial simulations using the local flux showed that the temperature variation along the y-direction of the cell surface was negligible, and using the average flux greatly simplified the computational process.

4.2.4 Performance of a CPVT system with & without an SBS

The CPVT system with and without an SBS was also compared. From Figure 4.11, only the flux within the spectral response of the cell will be utilised for electrical energy generation. The remaining flux will be converted to waste heat and contribute to an increase in cell temperature, which decreases PV cell efficiency. This analysis holds true for both systems.

The efficiencies between the two systems are compared in Table 4.2. With an SBS, the PV cell efficiency increased by 7.5%. This is because the SBS only allows a portion of the solar spectrum to be transmitted to the PV cells, where this portion is quantified by the transmitted fraction. Thus, with an SBS, the amount of flux converted to waste heat at the PV cells is considerably a lot less than that of a non-SBS system. This resulted in a lower cell temperature and a higher PV cell efficiency. At the same time, the thermal efficiency of the HTF also decreased. Should the total thermal efficiency be considered, which includes both the thermal output at the HTF and the thermal absorber, the SBS-CPVT system is expected to be superior in both efficiencies as compared to the CPVT system without an SBS.

A sensitivity analysis for the solar irradiance and mass flow rate of the HTF on the efficiencies of the non-SBS system was also conducted. For a change in parameter, the efficiencies exhibited a similar trend to that of an SBS-CPVT system.

Table 4.2: Performance of SBS & non-SBS CPVT systems

CPVT system	Without SBS	With SBS
T_{cell}	328 K (55 °C)	314 (41 °C)
T_{out} of HTF	324 K (50 °C)	311 (38 °C)
η_{PV}	17.3%	18.6%
η_{th}	65.4%	33.7%

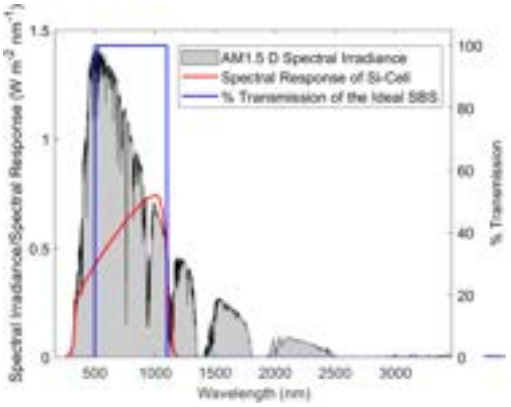


Figure 4.11: Spectral transmittance of the ideal SBS matched to the spectral response of the cells[30] under an AM1.5 D irradiance[30]

5. Conclusions

Numerical modelling of a CPVT system with an SBS was conducted. The SBS was specified to be a double-layer dichroic glass filter. The optimum SBS configuration is a convex one which minimised optical losses. The transmitted fraction of the SBS was calculated to be 0.52 under a 5-4500 nm solar irradiance, for Si-cells with a bandgap of 500-1100 nm. The SBS was modelled to transmit wavelengths within the bandgap to the Si-cells and reflect the remaining spectrum to a thermal absorber. The integration of an SBS lowered the cell temperature by 14 °C to 41 °C and achieved a HTF outlet temperature of 38 °C. Under the same operational parameters, this system was shown to have performed better than one without an SBS with a 7.5% increase in PV cell efficiency. The SBS-CPVT system also achieved an optical and thermal efficiency (WHR at the Si-cells only) of 90% and 33.7% respectively.

For an SBS-CPVT system, a lower solar irradiance, lower ambient temperature, higher windspeed, lower HTF inlet temperature and higher HTF mass flow rate enhances the PV cell efficiency. Whereas the thermal efficiency is improved with a higher solar irradiance, higher HTF mass flow rate, higher ambient temperature, lower HTF inlet temperature and lower windspeed. An 'optimum' set of operating parameter values may be specified according to system requirement.

6. Outlook

To further enhance system performance, we propose several target research directions. Firstly, the fluid flow in this study falls in the laminar flow regime, and it is speculated that increasing the flow rate to the turbulent flow regime will enhance heat transfer via convection in the collector. This will potentially further increase the PV cell efficiency. Furthermore, a glass evacuated tube concentrator (ETC) encasing the PV cells can be added to the system. ETCs are typically designed with air removed from within the tube, thus forming a vacuum within. This reduces conduction and convection heat losses, which can potentially enhance thermal energy collection at the HTF[31]. The effect of this on the PV cell temperature and hence the PV cell efficiency would have to be studied. Finally, an increase in curvature of both the parabolic concentrator and SBS can be considered, which will result in a shorter focal length[32]. Theoretically, this would lead to a more compact system and may further reduce capital costs.

To improve accuracy for future research on similar SBS-CPVT systems employing the current model setup, it is encouraged to incorporate a heat exchanger at the solid-based thermal absorber. This will allow for the total thermal efficiency to be calculated and thermal losses to be simulated. Thus, a complete evaluation of the system performance can be carried out.

Using the CFD model developed, ray splitting properties of various SBS material in similar CPVT systems can be evaluated. Different electrical and thermal energy absorber designs can also be explored to study their effects on the system efficiency.

Acknowledgements

The authors would like to thank Dr. Chandan Pandey for his continuous guidance and support throughout this study. We would also like to extend our gratitude to Professor Christos N. Markides and Dr. Gan Huang for helping us kick off our research in the right direction.

References

1. Hewett C, Cranston S. 2022-A bright future for solar 3. [cited 2022 Nov 8];Available from: https://solarenergyuk.org/resource/natural-capital/
2. Cornet C, da Silva M, Levallois C, Durand O. GaP/Si-Based Photovoltaic Devices Grown by Molecular Beam Epitaxy. Molecular Beam Epitaxy 2018;637–48.
3. Ho JKW, Yin H, So SK. From 33% to 57% – an elevated potential of efficiency limit for indoor photovoltaics. J Mater Chem A Mater [Internet] 2020 [cited 2022 Nov 9];8(4):1717–23. Available from: https://pubs.rsc.org/en/content/articlehtml/2020/ta/c9ta11894b
4. Schygulla P, Müller R, Lackner D, Höhn O, Hauser H, Bläsi B, et al. Two-terminal III–V//Si triple-junction solar cell with power conversion efficiency of 35.9 % at AM1.5g. Progress in Photovoltaics: Research and Applications 2022;30(8):869–79.
5. Xu S. The recent progress and state-of-art designs of Multi-junction Solar Cells. Highlights in Science, Engineering and Technology [Internet] 2022 [cited 2022 Nov 9];5:102–7. Available from: https://drpress.org/ojs/index.php/HSET/article/view/729
6. Jia Y, Alva G, Fang G. Development and applications of photovoltaic–thermal systems: A review. Renewable and Sustainable Energy Reviews 2019;102:249–65.
7. Liang H, Wang F, Cheng Z, Xu C, Li G, Shuai Y. Full-Spectrum Solar Energy Utilization and Enhanced Solar Energy Harvesting via Photon Anti-Reflection and Scattering Performance Using Biomimetic Nanophotonic Structure. 2020 [cited 2022 Nov 28];Available from: https://dx.doi.org/10.30919/esee8c456
8. Radziemska E. The effect of temperature on the power drop in crystalline silicon solar cells. Renew Energy 2003;28(1):1–12.
9. Ju X, Xu C, Han X, Du X, Wei G, Yang Y. A review of the concentrated photovoltaic/thermal (CPVT) hybrid solar systems based on the spectral beam splitting technology. Appl Energy 2017;187:534–63.
10. Yazdanifard F, Ameri M, Taylor RA. Numerical modeling of a concentrated photovoltaic/thermal system which utilizes a PCM and nanofluid spectral splitting. Energy Convers Manag 2020;215:112927.
11. Ju X, Xu C, Han X, Du X, Wei G, Yang Y. A review of the concentrated photovoltaic/thermal (CPVT) hybrid solar systems based on the spectral beam splitting technology. Appl Energy 2017;187:534–63.
12. Kandil AA, Awad MM, Sultan GI, Salem MS. Investigating the performance characteristics of low concentrated photovoltaic systems utilizing a beam splitting device under variable cutoff wavelengths. Renew Energy 2022;196:375–89.
13. Zhu T, Li Q, Yu A. Analysis of the solar spectrum allocation in a spectral-splitting photovoltaic-thermochemical hybrid system. Solar Energy 2022;232:63–72.
14. Gao Y, Yang X, Tan Z, Yang X, Zhang Y, Zhou H, et al. Effects of beam splitting on photovoltaic properties of monocrystalline silicon, multicrystalline silicon, GaAs, and perovskite solar cells for hybrid utilization. Int J Green Energy [Internet] 2022 [cited 2022 Dec 8];Available from: https://www.tandfonline.com/doi/abs/10.1080/15435075.2022.2119855
15. Djafar Z, Salsabila AZ, Piarah WH. Performance comparison between hot mirror and cold mirror as a beam splitter on photovoltaic-thermoelectric generator hybrid using labview simulator. International Journal of Heat and Technology 2021;39(5):1609–17.
16. Wang G, Yao Y, Wang B, Hu P. Design and thermodynamic analysis of an innovative hybrid solar PV-CT system with multi-segment PV panels. Sustainable Energy Technologies and Assessments 2020;37:100631.
17. Han X, Sun Y, Huang J, Zheng J. Design and analysis of a CPV/T solar receiver with volumetric absorption combined spectral splitter. Int J Energy Res [Internet] 2020 [cited 2022 Dec 8];44(6):4837–50. Available from: https://onlinelibrary-wiley-com.iclibezp1.cc.ic.ac.uk/doi/full/10.1002/er.5277
18. Wang K, Pantaleo AM, Herrando M, Pesmazoglou I, Franchetti BM, Markides CN. Thermoeconomic assessment of a spectral-splitting hybrid PVT system in dairy farms for combined heat and power. The 32nd International Conference on Efficiency, Cost, Optimization, Simulation and Environmental Impact of Energy Systems (ECOS 2019) [Internet] 2019 [cited 2022 Dec 8];Available from: https://www.researchgate.net/publication/333759478_Thermoeconomic_assessment_of_a_spectral-splitting_hybrid_PVT_system_in_dairy_farms_for_combined_heat_and_power
19. Peacock J, Huang G, Song J, Markides CN. Techno-economic assessment of integrated spectral-beam-splitting photovoltaic-thermal (PV-T) and organic Rankine cycle (ORC) systems. Energy Convers Manag 2022;269:116071.
20. Wang G, Yao Y, Lin J, Chen Z, Hu P. Design and thermodynamic analysis of a novel solar CPV and thermal combined system utilizing spectral beam splitter. Renew Energy 2020;155:1091–102.
21. Widyolar B, Jiang L, Winston R. Spectral beam splitting in hybrid PV/T parabolic trough systems for power generation. Appl Energy 2018;209:236–50.

22. Huaxu L, Fuqiang W, Ziming C, Yong S, Bo L, Yuzhai P. Performance study on optical splitting film-based spectral splitting concentrated photovoltaic/thermal applications under concentrated solar irradiation. Solar Energy 2020;206:84–91.
23. Calise F, Vanoli L. Parabolic Trough Photovoltaic/Thermal Collectors: Design and Simulation Model. Energies 2012, Vol 5, Pages 4186-4208 [Internet] 2012 [cited 2022 Dec 7];5(10):4186–208. Available from: https://www.mdpi.com/1996-1073/5/10/4186/htm
24. COMSOL. Ray Optics Module User's Guide. 2021 [cited 2022 Nov 23];Available from: www.comsol.com/blogs
25. Zhang JJ, Qu ZG, Wang Q, Zhang JF, He YL. Multiscale investigation of the plasmonic solar cell in the spectral splitting concentrating photovoltaic-thermal system. Energy Convers Manag 2021;250:114846.
26. ZONDAG H. Flat-plate PV-Thermal collectors and systems: A review. Renewable and Sustainable Energy Reviews [Internet] 2008 [cited 2022 Nov 29];12(4):891–959. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1364032107000020
27. Chandan, Suresh V, Iqbal SM, Reddy KS, Pesala B. 3-D numerical modelling and experimental investigation of coupled photovoltaic thermal and flat plate collector. Solar Energy 2021;224:195–209.
28. Chandan, Baig H, ali Tahir A, Reddy KS, Mallick TK, Pesala B. Performance improvement of a desiccant based cooling system by mitigation of non-uniform illumination on the coupled low concentrating photovoltaic thermal units. Energy Convers Manag 2022;257:115438.
29. Cope RC, Hanks RW. Transitional Flow in Isosceles Triangular Ducts [Internet]. UTC; 1972. Available from: https://pubs.acs.org/sharingguidelines
30. Wingert R, O'Hern H, Orosz M, Harikumar P, Roberts K, Otanicar T. Spectral beam splitting retrofit for hybrid PV/T using existing parabolic trough power plants for enhanced power output. Solar Energy 2020;202:1–9.
31. Hudon K. Solar Energy – Water Heating. Future Energy: Improved, Sustainable and Clean Options for our Planet 2014;433–51.
32. Liew NJY, Yu Z, Holman Z, Lee HJ. Parametric study about performances of a solar photovoltaic/thermal hybrid using a spectral beam splitting technique. Journal of Renewable and Sustainable Energy [Internet] 2022 [cited 2022 Nov 8];14(1):013701. Available from: https://aip.scitation.org/doi/abs/10.1063/5.0063382

High Resistance Metal Organic Frameworks (MOFs) Membranes for Organic Solvents

Investigating the effects of nanoparticles (UiO-66 and OPA-UiO-66) on membrane performance

Nisada Sila-On* and Harit Thumrongwongkawin*

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

**All authors contributed equally*

Abstract

Conventional water membranes are easily dissolved by organic solvents. In this paper, this issue was tackled by evaluating polymeric membranes as substitutes, specifically polyether ether ketone (PEEK) membranes. Moreover, nanoparticles were integrated into the PEEK membranes to enhance their performance. Two types of nanoparticles were synthesised: hydrophilic UiO-66 and hydrophobic OPA-UiO-66. Three types of membranes were fabricated, including pristine PEEK membrane, PEEK membrane with UiO-66, and PEEK membrane with OPA-UiO-66. The characteristics of the nanoparticles and membranes were studied using various characterisation techniques such as X-ray diffraction (XRD), X-ray photoelectron spectroscopy (XPS), attenuated total reflectance Fourier transform infrared spectroscopy (ATR-FTIR), field-emission scanning electron microscope (FE-SEM), contact angle, and porosity analysis. Their performances were then tested by permeance and rejection tests against multiple inorganic, organic, polar, and nonpolar solvents. Interestingly, the permeance of polar organic solvent and nonpolar organic solvent were highest for the hydrophilic and hydrophobic membrane, respectively. Modifying PEEK membrane with nanoparticles also proved to increase the membrane rejection drastically from 75.9% to 88.4% and 84.7% by integrating UiO-66 and OPA-UiO-66, respectively, in organic solvent.

Keywords — *Metal organic frameworks, Polymeric membrane, Polyether ether ketone, PEEK, Mixed matrix membrane, Nanoparticles, UiO-66, OPA-UiO-66, Permeance, Rejection*

1. Introduction

Organic solvents are commonly used in many industries including pharmaceutical, petrochemical, and agricultural; these solvents can be carcinogenic or have reproductive hazards and therefore should be separated from the industrial products [1]. Industrial chemical separation accounts for 10-15% of the total energy consumption worldwide, and 80% of the industrial separation energy is the energy-intensive thermal separation processes such as distillation, evaporation, and drying [2] [3] [4]. In this paper, the separation of organic solvents will be explored by membrane technology which has a potential in massively reducing the energy requirement in industrial processes, owing to their lower energy requirement for separation since there are no phase changes occurring apart from pervaporation [5].

Commercial water membranes like polyether sulfone (PES) can be easily dissolved by polar organic solvents, as shown in Fig. 1 [6]. This study will therefore be focused on polymeric membranes, specifically polyether ether ketone (PEEK) membrane, due to its high organic solvent resistance [7]. Polymeric membranes have emerged as a major study area in recent years compared to inorganic membranes because of their lower cost and higher flexibility [8]. The resources required to produce polymeric membranes have become more abundant and therefore cheaper compared to inorganic membranes such as ceramic membranes, water membranes, and zeolite membranes [9].

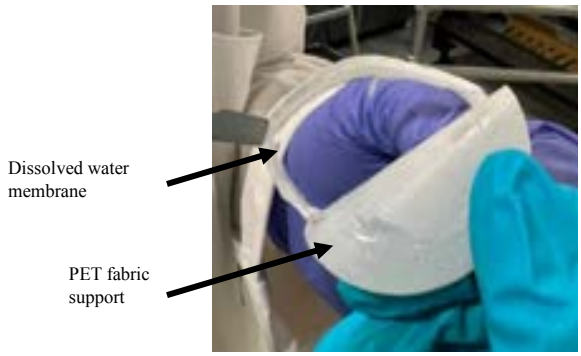


Fig. 1. Image showing water membrane dissolved by organic solvent leaving only PET fabric support layer.

2. Background/Theory

2.1 Inorganic membranes

Inorganic membranes possess advantages such as thermal and chemical stability, and the ability to withstand extreme operating conditions. However, they are not flexible and are highly fragile [8]; making them a poor candidate for organic solvent separation processes.

A possible method of strengthening polymeric membranes to become more resistant to organic solvents is cross-linking of polymers. An example of this modification had been carried out using polyimide, the organic solvent nanofiltration (OSN) membrane, and polybenzimidazole. Moreover, it is possible to use a fundamentally resistant material such as PEEK or poly (ether ketone) (PEK) [10]. Cross-linking was, however, not chosen as the approach to strengthen membranes in this study because the resulting membranes would not be

appropriate to use with chlorinated solvents such as strong amines, strong acids, and bases [10].

2.2 Polyether ether ketone (PEEK) membrane

PEEK membrane was chosen as the focus of this research because of its excellence in resistance towards most organic solvents. PEEK is a semi-crystalline engineered thermal plastic with melting temperature of 340 °C and glass transition temperature of 145 °C. Hydroquinone and benzophenone are two components that combine to create the rigid aromatic backbone structure of PEEK. Due to its resistance, PEEK can only be dissolved by sulphuric acid and methane sulphonic acid at room temperature, making it a suitable material for membrane fabrication but reduces the processability. Therefore, PEEK undergoes a sulphonation reaction (Fig. 2) after being dissolved in sulphuric acid, which modifies its chemical structure in preparation for phase inversion process to fabricate the membrane [10].

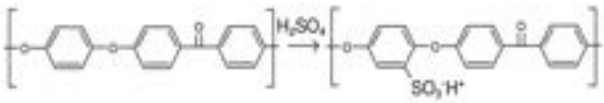


Fig. 2. Schematic principal for the sulphonation of PEEK [10].

2.3 Metal organic frameworks (MOFs)

MOFs are a class of porous crystalline materials made up of inorganic metal ions connected by organic ligands via coordination bonds (linkers) [11]. MOFs have an outstandingly large surface area as they have a hollow, cage-like structure resulting from formation of nodes by metal ions that bind the arms of the linkers together [12]. Moreover, they are superior to other porous nanomaterials owing to their properties such as controllable pore size, tuneable surface chemistry, and adaptable functionalities [13].

Among various MOFs, UiO-66, a Zr-based MOF, has become popular in the research field because of its ability to enhance the membrane hydrophilicity and permeability [14]. UiO-66 possesses various attractive properties such as great chemical resistance towards organic solvents such as benzene and acetone, superior thermal stability, exceptional chemical stability, and excellent resistance to high external pressure [15]. These properties arise due to the existence of the strong Zr-O bond and high coordination number between the Zr clusters and organic ligands [15].

N-Octadecyl phosphonic acid (OPA) is an alkyl phosphonic acid which can interact with the zirconium oxide clusters on the MOF surface by chemisorption through bidentate bonding, as shown in Fig. 3. OPA can integrate the property of super-hydrophobicity to Zr-based MOFs such as UiO-66, without altering the high porosity of the MOFs [16].

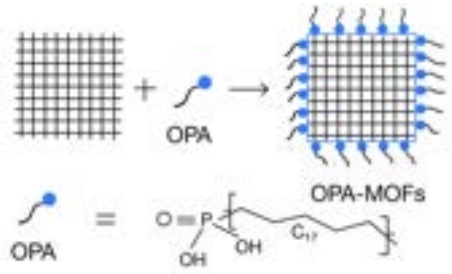


Fig. 3. Molecular level diagram of the addition of OPA to UiO-66 MOFs by chemisorption through bidentate bonding [16].

Furthermore, the two types of nanoparticles were studied because of their different properties: UiO-66 is hydrophilic, and OPA-UiO-66 is hydrophobic.

Five types of solvents were used in this study because of their varying properties. Water is a polar nonorganic solvent, methanol and acetonitrile are strong polar organic solvents, toluene is a non-polar organic solvent, and 1-methyl-2-pyrrolidine (NMP) is a weak polar aprotic organic solvent.

It is expected that polar solvents will have a high permeance through hydrophilic membranes, and nonpolar solvents will have a high permeance through hydrophobic membranes. This is because polar solvents are usually charge-polarized and capable of hydrogen bonding, and nonpolar solvents does not contain a complete or partial charges on their molecules [17].

In this study, organic solvents with varying viscosity were used to test the membrane stability and permeance alongside water.

2.4 Mixed matrix membranes (MMMs)

MMMs are fabricated by dispersing nanoparticles into the polymer membrane matrix. The inorganic nanoparticles allow the improvement of the membrane's mechanical, physical, and thermal properties. MOFs are commonly used fillers to prepare MMMs as they allow the characteristics and performances of the membranes to be modified. They can act as molecular sieves which selectively adsorb solvents and changes the permeability of the membrane depending on the molecular size. MMMs are superior to conventional polymer membranes owing to their improved robustness, permeability, and selectivity [18].

In this study, MOFs were selected as fillers instead of their alternatives such as zeolite and activated carbon [18]. This is due to their ability to prevent internal defects from forming and their compatibility with the polymer matrix [19].

2.5 Membrane fabrication

Phase inversion casting is a widely used method of membranes fabrication. It is a rapid demixing process where a homogeneous liquid state polymer solution is transformed in a controlled manner to a solid state,

forming a solid matrix. There are various techniques of phase inversion such as solvent evaporation, precipitation by controlled evaporation, thermal precipitation, precipitation from the vapour phase, and immersion precipitation. In this study, the technique of immersion precipitation was employed [21].

When the cast film is immersed in a coagulation bath containing a nonsolvent such as water, a film of polymer precipitates instantaneously because of the loss of solvent due to absorption of water by the polymer [20] [21].

The membrane morphology can be controlled by changing the initial stage of phase transition. Whether the membranes are porous or nonporous depends on the type of formation mechanism; instantaneous demixing or delayed onset of demixing, respectively. There are several factors which can affect the diffusion and demixing processes, and hence affect membrane morphology. These factors include the choice of solvent and nonsolvent system, the concentration and composition of the polymer solution, and the composition of the coagulation bath [21].

3. Experimental design

PEEK membrane was studied and further enhanced by adding MOF nanoparticles. The objective was to design membranes that are resistant to organic solvents. Three different types of membranes, pristine PEEK, PEEK modified with UiO-66, and PEEK modified with OPA-UiO-66, were synthesised. They were denoted as PEEK0, PEEK1, and PEEK2, respectively. The membranes fabricated were characterised by various methods to ensure successful modification with both nanoparticles; UiO-66 and OPA-UiO-66. These membranes were then compared in terms of performance (permeance and rejection) in five different solvents with different properties (water, methanol, acetonitrile, toluene, and NMP). For each of the membranes, independent repetitions were conducted three times for improved accuracy and reliability of results.

4. Methods

4.1 Materials

Zirconium (IV) chloride ($ZrCl_4$, Sigma-Aldrich) and terephthalic acid (TPA, Sigma-Aldrich) were used to synthesise UiO-66 nanoparticles. Additionally, OPA (Apollo Scientific Ltd) was used to synthesise OPA-UiO-66 nanoparticles. Dimethylformamide (DMF) and ethanol from VWR Chemicals were used as solvents for the synthesis of UiO-66 and OPA-UiO-66 nanoparticles, respectively. Polyethylene terephthalate (PET) nonwoven fabric (AMFOR Inc., USA) was used as a support layer for the PEEK membrane to withstand the pressure during performance test. Methanol, acetonitrile, toluene, and NMP from VWR Chemicals were used as solvents for the performance test and the rejection test. Polyethylene oxide (PEO, 100,000 Da, Thermo Scientific) was used as rejection material for the rejection test.

4.2 Synthesis of nanoparticles

4.2.1 Synthesis of UiO-66

The UiO-66 nanoparticles were prepared by solvothermal synthesis method. 900 mg of $ZrCl_4$ and 646 mg of TPA were dispersed in 150 mL of DMF in a sonication bath for 1 h. The resulting solution was heated at 100 °C for 24 h in an oven. The solvent was separated from the nanoparticles by vacuum filtration using nylon membrane filters (Sterlitech, 0.1 μm) and washed with DMF, ethanol, and DI water, respectively.

4.2.2 Synthesis of OPA-UiO-66

The OPA-UiO-66 nanoparticles were prepared using 400 mg of the synthesised UiO-66 nanoparticles and 334.5 mg of OPA dispersed in 400 mL of ethanol and sonicated for 10 min. The solution was stirred at room temperature for 24 h. Similarly, the solvent was separated from the nanoparticles by vacuum filtration and washed by ethanol and DI water. The resulting product was put in a vacuum oven at 30 °C for 24 h and then at 100 °C for 24 h connect UiO-66 with OPA by heating treatment.

Both UiO-66 and OPA-UiO-66 were put in a desiccator to remove any moisture. The dried samples were then grinded using a pestle and mortar to achieve fine nanoparticles.

4.3 Membrane fabrication

4.3.1 Casting solution preparation

The solutions were made using 22 wt% of sulfuric acid and 66 wt% of methane sulphonic acid. Three casting solutions differ in the nanoparticle composition with PEEK0 solution consisting of 12 wt% of PEEK, PEEK1 and PEEK2 solution consisting of 11.5 wt% of PEEK and 0.5 wt% of their respective nanoparticles.

For thorough mixing, the solutions were placed in a sonication bath for 24 h to disperse the nanoparticles uniformly and placed in the rolling machine for 48 h to completely dissolve the PEEK powder. The solutions then went through a degassing process to reduce the air bubbles which can affect the membrane morphology.

4.3.2 Membrane casting and phase inversion

The casting solutions were poured onto a PET fabric which was placed on a glass plate, with the casting bar set to thickness of 250 μm. Phase inversion via immersion precipitation occurred when the plate was submerged in coagulating bath containing non-solvent (DI water) for phase inversion to occur. The fabricated membranes were kept in the water container until the performance test.

4.4 Membrane characterisation

4.4.1 X-ray diffraction (XRD)

XRD is a technique employed to analyse the nanoparticle's crystal structure. The XRD patterns were produced using X'pert Pro, PANalytical diffractometer with Cu Kα radiation of the wavelength λ = 1.54178 Å. Additionally, XRD can identify chemical compounds presented in the sample to reassure sample purity [22].

4.4.2 X-ray photoelectron spectroscopy (XPS)

XPS (K-Alpha+, Thermo Scientific, UK) was used for chemical characterisation of membrane nanoparticles by the determination of atomic percentages of each element in nanoparticles. For the XPS analysis, carbon tape was covered completely with each of the nanoparticles before the samples were analysed to prevent getting a reading for carbon from the carbon tape. The monochromated Al Ka Micro-focused x-ray source was operated at around 100-4000 eV, at the resolution of 0.5 eV, and a 400 μm spot size was used.

4.4.3 Attenuated total reflectance Fourier transform infrared spectroscopy (ATR-FTIR)

The ATR-FTIR spectrometer (Perkin-Elmer Spectrum 100) equipped with a universal attenuated total reflectance (UATR) sampling attachment, middle infrared triglycine sulphate (MIR TGS), and a red laser with a wavelength of 633 nm as the excitation source, was used to determine the characteristic peaks for the samples. The samples include nanoparticles, PEEK0, PEEK1, and PEEK2 membranes. The changes in functional groups resulting from addition of UiO-66 and OPA-UiO-66 nanoparticles can be analysed. The spectral range of 4000-500 cm^{-1} was used for each scan sample [10].

Before measuring each sample, isopropyl alcohol was used to clean the ATR crystal.

4.4.4 Field emission scanning electron microscope (FE-SEM)

The changes in structures of the MOF-modified PEEK membranes compared to the PEEK0 membrane were studied using SEM images of the membranes' surfaces and cross-sections. For membrane with PET fabric support layer a sharp blade was used to precisely cut the membrane for cross sectional imaging as liquid nitrogen was not strong enough to cut the PET fabric. Carbon tape was used to keep all the samples in place, then a silver paste was applied on all the samples so that electrons can move through the silver conductor to achieve clearer images. The sputter coater used chromium to coat the samples with the coating thickness of 15 nm. For cross sectional imaging, the membranes were placed vertically onto the SEM grid, and for surface imaging the membranes were placed horizontally.

4.4.5 Contact angle

Ramé-hart instrument co. was used to measure the contact angle using drop method at room temperature. The membrane was taped onto a glass slide to ensure flat surface and a drop of DI water was place onto the membrane surface via micropipette. A video camera was used to capture the shape of the droplet and automatically analysed the contact angle. On different membrane samples, at least six independent measurements were recorded [10].

4.4.6 Porosity and pore size

Zinadini's work suggested that the membranes' overall porosity, ε (%) can be estimate by using the equation (1):

$$\varepsilon = \frac{\omega_1 - \omega_2}{A \times l \times \rho_w} \quad (1)$$

where ω_1 is the weight of wet membrane (kg), ω_2 is the weight of dry membrane (kg), A is the effective area of the membrane (m^2), ρ_w is the ethanol density (789 kg/m^3), and l is the membrane thickness (m) [23].

Membranes without the PET fabric was cut into 1 × 1 cm^2. The dry and wet (soaked in 10 mL of ethanol for 24 h) weights were recorded.

The mean pore radius (r_m) of the membranes were calculated using Guerout-Elford-Ferry equation (2):

$$r_m = \sqrt{\frac{(2.9 - 1.75\varepsilon) \times 8\eta l \dot{Q}}{\varepsilon \times A \times \Delta P}} \quad (2)$$

where $\dot{Q}$ is the amount of water collected per unit time (m^3/s), A is the membrane area (m^2), ε is the overall porosity, η is the viscosity of water (Pa.s) at 25°C, ΔP is the operational pressure (Pa), and l is the membrane thickness (m).

4.5 Performance test

4.5.1 Permeance test

A dead-end cell (Sterlitech, HP4750 High Pressure Stirred cell) was used to measure the membrane permeance. The compression of membrane at 2 bar using nitrogen gas for at least 1 h was necessary to obtain steady state during data collection. The permeance of the membranes against water, methanol, toluene and NMP were recorded for 1 h and repeated three times with different membrane samples at 1 bar. The permeance (P) of membrane in LMH/bar was calculated using equation (3):

$$P = \frac{\Delta m}{\rho_s \times A \times \Delta t \times \Delta p} \quad (3)$$

where Δm is the amount of solvent permeation (kg) for a certain time Δt (h), ρ_s is the density of the solvent (kg/m^3), A is the effective membrane area (m^2), and Δp is the transmembrane pressure.

4.5.2 Rejection test

The rejection tests using dead-end cell at 1 bar were performed by dissolving PEO in water, acetonitrile, toluene and NMP. Since PEO did not dissolve in methanol, acetonitrile was used as a strong polar organic solvent instead. An Agilent High-performance liquid chromatography (HPLC) and an evaporative light scattering detector (Varian 385-LC ELSD) were used to analyse the samples. The C18-300 Hichrom reversed phase Ace column with length of 250 mm and interior diameter of 4.6 mm was used in HPLC. The two mobile phases were a 5mM aqueous solution of ammonium acetate (mobile phase A) and a 4:1 (v:v) mixture of acetonitrile and methanol (mobile phase B). The initial gradient was isocratic 90% A for 2 min, followed by a 23 min linear rise to 5% A. After that, it was reduced to 90% A for more than 2 min and was allowed to re-equilibrate before the next injection with 90% A for 5 min. The column temperature was set to 30 °C and the flowrate of 1 mL/min were used with the nebulizer set to operate at 55 °C and the ELSD evaporator at 40 °C [24].

This allows the feed and permeate concentrations of PEO to be determined and the rejection (R) was calculated using equation (4):

$$R = 1 - \frac{C_p}{C_f} \quad (4)$$

where C_f and C_p are the concentration of feed solutions and permeate solutions, respectively.

5. Results and discussion

5.1. Characterisation of nanoparticles

5.1.1 Attenuated total reflectance Fourier transform infrared spectroscopy (ATR-FTIR)

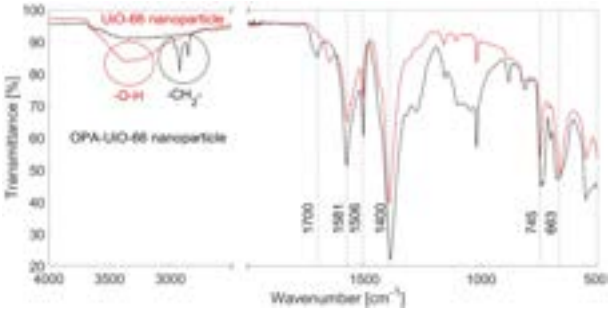


Fig. 4. ATR-FTIR spectra of UiO-66 and OPA-UiO-66 nanoparticles.

Fig. 4 shows the ATR-FTIR spectra of UiO-66 (red) and OPA-UiO-66 (black). The peaks at 1700 cm^{-1} and 1400 cm^{-1} belong to symmetrical stretching vibrations of C=O bond in the carboxyl group (-COOH). The C=C stretching vibration in phenyl ring corresponds to peaks at 1506 cm^{-1} and 1581 cm^{-1}. The peaks at 745 cm^{-1} and 663 cm^{-1} are attributed to O-Zr-O symmetric vibration. These peaks are present in both graphs which confirms that the OPA-UiO-66 nanoparticles contains the UiO-66 MOFs [25]. In addition, OPA-UiO-66 have characteristic peaks at 2920 and 2850 cm^{-1} resulting from symmetric and asymmetric stretching of alkyl -CH_2- groups [16]. The characteristic peak of UiO-66 due to -OH group is the broad peak at 3350 cm^{-1}.

The OPA-UiO-66 nanoparticles sample have a relatively weak vibration of the -OH bond in comparison to -CH_2- because OPA is made up of long -CH_2- chains. As a result, the -OH peak was not detected in the ATR-FTIR spectrum of the OPA-UiO-66 nanoparticle.

5.1.2 Field emission scanning electron microscope (FE-SEM)

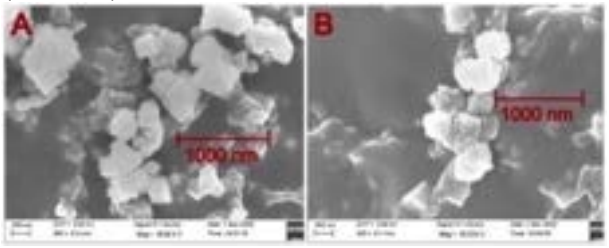


Fig. 5. SEM images of (A) UiO-66 and (B) OPA-UiO-66 nanoparticles.

UiO-66 nanoparticles depict octahedron structure in Fig. 5 (A) with the size of approximately 200 nm [26]. The surface of OPA-UiO-66 nanoparticles appears to be rougher; this confirms the intact morphologies of the OPA-UiO-66 crystals caused by the chemical addition of OPA into the originally synthesised UiO-66 nanoparticles.

5.1.3 X-ray diffraction (XRD)

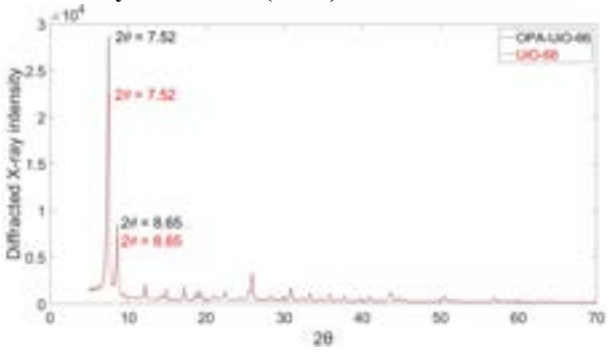


Fig. 6. The XRD patterns of UiO-66 and OPA-UiO-66 nanoparticles.

Fig. 6 depicts the XRD patterns of OPA-UiO-66 and UiO-66 nanoparticles. The diffraction peaks at 2θ = 7.5° and 8.6° attributed to the crystal faces (111) and (200) of UiO-66 verify the presence of the MOFs in both nanoparticles. The XRD shapes of both nanoparticles are similar, indicating that the intrinsic crystal structure was well preserved [14].

5.1.4 X-ray photoelectron spectroscopy (XPS)

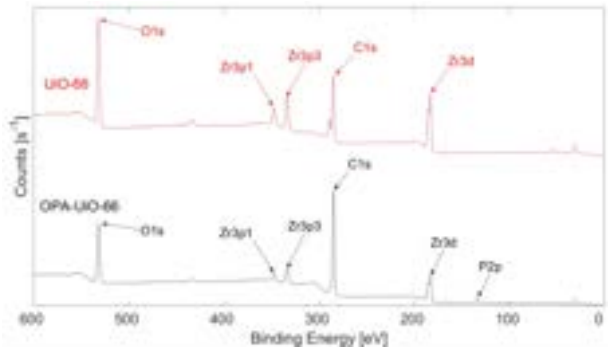


Fig. 7. XPS survey spectrum of UiO-66 and OPA-UiO-66

Table 1. Atomic compositions of each element in UiO-66 and OPA-UiO-66

	Atomic composition (%)			
	C	O	Zr	P
UiO-66	51.5	33.3	15.2	-
OPA-UiO-66	70.3	19.7	7.5	2.5

The red and black line in Fig. 7 shows the XPS spectra of survey for UiO-66 and OPA-UiO-66 nanoparticles, respectively. Both graphs contain O1s, C1s, Zr3p1, Zr3p3, and Zr3d. However, only the black graph contains P2p. This suggests that the elements oxygen, carbon, and zirconium were observed in both spectra, but only phosphorus was observed in the OPA-UiO-66 spectra, as shown by the peak appearing at 133.3 eV. It can be concluded that OPA was added to UiO-66 in the OPA-UiO-66 sample.

The compositions of each element in the two types of nanoparticles are listed in Table 1. The existence of phosphorus in the OPA-UiO-66 confirmed the successful addition of OPA into the originally synthesised UiO-66 nanoparticles. Furthermore, OPA-UiO-66 consisted of a higher percentage of carbon due to the addition of the long hydrophobic chain from the OPA.

5.2. Membranes characterisation

5.2.1 Attenuated total reflectance Fourier transform infrared spectroscopy (ATR-FTIR)

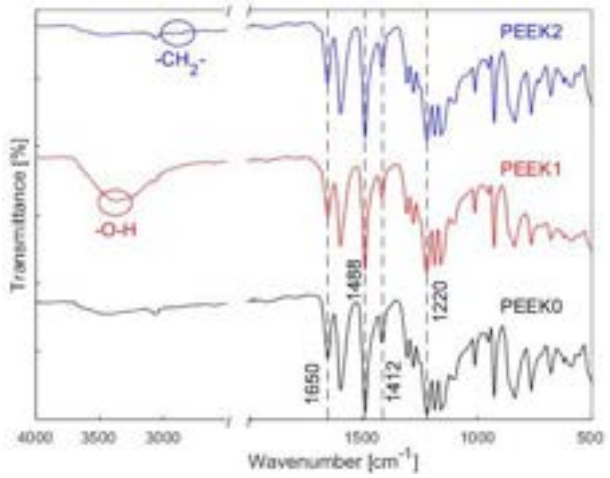


Fig. 8. ATR-FTIR spectra of PEEK0, PEEK1, and PEEK2 membranes.

There is a strong correlation between ATR-FTIR of the membranes (Fig. 8) and that of the nanoparticles (Fig. 4).

The characteristic peaks observed in the ATR-FTIR of UiO-66 and OPA-UiO-66 nanoparticles are also present in the ATR-FTIR of PEEK1 and PEEK2 membranes respectively. This indicates the presence of both nanoparticles on the membrane.

For PEEK0, PEEK1 and PEEK2, the peak at 1650 cm^{-1} corresponds to C=O of the carboxyl group (-COOH). The peak at 1488 cm^{-1} is due to the stretching vibration of aromatic C-C. Moreover, at 1220 cm^{-1} and 1412 cm^{-1} correspond to symmetric and asymmetric stretching vibration of O=S=O, respectively [10].

5.2.2 Field emission scanning electron microscope (FE-SEM)

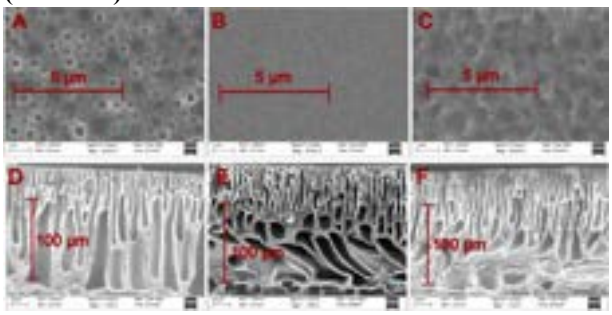


Fig. 9. SEM images of membrane surface and cross-section (A, D) PEEK0, (B, E) PEEK1, (C, F) PEEK2.

Fig. 9 shows the surface and cross-sectional images of the membranes, PEEK0, PEEK1, and PEEK2. The structure of each membranes contains a finger-like sublayer on top of the PET support layer with macrovoids [14]. These images (D to F) show the varying pore structure: straight, curved, and rounded for PEEK0, PEEK1, and PEEK2 membranes, respectively. The curved and round nature of the pore in Fig. 9 (E) and (F) show that the nanoparticles can affect the phase inversion reaction. The hydrophilic UiO-66 nanoparticles incorporated membrane exhibits curved and larger finger-like pores (E). The UiO-66 nanoparticles can result in thermodynamic instability, thus accelerating the exchange between the solvent and the nonsolvent [27] [14]. Additionally, the hydrophobic PEEK2 membrane (F) did not show finger-like voids on the bottom surface like the PEEK0 membrane, but instead exhibits smaller pores. This is due to the hydrophobicity of the OPA-UiO-66 nanoparticles causing more difficult exchange between the solvent and the nonsolvent in the phase inversion process, resulting in smaller pores [28]. This proves that UiO-66 nanoparticles increased the pore size of the membrane and OPA-UiO-66 nanoparticles reduced the pore size of the membrane.

It is clear there were pores on the surface of PEEK0 membrane and PEEK2 membrane. However, in Fig. 9 (B), it is not clear the pores were present; this is because of the formation of the hydration film by the hydrophilic

UiO-66 nanoparticles on the membrane surface [14]. On the other hand, PEEK2 membrane pores showed a rough texture, compared to the PEEK0 membrane pores, which is an indication of nanoparticles situating in the pores, confirming the addition of the nanoparticles.

5.2.3 Contact Angle

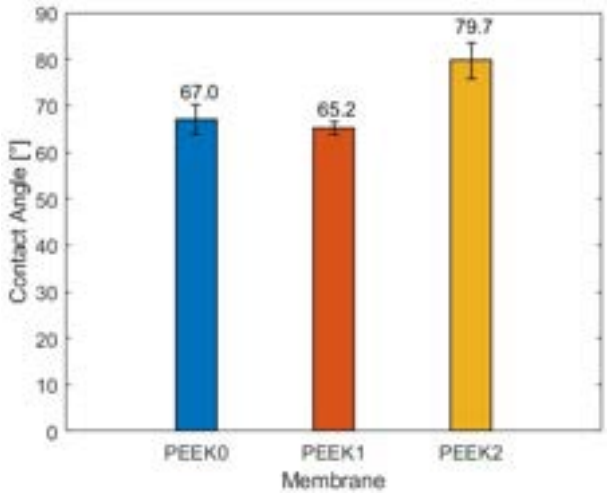


Fig. 10. Contact angle of PEEK0, PEEK1, and PEEK2. The comparison between the contact angle of each membrane and the standard deviation of the mean is shown by the black bar (from six independent measurements).

The contact angle was measured to determine the wettability of each membrane. The difference in the contact angle was observed when comparing each membrane due to hydrophobic and hydrophilic behaviour. The addition of OPA-UiO-66 nanoparticles increased the contact angle of the membrane, hence enhancing the hydrophobicity of the membrane. The addition of UiO-66 nanoparticles, on the other hand, lowered the contact angle and hence enhanced the hydrophilicity of the membrane. This finding indicates that PEEK2 membrane exhibit hydrophobic characteristic while PEEK1 membrane exhibited hydrophilic characteristics.

The standard deviation of the mean of PEEK2 (black bar) shows that the change from PEEK0 might be statistically significant. The error bars for PEEK1 and PEEK0, however, marginally overlapped, indicating that it is not statistically significant. As a result, additional data is needed to draw a more definite conclusion.

5.2.4 Porosity and pore size

Table 2. Calculated values of the overall porosity, mean pore radius, and pore size of membrane.

Membrane	Overall porosity ε (%)	Mean pore radius (nm)	Pore size (nm)
PEEK0	59.6	50.5	101.0
PEEK1	51.7	52.5	105.0
PEEK2	79.1	39.2	78.3

In Table 2, PEEK2 has the highest overall porosity, followed by PEEK0, and then PEEK1. With the addition of UiO-66 nanoparticles, PEEK1 membrane porosity was reduced due to blockage of the membrane pore caused by aggregation of the UiO-66 nanoparticles [29].

The mean pore radius aligns with the conclusion from the cross-sectional SEM images. It is confirmed that the hydrophilic UiO-66 nanoparticles enlarged the membrane pore size, while the hydrophobic OPA-UiO-66 nanoparticles reduced the membrane pore size, compared to PEEK0 membrane.

All three membranes' mean pore size are approximately in between 50 nm and 100 nm, suggesting PEEK0, PEEK1, and PEEK2, can be classified between microfiltration (MF) and ultrafiltration (UF) membranes [30].

5.3. Membrane performance test

5.3.1 Permeance test

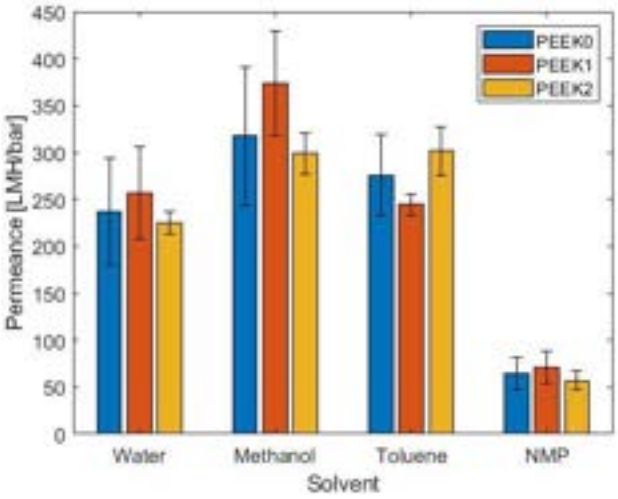


Fig. 11. Permeance result of PEEK0, PEEK1, and PEEK2 with water, methanol, toluene, and NMP and the standard deviation of the mean is shown by the black bar (from three independent measurements).

Fig. 11 shows the comparison of each membrane's permeance in each of the solvents including water, methanol, toluene, and NMP. The permeance trend of membrane can be explained by the solvent viscosity. At standard temperature and pressure, NMP has the highest viscosity (1.7 cP), followed by water (1 cP), toluene (0.56 cP), and methanol (0.5 cP). PEEK1 membrane exhibited the highest permeance for polar solvents (water, methanol, and NMP) due to its hydrophilic characteristic. On the contrary, PEEK2 membrane showed the highest permeance for a non-polar solvent (toluene) due to its hydrophobicity nature. This suggests that the permeance of the polar and non-polar solvents can be enhanced compared to pure PEEK depending on the characteristic of nanoparticles.

Each membrane samples were inspected after the permeance tests to identify damages to the structural integrity. The observations were that all three types of

membranes could withstand 1 h of usage with organic solvents.

However, the standard deviation of the mean error bar for membranes in each solvent overlapped indicating that it might not be statistically different enough to draw a definite conclusion.

5.3.2 Rejection test

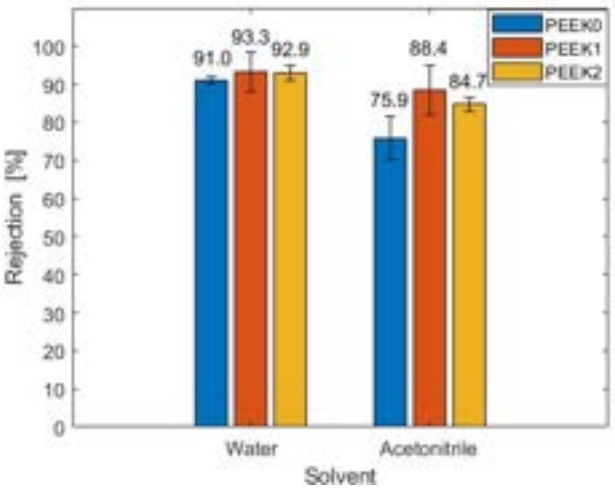


Fig. 12. Rejection of PEO on each membrane in each solvent and the standard deviation of the mean is shown by the black bar (from three independent measurements).

Initially, the membrane rejection was calculated by dissolving PEG of molecular weights 1,000, 2,000, 6,000, and 20,000 in 4 solvents (water, methanol, toluene, and NMP). However, the HPLC gave meaningless results which means the PEG used were too small for the synthesised membranes and cannot be detected by HPLC. The HPLC was then operated using PEO of molecular weight 100,000 Da. This only gave meaningful results for water and acetonitrile; this was caused by the insufficient time taken to dissolve the PEO particles in toluene and NMP.

According to Fig. 12, the PEEK0 membrane showed good rejection in an inorganic solvent (water), but not in an organic solvent (acetonitrile). This suggests the investigation of adding nanoparticles to improve the membrane performance in organic solvents is necessary.

When using water, the rejection of PEEK1 and PEEK2 were slightly better than PEEK0. However, this difference is not statistically significant as the standard deviation of the mean bar (black error bar) of PEEK1 and PEEK2 overlap with the PEEK0 value. Therefore, further tests are required before a conclusion can be drawn for water.

In acetonitrile, PEEK1 had the highest rejection, and since it is a polar solvent, this aligns with the conclusion drawn from the permeance test, that hydrophilic membranes are suitable for the use of polar organic solvent separation (Fig. 11). This also suggests the pore size could be a dominant factor which controls the membrane rejection, as PEEK1 has the largest pore size (Table 2). PEEK1 and PEEK2 membranes showed improved rejection values in acetonitrile, compared to PEEK0 without nanoparticles in the membrane matrix. This implies the addition of both the UiO-66 and the OPA-UiO-66 nanoparticles led to a successful increase of membrane rejection and therefore membrane performance.

6. Conclusion

In this study, PEEK membranes with UiO-66 (hydrophilic) and OPA-UiO-66 (hydrophobic) nanoparticles incorporated into the membrane matrix were synthesised. This was proved by the various characterisation methods performed as each type of membrane showed distinguishable qualities which are suitable for different types of solvents. Both the UiO-66 and OPA-UiO-66 modified membranes showed improved permeance in polar organic solvents and nonpolar organic solvents, respectively, compared to the pristine PEEK membrane. Moreover, modifying polymeric membranes with nanoparticles proved to increase the membrane rejection in both inorganic solvents and organic solvents. However, due to the time constraint of the project, rejection tests with organic nonpolar solvents were not carried out as operating conditions were needed to be varied for PEO to dissolve in each of the solvents.

Overall, the modification of PEEK membrane with both UiO-66 and OPA-UiO-66 nanoparticles successfully showed improvement to the PEEK membrane's performance with organic solvents.

7. Outlook

MOFs are powerful new devices for the more sustainable future of chemical separations by polymeric membranes. For further studies, the effects of the amount of MOFs added to polymeric membranes on membrane permeance and rejection could be investigated to improve the current understanding regarding the amount of MOFs which is required to optimise the membrane performance.

Performance tests with a higher variety of operating conditions should be explored when different solvents are used, as it will be more comparable to real-life industrial situations. Furthermore, more tests over a longer period can be conducted to evaluate the long-term effect of adding nanoparticles on the permeance and structural integrity of the membranes.

More study could be carried out regarding whether there are biodegradable alternatives for PEEK membranes, and whether polymeric wastes could be derived from landfills to fabricate membranes to reduce environmental waste.

A cost-benefit analysis can be conducted to assess the cost efficiency of using mixed matrix membranes instead

of commercial polymeric membranes; this will highlight how largely chemical industries can benefit from this new separation method monetarily and environmentally.

8. Acknowledgement

This research project was funded by a donation to the Department of Chemical Engineering, Imperial College London by Mr. Mark Richardson.

9. References

[1] The National Institute for Occupational Safety and Health (NOISH), "Centers for Disease Control and Prevention," 2 November 2018. [Online]. Available: https://www.cdc.gov/niosh/topics/organsolv/default.html. [Accessed 9 December 2022].

[2] Oak Ridge National Laboratory (ORNL), "Materials for Separation Technologies: Energy and Emission Reduction Opportunities," U.S.Department of Energy, 2005.

[3] W. G. K. B. L. H. M. G. a. H. L. Ameya Manoj Tandel, "Designing organic solvent separation membranes: polymers, porous structures, 2D materials, and their combinations," *Materials Advances,* 2021.

[4] G. D. T. T. H. M. Cuijing Liu, "Organic solvent reverse osmosis membranes for organic liquid mixture separation: A review," *Journal of Membrane Science,* vol. 626, no. 15, 2021.

[5] M. Mulder, "Energy Requirements in Membrane Separation Processes," *Membrane Processes in Separation and Purification,* vol. 272, pp. 445-475, 1994.

[6] A. S. Khanna, "Natural Degradation on Plastics and Corrosion of Plastics in Industrial Environment," *Encyclopedia of Materials: Plastics and Polymers,* pp. 956-986, 2022.

[7] S. C. B. A. P. a. S. P. N. Sandra L. Aristizábal, "Preparation of PEEK Membranes with Excellent Stability Using Common Organic Solvents," *Industrial & Engineering Chemistry Research,* 2020.

[8] G. M. A. B. H. A. S. F. M. M. K. a. M. A. Ahmad Kayvani Fard, "Inorganic Membranes: Preparation and Application for Water Treatment and Desalination," PubMed Central, 2018.

[9] P. A. M. Majumder, "1.14 - Carbon Nanotube Membranes: A New Frontier in Membrane Science," *Comprehensive Membrane Science and Engineering,* pp. 291-310, 2010.

[10] L. G. P. S. K. A. L. João da Silva Burgal, "Organic solvent resistant poly(ether-ether-ketone) nanofiltration membranes," Journal of Membrane Science, London, 2015.

[11] Y. Y. J. W. Z. W. F. K. a. X. L. Shenzhen Cong, "Highly Water-Permeable Metal−Organic Framework MOF-303 Membranes for Desalination," Journal of the American Chemical Society, 2021.

[12] M. Berger, "What is a MOF (metal organic framework)?," Nanowerk, [Online]. Available: https://www.nanowerk.com/mof-metal-organic-framework.php. [Accessed 13 November 2022].

[13] Q. A. M. H. A. A. O. Mohammad Ali Abdelkareem, "High-performance effective metal–organic frameworks for electrochemical applications," *Journal of Science: Advanced Materials and Devices,* vol. 7, no. 3, 2022.

[14] D. L. J. L. J. L. M. F. M. Z. L. Z. L. F. K. Yi Wang, "Metal organic framework UiO-66 incorporated ultrafiltration membranes for simultaneous natural organic matter and heavy metal ions removal," *Environmental Research,* vol. 208, 2022.

[15] H. M. M. R. S. T. A. B. F. K. T. M. A. J.-R. L. M. A. Farhad Ahmadijokani, "UiO-66 metal–organic frameworks in water treatment: A critical review," *Progress in Materials Science,* vol. 125, 2022.

[16] Q. S. H. H. B. A. Z. N. J. A. P. a. S. M. Yuxiu Sun, "A molecular-level superhydrophobic external surface to improve the stability of metal–organic frameworks," *Journal of Materials Chemistry A,* 2017.

[17] A. I. E. W. Q. I. P. S. Y. K. Toshiaki Kabe, "2 - Chemical and Macromolecular Structure of Coal," *Studies in Surface Science and Catalysis,* vol. 150, 2004.

[18] K. S. K. R. G. K. K. W. Zhou He, "Chapter 5 - CO2/CH4 Separation (Natural Gas Purification) by Using Mixed Matrix Membranes," *Current Trends and Future Developments on (Bio-) Membranes,* pp. 155-181, 2018.

[19] A. A. M.M.H. Shah Buddin, "A review on metal-organic frameworks as filler in mixed matrix membrane: Recent strategies to surpass upper

bound for CO2 separation," *Journal of CO2 Utilization,* vol. 51, 2021.

[20] M. K. S. P. M. R. S. Mihir Kumar Purkait, "Chapter 1 - Introduction to Membranes," in *Interface Science and Technology*, Elsevier, 2018, pp. 1-37.

[21] M. Mulder, Basic Principles of Membrane Technology, Kluwer Academic Publishers, 1996.

[22] Y. A. a. A. A. Alomair, "Microscopy and Spectroscopy Techniques for Characterization of Polymeric Membranes," *Membranes for Gas Separation,* vol. 10, no. 2, 2020.

[23] A. A. Z. M. R. V. V. H. Z. Sirus Zinadini, "Preparation of a novel antifouling mixed matrix PES membrane," *Membrane Science,* 2013.

[24] P. R. G. D. K. P. M. A. G. L. Adam Oxley, "Graft modification of polybenzimidazole membranes for organic solvent," *Journal of Membrane Science ,* 2021.

[25] M. D. L. D. *. a. X. Z. *. Weiwei Xu, "A Facile Method for Preparing UiO-66 Encapsulated," *nanomaterials,* 2019.

[26] M. M. C. G. H. L. L. L. F. D. a. Y. X. Liangyu Lu, "Metal Organic Framework@PolysilsesequioxaneCore/Shell-Structured Nanoplatform forDrug Delivery," *Pharmaceutics,* vol. 12, no. 98, 2020.

[27] T.-S. C. Panu Sukitpaneenit, "Molecular elucidation of morphology and mechanical properties of PVDF hollow fiber membranes from aspects of phase inversion, crystallization and rheology," *Journal of Membrane Science,* vol. 340, no. 1-2, pp. 192-205, 2009.

[28] Y. L. D. L. B. L. J. Y. Hao Sun, "Hydrophobic SiO2 nanoparticle-induced polyvinylidene fluoride crystal phase inversion to enhance permeability of thin film composite membrane," *Journal of Applied Polymer Science,* vol. 136, no. 45, 2019.

[29] J. G. a. J. Kim, "Modifications of polyethersulfone membrane by doping sulfated-TiO2 nanoparticles for improving anti-fouling property in wastewater treatment," *The Royal Society of Chemistry 2017,* 2017.

[30] S. C. P. D. X. V. Curtis D. Roth, "Chapter 13 - Customization and Multistage Nanofiltration Applications for Potable Water, Treatment, and Reuse," *Nanotechnology Applications for Clean Water (Second Edition),* pp. 201-207, 2014.

[31] A. G. A. I. P. M. A. Basile, "5 - Membrane technology for carbon dioxide (CO2) capture in power plants," *Advanced Membrane Science and Technology for Sustainable Energy and Environmental Applications,* pp. 113-159, 2011.

Improving the accuracy of enzyme capacity constrained metabolic models of CHO cells for biopharmaceutical production

Matthew Brown and Michael Zhou

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

Abstract

The biopharmaceutical industry widely uses Chinese Hamster Ovary (CHO) cells for antibody production; an approach to modelling CHO cell metabolism is to use constraint-based models of genome scale metabolic networks. Flux Balance Analysis (FBA) - a methodology used to solve genome-scale metabolic models to estimate intracellular reaction fluxes - has been found to lack quantitative accuracy in predicting intracellular fluxes. In this investigation we show that the use of enzyme capacity constrained FBA (ecFBA), updated for use with the genome-scale metabolic model iCHO2441 and adapted to Python, improves the accuracy of intracellular flux predictions when compared to FBA and other constraint-based modelling methods. We found that on average, ecFBA predicts cellular growth and Immunoglobulin G (IgG) production rates 9.3% and 3% closer to experimental values respectively. Additionally, with ecFBA 36.6% of predictions for all reactions in all samples are found to be within the error of the measured experimental value, an increase of 8.96% compared to FBA. Furthermore, we have identified 5 reactions as potential bottlenecks in the IgG secretory pathway that could be investigated as targets for genetic engineering; of these 5, BiP_ATPase is the most likely to be the bottleneck. We anticipate that the ecFBA model developed can be improved by the addition of more enzyme kinetic data and the expansion to additional experimental data sets. The approach outlined in this paper could also serve as a starting point for investigations in cellular engineering of CHO cells.

1. Introduction

In the biopharmaceutical industry, the primary production methods for antibody-based therapeutics involve genetic engineering of mammalian cells, particularly Chinese-hamster ovary (CHO) cells [1]. Considering this, understanding and modelling the metabolic behaviour and intracellular reactions of such cells could be vital for progression in this field. One such method is genome-scale models of metabolism (GeM). GeMs use knowledge of metabolic reactions and genes, the gene-protein-reaction (GPR) relations linking the two, the metabolites participating in these reactions as well as the reactions governing the transport of these metabolites in order to construct a mathematical expression of cellular metabolism [2,3]. The advantage of this is that a genome-scale model of intracellular reactions can be recreated for any cell whose parent species has a completely mapped genome.

Flux balance analysis (FBA) is a constraint-based method for analysing GeMs, using reaction stoichiometries, metabolite mass balances and thermodynamic constraints [4,5]. While a useful tool for studying cell behaviour, FBA results in an underspecified and under-constrained system which limits its potential for accurate metabolic modelling. To improve this, further constraints can be employed in attempts to narrow the solution space. Examples of this include parsimonious FBA (pFBA), carbon constrained FBA (ccFBA) and enzyme capacity constrained FBA (ecFBA) which constrain the total sum of fluxes, carbon availability, and enzyme availability respectively.

This investigation attempts to improve the accuracy of FBA models by applying enzyme capacity constrained FBA (ecFBA) to analyse an expanded CHO cell GeM, iCHO2441 [6]. Reaction fluxes obtained from the analyses are mapped to reactions with experimentally available data from Yeo *et al.* [7]; results are compared against alternative constraint-based models and experimental data. Relevant enzyme data from Yeo *et al.* [4] is used and built upon with additional data extracted from the BRENDA database [8] to constrain enzyme capacity more accurately.

2. Background

GeMs have the advantage of not requiring cellular kinetic data, being based on the mass balance of metabolites across the cell [9]. The only knowledge required is the mapping of the cell genome, and the stoichiometries of the metabolic reactions. Additionally, as GeMs explicitly include the GPR relations, the effects of individual genes can be determined; a useful property for investigating cellular engineering targets [10,11].

The incorporation of enzyme kinetic data to FBA models originates from Beg *et al.* [12], with the creation of FBA with molecular crowding, introducing the idea of limiting the volume of certain substances in a cell using a capacity constraint. The

incorporation of enzyme kinetic data into this framework was added by Sánchez *et al.* [13] through the incorporation of enzyme turnover number and concentration into the stoichiometric matrix to limit molecular fluxes based on enzyme activity. ecFBA, developed by [8], is a different approach which limits molecular fluxes based on enzyme mass in the cell. The secretory pathway was investigated in detail, as it has been identified as a bottleneck for IgG production [14,15], and work has been done to expand GeMs to include the secretory pathway [6]. This investigation is the first work applying ecFBA to the iCHO2441 model and investigating its impact on the accuracy of the secretory pathway.

3. Methods

Flux Balance Analysis (FBA)

FBA is a method used to analyse GeMs and the flow of metabolites through a metabolic network. A steady state mass balance of metabolites and constraints on reaction fluxes are used to create a system of linear equations; a particular reaction flux can be chosen as an objective function, and the system can be solved using linear programming (Fig. 1).

This system created by FBA is underdefined, as there are more reactions than metabolites (most metabolites will participate in many reactions), so to further constrain the solution space several approaches have been taken. FBA, pFBA, and ccFBA are used as benchmarks to compare the performance of ecFBA. pFBA is a bilevel optimisation optimising a metabolic reaction, usually growth, as the outer optimisation and minimising total cellular flux as the inner optimisation [16]. ccFBA in contrast constrains the total elemental carbon flux into the cell [17].

Enzyme Capacity Constrained FBA (ecFBA)

ecFBA [5] constrains the allowable mass of enzymes in the cell; as metabolic reactions require enzymes to occur at non-negligible rates this effectively constrains reaction fluxes. Each flux is assigned a coefficient which relates reaction flux to the mass of enzyme present, assuming all enzymes operate at their turnover number. The sum of these capacity coefficients multiplied by the flux must be less than or equal to the total mass of enzyme per unit mass dry cell weight, which is the enzyme capacity constraint C (Eqn. 01)

$$\sum_{j=1}^{n} \frac{M_{w,j}}{k_{c,j}} * v_j \leq C \qquad \text{Eqn. 1}$$

FBA and its more advanced variants all result in large solution spaces with an infinite number of solutions and therefore require at least one reaction flux to be chosen as an objective function, which poses three problems. The optimal choice for the objective function depends on the conditions that the modelled cell is in, and is an ongoing area of research [18]. Additionally, it has been demonstrated that two of the most popular choices of objective function - maximising biomass growth/ATP production - do not always accurately predict intracellular fluxes [19]. Finally, assuming a single, unchanging objective function leaves models unable to accurately model cells in changing conditions [20].

Flux Sampling

Markov-chain Monte Carlo methods, also known as flux sampling when applied to GeMs, remove the need for an objective function [20]. FBA methods can be summarised as solving a linear optimisation problem subject to a steady state mass balance, and flux feasibility constraints. Flux sampling is the random generation of sets of fluxes which sample the solution space defined by the same set of constraints as FBA. In this investigation, sampling is set to 5000 with a thinning of 10000 meaning one sample is taken for every 10000 potential solutions across 50,000,000 solutions to give a good representation of the entire solution space. COBRApy [21] was used to implement Flux sampling with ecFBA constraints. The value of each flux is calculated as the average flux of the 5000 samples.

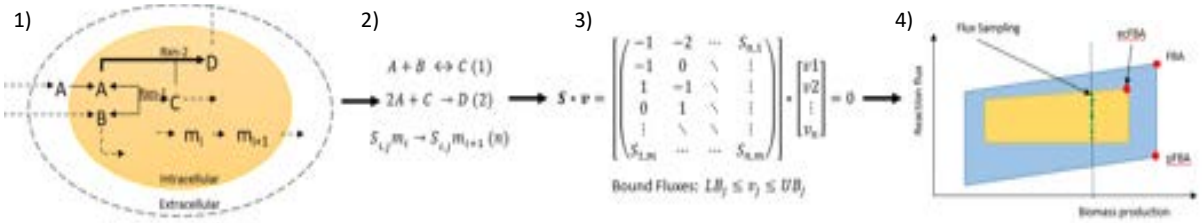


Figure 1. Representation of steps involved in Flux Balance Analysis. 1) Intracellular reactions are mapped based on GPR in the GeM. 2) Reactions with stoichiometry are extracted to generate stoichiometric matrix S of n reaction coefficients for each metabolite (m) 3) Mass balance on equations with steady state condition assumed for all intracellular metabolites gives that S*v = 0 generating a system of linear equations. 4) Objective functions are set to solve the linear programming problem and find an optimum solution within the solution space.

This work develops ecFBA models by investigating the effects of changing the value of the enzyme capacity constraint C [Eqn. 1], the addition of estimated enzyme data, and expanding the set of enzyme constrained reactions to the secretory pathway. In this work values of C of 0.05, 0.09, 0.11, 0.13, 0.15, and 0.20 g_{enzyme}/g_{DCW} were investigated. Required enzyme data not present in Yeo *et al.* were first searched for using the GPR from the model to extract gene numbers. These were then converted to EC numbers using the UniProt database [22], which in turn were used to extract enzyme parameters from the BRENDA database [8]. If the required data could not be found, they were estimated using Eqn. 2.

$$Coef = \frac{min\ M_r}{max\ k_c} \qquad \text{Eqn. 2}$$

Reaction Mapping

Due to experimental limitations, reaction fluxes for the exact individual reactions in the model cannot be measured experimentally and therefore groups of predicted reaction fluxes must be mapped to experimental fluxes to analyse results. FBA and flux sampling result in reaction fluxes for all 6337 reactions present in the iCHO2441 GeM. The experimental data meanwhile has intracellular fluxes for only 56 major metabolic pathways measured using ^{13}C labelling [7]. To compare these, a mapping process where the series of sequential and/or parallel reactions from the model corresponding to the experimentally measured metabolic pathways are combined to generate an overall flux for each pathway. If there are parallel reactions, the flux of each is summed to give a total flux while if there are sequential reactions the absolute minimum of the fluxes is taken.

Evaluation methods

Our model was evaluated against experimental data covering a wide range of conditions, thus testing the robustness of the ecFBA model. Experimental intracellular flux data obtained from ^{13}C flux tracing from 31 experiments were used to constrain exchange and biomass reaction fluxes for the ecFBA model, and to compare to the predictions made by our ecFBA model. The accuracy of our model to this experimental data was evaluated using three methods; the root mean squared error (RMSE), Pearson correlation coefficient, and capability. The RMSE and Pearson correlation coefficients were taken between the average of the flux samples for each reaction, and the experimental value for each reaction, for each experiment. The percentage capability of each reaction is defined as the percentage of reaction fluxes from all flux samples that fall within the experimental error for that flux.

4. Results and discussion

Optimal value of C

The value of the enzyme capacity constraint, C, was investigated to both determine its impact on the accuracy of the ecFBA model, and assess the accuracy of the literature value. The literature value of C=0.11 g_{enzyme}/g_{DCW} [7] was estimated by multiplying the measured value of 0.702 $g_{protein}/g_{DCW}$ [23] by the proportion of expressed genes which relate to cell metabolism (0.158) [7]. This value of C is not certain, as the value of 0.702 $g_{protein}/g_{DCW}$ itself is an average of two data sources [23]. Despite this, using the proportion of expressed genes to estimate the fraction of the proteome relating to metabolic enzymes has been found to be quite an accurate method [24].

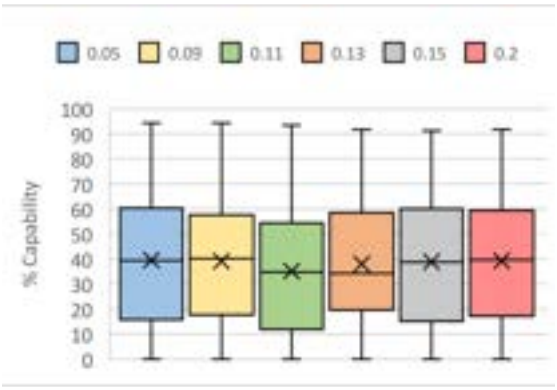


Figure 2. % Capability observed from flux sampling with changing values of C for ecFBA. Smaller values of C were investigated but not included in the figure as they were found to be infeasible with flux sampling

It was determined that the precise value of C does not greatly affect the accuracy of the ecFBA model, in the range around the literature value, using all three measures of accuracy (Figs. 2-4). No clear trends in mean, median, or interquartile range for any of the three measures are observed (Fig. 2), with mean spreads being ±0.058 for Pearson, ±1.29 for RMSE, and ±2.20% for capability. This is likely because only 62.3% of the mapped reactions have enzyme kinetic data to constrain the fluxes. As only a small subset of the reactions is constrained, changing the total enzyme capacity will not greatly affect the system.

Effect of additional estimated enzyme parameters

In contrast, the addition of the estimated enzyme kinetic data decreases the RMSE by an average of 18.0 (Fig. 3), with the trade-off of slight decreases in Pearson coefficient and average capability of 0.0287 and 3.12% respectively (Fig. 4). Adding any non-

zero constraint for each flux will constrain the system from predicting fluxes of unphysical magnitudes. Given that most cellular fluxes are of the order of magnitude of 10^{-3}, reducing the magnitude of predicted fluxes will decrease the RMSE. This is an approximate method, as these parameters are simply estimated with no biological basis. It is therefore likely they are extremely far from the true value, for example the minimum coefficient for reactions with enzyme data was 0.00129 compared to the value 0.000245 obtained using estimates. This could explain the marginal decrease in Pearson coefficient and capability, but as the estimated enzyme parameters are almost certainly lower than their true values, they will not over-constrain the system.

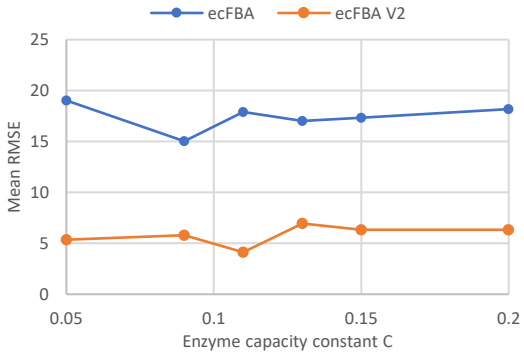


Figure 3. RMSE observed from flux sampling with changing values of C for ecFBA with estimated enzyme parameters (V2) and without.

This is demonstrated by the large decrease in the Pearson coefficient for C=0.05. This is caused by the small value of C and the estimated enzyme kinetic parameters over- constraining the system leading to excessively small fluxes, hence the minimal impact on RMSE.

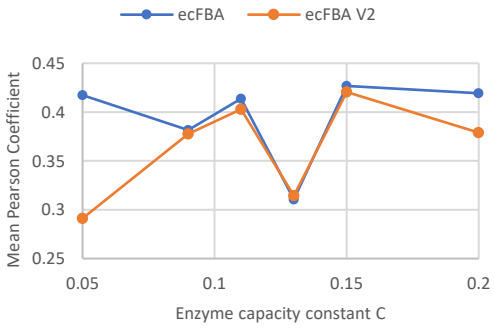


Figure 4. Mean Pearson coefficient observed from flux sampling with changing values of the enzyme capacity C for ecFBA with estimated enzyme parameters (V2) and without.

From this analysis, we conclude that moving forward we would use a value of C equal to 0.11 and add estimated enzyme kinetic parameters. This was deemed to be optimal for not over- or under-constraining the model, maintaining Pearson correlation coefficient and percentage capability high, while also limiting RMSEs.

Performance compared to benchmark methods

Using capability analysis, it was determined that ecFBA was more accurate than the benchmark methods of FBA, pFBA, and ccFBA (Fig. 5). The increase in accuracy compared to FBA can be explained by the addition of physical data to an underdefined system which improves its accuracy. The increase in accuracy when compared to pFBA as the constraint on the sum of fluxes to the minimum, as found in pFBA, implicitly assumes total flux minimisation as an objective, which introduces inaccuracies [18]. ccFBA with flux sampling is closer to ecFBA in terms of numerical performance; these two methods are the most similar in trying to constrain the solution space. ccFBA only constrains elemental carbon flux into the cell. This, however, does not offer the more detailed control over reaction fluxes as achieved by ecFBA.

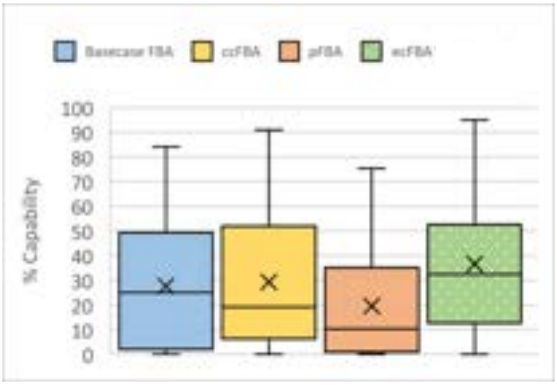


Figure 5. % Capability observed from flux sampling with ecFBA compared to benchmark methods.

The mean RMSE for ecFBA is 4.13, compared to 40.47 for FBA; both values are much higher, however, than the values of 0.13 and 0.11 obtained for ccFBA and pFBA respectively. The lower accuracy of ecFBA can be explained through an analysis of the mitochondrial transport reactions akG.m and Glu.m, both anaplerotic reactions. As only estimated enzyme parameters are available for these reactions, they are massively under-constrained compared to ccFBA and pFBA (Fig. 6) resulting in a much higher RMSE.

Anaplerotic reactions play a key role in the citric acid cycle [25], so increasing the fluxes of these reactions would increase the energy available to the cell. A wider range of fluxes are feasible with more energy available, thus representing a greater proportion of flux samples, meaning unconstrained

anaplerotic fluxes are likely to be larger in magnitude. Given the key role of anaplerotic reactions, it is vital they are more accurately modelled in future work to capture CHO cell metabolic behaviour more accurately.

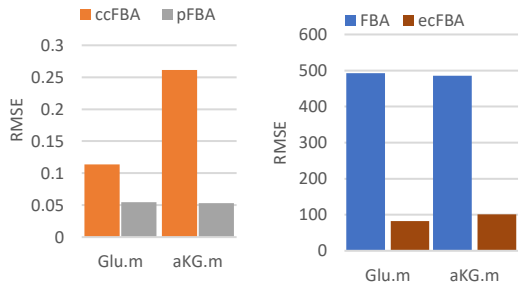


Figure 6. RMSE of anaplerotic reactions. Note the two different axes scales due to the wide range of values.

Qualitative analysis

The quantitative analysis carried out has large uncertainties; in both the values we generate from flux sampling as well as the experimental data we compare to. Each flux was calculated as an average of the flux samples for that flux, but the random nature of flux sampling results in a large uncertainty in that value. Additionally, the experiments we use to evaluate the accuracy of our results also have large uncertainties. As an example, for experiment 'stat', the experimental error on the LDH flux was 67%, not an insignificant amount.

A novel method we created to evaluate how accurately our ecFBA model emulates real-life CHO cell metabolism is to determine if general trends in fluxes as cell genomes and growth phase change as described in the literature are followed by our ecFBA model. From the literature, three pairs of experiments were chosen, to observe the change in flux of a particular reaction between the experiments in each pair. The rationale for choosing these experiments and fluxes is explained in Table 1.

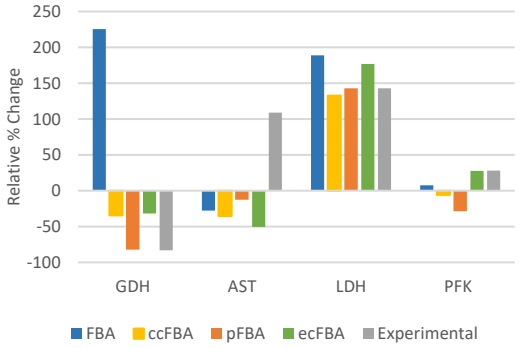


Figure 7. Relative % change in fluxes for changes in cell condition/growth stage/genome. Data for C=0.15 is used as flux sampling for all relevant experiments was feasible.

Our results determine our model is more accurate than the benchmarking methods at predicting qualitative trends (Fig. 7). We can observe that all the models correctly predict the trend in changing cell growth phase, but can only correctly predict 1 out of 2 trends in changing growth medium. Only ecFBA and FBA, however, correctly predict the trend in changing cell gene expression. This shows that ecFBA is not only quantitatively more accurate, but can also more accurately predict broad trends in cellular activity compared to the benchmark methods.

Predicting growth rate and IgG secretion with ecFBA

The accuracy of the ecFBA model's ability to predict the growth rate of CHO cells was compared to that of FBA. To investigate growth rate predictions, ecFBA and FBA (unconstrained) were run with the objective function of maximising biomass production. Comparing the two methods to experimental values, we can see ecFBA predicts growth rate more accurately compared to FBA (Fig. 8). This is demonstrated by the trendline for ecFBA being closer to y=x which is the perfect result of

Table 1. Reactions and experiments analysed qualitatively and their expected trend

Reactions and Experiments	Assessed feature of ecFBA	Expected trend and biological explanation
GDH (Glutamine dehydrogenase) and AST (Aspartate transaminase) flux for cells from standard (CM) media to low NH_3 media (LA)	Model trends in changing cell growth media	Positive change in GDH and negative change in AST - NH_3 deficient media causes change in amino acid metabolism [26]
LDH (lactate dehydrogenase) flux for cells from late→stat	Model trends in changing cell growth phase	Positive relative change - cells increase lactate consumption to synthesise NADH, for ATP synthesis, as cells transition from peak growth to peak IgG production [27]
PFK (phosphofructokinase) flux for SVGS→BCL2	Model trends in changing cell gene expression	Positive change in PFK – PFK is determining step in glycolysis, which increases in cells expressing the BCL-2Δ gene affecting apoptosis [28].

exactly predicting the experimental value. Additionally, ecFBA's data points fit closer to the trendline by measure of R^2, demonstrating there are fewer outliers in ecFBA's predictions. On average, ecFBA's predicted growth rates deviate 46.4% from experimental values compared to 55.7% for FBA. This is because FBA predicts excessively large growth rates for some experiments, which become constrained when modelled by ecFBA, and are brought closer to the experimental value.

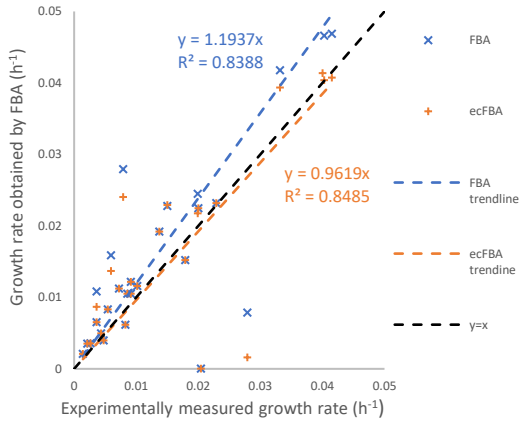


Figure 8. Comparison of growth rate predictions using FBA and ecFBA to experimental data.

To evaluate the effect of the additional enzyme kinetic parameters on the secretory pathway, the accuracy of the ecFBA model's ability to predict the IgG secretion rate was compared to that of ecFBA without the additional secretory pathway enzyme kinetic parameters. It was determined that the additional enzyme parameters increased the accuracy of the model (Fig. 9); ecFBA with the additional enzyme kinetic parameters yields results deviating 31.2% on average from experimental values compared to 34.2% without. This improvement in accuracy is due to the addition of enzyme parameters which resulted in the activation of the enzyme capacity constraint for excessively large IgG flux predictions (namely "SVM3", "SVM4", and "late"). The prediction for "fed-batch", however, became over-constrained and its accuracy decreased.

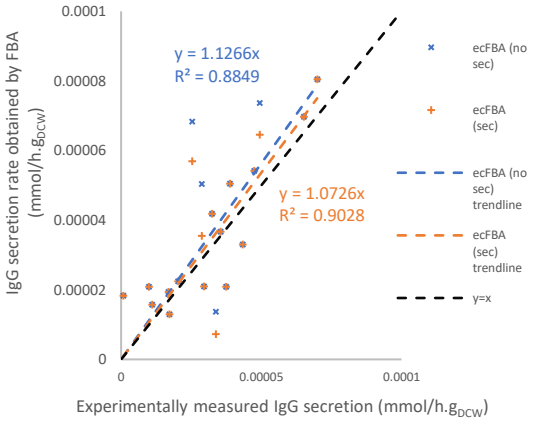


Figure 9. Predicted IgG secretion vs experimental for ecFBA with and without secretory pathway enzyme data.

Secretory pathway investigation

ecFBA with an objective function of maximising IgG secretion was further investigated to identify potential bottlenecks in the secretory pathway. When comparing predictions with experimental data sets, the experiments 'late', 'SVM3' and 'SVM4' showed the largest improvements in the accuracy of IgG secretion rate compared to FBA. These data sets were isolated and reaction fluxes for FBA and ecFBA compared to investigate for which reactions the enzyme capacity constraint has the greatest effect.

Five reactions were identified as potential bottlenecks for the IgG secretory pathway. These reactions were identified using two methods. Firstly, the difference in the Pearson coefficient between a reaction's flux and the IgG secretion rate, when comparing FBA and ecFBA was analysed. A significant increase in the Pearson coefficient suggests a previously insignificant reaction in FBA now has a large impact on the IgG secretion rate. This indicates that the flux was not previously limiting the IgG production but becomes limiting when constrained. The second method was using the reduced costs of reaction fluxes when using ecFBA which are extracted from the model using COBRApy. The reduced cost of a reaction flux is the rate at which the objective value (in this case IgG secretion) changes when the value of that reaction flux changes. Therefore, positive reduced costs could indicate that a reaction has some limiting effect on IgG secretion. The constrained secretory pathway reactions with the greatest increases in Pearson coefficient and reduced costs are presented in Table 2.

Table 2. Secretory pathway reactions identified as the most likely bottlenecks limiting IgG production

	Reaction ID	% Reduction in reaction flux with ecFBA	Pearson coefficient with IgG production (FBA)	Pearson coefficient with IgG production (ecFBA)	Reduced Cost of reaction (ecFBA)
1	**ICproduct_co_TRANSLOC_6**	47.989	0.770	0.995	0.00291
2	**ICproduct_BiP_release**	47.989	0.770	0.995	0.00291
3	**ICproduct_PDI_2**	47.989	0.770	0.995	4.23×10^{-5}
4	**ICproduct_GOLGI_MGAT2**	19.549	1.000	1.000	1.69559×10^{-6}
5	**BiP_ATPase**	47.989	0.770	0.995	0.00292
6	**Average**	40.078	0.687	0.837	-2.27×10^{-5}

Reactions 1, 2 and 5 in Table 2 were identified as the most likely to be bottlenecks of the secretory pathway. This is determined by their significant increase in Pearson coefficient when changing from FBA to ecFBA, as well as having the greatest reduced cost of all secretory pathway reactions, suggesting changes in these reaction fluxes would have a significant effect on IgG production.

From plotting the metabolic network of these reactions (Fig. 10), it is apparent BiP_ATPase plays a key role. BiP (Binding immunoglobulin protein) is a molecular chaperone playing a key role in post-translational protein folding [29] as well as translocation of proteins such as IgG into the endoplasmic reticulum [30]. The BiP_ATP cycle regulates these interactions between BiP and its substrate [31], thus constraining this key flux in the secretory pathway would have a substantial impact on IgG secretion. The key role of BiP is reflected in the stoichiometry of the reactions in the network linking these reactions, which necessitates the flux of BiP_ATPase be significantly greater (x35) than the other reactions (Fig. 10). BiP_ATPase has the greatest flux, meaning it is more affected by ecFBA's enzyme capacity constraint, and thus be the bottleneck in the IgG secretory pathway. The enzyme catalysing this reaction is thus a potential target for cell engineering; specifically, enzyme engineering to increase its activity [32,33] and subsequently IgG secretion.

While our results suggest BiP_ATPase is the most likely bottleneck, the other reactions present in Table 2 have large Pearson coefficients (between their fluxes and IgG secretion), and reduced costs. Additionally, only 15 of the 101 reactions in the secretory pathway have enzyme kinetic data and are thus constrained by ecFBA. Given this lack of data, it is highly likely that the reaction which is the bottleneck has not been constrained at all. Further investigation is needed to determine the true bottleneck of the secretory pathway.

Figure 10. Escher map of reactions 1,2 and 5 in table 2. Metabolite *ICproduct-SEC61-SPC[r]* (left) comes from a direct line of reactions from the start of the secretory pathway. Metabolite *ICproduct[r]* (right) leads to a series of reactions and protein folding steps that result in IgG secretion. Thick red lines correspond to larger reaction fluxes with the numerical value following the reaction name.

5. Conclusion and Outlook

This work concludes ecFBA models CHO cell metabolism more accurately than the benchmark methods, but can be improved with additional enzymatic data. It was found the value of C had a minor impact on the accuracy of ecFBA, with the ranges of Pearson coefficient, RMSE, and capability being ±0.0581, ±1.29, and ±2.20% respectively. The addition of estimated enzyme parameters, however, greatly improved the accuracy of ecFBA, decreasing the average RMSE by 11.6. Finally, the addition of enzyme parameters for the secretory pathway increased the accuracy of IgG flux prediction by 3% on average, and allowed for the identification of BiP_ATPase as a potential bottleneck in the secretory pathway.

To further validate these findings more experimental datasets, covering a wider range of cellular conditions and gene expression could be included. An obvious area in which ecFBA can be improved is with the addition of more enzyme kinetic data. Currently less than 45% of reactions in the GeM have complete enzyme data, with the rest reliant on estimated enzyme parameters which almost certainly under-constrain the relevant fluxes. The addition of more enzyme kinetic data will likely make ecFBA's flux predictions more accurate. This data will also help in conclusively determining the bottleneck of the IgG secretory pathway, which could be subsequently verified by experiment. This increased understanding of the IgG secretory pathway could be utilised to increase CHO cell IgG production.

A feature of GeMs which this work has not fully utilised are the gene-protein-reaction relations. A new area of research in systems biology is ME models, which incorporate gene expression data into the model along with metabolic data [34]. These models, being more comprehensive, are able to model cell phenotypes more accurately [35] and the development of this model to an ME model would be a natural evolution of the GeM after the development of a more comprehensive ecFBA model.

6. Acknowledgements

The Authors would like to thank James Morrissey and Ben Strain (Imperial College London) for their support and guidance throughout this project.

7. References

[1] Jayapal, KP, Wlaschin, KF, Hu, WS & Yap, MGS 2007, "Recombinant protein therapeutics from CHO Cells - 20 years and counting", *Chemical Engineering Progress*, vol. 103, no. 10, pp. 40-47.

[2] O'Brien, E.J., Monk, J.M. and Palsson, B.O. (2015) "Using genome-scale models to predict biological capabilities," *Cell*, 161(5), pp. 971–987. Available at: https://doi.org/10.1016/j.cell.2015.05.019.

[3] Antonakoudis, A. *et al.* (2020) "The Era of Big Data: Genome-scale modelling meets machine learning," *Computational and Structural Biotechnology Journal*, 18, pp. 3287–3300. Available at: https://doi.org/10.1016/j.csbj.2020.10.011.

[4] Maranas, C.D. and Zomorrodi, A.R. (2016) *Optimization methods in metabolic networks*. Hoboken, NJ: Wiley.

[5] Orth, J.D., Thiele, I. and Palsson, B.Ø. (2010) "What is Flux Balance Analysis?," *Nature Biotechnology*, 28(3), pp. 245–248. Available at: https://doi.org/10.1038/nbt.1614.

[6] Kontoravdi, Cleo; Strain, Benjamin; Morrissey, James; Antonakoudis, Athanasios (2022), "iCHO2441 genome-scale metabolic model", Mendeley Data, V1, doi: 10.17632/73cmrfk8x9.1

[7] Yeo, H.C. *et al.* (2020) "Enzyme capacity-based genome scale modelling of Cho Cells," *Metabolic Engineering*, 60, pp. 138–147. Available at: https://doi.org/10.1016/j.ymben.2020.04.005.

[8] Chang, A. *et al.* (2020) "Brenda, the Elixir Core Data Resource in 2021: New Developments and updates," *Nucleic Acids Research*, 49(D1). Available at: https://doi.org/10.1093/nar/gkaa1025.

[9] Angione, C. (2019) "Human Systems Biology and metabolic modelling: A review—from disease metabolism to precision medicine," *BioMed Research International*, 2019, pp. 1–16. Available at: https://doi.org/10.1155/2019/8304260.

[10] Kuo, C.-C. *et al.* (2018) "The emerging role of systems biology for engineering protein production in Cho Cells," *Current Opinion in Biotechnology*, 51, pp. 64–69. Available at: https://doi.org/10.1016/j.copbio.2017.11.015.

[11] Lewis, N.E., Nagarajan, H. and Palsson, B.O. (2012) "Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods," *Nature Reviews Microbiology*, 10(4), pp. 291–305. Available at: https://doi.org/10.1038/nrmicro2737.

[12] Beg, Q.K. *et al.* (2007) "Intracellular crowding defines the mode and sequence of substrate uptake by *escherichia coli* and constrains its metabolic activity," *Proceedings of the National Academy of Sciences*, 104(31), pp. 12663–12668. Available at: https://doi.org/10.1073/pnas.0609845104.

[13] Sánchez, B.J. *et al.* (2017) "Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints," *Molecular Systems Biology*, 13(8), p. 935. Available at: https://doi.org/10.15252/msb.20167411.

[14] Mathias, S. *et al.* (2018) "Visualisation of intracellular production bottlenecks in suspension-adapted CHO cells producing complex biopharmaceuticals using fluorescence microscopy," *Journal of Biotechnology*, 271, pp. 47–55. Available at: https://doi.org/10.1016/j.jbiotec.2018.02.009

[15] Reinhart, D. *et al.* (2014) "In search of expression bottlenecks in recombinant cho cell lines—a case study," *Applied Microbiology and Biotechnology*, 98(13), pp. 5959–5965. Available at: https://doi.org/10.1007/s00253-014-5584-z.

[16] Lewis, N.E. *et al.* (2010) "OMIC data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models," *Molecular Systems Biology*, 6(1), p. 390. Available at: https://doi.org/10.1038/msb.2010.47.

[17] Lularevic, M. *et al.* (2019) "Improving the accuracy of flux balance analysis through the implementation of carbon availability constraints for intracellular reactions," *Biotechnology and Bioengineering*, 116(9), pp. 2339–2352. Available at: https://doi.org/10.1002/bit.27025.

[18] García Sánchez, C.E. and Torres Sáez, R.G. (2014) "Comparison and analysis of objective functions in flux balance analysis," *Biotechnology Progress*, 30(5), pp. 985–991. Available at: https://doi.org/10.1002/btpr.1949.

[19] Boyle, N.R., Sengupta, N. and Morgan, J.A. (2017) "Metabolic flux analysis of heterotrophic growth in Chlamydomonas reinhardtii," *PLOS ONE*, 12(5). Available at: https://doi.org/10.1371/journal.pone.0177292.

[20] Herrmann, H.A. *et al.* (2019) "Flux sampling is a powerful tool to study metabolism under changing environmental conditions," *npj Systems Biology and Applications*, 5(1). Available at: https://doi.org/10.1038/s41540-019-0109-0.

[21] Bateman, A. *et al.* (2020) "Uniprot: The Universal Protein Knowledgebase in 2021," *Nucleic Acids Research*, 49(D1). Available at: https://doi.org/10.1093/nar/gkaa1100.

[22] Ebrahim, A. *et al.* (2013) "COBRApy: Constraints-based reconstruction and analysis for Python," *BMC Systems Biology*, 7(1). Available at: https://doi.org/10.1186/1752-0509-7-74.

[23] Hefzi, H. *et al.* (2016) "A consensus genome-scale reconstruction of Chinese hamster ovary cell metabolism," *Cell Systems*, 3(5). Available at: https://doi.org/10.1016/j.cels.2016.10.020.

[24] Shlomi, T. *et al.* (2011) "Genome-scale metabolic modeling elucidates the role of proliferative adaptation in causing the Warburg effect," *PLoS Computational Biology*, 7(3). Available at: https://doi.org/10.1371/journal.pcbi.1002018.

[25] Owen, O.E., Kalhan, S.C. and Hanson, R.W. (2002) "The key role of anaplerosis and cataplerosis for citric acid cycle function," *Journal of Biological Chemistry*, 277(34), pp. 30409–30412. Available at: https://doi.org/10.1074/jbc.r200006200.

[26] McAtee Pereira, A.G. *et al.* (2018) "13C flux analysis reveals that rebalancing medium amino acid composition can reduce ammonia production while preserving central carbon metabolism of cho cell cultures," *Biotechnology Journal*, 13(10), p. 1700518. Available at: https://doi.org/10.1002/biot.201700518.

[27] Templeton, N. *et al.* (2013) "Peak antibody production is associated with increased oxidative metabolism in an industrially relevant fed-batch cho cell culture," *Biotechnology and Bioengineering*, 110(7), pp. 2013–2024. Available at: https://doi.org/10.1002/bit.24858.

[28] Templeton, N. et al. (2017) “Application of 13C flux analysis to identify high-productivity cho metabolic phenotypes,” Metabolic Engineering, 43, pp. 218–225. Available at: https://doi.org/10.1016/j.ymben.2017.01.008

[29] Torres, M., Hussain, H. and Dickson, A.J. (2022) “The secretory pathway – the key for unlocking the potential of Chinese hamster ovary cell factories for manufacturing therapeutic proteins,” *Critical Reviews in Biotechnology*, pp. 1–18. Available at: https://doi.org/10.1080/07388551.2022.2047004.

[30] Nguyen, T.H., Law, D.T. and Williams, D.B. (1991) “Binding protein bip is required for translocation of secretory proteins into the endoplasmic reticulum in saccharomyces cerevisiae.,” *Proceedings of the National Academy of Sciences*, 88(4), pp. 1565–1569. Available at: https://doi.org/10.1073/pnas.88.4.1565.

[31] Pobre, K.F., Poet, G.J. and Hendershot, L.M. (2019) “The endoplasmic reticulum (ER) chaperone BiP is a master regulator of ER functions: Getting by with a little help from Erdj Friends,” *Journal of Biological Chemistry*, 294(6), pp. 2098–2108. Available at: https://doi.org/10.1074/jbc.rev118.002804.

[32] Fisher, A.K. *et al.* (2014) “A review of metabolic and enzymatic engineering strategies for designing and optimizing performance of microbial cell factories,” *Computational and Structural Biotechnology Journal*, 11(18), pp. 91–99. Available at: https://doi.org/10.1016/j.csbj.2014.08.010.

[33] Basheer, S.M. and Chellappan, S. (2017) “Enzyme engineering,” *Bioresources and Bioprocess in Biotechnology*, pp. 151–168. Available at: https://doi.org/10.1007/978-981-10-4284-3_6.

[34] Thiele, I. *et al.* (2012) “Multiscale modeling of metabolism and macromolecular synthesis in E. coli and its application to the evolution of codon usage,” *PLoS ONE*, 7(9). Available at: https://doi.org/10.1371/journal.pone.0045635.

[35] O'Brien, E.J. *et al.* (2013) “Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction,” *Molecular Systems Biology*, 9(1), p. 693. Available at: https://doi.org/10.1038/msb.2013.52.

A Dual Polymer Chemical Consolidation Approach for the Structural Reinforcement of Calcium Carbonate Reservoirs

Paulina Gordina and Katya Longinova

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

Abstract Fines migration in carbonate reservoirs presents numerous operational challenges and consequences within the oil and gas industries. This study explores the use of a dual polymer chemical consolidation approach to strengthen carbonate reservoirs and control fines migration and compares it to the previously studied single polymer approach. For 60g of $CaCO_3$, the optimal secondary cationic polymer concentration to be used in conjunction with 90g of a 2000 ppm solution of the primary anionic polymer, a polyacrylamide (PAM) called FLOPAAM 3330S, was optimized on a range of concentrations from 2000 ppm to 7000 ppm. The optimal one was found to be 30g of a 5000 ppm solution, which yielded a 40% increase in unconfined compressive strength (UCS) when compared to consolidation using only FLOPAAM 3330S. When the optimal concentration was tested on a broad range of potential secondary polymers, the greatest improvement in compressive strength was obtained from FLOPAM FO 4650 VHM, a high molecular weight (MW) cationic PAM, followed closely by FLOPAM FO 4698 SSH, a slightly lower MW cationic PAM. The high MWs of the secondary PAMs serve to enhance bridging interactions between polymers and $CaCO_3$ particles, leading to increased flocculation and improved compressive strength. The increased compressive strength came at the expense of porosity, as consolidated sample porosity was shown to be 21% lower than that of the untreated $CaCO_3$. Lastly, a temperature degradation experiment performed on the best two secondary polymers showed that polymer bonds break at reservoir temperature (100 °C), reducing the MW of the PAMs and leading to lower flocculation ability. The conclusion of the study showed that a dual polymer approach can lead to an improvement in UCS of a $CaCO_3$ sample, giving good results at room temperature and ambient pressure. However, further testing is required to give concrete results as to what combination will work best at true reservoir conditions and give the least porosity drop.

***Keywords**:* Carbonate reservoirs, Polyacrylamides, UCS, Flocculation, Porosity

Introduction & Background

Industrial Context

Carbonate reservoirs have been a dominating source of oil and gas for many years. They will likely continue to be so, as it is estimated that 60% and 40% of the world's remaining oil and gas respectively are stored in reservoirs such as these [1].

These reservoirs are composed of primarily calcite ($CaCO_3$) and dolomite ($CaMg(CO_3)_2$) and typically characterized by a heterogeneous pore structure with relatively high porosity [2]. Usually, due to their young geological age, many of these reservoirs are weakly consolidated [3] and thus prone to shear breakage and a phenomenon called pore collapse. As more and more oil and gas are extracted, the reservoir pressure decreases, consequently increasing the effective stress on the rock, leading to shear failure and pore collapse [4,5]. While this behavior is observed in many different rock structures, it is particularly a danger for highly porous rocks, such as those making up carbonate reservoirs, as the stress level needed to cause shear failure and pore collapse is much lower than rocks with a lower porosity [6].

Following either of these types of failure, the production of "fines" is observed from the crushed rock. Fines migration, the movement of these particles through the pores of the rock structure, is notorious in the oil, gas and carbon capture industries as it often leads to the plugging up of pores and consequently, a potentially irreversible decrease in the permeability of the rock structure, as shown in Figure 1 [7,8]. Well productivity losses of up to 100% have been reported as a direct consequence of this phenomenon [9]. Furthermore, the fines can be damaging to equipment used in extraction, incurring both unwanted maintenance costs and safety concerns [5,10]. Mitigation of pore collapse and fines migration are therefore critical to the continued efficiency of carbonate reservoirs.

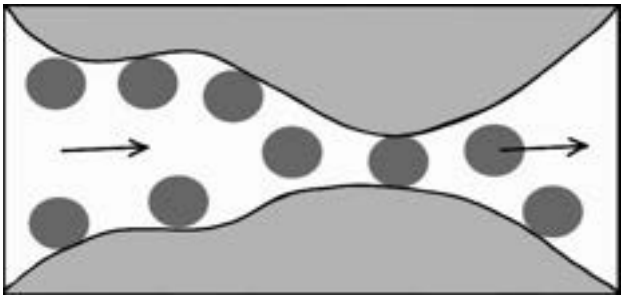


Figure 1: Fines migration leading to pore plugging and permeability decrease in carbonate reservoirs

Mechanical vs Chemical Consolidation

A great many techniques for the control of fines migration have been developed over the last century, all of them

falling chiefly into two categories: mechanical and chemical. Mechanical methods generally employ screens, filters, or gravel packing to block fines migration, while chemical methods involve the injection of a chemical consolidant into the well with the aim of both catching loose particles and reinforcing the overall rock structure by improving its compressive strength [3,10,11]. It is generally acknowledged that mechanical methods are more costly, time-consuming and have more associated issues, including equipment erosion, installation damage, and ineffectiveness for small particles, making chemical techniques a favorable choice [10,11].

Even among chemical methods, there are countless choices of consolidant, each presenting a specific set of advantages and disadvantages. A review of chemical consolidation methods for fines control conducted and compiled by Alakbari et al. [10] found that many of the resins and polymers that have historically been used to control fines migration in sandstone (another widespread type of reservoir) have notable associated drawbacks such as insignificant compressive strength improvement and environmental harm. Furthermore, those that have successfully introduced considerable reinforcement to the rock structure have done so at the expense of rock permeability.

Relationship of Rock Strength and Permeability

As the chemical consolidant mixes with loose rock particles, a decrease in porosity is often observed due to the nature of the binding taking place. The relationship between rock porosity and strength is most often inverse and has been studied extensively even outside of the oil and gas industries due to its relevance in the conservation of sculptures, monuments, and buildings [12,13]. Furthermore, the positive relationship between permeability and porosity is described by the Carman-Kozeny equation [14], which can be found in Appendix 1, allowing the inference of an inverse relationship between rock strength and permeability. Too high a drop in reservoir permeability leads to fewer paths to the surface for extracted material, impeding extraction efficiency. Thus, preserving the reservoir's permeability is key. An ideal chemical consolidant would increase the compressive strength of the rock without compromising too much on permeability.

Single Polymer Consolidation

According to Alakbari et al. [10], one of the polymer types that had the fewest drawbacks was polyacrylamide (PAM). PAMs have been a popular choice of flocculant within the wastewater industry as well as within enhanced oil recovery (EOR) applications for many years due to their favorable price, water-solubility, and ease of modification to different molecular weights (MW) and charges [15,16]. Their ability to cross-link makes them highly effective consolidants as well [10]. A study by Salehi et al. [17] found that a sulfonated PAM cross-linked with $Cr(OAc)_3$ improved the compressive strength of an unconsolidated sandstone thirty-fold. No permeability data was presented from this study, however.

The two main mechanisms governing particle flocculation following the addition of a polymer to the system are charge neutralization and bridging interactions. The first is straightforward in principle: polymer molecules of a certain charge adsorb onto particles of the opposite charge and aggregate them. In this case, charge density plays a deciding role in the effectiveness of the flocculant and often governs the optimal dosage of a particular polymer [18]. For high MW polymers, bridging interactions, in which one polymer molecule can adsorb onto multiple particles are observed [19].

Most of the existing research in chemical consolidation using PAMs has been conducted on sandstone reservoirs, which are mostly made up of negatively charged silica particles, and thus the applicable PAMs are cationic (C-PAM) [20]. Carbonate fines are most often positively charged under reservoir conditions [21], however there is limited literature examining the application of an anionic PAM to aggregate them [11,13].

A recent study by Lew et al. [11] examined the adsorption of three hydrolyzed polyacrylamides (HPAM), which are anionic by nature, onto calcium carbonate ($CaCO_3$). Their findings indicated that the equilibrium amount of polymer adsorbed onto the solids increased with increasing MW due to the different natures of the adsorption processes for different MWs. The conceptual difference between low and high MW polymer adsorption onto a particle surface is depicted in Figure 2.

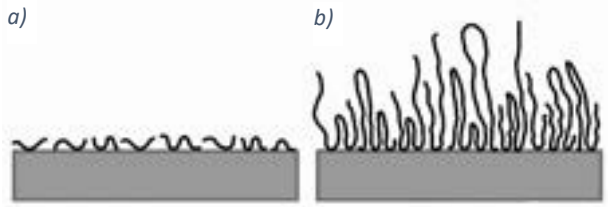


Figure 2a) depiction of a low MW polymer adsorbed onto particle surface 2b) depiction of a high MW polymer adsorbed onto particle surface [11]

A higher MW polymer forms "loops and tails" on the particle surface (Figure 2b), while a lower MW polymer adopts a much flatter configuration (Figure 2a). Using the formed loops and tails, the high MW polyacrylamides have the capacity to form larger flocs through bridging, as shown in Figure 3 [11,19]. This indicates that for the chemical consolidation of $CaCO_3$ particles, a higher MW polymer might be preferred in order to enhance the bridging interactions between molecules and form larger flocs that are more resistant to high shear [18].

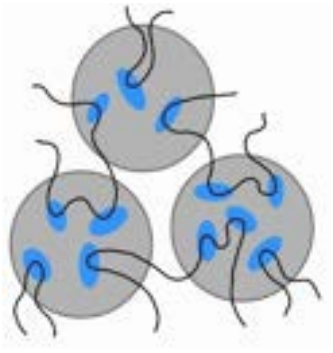


Figure 3: Bridging interactions for a single polymer system [18]

With the addition of enough polymer, it is possible for the particle system's overall charge to change [18]. In fact, through a zeta potential analysis in their study, Lew et al. [11] also demonstrated that after the addition of HPAM to the $CaCO_3$, the charge of the system becomes overall negative. This invites the possibility of using a dual polymer approach to further flocculate the $CaCO_3$ particles in the system.

Dual Polymer Consolidation

Dual polymer flocculation has found uses in several industries, most notably dewatering of activated sludge and papermaking [18]. The addition of a secondary polymer to a colloidal system has been shown to significantly improve flocculation in certain systems. A study done on the flocculation of alumina particles by Yu and Somasundaran [18] demonstrated that after the addition of a secondary anionic polymer (polyacrylic acid) to a pre-adsorbed mixture of alumina and cationic polymer (PDADMAC), a significantly improved flocculation response was obtained, even at lower dosages than the original single polymer system. This result was attributed to the enhanced bridging interactions caused by polymer chain interactions of the polyacrylic acid with the PDADMAC [18]. A conceptualization of dual polymer bridging interactions can be seen in Figure 4.

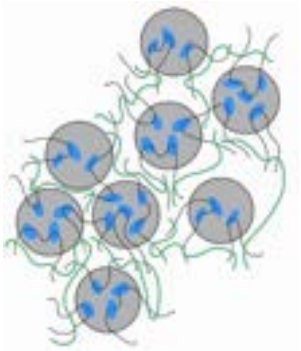


Figure 4: Dual polymer chain interactions leading to enhanced bridging between particle flocs, inferred by the authors

Objectives

This study aims to discover whether a dual polymer approach could be more successful than a single polymer one in flocculating $CaCO_3$ with the aim of consolidating carbonate reservoirs. The work is meant to be exploratory and either confirm or deny the need for further investigation into the subject. Within the scope of polymer availability, the work specifically aims to find both the optimal secondary cationic polymer for use with the set anionic primary polymer and the optimal dosage of said secondary polymer. Improvements in compressive strength and changes in porosity of samples treated using a dual polymer approach, as opposed to a single polymer approach, are considered as indicators of success.

Materials and Methodology

Assumptions

All experiments, unless otherwise specified, were performed at room temperature and pressure (RTP). These conditions are not reflective of reservoir conditions and are meant only to give an indication of polymer performance to determine whether further study is necessary.

The polymers studied were chosen based on industrial relevance and availability. The density of polymer solution was assumed to be equal to water density based off previous work done by the Luckham research group. All variables tested were assumed to be independent of each other.

Pre-Testing & Sample Preparation

Each sample was prepared with 60g of $CaCO_3$. The sample preparation sequence was dictated by several parameters: primary and secondary polymer amounts, water mass in sample, total mass of sample, and final sample texture and mixability.

After preliminary testing (Appendix 2) and experimentation with these parameters, the following sample preparation sequence was derived:

1. 60g of $CaCO_3$ was mixed with 90g of 2000 parts per million (ppm) of an anionic PAM, FLOPAAM 3330S, for 1 hour.
2. 30g of a secondary cationic polymer solution was mixed into the sample for 30 minutes.

The full step-by-step procedure for sample preparation can be found in Appendix 3. All components of the first step were based off previous research within the Luckham group. Their work has indicated that the anionic PAM in this amount gave strong single polymer consolidation results. This step was fixed as testing for other primary polymers was outside the scope of this study.

The total "wet" mass of each sample was therefore approximately 180g. The samples were separated into three cylindrical drying tubes and set on a drying rack next to a fan. Sample drying took on average five to ten days. A sample was considered dry once there was less than 5% mass change two days in a row.

Unconfined Compressive Strength Testing

The instrument used to gauge the compressive strength of dried samples was Lloyd EZ-50. Samples were compressed at a constant deformation rate of $1mm \cdot s^{-1}$ until a sharp drop-off in unconfined compressive strength (UCS) was observed following the sample breaking.

Details on sample preparation for UCS testing and machine settings are in Appendices 4 and 5.

UCS Data Processing

The target data obtained from compressive strength testing was the peak load which was recorded for each sample tested. All samples for which the length to diameter ratio ranged from 1 to 3, i.e., $1 \leq L \cdot D^{-1} \leq 3$, an adjustment formula was used to obtain an accurate value for true peak load [22] (Eq. 1).

$$C = \frac{C_a}{0.88+(0.24 \cdot \frac{D}{L})} \qquad \text{(Eq. 1)}$$

Where L is the height and D is the average diameter of the sample core, in mm, C is the calculated compressive strength of an equivalent 2:1 length/diameter sample or the "adjusted" UCS, in N, and C_a is the measured compressive strength of the sample tested, in N, obtained from EZ-50 results. Sample calculations of raw UCS data processing can be found in Appendix 6.

Porosity Testing

To measure the porosities and pore size distributions of the samples, mercury intrusion porosimetry (MIP) was used. Low- and high-pressure intrusion tests were conducted on Micrometrics' Autopore IV 9500 and data for the porosity and pore size distribution with respect to the differential intrusion for each sample were extracted from the summary report produced by the machine. Details on sample preparation for MIP can be found in Appendix 7.

General Experimental Sequence

Concentration Profile of Secondary Polymer

The first set of experiments focused on determining the best secondary polymer concentration within the scope of the sample preparation method.

As previously mentioned, the primary polymer type and concentration were fixed for all experiments, and it was assumed that the optimal concentration of secondary polymer was independent of polymer type.

The secondary polymer for this set of experiments, FLOPAM FO 4650 VHM, was chosen arbitrarily from those selected for testing. The set of testing concentrations ranged from 2000 to 7000 ppm, as seen in Table 1. All samples were prepared as outlined in the general sample preparation method, with one sample for each secondary polymer concentration. Once dry, the UCS was measured for each sample.

Table 1: Concentrations of secondary polymer tested for a 30g solution

Mass of secondary polymer solution (g)	Concentration of solution (ppm)
30	2000
	3000
	4000
	5000
	6000
	7000

Polymer Type Testing

Following the concentration profile testing, different secondary polymers were tested at the determined optimal concentration. This approach was based on the assumption that the optimal concentration was independent of polymer type. The primary polymer type and concentration were, once again, fixed for all experiments.

The secondary polymers investigated can be found in Table 2. These were chosen to test different polymer types (both PAMs and non-PAMs) and a broad scale of MWs. All options were specifically cationic to leverage the charge difference between the primary and secondary polymers for enhanced flocculation. The goal was to see which polymer would perform best and thus give an indication of what polymer properties should be at the focus of further investigations, if applicable.

All samples were prepared as outlined in the general sample preparation method and once dry, the UCS for each sample was measured.

Temperature Degradation of Polymers

All concentration profile and secondary polymer type experiments were performed at RTP, whereas reservoir conditions are closer to 100 °C [23]. To gain an understanding of how results may change under reservoir

Table 2: Secondary polymers tested for dual polymer consolidation approach. All information is from manufacturer description on packaging

Producer	Polymer Name	Polymer Type	Average MW	Charge Density
SNF	FLOPAM FO 4650 VHM	Polyacrylamide	Ultra-high	High
SNF	FLOPAM FO 4698 SSH	Polyacrylamide	High	Medium
BASF	Alcomer SK	Polyethylenimine modified	Low	High
BASF	Alcomer 819	Polyacrylate ester quat	Medium	High
Sigma Aldrich	DAC	Poly(diallyldimethylammonium chloride)	Low	High

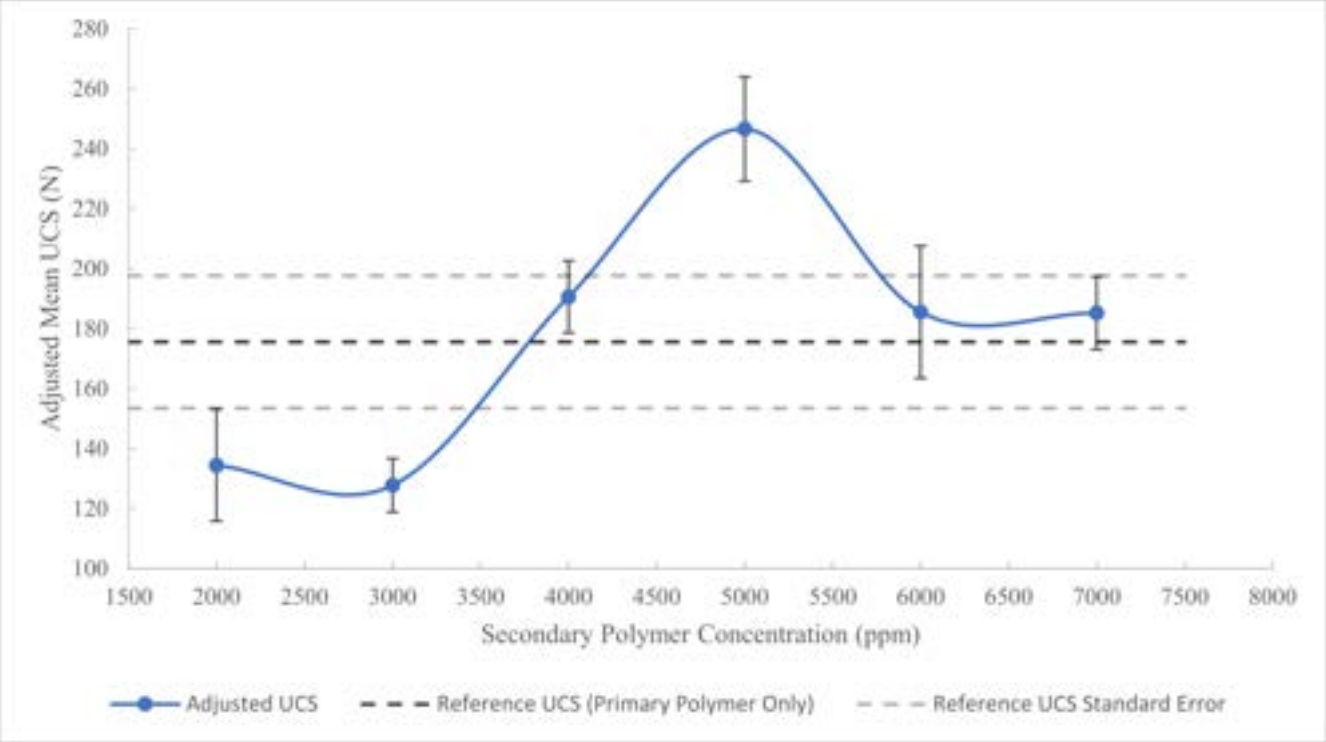


Figure 5: Adjusted mean UCS for a concentration profile of FLOPAM FO 4650 VHM relative to UCS for a single polymer consolidation

conditions, a temperature degradation test was performed for the two top-performing secondary polymers.

Two solutions of these polymers were made at optimal concentration. Using a Thermo Scientific HAAKE MARS 60 rheometer, the viscosity measurement, in $Pa \cdot s$, over shear strain, in s^{-1}, was found for each sample at 25 °C. The two samples were then heated in an oven at 100 °C for 24 hours. The viscosity over shear strain of the heat-treated samples was then measured. Details of the machine settings for the rheometer are outlined in Appendix 8.

Results & Discussion

Concentration Profile of FLOPAM FO 4650 VHM

A range of concentrations of a 30g FLOPAM FO 4650 VHM dilution was added to a 90g solution of primary polymer at 2000 ppm and 60g $CaCO_3$. The UCS results were then compared to a sample consolidated using only 120g of the primary polymer at a concentration 1500 ppm. The total masses of the compared samples are thus kept constant at 180g.

A general positive trend was observed between the concentration of added secondary polymer and the resulting UCS of the sample. As shown in Figure 5, this relationship peaks at a secondary polymer concentration of 5000 ppm with an adjusted UCS of about 245 N. After this, with a higher concentration of secondary polymer, the compressive strength declines, before plateauing around the UCS of the primary polymer sample, represented by the black dotted line at approximately 175 N. The grey dotted lines on either side of this reference value represent the standard error of the primary polymer sample, calculated based on triplicated sample measurements on Lloyd EZ-50. For a detailed description of the calculation of the measurement error associated with triplicated UCS results, see Appendix 9.

The optimal amount of FLOPAM FO 4650 VHM shows a marked improvement on the compressive strength of the consolidated $CaCO_3$ when compared to the single polymer sample, by approximately 40%. Concentrations around the optimal, namely 4000 ppm, 6000 ppm, and 7000 ppm, also give minor improvements on the reference single polymer value. However, since the 4000 ppm, 6000 ppm and 7000 ppm data points fall within the standard error margin of the reference, their improvement is not considered definitive.

As previous studies have shown that the flocculation in a dual polymer system occurs mainly due to polymer bridging, the optimum concentration corresponds to the best conditions for bridging to occur [18,19]. Following this concentration, polymer "overdosing" occurs, leading to a decrease in performance as has been reported by several studies on single polymer flocculation systems [18]. Furthermore, in their study on the effects of shear rate and polymer overdosing on floc formation, Blanco et al. [25] showed that at very high doses of a cationic PAM, if flocs are sheared by the mixing instrument, they do not reform, potentially accounting for the comparatively

Figure 6a) Wet samples mixed with 3000 ppm FLOPAM FO 4650 VHM (left) and 2000 ppm FLOPAM FO 4650 VHM (right) b) Wet sample mixed with 7000 ppm FLOPAM FO 4650 VHM

poorer consolidation of the $CaCO_3$ at high FLOPAM FO 4650 VHM doses. Qualitative observation reinforces this hypothesis, as samples with higher concentrations of secondary polymer were creamier and more homogeneous than those at lower dosages, indicating a breakdown of flocs during the high shear mixing process. A comparison of these samples before they were left to dry can be observed in Figure 6. The samples in Figure 6a have a lower amount of secondary cationic polymer (3000 and 2000 ppm from left to right) and are seen to still contain macroscopic flocs of aggregated $CaCO_3$. The sample in Figure 6b contains 7000 ppm of secondary polymer, equal, which is well into the overdosing region of Figure 5 and its texture is much more uniform with no visible clumps of particles.

Secondary Polymer Type Testing

Based on the results of the concentration profile, a 5000 ppm 30g solution of secondary polymer was accepted as optimal. This concentration was then further tested on a broader range of secondary polymers.

As seen in Figure 7, FLOPAM FO 4650 VHM, the initially selected secondary polymer, gave the best result, followed by a slightly lower MW cationic PAM, FLOPAM FO 4698 SSH. These are the only two secondary polymers tested which resulted in an improvement of compressive strength in the consolidated $CaCO_3$ samples as compared to the single polymer system.

This result is in agreement with Lew et al. [11], who showed that the MW of polymers plays a significant role in the pervasion of the bridging flocculation mechanism, as discussed in the introduction. Among the polymers tested, FLOPAM FO 4650 VHM has the highest MW, allowing for enhanced polymer chain interactions and the formation of stronger flocs. This result also seems to reinforce PAMs as an effective choice as a secondary polymer as well as a primary one. Compared to the two PAMs, DAC, Alcomer 819, and Alcomer SK have very low MWs. This is most likely the reason they do not surpass the primary polymer benchmark.

Since most of the studied polymers have a high charge density, the effect this parameter has on results is unclear, particularly since FLOPAM FO 4698 SSH, the runner-up, is only classified as moderately cationic. In future studies, charge density can be further investigated as a variable parameter and more conclusive results obtained.

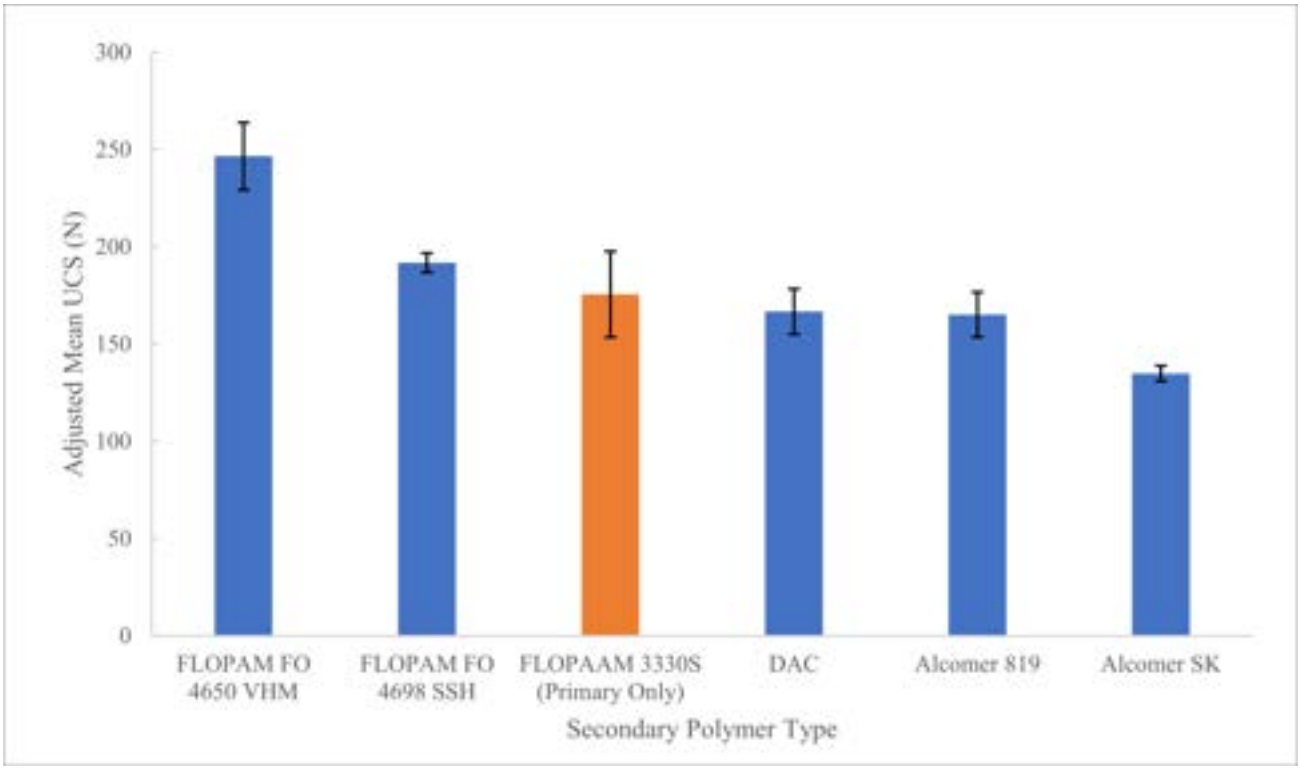


Figure 7: Adjusted mean UCS for a range of cationic secondary polymers relative to UCS for a single polymer consolidation

Since the optimal concentration of secondary polymer was determined through experiment on the polymer which turned out to be the highest performing, the question of the validity of its optimality across the entire concentration spectrum is raised. To test this, a full matrix of polymers and concentrations would need to be considered, which was not possible in this study due to time constraints. Therefore, these results can be taken as an indication of optimality. It can be said with reasonable confidence that secondary high MW cationic PAMs in conjunction with a primary high MW anionic PAM, give the best results in terms of UCS improvement in consolidated $CaCO_3$.

Porosity Analysis

To assess the effects of dual polymer consolidation on the porosity of the $CaCO_3$, the pore size distribution of the strongest sample (flocculated using a 2000 ppm 90g FLOPAAM 3330S solution and a 5000 ppm 30g FLOPAM FO 4650 VHM solution) was quantified using MIP. The resulting distribution and porosity values were compared to the results from an unconsolidated sample of $CaCO_3$ as well as a sample treated with just the primary polymer. The data for the comparison was obtained previously by researchers in the group.

The data for the porosity of each sample can be seen in Table 3. While a minimal decrease in porosity (2%) is observed for the single polymer consolidation using a 1500 ppm 120g FLOPAAM 3330S solution, a significant further drop of approximately 19% occurs upon the addition of the secondary polymer FLOPAM FO 4650 VHM. While porosity is an expected casualty of increased compressive strength through chemical consolidation [10], as described in the introduction, the magnitude of the value is surprising when compared to the initial drop after just single polymer consolidation.

Table 3: Porosities for untreated $CaCO_3$, $CaCO_3$ consolidated using a single polymer approach, and $CaCO_3$ consolidated using a dual polymer approach

Sample	$CaCO_3$	$CaCO_3$ + FLOPAAM 3330S	$CaCO_3$ + FLOPAAM 3330S + FLOPAM FO 4650 VHM
Porosity	80.4%	78.5%	59.1%

This decrease in sample porosity is also reflected in Figure 8, where the differential intrusion peak of pure $CaCO_3$ is notably higher than those of the other samples. The differential intrusion correlates with the total intrusion of mercury in the sample, in $mL \cdot g^{-1}$, and therefore with the porosity. However, Figure 8 also shows that while the three samples all have the same general shape, the pore size distribution becomes broader with more consolidation. The rightward shift of the differential intrusion peak demonstrates that, as more polymer is added to the system, the pores accounting for the most intrusion become bigger in diameter. The "pores" formed between flocs of particles likely account for both the less uniform distribution and the size increase.

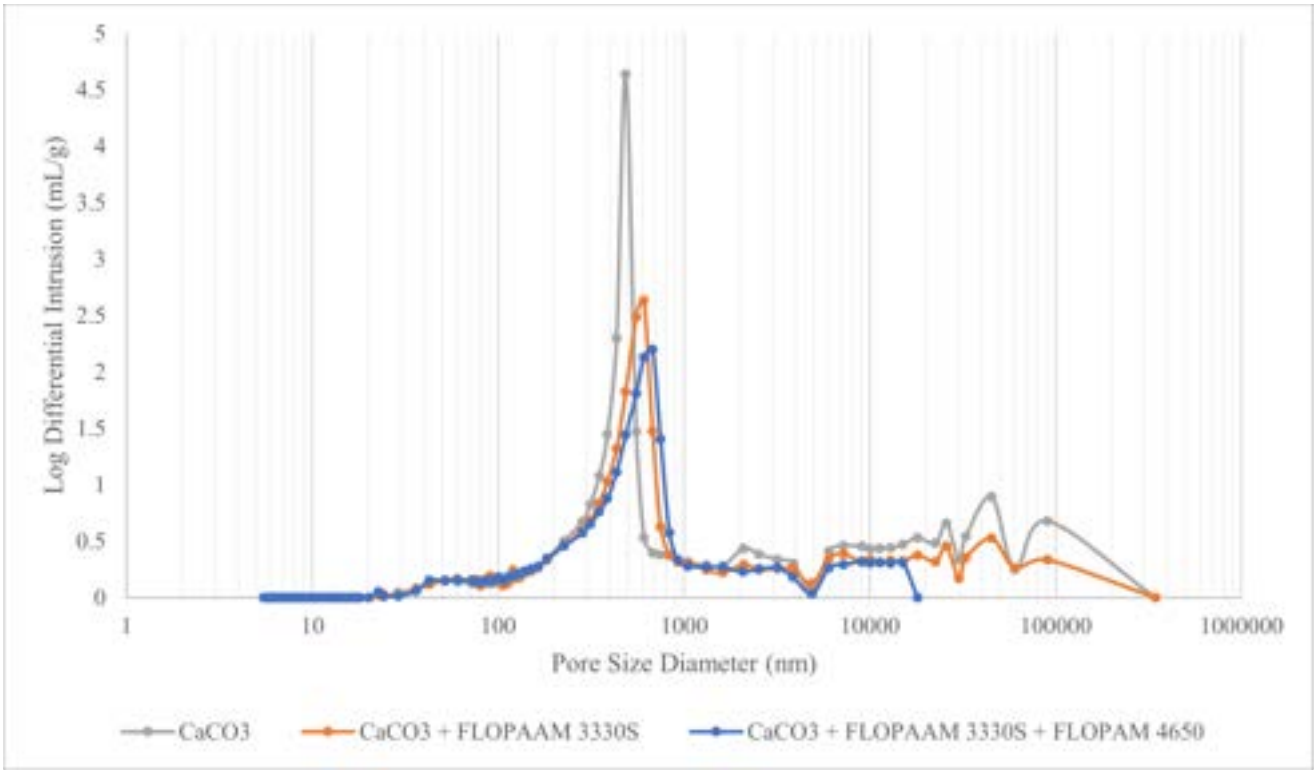


Figure 8: Pore size distributions (on a log scale) for untreated $CaCO_3$, $CaCO_3$ consolidated using a single polymer approach, and $CaCO_3$ consolidated using a dual polymer approach

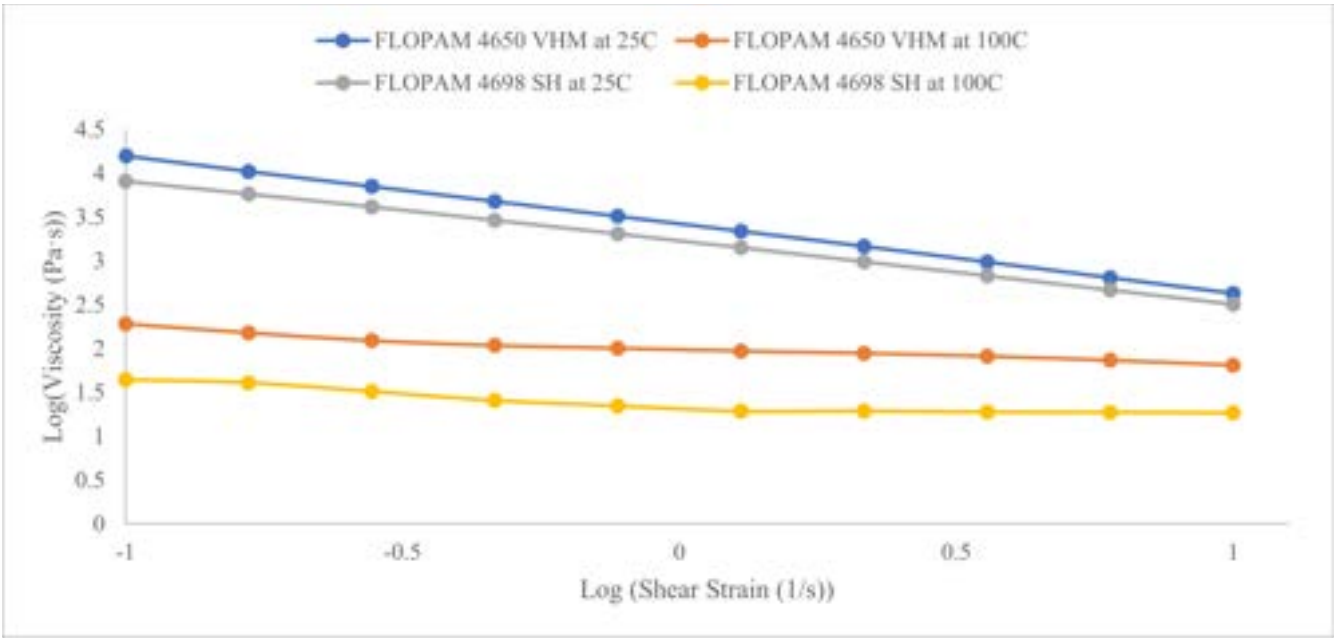


Figure 9: Viscosity over shear strain in a log-log scale, for secondary polymers FLOPAM FO 4650 VHM and FLOPAM FO 4698 SSH at 25 °C and 100 °C

Overall, the porosimetry results indicate that the porosity drop following chemical consolidation using an anionic PAM followed by a cationic PAM is significant, which counts against the increase in compressive strength. Further study into the extent to which porosity affects the efficiency of oil and gas extraction would be necessary to assess this combination's viability for enhanced consolidation of carbonate reservoirs.

Temperature Degradation of FLOPAM FO 4650 VHM and FLOPAM FO 4698 SSH

The effect of temperature on FLOPAM FO 4650 VHM and FLOPAM FO 4698 SSH at 5000 ppm was measured via change of polymer viscosity over shear strain. A log-log (base 10) plot of viscosity vs shear strain can be seen in Figure 9. The overall viscosity of both polymers is higher at 25 °C than at 100 °C. Additionally, the PAM found to be the best option, FLOPAM FO 4650 VHM, has a higher viscosity over all shear strain values, compared to the second-best performing polymer, FLOPAM FO 4698 SSH.

For PAMs, a known relationship between viscosity and MW [25] can be seen in Eq. 2.

$$\eta = 9.33 \times 10^{-2} \times MW^{0.75} \quad \text{(Eq. 2)}$$

Where η is viscosity in $Pa \cdot s$.

This relationship is further reinforced by the higher viscosity polymer being the one of higher MW. This means that the reduction in viscosity of the sample after heating can be linked to reduction in MW of the polymer. This would occur in the case of polymer bonds breaking due to the heat, leading to shorter chains. For the shorter polymer chains, it can then be inferred that there will be fewer bridging interactions to flocculate the $CaCO_3$, leading to a less consolidated sample.

These results show that high MW polymers, which have been recommended for further investigation, experience a reduction in MW due to polymer bond breakage when reservoir conditions are applied. Thus, a lower effectiveness in consolidation of $CaCO_3$ is expected. In any further research, the temperature stability of high MW polymers should be investigated through degradation testing. This will help determine if the polymer in question would perform similarly under real reservoir conditions as it does at RTP.

In the case of this study, the temperature degradation results imply a significant drop in MW of polymer, posing a question regarding the validity of optimality results under reservoir conditions.

Conclusions

The primary objective of this study was to determine preliminarily whether a dual polymer chemical consolidation approach for a carbonate reservoir application could be more successful than a single polymer approach using a primary anionic PAM, FLOPAAM 3330S. Further to this, the study aimed to determine the optimal secondary polymer and its optimal dosage. The main success parameters under consideration were the unconfined compressive strengths (UCS) of the samples and their porosities. Based on the results of this exploration, a decision regarding the need for further investigation of this consolidation technique could be made.

At RTP, a series of experiments adding 30g of FLOPAM FO 4650 VHM at different concentrations to 90g of a 2000 ppm solution of FLOPAAM 3330S mixed with 60g of $CaCO_3$ showed that the optimal concentration of FLOPAM FO 4650 VHM was 5000 ppm. This yielded the greatest increase in compressive strength compared to consolidation by only the primary polymer, by approximately 40%.

As the optimal secondary polymer concentration was considered to be independent of polymer type, this optimal concentration was tested on several other secondary polymer options, representing a large scale of MWs. The highest MW polymer, the initially selected FLOPAM FO 4650 VHM, performed best, followed by the second-highest MW polymer, FLOPAM FO 4698 SSH. The examination of different secondary polymers indicates that high MW is an important parameter when it comes to improving UCS, due to the enhanced bridging interactions that it allows. This result is in line with several similar studies performed on single polymer systems.

The increase in UCS of the strongest sample presented a trade-off with the sample porosity. A MIP analysis showed that, while a 2% porosity decrease is observed after the addition of the primary polymer to the $CaCO_3$, a further 19% drop follows the addition of a second polymer to the system. This result reflects the theoretical understanding that porosity is negatively correlated with strength. It presents a significant drawback in the use of a secondary polymer in carbonate reservoir consolidation, as it could lead to a high permeability reduction of the carbonate rock, impeding reservoir productivity.

Thus, it can be said that the dual polymer consolidation technique is effective, however it comes at a price. Further investigation is highly recommended to determine whether its advantages outweigh its disadvantages.

Lastly, a temperature degradation test of the top two performing secondary polymers showed that at reservoir temperature, 100 °C, the polymer bonds break. The broken polymer chains are shorter, reducing the bridging properties of the polymer and thus reducing the strength of the sample. Thus, when further testing polymers, reservoir conditions must be considered, as degradation due to temperature is significant.

Outlook

The results obtained from this study indicate that the high MW cationic FLOPAM FO 4650 VHM is the optimal secondary polymer to use for carbonate reservoir consolidation and that the optimal dose is 30g of a 5000 ppm dilution for 60g of $CaCO_3$. However, since the dosage was optimized on this particular polymer and later tested on others, it cannot be said with absolute certainty that no better combination exists.

Experimental examination of the full polymer matrix (i.e., every tested polymer at all possible concentrations) would be necessary to verify the results of this study. Furthermore, additional research into other high MW cationic PAMs would be beneficial to capitalize on the evidence that MW drives increased consolidation through polymer bridging. A comprehensive examination of other high MW PAMs could also shed light on whether there are any that can give the same improved UCS but at less of a cost to porosity and thermal stability. Effects of charge density on sample strength could also be quantified through an examination of polymers with similar molecular weights but varying charge densities.

Results from the preliminary temperature degradation experiment indicate that polymer performance worsens with increasing temperature. It is recommended to also examine changes in polymer behavior at other reservoir conditions including the presence of brine, and higher pH.

Lastly, the crux of deciding whether a dual polymer approach to carbonate reservoir consolidation is worth it over a single polymer approach lies in accepting the sacrifice of porosity for improved compressive strength. With the addition of a secondary polymer, as some pores are blocked and some formed between the flocs of particles, a more complex model of the change in porosity and permeability in a sample is necessary to understand the level of interference the addition of the polymer introduces to the practical application of the study: oil and gas extraction. Following this investigation, a final decision regarding the superiority of the dual polymer approach can be made.

Acknowledgements

The authors would like to thank their supervisors Professor Paul Luckham and Shawn Lew for their unwavering support and guidance throughout this project. They would also like to recognize the contributions of Patricia Carry and Kaho Cheung, who ran all mercury intrusion porosimetry measurements for this study as well as providing comprehensive training on Lloyd EZ-50.

References

[1] Bristol U of. Carbonate reservoirs | Research | University of Bristol [Internet]. www.bristol.ac.uk. Available from: http://www.bristol.ac.uk/research/impact/carbonate-reservoirs/#:~:text=More%20than%2060%20per%20cent

[2] Chang FF. Acid fracturing stimulation. Fluid Chemistry, Drilling and Completion. 2022;387–419.

[3] Dees JM. Method of sand consolidation with resin [Internet]. 1993. Available from: https://patents.google.com/patent/US5178218A/en

[4] Zaman M, Roegiers J-C ., Abdulraheem A, Azeemuddin M. Pore Collapse in Weakly Cemented and Porous Rocks. Journal of Energy Resources Technology. 1994 Jun 1;116(2):97–103.

[5] Talaghat MR, Esmaeilzadeh F, Mowla D. Sand production control by chemical consolidation. Journal of Petroleum Science and Engineering. 2009 Jul;67(1-2):34–40.

[6] Smits RMM, de Waal JA, van Kooten JFC. Prediction of Abrupt Reservoir Compaction and Surface Subsidence Caused By Pore Collapse in Carbonates. SPE Formation Evaluation. 1988 Jun 1;3(02):340–6.

[7] Loi GM, Chequer L, Nguyen CC, Zeinijahromi A, Bedrikovetsky P. Well Inflow Performance Under Fines Migration: Analytical Model, Production Data Treatment. Day 1 Tue, November 17, 2020. 2020 Nov 12;

[8] Othman F, Wang Y, Hussain F. The Effect of Fines Migration During CO2 Injection Using Pore Scale Characterization. Day 2 Wed, October 24, 2018. 2018 Oct 23;

[9] Ayopo K, Eze C, Dokubo R, Oluwatobiloba O. Impact of Fines Migration on Well XX and Lessons Learnt from Stimulation Exercise. All Days. 2018 Aug 6;

[10] Alakbari FS, Mohyaldinn ME, Muhsan AS, Hasan N, Ganat T. Chemical Sand Consolidation: From Polymers to Nanoparticles. Polymers. 2020 May 7;12(5):1069.

[11] Lew JH, Matar OK, Müller EA, Maung MTM, Luckham PF. Adsorption of Hydrolysed Polyacrylamide onto Calcium Carbonate. Polymers. 2022 Jan 20;14(3):405.

[12] Benavente D, Martinez-Martinez J, Galiana-Merino JJ, Pla C, de Jongh M, Garcia-Martinez N. Estimation of uniaxial compressive strength and intrinsic permeability from ultrasounds in sedimentary stones used as heritage building materials. Journal of Cultural Heritage. 2022 May;55:346–55.

[13] Samarkin Y, Aljawad MS, Amao A, Solling T, Abu-Khamsin SA, Patil S, et al. Carbonate Rock Chemical Consolidation Methods: Advancement and Applications. Energy & Fuels. 2022 Apr 4;36(8):4186–97.

[14] Jon Jincai Zhang. Applied Petroleum Geomechanics. Gulf Professional Publishing; 2019.

[15] Ahmad AL, Wong SS. Improvement of alum and PACl coagulation by polyacrylamides (PAMs) for the treatment of pulp and paper mill wastewater. Chemical Engineering Journal. 2008 Apr 15;137(3):510–7.

[16] Gbadamosi A, Patil S, Kamal MS, Adewunmi AA, Yusuff AS, Agi A, et al. Application of Polymers for Chemical Enhanced Oil Recovery: A Review. Polymers. 2022 Mar 31;14(7):1433.

[17] Salehi MB, Moghadam AM, Marandi SZ. Polyacrylamide hydrogel application in sand control with compressive strength testing. Petroleum Science. 2018 Sep 4;16(1):94–104.

[18] Gregory J, Barany S. Adsorption and flocculation by polymers and polymer mixtures. Advances in Colloid and Interface Science. 2011 Nov;169(1):1–12.

[19] Lee RY. Development of sand agglomeration formulation for oil and gas well applications to reduce the production of fine particulates [Internet]. [Imperial College London]; 2020. Available from: https://doi.org/10.25560/95895

[20] The formation of sand from quartz | Britannica [Internet]. www.britannica.com. Available from: https://www.britannica.com/video/185632/formation-sand-quartz-role-processes-weathering-grains#:~:text=Well%2C%20much%20of%20the%20world

[21] Hou J, Han M, Wang J. Manipulation of surface charges of oil droplets and carbonate rocks to improve oil recovery. Scientific Reports. 2021 Jul 15;11(1).

[22] Thuro K, Plinninger RJ. Scale effects in rock strength properties. Part 1: Unconfined compressive test and Brazilian test. In: Rock Mechanics – a Challenge for Society [Internet]. Swets & Zeitlinger Lisse; 2001. p. 169–71. Available from: http://geomess.com/images/pdfs/2001_eurock_espoo_1.pdf

[23] Skauge T, Ormehaug PA, Alsumaiti A, Masalmeh S, Skauge A. Polymer Stability at Harsh Temperature and Salinity Conditions. In: SPE Conference [Internet]. 2022 [cited 2022 Dec 15]. Available from: https://doi.org/10.2118/200178-MS

[24] Blanco A, Negro C, Fuente E, Tijero J. Effect of Shearing Forces and Flocculant Overdose on Filler Flocculation Mechanisms and Floc Properties. Industrial & Engineering Chemistry Research. 2005 Oct 26;44(24):9105–12.

[25] François J, Sarazin D, Schwartz T, Weill G. Polyacrylamide in water: molecular weight dependence of 〈R2〉 and [η] and the problem of the excluded volume exponent. Polymer. 1979 Aug;20(8):969–75.

Design of *Escherichia coli* Lactate Biosensors with the Insertion of L-Lactate Oxidase

Tobi Ho and Yunmin Lee
Department of Chemical Engineering, Imperial College London, U.K.

Abstract
Lactate is a key metabolite of the anaerobic metabolic pathway within the human body, and increased lactate levels have been linked to several life-threatening medical conditions. As a result, there has been a growing interest in developing biosensors to monitor lactate levels. Several lactate biosensors that have been developed are able to reflect lactate concentration through lactate induced GFP expression but are unable to alter the amount of lactate that is present in solution. This study analyses the diffusion and uptake of lactate into an *Escherichia coli* (*E. coli*) lactate biosensor, then examines the design and cloning of a modified lactate biosensor that contains L-lactate oxidase which can consume lactate. Several methods were employed, including the formation of hydrogel beads and polymerase chain reaction (PCR) with the assumption of Gibson assembly to follow. The results of this study showed that it is difficult to quantitatively model the reaction kinetics of the biosensor unless there is data on the lactate consumption rate from the enzyme, and further experiments are needed to optimise the cloning and transformation of the new biosensor to obtain this consumption rate.

Keywords: L-lactate, L-lactate oxidase, biosensor, *E. coli*, PCR, hydrogel

1 Introduction

L-lactate, commonly known as lactic acid, is a key metabolite that is formed as a by-product of glycolysis (Goers et al., 2017). Increased L-lactate levels can act as an alarm signal for the diagnosis of several pathological conditions such as liver disease and renal failure (Rassaei et al., 2014). Higher L-lactate levels were also linked to greater likelihood of metastases and recurrence of cervical tumours (Walenta et al., 2000). In addition to applications within the medical field, lactate levels are relevant to the food (Kriz et al., 2002) and wine (Lonvaud-Funel, 1999) industries as it is a crucial component in fermentation, as well as in sports where the lactate threshold can be used to evaluate the performance endurance of athletes (GA, 1985). Thus, there has been a growing interest in developing biosensors that can monitor lactate concentration, as they have potential to be used for metabolite control in biomanufacturing (Moya-Ramírez et al., 2022).

This paper explores the potential of L-lactate biosensors to alter the concentration of L-lactate in a solution through two parts. By investigating the diffusion characteristics of L-lactate into biosensors within alginate hydrogel beads, predictions were made for the rate of change in lactate concentration after the insertion of L-lactate oxidase (LOx), an enzyme which catalyses the oxidation of L-lactate. Then, after initial predictions were made, a full genetic sequence containing the LOx gene was created, and cloning could be conducted to experimentally measure the changes in lactate concentration and verify the observations made from the analysis.

2 Background

Whole cell *Escherichia coli (E. coli)* L-lactate biosensors have previously been developed by Goers et al. (2017), while Moya-Ramírez et al. (2022) have encapsulated a hydrogel core containing the biosensor within multiple layers of polymeric shells, known as living analytics in a multilayer polymer shell (LAMPS). However, these biosensor designs mainly focused on monitoring the lactate concentration in a cell culture through the expression of green fluorescent protein (GFP). The insertion of the LOx enzyme into the existing L-lactate biosensor design optimised by Moya-Ramírez et al. (2022) aims to alter the L-lactate concentration in solution through the consumption of L-lactate, in addition to sensing the current levels.

Enzymes including LOx and L-lactate dehydrogenase (LDH) have previously been used in other lactate biosensors due to their simple enzymatic reactions (Goers et al., 2017). The two enzymes above have been used extensively in fluorometric, electrochemical and chemiluminescent biosensors among others. Lactate dehydrogenase catalyses the conversion of L-lactate to pyruvate through a coenzyme (NADH or NADPH) and is commonly used in fluorometric sensors due to fluorescence properties NADH (Rassaei et al., 2014). The NADH fluoresces strongly around the wavelengths of 450-460nm, and the light intensity measured at this wavelength is proportional to the concentration of NADH and therefore the concentration of substrate (McComb et al., 1976). Conversely, L-lactate oxidase has been more widely studied in electrochemical and chemiluminescent

biosensors. L-lactate oxidase catalyses the oxidation of L-lactate in the presence of dissolved oxygen, producing pyruvate and hydrogen peroxide. The electrochemically active hydrogen peroxide can go through an oxidation-reduction reaction to give a current proportional to the L-lactate concentration (Rassaei et al., 2014). Hydrogen peroxide can also react with hydroxide and luminol to produce an electrochemiluminescence that proportionally corresponds to lactate levels (Rassaei et al., 2014).

In biosensors that contain reporter proteins such as GFP, the effects of inserting L-lactate oxidase can be analysed through changes in the fluorescence of the biosensor, which are correlated with lactate concentration levels. As a result, when L-lactate is converted into pyruvate through oxidation, the fluorescence of the biosensor will decrease relative to decreased concentrations of L-lactate.

3 Methods

Unless otherwise specified, salts and other ingredients for buffers and bacteria culture media were obtained from Sigma-Aldrich (St. Louis, MO). Kanamycin sulphate, alginic acid sodium salt from brown algae (ref 71238) and poly-L-lysine hydrochloride (MW 15,000-30,000 Da) were also obtained from Sigma-Aldrich. Chemically competent *E. coli* NEB5α cells were purchased from New England Biolabs (NEB, MA).

3.1 Living Analytic Biosensor (LAB) Bead Preparation

3.1.1 Solutions Preparation

M9 with Kanamycin

Bacterial growth medium (M9) was prepared with 33.7 mM Na_2HPO_4, 22 mM KH_2PO_4, 8.55 mM NaCl and 9.35 mM NH_4Cl, and supplemented with 0.4% D-glucose (or glycerol), 1 mM $MgSO_4$, 0.3 mM $CaCl_2$(VWR Chemicals BDH), and 1 mg/L thiamine as described in Moya-Ramírez et al. (2022). 37.5 mg/L of kanamycin was then added, and the solution was filter-sterilized.

Krebs−Ringer N-(2-hydroxyethyl)piperazine-N′-ethanesulfonic acid (HEPES) buffer (KRH buffer)

100 mL of buffer was prepared with 20 mM HEPES (Sigma-Aldrich), 135 mM NaCl (VWR Chemicals BDH), 5 mM KCl (VWR Chemicals BDH) and 0.4 mM K_2HPO_4 (Sigma-Aldrich). The buffer pH was then measured using a pH meter (Fisher Scientific accumet® AE150) and was adjusted to pH 7.4 through the addition of 5 M HCl. After supplementing 1 mM $MgSO_4$ and 1 mM $CaCl_2$ from separately prepared stocks, the buffer was filter-sterilized using 0.45 μm and 0.20 μm filter units (sartorius).

1 mg/mL Poly-L-Lysine (PLL) in KRH buffer

From the prepared KRH buffer, 50 mL of a 1 mg/mL Poly-L-Lysine (Sigma) solution was prepared.

10 mM Tris Buffer pH 8.5

A 1 M stock solution of Tris buffer was prepared with ultra pure water (UPW), then diluted to 10 mM with additional UPW to a total volume of 1 L. The pH of the solution was then measured using a pH meter (Fisher Scientific accumet® AE150) and balanced to pH 8.5 through the addition of 5 M NaOH.

Solutions in Tris Buffer pH 8.5

From the prepared Tris Buffer, 100 mL of a 2% (w/v) sodium alginate solution in Tris, and two 300 mL bottles of 100 mM $CaCl_2$ in Tris were prepared as described in Kim et al. (2014). The sodium alginate solution was stirred at a low speed overnight to ensure thorough mixing, while the two 100 mM $CaCl_2$ in Tris solutions were filter-sterilized using a 0.2 μm filter membrane (fisherbrand) into sterile glass bottles.

3.1.2 *E. coli* Suspension Preparation

E. coli cells were pelleted from precultures grown following the method in Moya-Ramírez et al. (2022) at 2300 g for 10 min, then resuspended in sterile M9. The optical density at 600 nm (OD_{600}) was measured for the suspension using a spectrophotometer (Eppendorf), and the suspension was then diluted to two separate 1 mL samples of OD_{600} of 1 and 2.

3.1.3 Formation of LAB beads

Using a sterile Eppendorf tube, the alginate solution and *E. coli* suspension were mixed at a 1:3 volume ratio as outlined in Moya-Ramírez et al. (2022). A total volume of 1 mL each was prepared for the OD1 and OD2 samples of *E. coli* suspension. This mixture was drawn into a 1 mL syringe, and a sterile 30G- blunt end needle (Weller) was added to the tip prior to the bead formation. To form the beads, one bottle of the $CaCl_2$ solution was gently agitated with a magnetic stirrer and drops of the alginate-*E. coli* mixture was added from a height of 3 cm. The alginate hydrogel beads were then crosslinked for 30 min, and after 30 min the excess $CaCl_2$ was drained from the bottle. The remaining $CaCl_2$ along with the beads were transferred to a Petri dish, where they were incubated for 1 min in 10 mL of KRH buffer. This process was repeated twice to obtain beads with the OD1 and OD2 *E. coli* suspensions, and the completed alginate-bacteria are hereafter referred to as living analytic biosensors (LABs) in this report.

A minimum of 12 LABs for each OD type were used to create LAMPS, where the LABs were coated with one additional layer of Poly-L-Lysine (PLL). The LABs were transferred from the Petri dish to a 15 mL Falcon tube containing 5 mL of the PLL in KRH buffer and were gently agitated while incubating for 10 min. The beads and buffer solution were then transferred to a sterile Petri dish.

3.2 LAMPS Fluorescence Measurement

Fluorescence measurements were taken using a CLARIOstar plate reader (BMG Labtech). A 96-well spheroid microplate (Corning) was used for the measurement, as the spherical indent in each well helped the bead settle near the centre while fluorescence was measured. Due to an error in the script, data from four time points were collected: 0 min, 20 min, 40 min and 15 hrs. The measurements at 0, 20 and 40 min were 7x7 matrix scans with a 2 mm scan width, while the reading at 15 hrs was a 20x20 matrix scan, also with a 2 mm scan width. The LABs and LAMPS were incubated in 200 µL of M9 culture medium, and a gain value of 2424 was used. The three parameters that were varied and analysed through the CLARIOstar are shown in the table below, and triplicates of each condition were collected for analysis.

Table 1: Varied Parameters for LAMPS Formation

Parameter	Variations
Hydrogel bead type	Alginate only
	Alginate + one Poly-L-Lysine (PLL) layer
Initial Lactate Concentration in Solution	0 mM
	5 mM
	10 mM
Optical Density at 600nm (OD_{600})	OD 1
	OD 2

3.3 Genetic Design

The lactate oxidase gene sequence was selected from *Aerococcus viridans* as it is the most widely characterised lactate oxidase. The open reading frame of the nucleotide sequence was identified, and the sequence was optimised for *E. coli*. The LldR biosensor developed by Moya-Ramírez et al. (2022) was used as the backbone, and the codon-optimised sequence for LOx was inserted in the location before the GFP gene for ease of insertion. The primers for the amplification of the backbone as well as the LOx insert were then designed using the NEBuilder tool online. Benchling was used to design the cloning strategy, and a schematic of the completed design is presented in Section 4.4. Since the inserted LOx gene and GFP gene are next to each other in the genetic sequence, the stop codons after both genes were kept, ensuring that a fusion protein was not created. The following configuration was then used in the final genetic design: *Promoter*-ribosome binding site (RBS) – Spacer- Start codon- Lactate oxidase open reading frame- stop – *RBS- Spacer- start codon – GFP open reading frame- stop – terminator,* where the italicised regions were present in the plasmid backbone.

3.4 Glycerol Stock, Culture Preparation and DNA Extraction

Glycerol Stock

0.5 mL of sterile glycerol (100%) was added to a 2 mL screw-cap cryogenic storage vial. 0.5 mL of *E. coli* from a logarithmic-phase broth culture was then added. The vial was vortexed vigorously to ensure even mixing of the bacterial culture and glycerol, and the vial was frozen in ethanol-dry ice or liquid nitrogen. The glycerol stock was stored at -80 °C until taken out for use.

***E. Coli* Overnight Culture Preparation**

Using a portable Bunsen burner to maintain sterility, 20 µL of the prepared glycerol stock was transferred to an agar plate containing Luria-Bertani (LB) broth and 37.5 mg/L kanamycin. A sterile inoculating loop was used to spread the bacteria, and the plate was placed in the incubator overnight at 37 °C. The following day, again working next to the flame from the Bunsen burner, one colony was picked from the plate using a sterile pipette tip. This was transferred to a 15 mL culture tube that contained 5 mL of LB and 37.5 mg/L kanamycin, to ensure that only the target bacteria with kanamycin resistance survived. The culture tube was then placed in an agitator for 15 hours at 37 °C.

DNA Extraction Miniprep

DNA was extracted from the overnight *E. coli* cultures by following the manufacturer method given for the QIAprep® Spin Miniprep Kit (Qiagen, 2021). To measure the concentration of the extracted DNA prior to PCR, 0.5 µL was placed into a spectrophotometer (BioDrop).

3.5 Polymerase Chain Reaction (PCR)

PCR was used to amplify the target DNA fragments, the biosensor backbone and the LOx gene insert. A high-fidelity DNA polymerase, either Q5 or Phusion, was chosen for the amplification. Experiments were carried out with and without the GC enhancer (Q5) and DMSO (Phusion), and an additional 10% of each component was added to the PCR mixture for improved accuracy, bringing the total volume to 55 µL for each sample. 50 µL was then transferred to a new PCR tube which was put in the thermocycler.

Tables 2 and 3 outline the reaction setup and the thermocycling conditions used for PCR. The actual amplification of the DNA occurred in three steps: denaturation, annealing and extension. 30 cycles were completed for each PCR, and the temperatures and durations of the annealing and extension steps were specific to the polymerase used and the length of the fragment to be amplified. The denaturation occurred at 98 °C for 7 seconds, and the annealing temperatures were as follows: 68 °C (Q5, backbone), 63 °C (Q5, LOx insert) and 55 °C (Phusion, backbone). Finally, the extension time occurred at 72 °C and was dependent on the length of the fragment. The rate of 25 seconds per kb was used to determine the length of this step, where the

template was set for 1 minute 43 seconds, and the LOx insert was set for 30 seconds.

Table 2: Reaction Setup for PCR using Q5® and Phusion® high-fidelity DNA Polymerase (New England Biolabs)

Component	Quantity (µL)	Final Concentration
5X Q5 Reaction Buffer/ 5X Phusion HF buffer	11	1X
10 mM dNTPs	1.1	200 µM
10 µM Forward Primer	2.75	0.5 µM
10 µM Reverse Primer	2.75	0.5 µM
Template DNA	1.1	< 1,000 ng
Q5 High-Fidelity DNA Polymerase/ Phusion Polymerase	0.55	0.02 U/µl
5X Q5 High GC Enhancer / DMSO (optional)	(11) / (1.65)	(1X)
Nuclease-Free Water	to 55	

Table 3: Thermocycling conditions for PCR using Q5® and Phusion® high-fidelity DNA Polymerase (New England Biolabs)

Step	Temperature (°C)	Time (s)
Initial Denaturation	98	30
30 Cycles	* See paragraph above	
Final Extension	72	120
Hold	10	∞

3.5.1 Gradient PCR

Gradient PCR was carried out for the biosensor backbone to analyse the effects of the annealing temperature on the concentration of primer dimers and the target sequence. An 8.8 x mastermix of the reaction setup for PCR using Phusion was prepared and aliquoted into 8 separate PCR tubes. The annealing temperatures for the samples were set at 51 °C, 53 °C, 55 °C, 58 °C, 60 °C, 63 °C, 65 °C, and 70 °C.

3.6 Gel Electrophoresis

To measure and analyse the size of DNA fragments, gel electrophoresis was performed. The agarose gels were made using a 1% (w/v) solution of agarose in TAE (40 mM Tris-acetate, 1 mM EDTA), and was heated until the agarose was completely dissolved without clumps. Using the casting apparatus and either a small or large gel tray, 3 µL (small tray) or 5 µL (large tray) of SYBR Safe DNA Gel Stain (Life Technologies, S33102) was added to the bottom of the tray. An appropriate comb was selected based on the number of wells that were needed for that gel. For samples larger than 50 µL, two or three wells were taped together using autoclave tape.

The heated agarose gel solution was poured into the tray and was stirred with the pipette tip to ensure even distribution of the gel stain and removal of any bubbles. After solidifying, the gel was removed from the apparatus and placed in an electrophoresis tank containing 1X TAE buffer. All samples for gel electrophoresis were prepared with a loading dye, such that the dye had a final concentration of 1X.

The Hyperladder 1kb (SLS) was added to the first well in each gel as a means of verification for band size. Wells that did not contain the ladder or DNA samples were filled with 1X loading dye to ensure the bands would run in a straight line down the gel. Each gel was run at 90 Volts for 1 hour, and images of the gel were captured using a gel imaging system with the UV light setting and were exported for further analysis.

4 Results and Discussion

4.1 LAB and LAMPS Formation

Though the methods and process for the LAB and LAMPS formation were straightforward, there were two aspects that potentially affected the results of the fluorescence data or the replicability of the experiment.

The first aspect was the amount of alginate that was used to alginate-*E. coli* mixture. Due to the high viscosity, a significant amount of alginate was stuck to the inside of the pipette tip while creating the 1 mL mixture, especially while using the 200 µL pipette tips. This resulted in high errors while preparing the experiment and would ultimately affect the concentration of *E. coli* contained within each bead, as there would be a higher starting concentration of *E. coli* than calculated. To ensure that the right amount of alginate solution was added to the mixture, a larger pipette tip was used to minimise the number of times the alginate was pipetted to the Eppendorf tube. For future experiments, an average amount of alginate that remained in the pipette tip could be calculated. This could then be accounted for to obtain more accurate results.

The second aspect was the formation of the beads themselves through the dropwise addition into the $CaCl_2$ solution. Though this does not affect the results of the experiment, to help with replicability of the methods, the drops should be added away from the centre of the vortex created by the magnetic stir bar to avoid clumping of the LABs or LAMPS.

4.2 LAMPS Fluorescence Data

The fluorescence data of the GFP-only biosensor obtained from CLARIOstar were analysed to determine the effect of three different parameters and the diffusion characteristics of lactate. However, as mentioned earlier, data were only collected at four

time points: 0, 20 min, 40 min, and 15 hours. The dataset was much smaller than anticipated, so the trends over time could not be analysed properly, but the fluorescence data was still used to reinforce the understanding of the LAMPS and to allow for the optimisation of parameters for future experiments involving LAMPS.

The raw fluorescence data at 15 hours were plotted in box and whisker plots to compare the effect of the different parameters. The data from blank wells containing only M9 were also plotted for each of the graphs for better comparison.

As shown in Figure 1, the fluorescence values for alginate-only LABs were similar to the values of the blanks, while the samples with one PLL layer had much higher fluorescence values. This result showed that the alginate beads alone were not sufficient to contain the LAB.

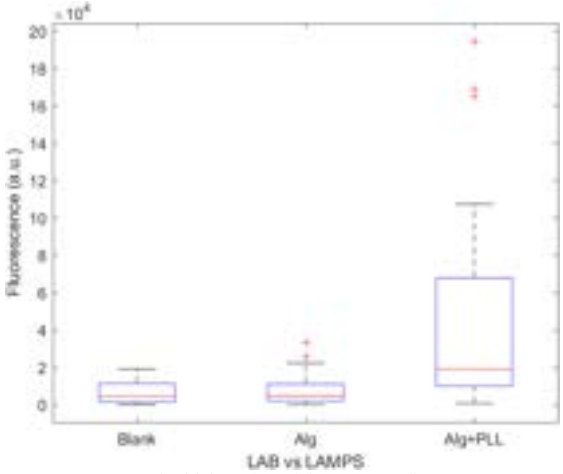


Figure 1. Box and whisker plot of the average fluorescence at 15 hours to compare alginate-only LABs and LAMPS with one layer of PLL. The red lines indicate the medians of the data set; the boxes represent the 25-75 percentiles; the whiskers include all data excluding the outliers; the red + signs indicate outliers of the dataset.

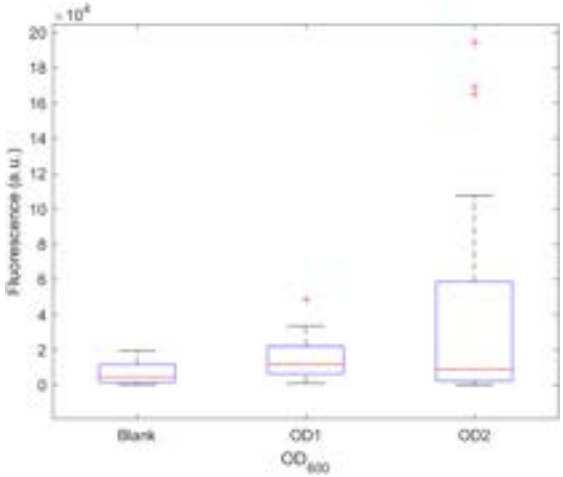


Figure 2. Box and whisker plot of the average fluorescence at 15 hours to compare the optical density (OD). The red lines indicate the medians of the data set; the boxes represent the 25-75 percentiles; the whiskers include all data excluding the outliers; the red + signs indicate outliers of the dataset.

The density of the *E. coli*, quantified by the OD_{600}, was then varied between the values of 1 and 2. The result showed that the fluorescence readings from OD1 were slightly higher than those of the blanks, but much lower than values observed from OD2 data. Therefore, it was concluded that a density of OD2 is more suitable for future experiments involving encapsulation (Figure 2).

Finally, the fluorescence level was found to increase with increasing lactate concentration as shown in Figure 3. Amongst the conditions that were experimented, the 10 mM sample had the highest fluorescence readings. The 0 mM data were considered to account for the auto-fluorescence rate due to E. coli production of lactate. However, the values obtained from 0 mM samples did not seem significantly different compared to the readings from the blanks. This indicated that the effect of auto-fluorescence was negligible to the overall GFP expression due to lactate.

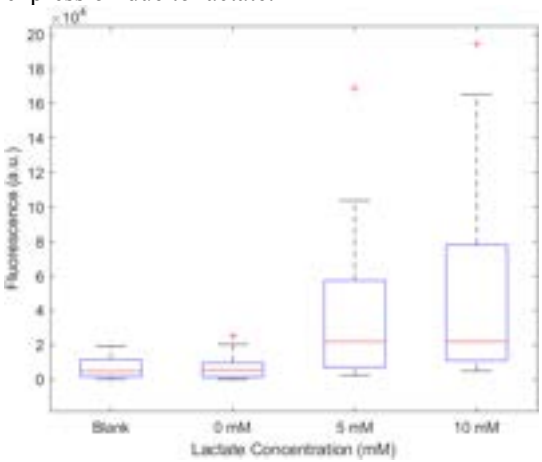


Figure 3. Box and whisker plot of the average fluorescence at 15 hours to compare the effect of different liquid lactate concentrations of the solution: 0 mM, 5 mM and 10 mM. The red lines indicate the medians of the data set; the boxes represent the 25-75 percentiles; the whiskers include all data excluding the outliers; the red + signs indicate outliers of the dataset.

In addition to the script error resulting in fewer data points, the location of the alginate bead in the well of the microplate also affected the accuracy of the fluorescence readings from the CLARIOstar. The dropwise addition method outlined by Moya-Ramírez et al. (2022) was used to create the LABs and LAMPS, so the average diameter of the LABs was taken to be 1.6 mm, while the scan width of the readings was 2 mm.

Since the readings are taken from the centre of the well plate, if the bead was not located perfectly in the centre, a portion of the bead would not be measured, which would result in a lower average well scan. There was no direct way to ensure that the beads were centred. However, the script was written such that the microplate was agitated after every reading to try to let the LABs or LAMPS move within the well and settle to the bottom prior to the next reading. Preparing triplicates of each condition also decreased the error, as the average of the three

readings was a better representation of the fluorescence for a specific set of conditions. Despite this, the only way to account for this inaccuracy would be to collect readings for multiple cells with the same conditions and analyse the trends over time.

Due to the lack of certainty in the data collected for the four time points, the dataset obtained by Moya-Ramírez et al. (2022) was analysed. This dataset was used to determine the proportions of *E. coli* cellular autofluorescence compared to the overall fluorescence readings, and the diffusion characteristics of lactate through LAMPS.

4.3 Lactate Diffusion Rate and LOx Reaction Kinetics Prediction

The insertion of LOx was considered to improve the existing design of the biosensor. Before actually inserting the LOx, however, the reaction kinetics of the enzyme were considered using the Michaelis-Menten equation:

$$v = \frac{k_2[E]_t[S]}{k_M+[S]} \qquad (1)$$

where $[E]$ is the concentration of the enzyme, which was the LOx, $[S]$ was the concentration of substrate, which was lactate, k_M was the Michaelis Menten constant that accounts for all the reaction kinetics constants, and k_2 or k_{cat} was the turnover number which determined the amount of substrate an enzyme could catalyse.

To find the concentration of lactate, the diffusion rate and rate of consumption by LOx had to be considered. Since there was no data on the consumption rate of lactate by LOx in LAMPS at this stage, qualitative predictions of the reaction kinetics were made.

Initially, Fick's law of diffusion was considered to model the diffusion rate of lactate:

$$J = -D\frac{dC}{dx} \qquad (2)$$

where J is the diffusive flux, D is the diffusion coefficient, C is the concentration, and x is the position.

However, there was not much literature data available on the diffusion of lactate through a PLL layer. Therefore, the diffusion through the PLL layer and the alginate bead, along with the uptake of the cell, were considered all together using the rate of change of fluorescence at different time intervals.

The dataset from the previous work by Moya-Ramírez et al. (2022), collected over 15.7 hours, was used in the diffusion characteristics analysis as stated in the previous section. In addition, since the best results were obtained from LAMPS with one PLL layer, an *E. coli* density of OD_{600} =2 and 10 mM lactate concentration, these conditions were used as a focus of the analysis.

To determine the biosensor fluorescence that was solely from the lactate diffusion from solution, the *E. coli* cellular autofluorescence was subtracted from the total fluorescence over time. The autofluorescence values were determined by considering the 0 mM lactate samples. The proportions of these values from autofluorescence in the total readings were then calculated and plotted over time (Figure 4).

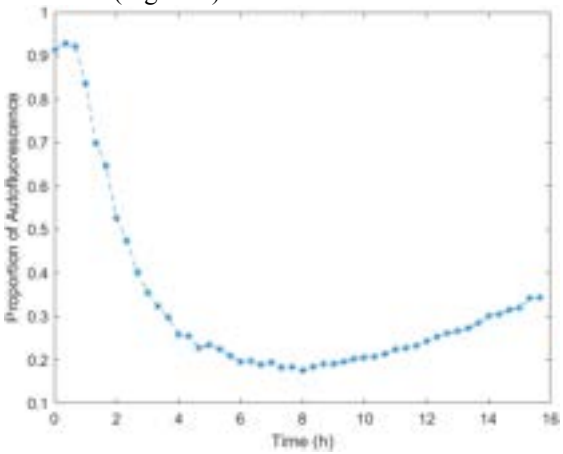


Figure 4. The proportion of E. coli cellular autofluorescence in the total fluorescence reading for 10 mM lactate concentration sample. The autofluorescence initially dropped but started to increase at around 8 hours.

After accounting for the cellular autofluorescence, the rates of change in fluorescence due to diffused lactate were calculated for each time interval. Even though these values fluctuated, the overall fitted rate increased initially then decreased and displayed more negative values after around 10 hours (Figure 5).

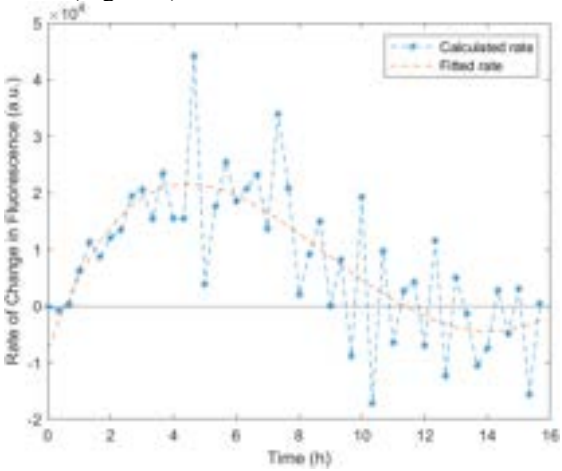


Figure 5. Rate of change in diffusion induced fluorescence over time for 10 mM lactate concentration sample. A fourth-degree polynomial was fitted to visualise the trend better.

The trend observed for autofluorescence and rate of change in fluorescence can be explained using Fick's law (equation 2), which states that the rate of diffusion is directly proportional to the concentration gradient. Initially, the concentration gradient between the lactate solution and the bead was higher, which led to relatively higher diffusion

rates. However, as more lactate diffused over time, the concentration gradient between the lactate solution and LAMPS decreased. As a result, diffusion rate also decreased, which led to a relative increase in autofluorescence proportion and decrease in the values and rate of change in diffusion-induced fluorescence.

The specific strain of the organism used for the LOx genetic design was required to determine the equation constant. However, since this information was unclear, a strain with the lowest k_M and the highest k_{cat} was selected to assume the best-case scenario. The strain selected for this model was the wild-type *Aerococcus viridans* at a pH of 7.0 and a temperature of 25 °C using the BRENDA Enzyme Database. The values found for k_M and k_{cat} were 0.87 and 283, respectively (Yorita et. al., 1996).

Finally, the initial density of *E. coli* and the change in LOx concentration over time were considered to estimate the concentration of LOx enzyme. The initial concentration of enzyme per *E. coli* was determined using the OD_{600} values obtained from the experiment, with the assumption that the LOx concentration is the same as the concentration of *E. coli*. Then, the change in enzyme concentration was determined using the growth rate of *E. coli* over time.

Considering all the parameters, the change in lactate concentration was predicted. The diffusion rate was predicted to initially dominate the system over the reaction rate. However, as more lactate is consumed over time, the reaction rate of lactate consumption would eventually dominate over the diffusion rate, which would decrease the overall lactate concentration and thus fluorescence values.

Since there was not enough information for a quantitative model, the empirical relationship of the reaction kinetics had to be investigated. The genetic sequence of the biosensor with LOx insert was designed as a first step, and PCR was carried out to amplify the genes to be used for Gibson assembly to combine the LOx insert with the biosensor backbone.

4.4 Genetic Design

Figure 6 shows the schematic for the final design of the biosensor, containing the lactate oxidase inserted before the GFP gene.

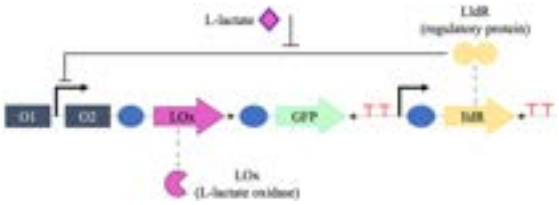


Figure 6. Adapted from Figure 1(b) in Moya-Ramírez et al. (2022). Schematic of the biosensor genetic design with LOx. O1 and O2 represent the operator sites, blue ovals represent ribosome binding sites, asterisks represent stop codons and dashed lines indicate protein synthesis.

In the absence of lactate, the dimers of LldR bind to operator sites O1 and O2 and form a tetramer that prevents the transcription of LOx and the GFP genes (Goers et al., 2017). However, when lactate is present, it binds to the LldR dimer that was bound to O2. This dissociates the bond previously formed by the dimer and O2. The lactate also binds to the dimer bound to O1, which forms a transcriptional activator (Goers et al., 2017) that allows for transcription of the LOx and GFP.

The LOx was chosen to be inserted before the GFP due to the complexities of the sequence after GFP, and the primers for amplification were designed with the assumption that Gibson assembly would occur to bind the fragments together. One of the three enzymatic activities that occurs in the Gibson assembly involves the 5' exonuclease, which exposes the complementary sequence for annealing by digesting part of the 5' end to produce sticky ends (Gibson et al., 2009). Thus, there were 30 base pairs (bp) of homology which was added to the standard primer length. Additionally, these regions of homology were added to both the 5' and 3' ends, for both the backbone and the insert. The overlapping ends would then be complementary, allowing the fragments to form a circular plasmid (Sinfield, 2014).

4.5 Biosensor Backbone and LOx PCR

A series of PCRs were carried out to amplify the biosensor backbone and the LOx gene, and gel electrophoresis was used to analyse the results of the PCR. The gene sequence of the biosensor backbone had 4115 base pairs (bp), so the expected fragment length on the gel was around 4000 bp.

The first attempt to amplify the backbone using Q5 DNA polymerase without the GC enhancer was unsuccessful, as the only visible band at the edge of the gel was less than 200 bp, indicating that the PCR only formed primer dimers (not pictured).

Since the use of Q5 polymerase without the GC enhancer was unsuccessful, the second PCR attempt included the use of Q5 polymerase with and without GC enhancer, as well as Phusion polymerase with and without DMSO.

As shown in Figure 7, there were again bands near the end of the gel less than 200 bp, which were the primer dimers. The high concentration of the primer dimers indicated that there were few copies of the desired template DNA.

There were also undesirable bands visible at around 1000 bp for Phusion and 2000 bp for Q5 polymerase. For Phusion, a higher concentration of the undesirable band was observed when DMSO was not added. Therefore, addition of DMSO was concluded to be effective, and future PCR trials were attempted with Phusion DNA polymerase with DMSO.

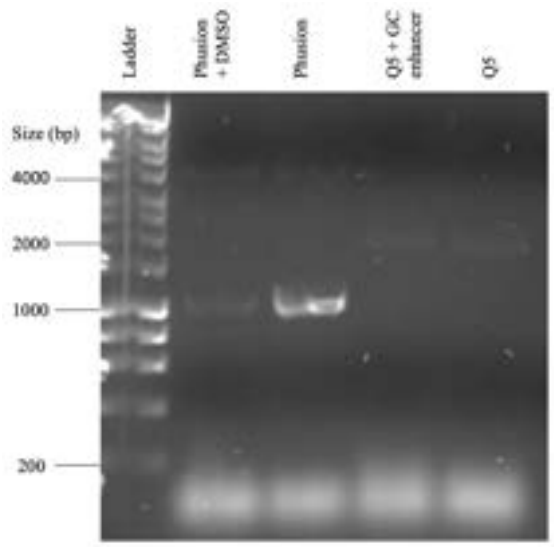


Figure 7. Gel analysis of biosensor backbone PCR using Phusion and Q5 polymerase with and without the DMSO and GC enhancer, respectively.

Lastly, there were faint 4000 bp bands for both samples that used Phusion. However, since the concentration was very low, the next step was to modify the PCR and optimise it by application of a gradient PCR and a DpnI digest.

The third gel with the gradient PCR and amplified LOx samples showed a similar clarity for all the bands of the gradient PCR (Figure 8), demonstrating that the annealing temperature did not have a significant effect on amplification. A closer examination of the gel results showed that the bands from the gradient PCR were between 5000-6000 bp, longer than the anticipated fragment size of 4000 bp. The nonspecific bands could be caused by excessive cycling, annealing time or extension time, but were more likely caused by impurities in the PCR components (Bio-Rad, n.d.).

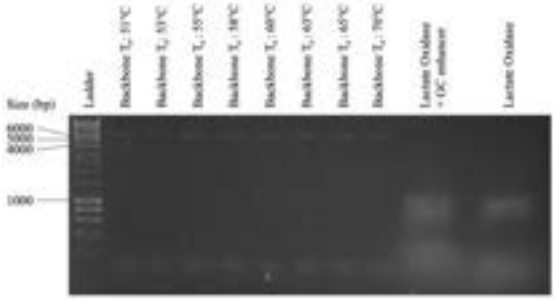


Figure 8. Gel analysis of biosensor backbone PCR using Phusion polymerase with DMSO at varying annealing temperatures (T_a) between 51°C and 70°C, and LOx insert PCR using Q5 with and without GC enhancer.

There are two potential methods to prevent the nonspecific bands. Since the template DNA concentration was increased, fewer cycles could be used, as nonspecific amplification and errors could occur from excessive cycling (Bio-Rad, n.d.). Conversely, the primer concentration could be decreased to minimise primer dimer formation and the binding of the primers to nonspecific sites on the template. If the experiment gave the same result of larger product sizes when repeated, the DNA could be extracted from the gels and sequenced to assess the cause more accurately.

The PCR for the LOx insert was also analysed simultaneously. Since the length of LOx insert was 1185 base pairs, a band around 1000 bp was expected. DNA was extracted from the resulting band using a gel DNA recovery kit (Zymoclean^{TM}) to preserve the LOx insert for further experiments and optimisation.

The fourth and the final attempt of the PCR included the DpnI digest and one well containing the original template DNA, to test if the bands that were observed in the second and the third attempt were from the amplified DNA or the original template. This was prepared following the manufacturer protocol (NEBcloner), with the following modification: the nuclease free water was replaced with the PCR sample. DpnI cleaves when the recognition site is methylated (Biolabs, n.d.), and since most *E. coli* strains are dam methylated, they are susceptible to DpnI digestion. The parental DNA would then be digested, leaving the synthesized DNA fragments (Jena Bioscience, n.d.).

The DpnI digest sample did not show any bands except the primer dimers, while the original template DNA had a clear band around the expected 4000 bp mark (not pictured). This analysis indicated that all the previous PCRs were unsuccessful at amplifying the backbone DNA.

Due to the time constraint, the PCR was not further optimised. Therefore, additional trials would be necessary to successfully amplify the backbone and the LOx gene for potential insertion using Gibson assembly and to test the new biosensor design.

6 Conclusion

The diffusion and uptake of L-lactate into the biosensor was calculated, and the reaction kinetics of lactate oxidase were analysed to determine their effect on lactate concentration over time. However, in this process, it was difficult to quantitatively model the reaction kinetics of the biosensor unless there was data on the lactate consumption rate from the enzyme. This data could be obtained after the optimisation of the cloning and transformation of the full genetic sequence containing the LOx insert.

These processes can be achieved through several steps. Firstly, amplification of the backbone and insert fragments must be optimised such that the target fragments are synthesised with high concentration. Next, Gibson assembly would be used to bind the two fragments together and form a single plasmid. This assembled plasmid would then contain the full genetic sequence and can be transformed and plated. Following the transformation, the lactate consumption could then be analysed by using colorimetry to measure the

lactate concentration over time. This information would then be fed back into the Michaelis-Menten equation to find the reaction kinetics model in LAMPS, which could potentially be used to optimise the LAMPS in other environments.

Due to the time constraint, the biosensor containing the lactate oxidase was not transformed. Despite this fact, the study still shows that there are many areas that can be explored regarding this biosensor design. Ultimately, a biosensor that could not only express the amount of lactate in solution but also alter it could eventually be used in a co-culture with mammalian cells through encapsulation, with the potential to function as a simple but efficient diagnostic tool to save lives.

7 Acknowledgement

The authors would like to thank Dr. Masue Marbiah and Dr. Zalike Keskin Erdoğan for their continuous support and guidance throughout this project.

8 Reference

Biolabs, N.E. (no date) *DpnI*, *NEB*. Available at: https://international.neb.com/products/r0176-dpni#Product%20Information (Accessed: December 14, 2022).

Biolabs, N.E. (no date) *Gibson Assembly®*, *NEB*. Available at: https://international.neb.com/applications/cloning-and-synthetic-biology/dna-assembly-and-cloning/gibson-assembly (Accessed: December 14, 2022).

DpnI (no date) *DpnI, Restriction Enzymes: "D" Enzymes - Jena Bioscience*. Available at: https://www.jenabioscience.com/molecular-biology/enzymes/restriction-enzymes/d-enzymes/en-160-dpni#:~:text=DpnI%20is%20specific%20for%20methylated,newly%20synthesized%20DNA%20containing%20mutations. (Accessed: December 14, 2022).

Duchen, M.R. and Biscoe, T.J. (1992) "Mitochondrial function in type I cells isolated from rabbit arterial chemoreceptors.," *The Journal of Physiology*, 450(1), pp. 13–31. Available at: https://doi.org/10.1113/jphysiol.1992.sp019114.

GA, B. (1985) "Anaerobic threshold: review of the concept and directions for future research.," *Medicine and Science in Sports and Exercise*, 17(1), pp. 22–34. Available at: https://europepmc.org/article/med/3884959 (Accessed: December 12, 2022).

Gibson, D.G. *et al.* (2009) "Enzymatic assembly of DNA molecules up to several hundred kilobases," *Nature Methods*, 6(5), pp. 343–345. Available at: https://doi.org/10.1038/nmeth.1318.

Goers, L. *et al.* (2017) "Whole-cell*escherichia coli*lactate biosensor for monitoring mammalian cell cultures during biopharmaceutical production," *Biotechnology and Bioengineering*, 114(6), pp. 1290–1300. Available at: https://doi.org/10.1002/bit.26254.

Information on EC 1.1.3.2 - L-lactate oxidase and organism(s) aerococcus+viridans (no date) *Information on EC 1.1.3.2 - L-lactate oxidase and Organism(s) Aerococcus+viridans - BRENDA Enzyme Database*. Available at: https://www.brenda-enzymes.org/enzyme.php?ecno=1.1.3.2&Suchword=&reference=&UniProtAcc=&organism%5B%5D=Aerococcus%2Bviridans&show_tm=0#REF. (Accessed: December 14, 2022).

Kim, B.J. *et al.* (2014) "Cytoprotective alginate/polydopamine core/shell microcapsules in microbial encapsulation," *Angewandte Chemie International Edition*, 53(52), pp. 14443–14446. Available at: https://doi.org/10.1002/anie.201408454.

Kriz, K. *et al.* (2002) "Amperometric determination of l-lactate based on entrapment of lactate oxidase on a transducer surface with a semi-permeable membrane using a sire technology based biosensor. application: tomato paste and baby food," *Journal of Agricultural and Food Chemistry*, 50(12), pp. 3419–3424. Available at: https://doi.org/10.1021/jf0114942.

Lonvaud-Funel, A. (1999) "Lactic acid bacteria in the quality improvement and depreciation of wine," *Lactic Acid Bacteria: Genetics, Metabolism and Applications*, pp. 317–331. Available at: https://doi.org/10.1007/978-94-017-2027-4_16.

McComb, R.B. *et al.* (1976) "Determination of the molar absorptivity of NADH.," *Clinical Chemistry*, 22(2), pp. 141–150. Available at: https://doi.org/10.1093/clinchem/22.2.141.

Moya-Ramírez, I. *et al.* (2022) "Polymer encapsulation of bacterial biosensors enables Coculture with mammalian cells," *ACS Synthetic Biology*, 11(3), pp. 1303–1312. Available at: https://doi.org/10.1021/acssynbio.1c00577.

PCR troubleshooting (no date) *Bio-Rad*. Available at: https://www.bio-rad.com/en-uk/applications-technologies/pcr-troubleshooting?ID=LUSO3HC4S (Accessed: December 14, 2022).

Rassaei, L. *et al.* (2013) "Lactate biosensors: Current status and outlook," *Analytical and Bioanalytical Chemistry*, 406(1), pp. 123–137. Available at: https://doi.org/10.1007/s00216-013-7307-1.

Sinfield, O. (2014) *A Guide to Gibson Assembly Design*. University of Warwick. Available at: https://warwick.ac.uk/study/csde/gsp/eportfolio/directory/pg/lsujcw/gibsonguide/ (Accessed: December 15, 2022).

Walenta, S. *et al.* (2000) "High Lactate Levels Predict Likelihood of Metastases, Tumor Recurrence, and Restricted Patient Survival in Human Cervical Cancers1.," *Cancer Research*, 60(4), pp. 916–921. Available at: https://pubmed.ncbi.nlm.nih.gov/10706105/ (Accessed: December 10, 2022).

Yorita, K. *et al.* (1996) "Conversion of L-lactate oxidase to a long chain α-hydroxyacid oxidase by site-directed mutagenesis of alanine 95 to glycine," *Journal of Biological Chemistry*, 271(45), pp. 28300–28305. Available at: https://doi.org/10.1074/jbc.271.45.28300.

Investigating thermo-responsiveness, reversibility and pH sensitivity in encapsulated AIE luminogens

Geoffrey Rwamakuba & Aaron Thayaparan

Department of Chemical Engineering, Imperial College London, UK

December 15, 2022

Abstract

Detection of various ions in the interstitial fluid (ISF) can lead to point-of-care treatments for different conditions like hypertension and kidney disfunction in the case of hypernatremia (excess sodium). This study investigates the production of a hydrogel-fluorophore matrix via emulsion polymerisation and the dexterity and compatibility of the resulting matrix in conditions that simulate the ISF in humans. The study is done to test the matrixes potential for biosensing purposes in the human body and ultimately its incorporation into functionalized tattoos. The following report describes the production of a P-NIPAM hydrogel with Berberine Chloride (BBR Chloride) used as the fluorophore. Using a microplate reader, it was found that although BBR Chloride is a known aggregation induced emission luminogen (AIEgen) the highest emission intensity was recorded at roughly 2au of a sample made with the least amount of fluorophore (1.7mg). The effects of dilution did decrease the global emission intensity however, did make the samples emission spectrum more distinguishable between 35°C to 40°C, the range of interest incorporating the extremes of body temperature. Further measurements taken on the pH effect on the thermos-response displayed an increasing emission intensity with higher values of pH with consistent trends in the thermal-responsivity. DLS analysis yielded an average particle size of 645nm, and a polydispersity index of 0.3, which with further purification are expected to be more consistent and appropriate for industrial use. The sample exhibited an okay degree of thermal reversibility however, this variable will require further study in conjunction with analyte concentration.

Keywords: Optical biosensor, Tattoo, Hydrogel, AIE Luminogens, P-NIPAM, BBR Chloride

1 Introduction

Tattoos have been around for millennia, with the oldest record of them dating back to roughly 3000 BC.[1] They serve various purposes including the upholding of tradition, expressing individualism or even something as simple as aesthetics. Regardless of the reason, tattoos are wide-spread and are quite prominent in todays' society, it has been estimated previously in 2019 using google trends that 10-29% of the world's population are tattooed.[2] However, due to a scarcity of data on regions outside of the western world, this number could potentially be higher or lower. With such a huge market, the opportunities for the development of humanity through the use of technology and engineering have risen. One such opportunity, which will also be the focal point of this report, is the functionalization of tattoos using optical biosensors for the monitoring of various biomarkers in the interstitial fluid (ISF).

Examples of observable biomarkers include but are not limited to glucose [3], albumin[4], sodium[5], oxygen[6] and pH[7]. There is also external biomarkers such as environmental contaminants[8] or toxic food additives as well as potential for multiplexing where multiple biomarkers can be assessed through one channel. Additionally there are applications in drug delivery, vaccination, alcohol levels and body temperature.

One of the major concerns with the use of biosensors in vivo is biocompatibility. Biocompatibility considers the impact on the body as well the functionality of the technology. Common luminescent nanoparticles like quantum dots contain heavy metals so require encapsulation to ensure long term use. To achieve biocompatibility materials that mimic biology are selected, for example polymeric hydrogels, like those considered in this paper, naturally occur in collagen and bone, are flexible and soft, and will biodegrade into non-toxic orthosilicic acids that are very easily removed from the body. More generally tattoo inflammation is the most common complication so anti-inflammatory drug loaded alginate microspheres have been used that release biosensors but also improve biocompatibility.

2 Background

2.1 Tattoos

Tattooing is the process of using needles to create holes in the skin, which upon removing creates a vacuum, drawing in ink into the cavity. The ink ideally should be injected into the dermis, the layer of skin found between in epidermis and the hypodermis depicted in figure 1. Needle depth is a key factor in the permanence of a tattoo, as too shallow can mean the tattoo dissipating quickly, while too deep can cause damage to the subcutaneous tissue. Inks are generally injected around 0.4 mm to 2.2 mm. With mean skin thickness being 1.5 mm to 2.5 mm it is generally accepted that the ideal needle penetration depth is 2.0 mm. The effects of a variety of factors such as ethnicity, age and hormones on tattoo implantation depth requires further studies. The uniformity of needle depth will affect the accuracy

and reliability of the colorimetric detection of biomarkers in the ISF.

When tattooing another point of consideration is the angle at which the needle is inserted into the dermis. The needle angle normal to the skin surface, depending on the desired look and aesthetics, can be varied between 45° and 80°. This difference in appearance is because the angle can alter the distribution of ink in the dermis. Needle angle can be controlled accurately for purposes of biosensor injections.

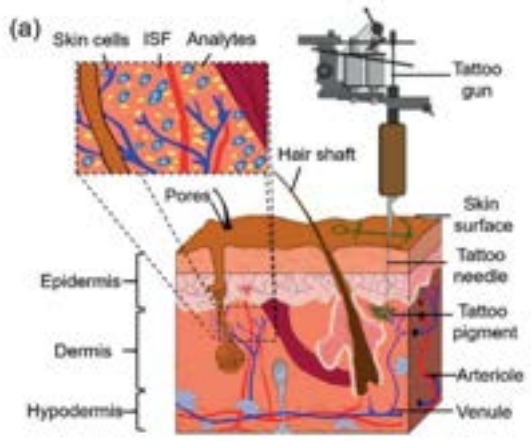


Figure 1: Layers of the skin and tattoo injection [5]

When determining the size of the biosensor the characterisation of tattoo particle size is a requirement. Tattoo particle size can affect the longevity of the tattoo itself. This is because if the particles are too small the particles can diffuse from the tattooed area. However, there must be a balance as too big of a particle would require a surgical procedure. The same ideology is applied to the size of the optical sensor. Pigments found in tattoo ink have been studied to show varying particles sizing ranging from 10-5000nm [38], these sizes vary with ink colour but are generally within said range.

Another factor to be considered is tattoo particle density. There have been cases of a person carrying up to 40g of tattoo pigment [39]. It can be noted that the particle density, size and colour of a tattoo can impact the amount of hazardous substances in the body which can subsequently increase health risks, for instance they can interact with lymph nodes mimicking the effects of metastatic melanoma [40].

2.2 Permanence of tattoos and Laser removal

The injection of tattoo ink into the dermis triggers the body's immune response to deal with the foreign bodies. One theory explaining the permanence of a tattoo is based around the ability of a macrophage to ingest and store material. Dermal macrophages at the location of the tattoo try to ingest the pigments. The small and soluble particles are taken away in the blood stream or via the lymphatic system where they are ultimately stored or destroyed in various organs. The larger particles cannot be broken down by the macrophages due to their size and as such are stored within cytoplasmic vacuoles [12]. Eventually when the macrophage life cycle comes to an end the pigment is released and again ingested and stored by macrophages of the next generation. Large groups of pigment particles have been observed in a single lysozyme of up to 100μm (Furguson et. al 1997).

Currently, laser surgery is the most prevalent method of tattoo removal. During the procedure a laser (typically Nd: YAG) is used up lyse macrophages and to heat tattoo particles in a matter of nanoseconds to deform them which in turn causes them to break down into smaller particles. The smaller particles can then be dealt with as describes previously. Additional considerations must be taken into account when dealing with the removal of colours, such as wavelength of the laser, pulse duration and the size of the laser spot or even the skins pigment. This procedure takes a while and requires many sessions to complete, even then the complete removal of the tattoo is not currently possible[36]. Some other non-laser techniques of tattoo removal discussed by Dash et al [37] to name a few include radio surgery, dermabrasion, cryotherapy even so called "home remedies" which come with their own complications.

2.3 Optical Biosensors

Biosensors are platforms that allow us to transform biological signals into those that can be interpreted quantitatively or qualitatively. They are made of 3 different components. Firstly, there is a biological recognition site which will interact with the analyte, which is a molecule of interest. This can be done similarly to an enzyme where the analyte is converted into a metabolite (Catalytic biosensor) or more like an antibody where the analyte binds without reaction (Affinity biosensor). The second component is a transducer which converts that biological response to a measurable one, and lastly a processing unit will interpret the signal. The ideal biosensor should display excellent sensitivity and selectivity for the biomarker of interest, be able to operate for long periods of time across the required concentration ranges and have multiplexing abilities. There are different approaches used for transducers which give rise to the different types of biosensors including, optical, thermal, magnetic and piezoelectric.

Optical sensors are more sensitive, selective and smaller than alternatives. They also aren't affected by electromagnetic signals or radio interference.[9] Tattoo functionalisation uses optical biosensors that monitor the optical characteristics of an analyte. The incident light that passes through an analyte will be affected by absorption, transmission, emission and elastic/inelastic scattering. A photodetector can convert these changes into electrical signals proportional to analyte

concentration.[13] Optical biosensors can be grouped further into labelled and non-labelled, i.e. does the label need to generate a signal or can the analyte binding produce a signal directly. Fluorophores and phosphorescent molecules are the most commonly used labels which use fluorescence- the excitation of electrons to higher energy states. Proteins, peptides, polymers and synthetic oligomer are potential fluorophores and phosphorescent molecules some of which require constant energy sources to emit light while others can continue emitting after the incident light stops. Generally, the excitation energies are in the UV/ Visible spectrum while emission energies range from visible light to the near-IR. Some examples of labelled sensors include Fluorophore resonance electron (FRET), Bioluminescent resonance electron transfer (BRET) and Chemiluminescent resonance electron transfer (CRET) sensors all of which require an analyte binding to an acceptor or cancelling the flow of electrons to a doner. [14,15] It's often better to avoid labels especially if the biomarkers is more complex, in these cases alternatives like SPR and LSPR are considered because they can offer reduced analysis time, consumption of solvents, cost and greater sensitivity. [16]

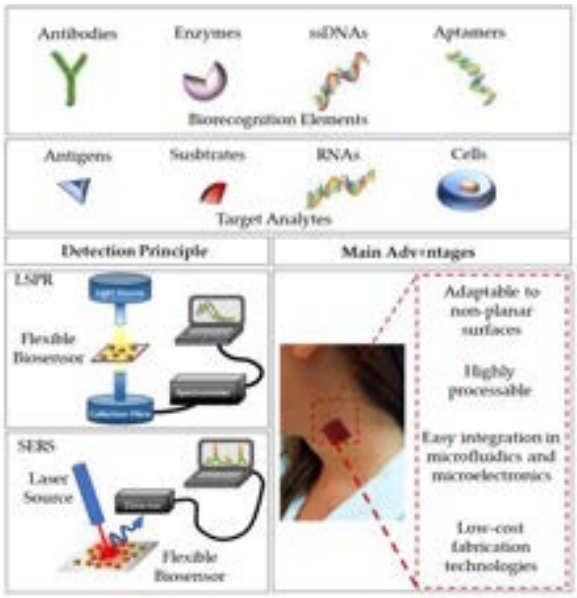


Figure 2: Illustration of biosensor elements and detection principles [11]

A key feature of biosensors is the large surface area to volume ratio. This allows for greater accessibility and binding to analyte but leads to reduced sensitivity, stability and reusability. One of the growing applications of biosensors is in non-invasive wearable sensors that are in contact with the skin to detect physiological changes. [17,18] If functioning perfectly these sensors will be resistant to mechanical stress and cause little inconvenience in everyday activities however they will always require electrical power and be subject to environmental interference. Those issues make high quality continuous real-time monitoring via biosensors in the body very attractive since they allow for Point-of-care (POC) use.

2.4 Construction and characterisation

In the construction and characterisation of biosensors nanomaterials have gained a lot of interest, this includes carbon-based nanomaterials (i.e. carbon nanotubes & graphene) to plasmonic (i.e. gold nanoparticles) and photoluminescence nanoparticles (i.e. quantum dots & upconverting nanoparticles). Carbon based nanomaterials have versatile surface properties as well as optical and electrical merits. [19] The production method paired with structural variations in carbon can lead to different nanomaterial properties. Graphene is one example, by using liquid phase exfoliation or chemical vapour decomposition you can introduce different defects in the structure and therefore different surface characteristics. Noble metals like gold and silver have the ability to maintain surface plasmon on their dielectric metal interfaces [19] which gives way to Surface Plasmon Resonance (SPR), Localized Surface Plasmon Resonance (LSPR), and Surface Plasmon Resonance Scattering (SERS). There are also quantum nanoparticles that can convert two or more photons to a higher energy level.[19] Nanoparticle surface modifications are achieved via thiol-thiol interactions, streptavidin-biotin interactions, p-stacking interactions and NHA-EDC chemistry. To validate surface changes the following techniques are employed: UV-vis Spectrometry, Circular Dichroism. Dynamic Light Scattering, and Gel Electrophoresis. [10] SDVD

2.5: ISF vs Blood

The best way to monitor biomarkers is by drawing blood and testing directly, however this has several downsides: it requires specialists, real time monitoring is impossible and the patient is often uncomfortable. Alternatives like sweat and urine are easier to measure but lack the necessary biomarkers and their concentrations can vary significantly. The balancing of the starling forces generates a net force [20] that can push uncharged molecules into the ISF (Interstitial Fluid) by simple diffusion while others require transport proteins and vesicles. As a result the ISF composition is around 60-79% similar [21, 22] to the blood and in some cases biomarkers are only present in the ISF due to metabolic processes or environmental exposure and can be particularly important in identifying skin diseases.

2.6 Hydrogel based nanoparticles and their fabrication (Emulsion polymerisation)

Hydrogels are 3-D networks of natural and sometimes synthetic polymers that can absorb large quantities of water while maintaining their shape. This ability and their permeability for small molecules makes them well suited to drug, protein and general biomolecule encapsulation.[27] The stiffness, swelling ratio, and processability of hydrogels can be adjusted by

introducing different functional groups like carboxylic acids, amino acids and hydroxyl moieties. Additionally the hydrogels can be thermos-responsive, pH-responsive, degradable and even magnetic with the right chemical modifications. There are several delicate biological molecules that can denature on surfaces but hydrogel matrices create a solution like environment the protects their structures. [27] A very common approach to encapsulation is by direct immobilisation onto substrates but hydrogels have several advantages including higher loading, better accessibility, and studies have shown they maintain protein conformation and therefore improve enzyme activity.[27] Emulsion polymerisation is common method to manufacture these hydrogels because it leads to high rates of polymerisation, has high heat removal rates and can produce particles with a higher average molecular weight with size 50-700nm. The process requires an initiator, a surfactant, a dispersion medium and a monomer. Additionally phosphates can be added to maintain pH and sequestering agents can be added to stabilise the initiator.[25]

At first the monomers will be held in micelles once the surfactant concentration is above the critical micelle concentration. When the initiator is introduced it will produce free radicals that will lead to propagation and termination steps producing polymer filled micelles. The Particle size distribution (PSD) of any product made via emulsion polymerisation will be dependent on the particle nucleation, particle growth and coagulation. Uncontrolled coagulation can lead to poor product quality, loss of product and issues scaling the process to the industrial scale. There are often a range of forces acting on particles like Van der Waals forces, electrostatic, steric and depletion forces. Coagulation in this process generally occurs via Brownian motion of particles (Perikinetic) or via the motion of fluids (Orthokinetic) but these systems can co-exist. Different types and concentrations of the key components as well as operating conditions will influence coagulation and ultimately PSD.[26]

2.7 Further developments and challenges

There are plenty of potential uses for biosensors in the future including as DNA vaccines [30] ,'solar freckles' that can prevent skin cancer [31] and graphene based electrical sensors that can act as fitness accessories tracking hydration, temperature and electrocardiogram. Several challenges for biosensors still remain. The ISF contains many similar biomarkers which can limit sensitivity and selectivity so blocking agents need to implemented. [29] Similarly there is a need for sensor coating to reduce the impact of sweat on sensor stability. When sensors are applied in vivo physiological molecules can block the sensor surface so these surfaces need to be modified or surrounded by antifouling membranes. That same pre-existing biological material can have optical properties so it's important to have a high signal to background ratio with the devices that are used. Looking at the existing approaches, enzyme based tattoo biosensors have shown stellar capabilities however they are still limited by their longevity, poor reversibility, low stability and low precision. Fluorophore-based sensors are still subject to photobleaching, light scattering and weak fluorescence [29] and label-free optical sensors have not been used for in vivo monitoring and have low sensitivity and selectivity especially in more complex solutions like the ISF.

3 Materials and Method

3.1 Materials

NIPAM N-isopropylacrylamide (≥99%) , BIS-acrylamide (≥99%), Tween 20 (Polysorbate 20), APS (≥98%), Phosophate buffered saline (PBS) tablets were all purchased from Sigma Aldrich. These commercially available reagents were used as received without further purification. Berberine Chloride (BBR) was synthesized internally and so was the Hydrochloric acid (HCL- 1M) and Sodium Hydroxide (NaOH- 0.1M) solutions. Additionally there was distilled water, a magnetic stirrer and oil as a medium to heat flasks. Key pieces of equipment included a pH probe with a heating element from Mettler Toledo, a large nitrogen tank, an Upright Microscope for Polarisation (DM2700 P, Leica), the Varioskan Lux Microplate reader by Thermoscientific, a Fisherbrand Classic Vortex Mixer and the Particle Analyzer Litesizer 500 Anton Paar.

3.2 Method

3.2.1 Emulsion Polymerisation for hydrogel synthesis

The procedures employed for the synthesis of the P(NIPAM-Nile Blue A) microgels was as follows. All components were added relative to the molar amount of NIPAM monomer. (BIS at 1%, Tween 20 at 1%, APS at 8%, BBR 0.3-1.8%). NIPAM (0.16 g, 1.41 mmol), BIS (2.17 mg, 14.1 μmol), Tween 20 (17.3 mg, 0.0141 mmol, and deionized water (20 mL) were added to a 50 ml round neck flask. After degassing by bubbling with nitrogen for 20 min and heating to 75-80°C, APS (25.7 mg, 112.8μmol) dissolved in 1.0 mL of deionized water was injected under mechanical stirring at 400 rpm. The fluorophore BBR (1.7mg, 5.1mg & 10.1mg) was fully dissolved in ethanol (amount depending on mass used)

then added into the flask. The polymerization was conducted under stirring for 1hr.

3.2.2 Phosphate buffer solution

Initially a 7.4 pH solution was created by dissolving a PBS tablet in 200 ml of deionized water. Using a Ph Probe and adding small amounts for Hydrochloric acid (HCL) and Sodium hydroxide (NaOH) it was possible to produce solutions with pH 6.0, 6.5, 7, 7.5 & 8. The pH of the interstitial fluid can varies between 7.1 and 7.4 much wider than the range of the blood (7.35-7.4) due to a lack of natural pH buffers. The range selected for this research included and extended these values.

Figure 3: Set-up of emulsion polymerization

3.2.3 Microplate Reader

The emissivity of the hydrogel was assessed with the help of the Varioskan Lux microplate reader and the Skanlt RE 6.1 Program. Using a precise pipette 100µL of the desired sample was collected and placed in each quadrant and repeated so that there were 3 repeats. The samples were placed in the centre of the well and at the same heights to ensure accuracy. The plate was placed into the reader and the appropriate protocol for absorbance and fluorescence spectrum was selected. For the fluorescence spectrum a range of 510-800 nm at an excitation wavelength of 488 nm was selected. For absorbance a range of 250-800nm was selected. The goal temperature was selected and the sample was incubated for 15 minutes to ensure it acclimated and reached this temperature. In situations where the temperature needed to be reduced, a cold microplate from the freezer was added to aid the equipment cooling, before the sample plate was re-introduced to acclimate. The time taken to complete the run depended on the number of operations selected and the number of samples selected. Afterwards the excel data was exported and processed.

3.2.4 Particle sizer

Depending on the amount of sample available select the appropriate cuvette. Factors affecting these measurements include the type of cuvette used, solvent present, concentration of the sample and the filling of the cuvette. All cuvettes used were reduced volume cuvettes to maintain consistency. Using a pipette fill up the cuvette with the appropriate sample to the same level each run. Once the cuvette is filled place it inside the apparatus and select the appropriate options along with the correct solvent, in most cases the solvent was water but in others PBS was used, this data can be entered into the system if the refractive index and viscosity is known. Each sample will have up to 60 measurements taken for each run, and each run was repeated 3 times. Because concentration is unknown at the time, this process was repeated with more dilute samples (x10,x20,x30 dilutions) in order to ensure the particles size information obtained was consistent.

3.2.5 Optical microscope

A Leica DM2700 P optical microscope alongside the Las core software helped collect quality images of the manufactured particles. Objective lens from 2.5x to 100 was used.

4 Results and Discussion

4.1 Berberine Chloride

Many well-known Luminogens are subject to the aggregation-caused quenching effect (ACQ) meaning they show little to no emissivity in aggregated solutions but this improves with dilution. Aggregation induced Emission Luminogens have the opposite behaviour and several types exist with colour tunability however they are expensive and complicated to synthesise. Berberine Chloride is a naturally occurring aggregation-induced luminogen (AIEgen) that can be extracted from herbal plants like Hydrastis canadensis and Rhizoma coptidis. BBR Chloride has been researched extensively and has shown pharmacological, antimicrobial and anti-inflammatory properties. A 2018 study by zhao, Gu et al investigated the optical properties of BBR Chloride. [28] The study investigated how the molecule's conformation in different concentrations can lead to intramolecular vibronic motion behaviour and the twisted intramolecular charge transfer effect explaining emission strength. AIE nature was proven by introducing poor solvent which resulted in greater and greater emissivity. Additionally BBR Chloride was identified as a great candidate for LD imaging and disease diagnosis due to its fluorescence. BBR chloride could also function within a tattoo biosensor with biological recognition elements.

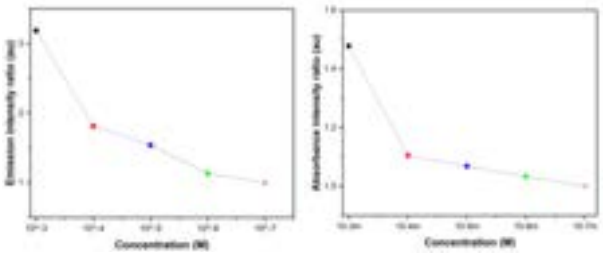


Figure 4: BBR Chloride peak emission (left) and absorbance (right) intensity at 25°C with varying concentrations

In this piece of research by weighing and then dissolving pure BBR Chloride in a 7.4pH Phosphate buffer a 10^{-3} M solution was produced. By diluting further continuously by 10x, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}M solutions were also produced. From figure 4 the Aggregation induced emission quality of Berberine chloride discussed is clear. As the concentration of BBR Chloride is increased the peak intensity also increases,

however the wavelength at which this occurs is relatively unchanged at around 515nm a slight shift from the around 530nm peak observed by Yuan Gu et al.[28] This could be explained by the use of different solvents—PBS vs water/ tetrahydrofuran (THF) mixtures.

4.2 Emulsion Polymerisation for hydrogel synthesis

Using the method outlined above it was possible to make hydrogels containing the fluorophore BBR Chloride at varying amounts i.e 1.7mg, 5.1mg and 10.2 mg. Figure 5 depicts the prepared samples with different amounts of BBR Chloride both at room temperature (25°C) and body temperature (37°C) under white light and blue light with a long pass filter attachment. The colour is darker both at room temperature and body temperature with greater masses of Berberine Chloride used in the polymerisation process which is expected. At room temperature the solutions exhibited fluorescence under

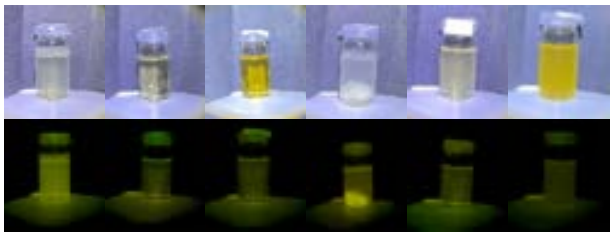


Figure 5: Photos of samples under white light (first row) and blue light with long pass filter (second row). First 3 columns taken at room temperature (25°C), last 3 column taken at body temperature (37°C).

blue light with sample 1, made with 1.7mg of BBR Chloride, appearing the most emissive. This could be the effect of to the presence of ions in the chosen buffer. When the samples were heated to body temperature, they became cloudier which was expected due the reduced solubility of the hydrogel at higher temperatures (see figure 6) resulting in precipitation. There was also a significant increase in emission intensity from sample 1 but not from sample 2 or 3.

On top of particle nucleation and particle growth from the polymerisation process, coagulation is another important factor in the particle size distribution of particles made. Undesired coagulation can lead to fouling of reactor internals, lower product quality, and product loss which will prevent processes from reaching the large scale.

There are van der Waals forces in play between particles like electrostatic attractions, steric forces and depletion forces to name a few. Perikinetic and orthokinetic coagulation describe the two most common mechanisms for coagulation in emulsion polymerisation which are by diffusive forces and convective forces respectively. All variables including the concentration and type of monomer, surfactant, initiator, and operating conditions will influence coagulation in emulsion polymerisation.[33]

As part of this process undesired aggregation needed to be dealt with. Aggregation can occur due to the formation of hotspots on the degassing nitrogen inlet needle, caused by insufficient mixing. The presence of these hotspots encourages the formation of a film that cannot be redistributed into the reaction mixture therefore the stirrer speed needed to be maximised to avoid this.[32] Additionally the amount of initiator added needed to be monitored. Work done by Baijun Liu et al on anionic KPS, very similar to APS, found that increased initiator concentration increased primary radical concentration and collision probability resulting in larger particle size and narrowly dispersed particles. Due to this an initiator of 4% relative to the monomer was selected in line with work by Jun Yin et al on fluorescent potassium ion sensors.[34]

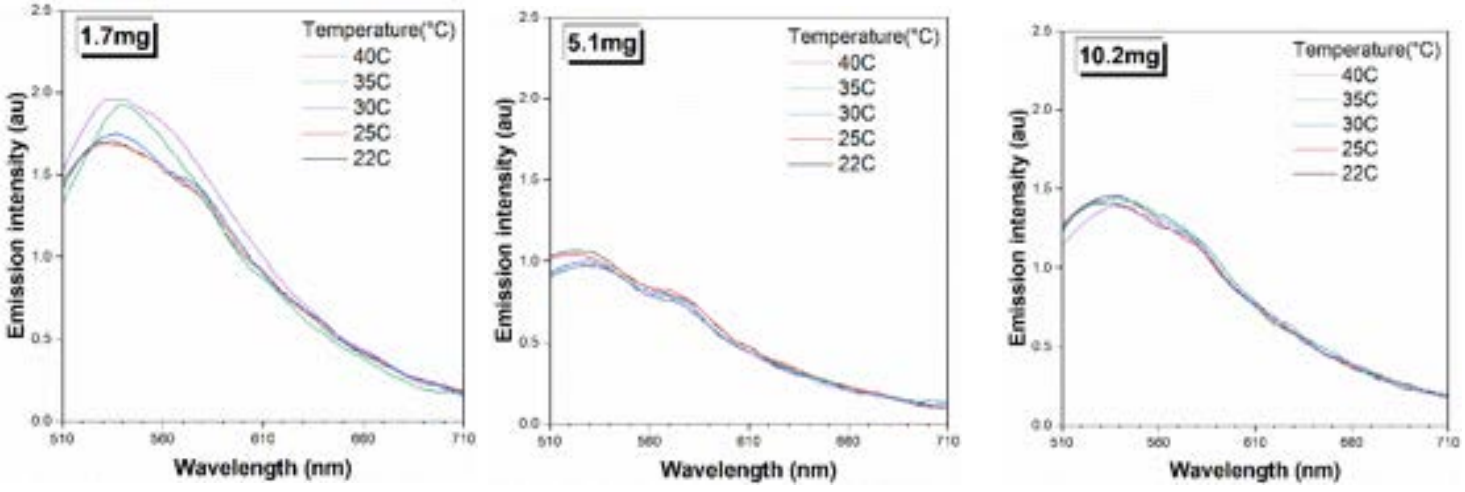


Figure 7: Fluorescence spectra of samples with varying masses of BBR Chloride used in synthesis

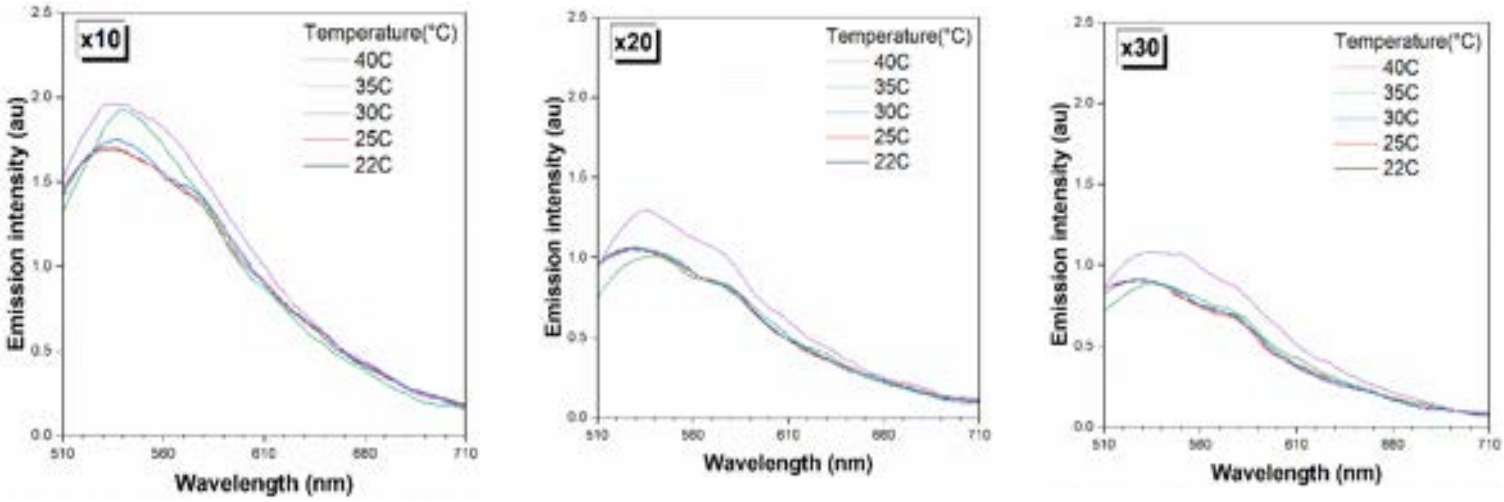


Figure 8: Fluorescence spectra of sample 1 (made with 1.7mg of BBR chloride) at varying levels of dilution.

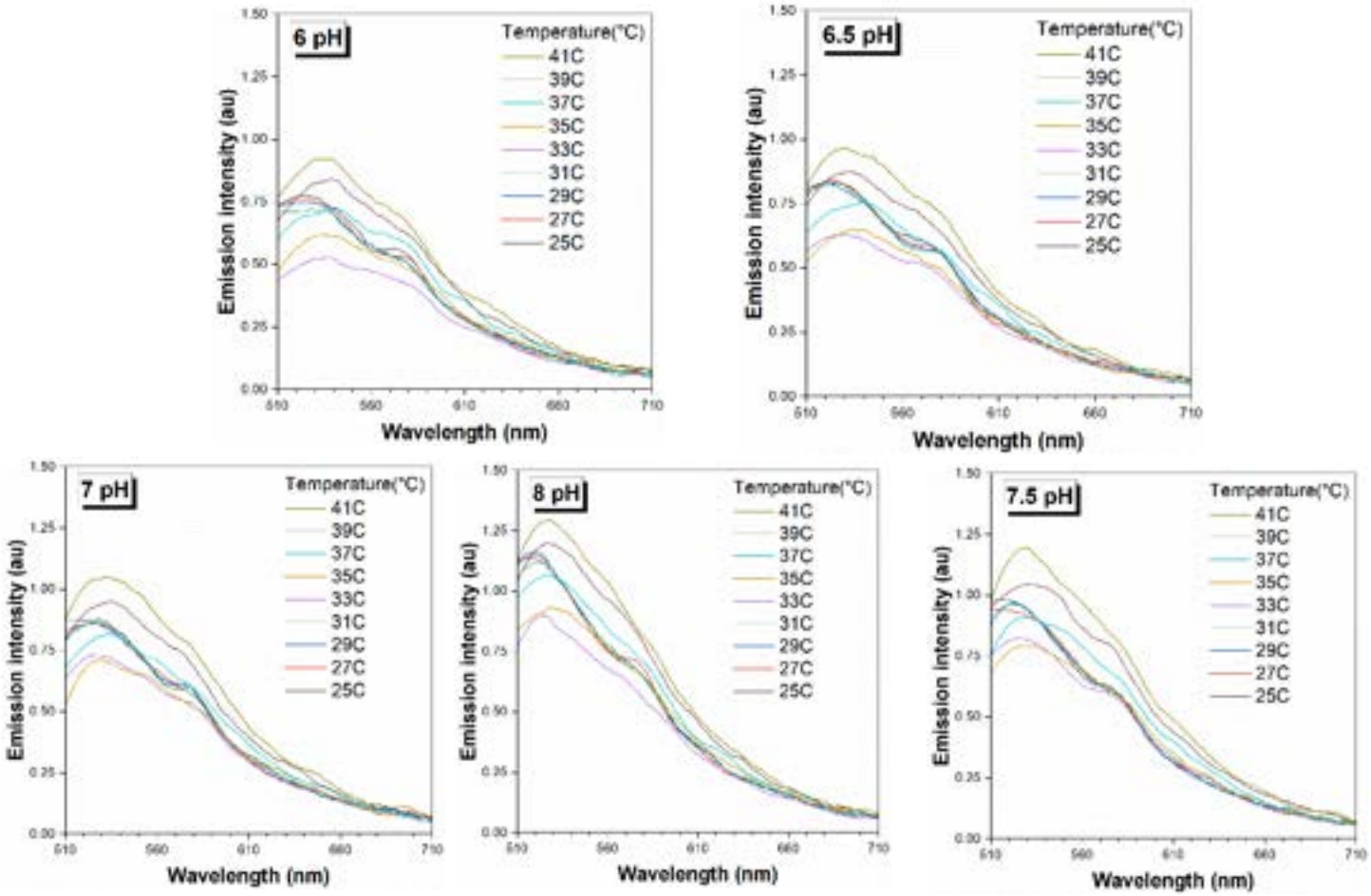


Figure 9: Fluorescence spectra of sample 1 at x30 dilution at various temperatures from 25°C - 41°C at specified pH levels.

4.3: Effect of fluorophore mass

Looking at the fluorescence spectra from figure 7 there are differences in the curves where different amounts of BBR Chloride was used in synthesis. The wavelength of peak intensity remains relatively unchanged at around 530nm. Due to the AIE nature of the fluorophore the expectation is that a larger amount would lead to greater emissivity which is observed from sample 3 (10.2mg) relative to sample 2(5.1mg). However, sample 1 with the smallest amount of fluorophore (1.7mg) had the greatest peak intensity overall of close to 2au. This could be explained by greater agglomeration in sample 1 which was prepared with lower stirrer speeds. Further investigation should be carried out into the effect of stirrer speed on particle size distribution for this particular fluorophore-hydrogel matrix. When it came to selecting a sample for further analysis, sample 1 was chosen for two reasons. 1) It had greater emissivity both visually and according to fluorescence data 2) It had clearer differentiation between fluorescence at increasing temperatures, particularly at the temperatures of interest (35-40°C i.e. body temperature).

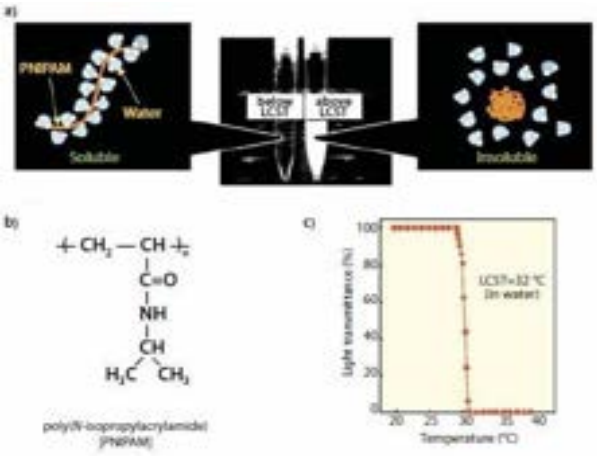


Figure 6: PNIPAM Solubility with temperature [35]

4.4: Effect of Dilution

Looking to figure 8 it is possible to discern the impact of dilution on the thermos-responsiveness of sample 1. Again, here the wavelength of peak intensity does not significantly shift with dilution. However, the peak intensity decreases from 10x dilution to 30x dilution. At 20x and 30x dilution the 40°C curve has the greatest emissivity across the observed range relative to other temperatures at the same level of dilution, especially around the 510-600nm range. This was less clear at 10x with the 35°C curve being much closer. The general behaviour of greater emissivity at greater temperature for the hydrogel-fluorophore matrix can be explained by figure 6. Above the lower critical solution temperature the hydrogel precipitates. This brings the fluorophore closer together and leads to greater emissivity due to its AIE active nature. The ISF will follow the body and blood closely in temperature therefore its range will be within 36-40°C and this the ultimate range of interest for our optical biosensor. The greatest distinction between 35°C and 40°C was seen at 30x dilution so this was selected for further analysis.

4.5: Effect of pH on thermos-response

As previously mentioned the pH of the interstitial fluid can vary between 7.1 and 7.4 but this was extended for investigation. With reference to figure 9 the peak wavelength and general shape of the fluorescence spectrums do not change significantly due to pH and are the same for each temperature. The same trend of greater emissivity with temperature is clear at all pH values. The peak intensity increases with pH for all the temperatures.

This is clearer to see with figure 10. There is a surprising decline in emissivity at 33°C and 35°C. Considering the range of importance is between 36°C and 40°C, for which there is a clear upward trend, this is not alarming.

4.6: Dynamic light scattering (DLS) data discussion

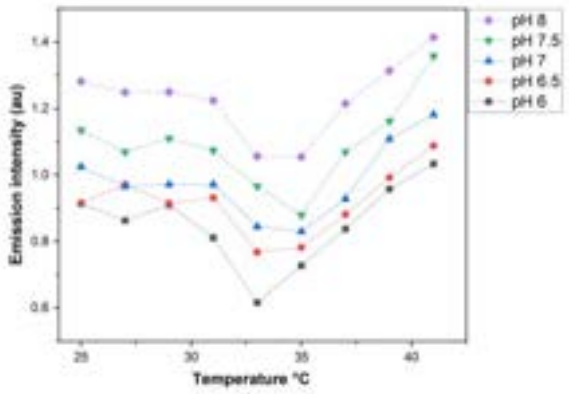


Figure 10: pH effect on peak intensity at various temperatures

DLS was used to inspect the size distribution of the hydrogel-fluorophore matrix. Key figures to note are the particle sizes and polydispersity index. Figure 12 depicts the particle size distribution of sample 1 with 1.7mg of BBR Chloride at x30 dilution. 3 peaks were observed with the greatest peak found at roughly 690nm. This is in line with data compiled by other researchers, who found P-NIPAM microgels to be in the range of 250-760nm [41]. Smaller peaks can be seen at both extremes at 59nm and 12,000nm, constituting to roughly 14% and 5% of the sample respectively. The smaller peak can be attributed to any unreacted elements in the mixture, while the greater peak is caused by the presence of smaller agglomerates formed during the reaction due to the formation of hotspots in the reaction vessel as explained earlier. With further purification of the reaction mixture, with techniques such as dialysis or centrifugation, these peaks are expected to disappear. With the majority of the sample being around 645nm, this matrix is promising for application in functionalization of tattoos and the replacement of tattoo pigments with biosensors, as this is on the lower end of the tattoo pigment size range of 10-5000nm. The

polydispersity index (PDI) of the sample is also of importance and averaged at roughly 30% or 0.3. This value in DLS measurements, denotes the size distribution of nanoparticles. PDI ranges from 0.0 to 1.0, with 0.0 displaying perfectly uniform particle size distribution and 1.0 being an extremely dispersed sample with various particle sizes present. As discussed by Danaei et al [42], acceptable values for polymer-based nanoparticles in industry will be 0.0-0.2. Although the obtained value is slightly higher, the PDI of the product can be brought lower with further purification of the sample as mentioned previously.

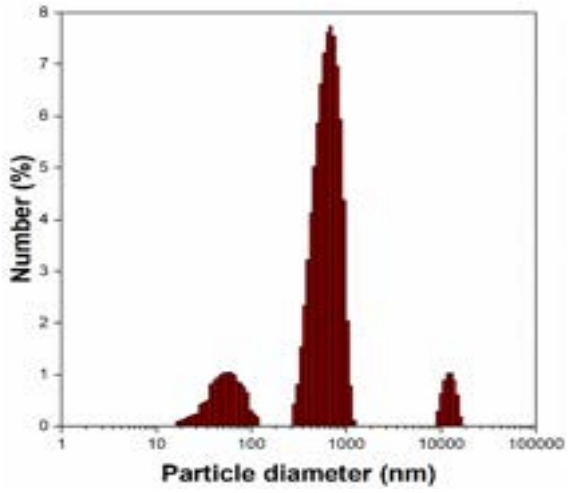


Figure 12: Particle size distribution for sample 1, 30x

4.7 Microscope images

Figure 11, displays the sample acclimatized at room temperature at x100 magnification under white light with no filters. The images obtained resemble PNIPAM microgels formed by Ruscito et al [43] using similar materials and ratios at room temperature. Particles in figure x exhibit a ring like structure which is typical of microgels with average diameters lower than 3 µm as discussed by Ruscito et al.

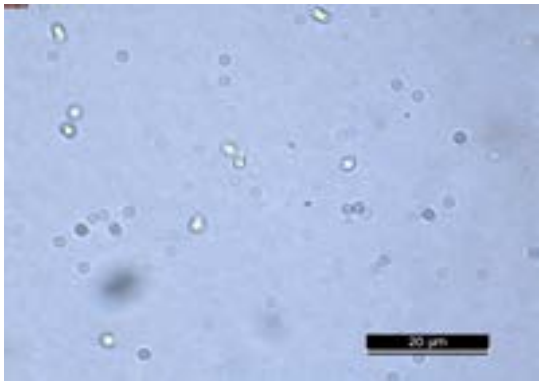


Figure 11: Optical microscope image of sample 1 at 100x magnification at room temperature. Scale bar = 20 µm

4.8: Thermal Reversibility

The ability of an optical biosensor to operate nimbly within the ISF temperature range is very desirable and it is important to understand how emissivity varies with cycles. Figure 13 shows the wavelength of peak intensity varies with temperature and can return to the same value over multiple cycles which is a positive indicator. However figure 13 also shows there is an upward trend in the peak intensity with cycles making it harder to differentiate between the two distinct temperatures and any in between. The peak at 40°C did remain above the peak at 36°C for all the cycles. The primary function of the biosensor is not to detect temperature changes but temperature contribution to emissivity will need to understood alongside analyte concentration.

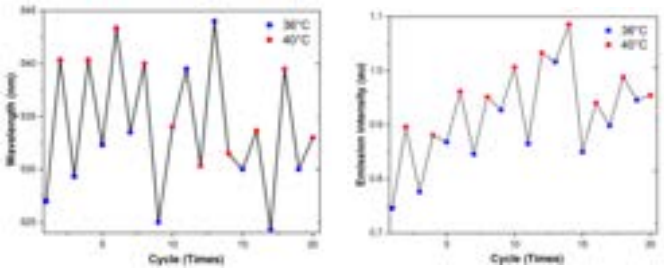


Figure 13: Peak intensity and wavelength reversibility

5 Conclusion and Outlook

Fully functional optical biosensors have the potential to improve the detection of diseases and quality of life via point of care testing. In this work we investigated the properties of a hydrogel formulation with an AIE fluorophore. The product showed responsiveness to temperature and pH changes as well as limited reversibility in the ISF range (36-40 °C)

5.1 Further improvements and experiments

Following on from this work there are several potential improvements and extensions. The samples made via emulsion polymerisation can be further purified either by centrifugation or using dialysis membranes. This would remove unreacted compounds and not only improve the DLS and emissivity data but also ensure biocompatibility. Additionally the consistency of the emulsion polymerisation process needs to be investigated and maintained since the difference in operating conditions could have contributed to the unexpected greater emissivity of sample 1 relative to 2 and 3. To realize the end goal of a functionalised tattoo a recognition element needs to be selected and added to the hydrogel and fluorophore formulation. Then any changes in temperature, pH responsiveness and reversibility should be investigated and explained. Finally the optical biosensor should be paired with a software application in an in vivo study too test its accuracy.

Acknowledgements

We would like to thanks the members of the Yetisen Biosensors group for advising us throughout this time. Particularly Yubing Hu and Dr Ali K. Yetisen.

References

1. https://www.bbc.co.uk/news/science-environment-43230202 (first recorded tattoo) [Accessed October 2022]
2. https://www.karger.com/Article/Pdf/496986 (tattooed population of the world. [Accessed October 2022]
3. H. Shibata, Y. J. Heo, T. Okitsu, Y. Matsunaga, T. Kawanishi, S. Takeuchi, Proc. Natl. Acad. Sci. USA 2010, 107. (Glucose Biomarker)
4. N. Jiang, A. K. Yetisen, N. Linhart, K. Flisikowski, J. Dong, X. Dong, H. Butt, M. Jakobi, A. Schnieke, A. W. Koch, Sens. Actuators, B 2020, 320, 128378. (Sodium Biomarker)
5. A. K. Yetisen, R. Moreddu, S. Seifi, N. Jiang, K. Vega, X. Dong, J. Dong, H. Butt, M. Jakobi, M. Elsner, A. W. Koch, Angew. Chem., Int. Ed. 2019, 58, 10506. (Albumin biomarker)
6. C. Samson, A. Koh, Front. Bioeng. Biotechnol. 2020, 8, 1037. (Oxygen Biomarker)
7. K. Vega, N. Jiang, X. Liu, V. Kan, N. Barry, P. Maes, A. Yetisen, J. Paradiso, ACM International Symposium on Wearable Computers 2017, 1, 138. (Ph Biomarker)
8. X. Liu, H. Yuk, S. Lin, G. A. Parada, T.-C. Tang, E. Tham, C. de la Fuente-Nunez, T. K. Lu, X. Zhao, Adv. Mater. 2018, 30, 1704821. (Environmental contaminants biomarker)
9. Sabri, N.; Aljunid, S.A.; Salim, M.S.; Ahmad, R.B.; Kamaruddin, R. Toward optical sensors: Review and applications. J. Phys. Conf. Ser. 2013, 423, 012064.
10. How to make nanobiosensors: surface modification and characterisation of nanomaterials for biosensing applications Meral Yuce ¨ *a and Hasan Kurtb
11. Biosensor nanoengineering: Design, operation, and implementation for biomolecular analysis Buddhadev Purohit b , Pramod R. Vernekar c , Nagaraj P. Shetti c , Pranjal Chandra a,b
12. A. Baranska, A. Shawket, M. Jouve, M. Baratin, C. Malosse, O. Voluzan, T.-P. Vu Manh, F. Fiore, M. Bajénoff, P. Benaroch, M. Dalod, M. Malissen, S. Henri, B. Malissen, J. Exp. Med. 2018, 215, 1115.
13. D. Rodrigues, A. I. Barbosa, R. Rebelo, I. K. Kwon, R. L. Reis, V. M. Correlo, Biosensors 2020, 10, 79
14. J. R. Ott, M. Heuck, C. Agger, P. D. Rasmussen, O. Bang, Opt. Express 2008, 16, 20834
15. Z. Xia, J. Rao, Curr. Opin. Biotechnol. 2009, 20, 37.
16. B. G. Andryukov, N. N. Besednova, R. V. Romashko, T. S. Zaporozhets, T. A. Efimov, Biosensors 2020, 10, 11
17. T. R. Ray, J. Choi, A. J. Bandodkar, S. Krishnan, P. Gutruf, L. Tian, R. Ghaffari, J. A. Rogers, Chem. Rev. 2019, 119, 5461.
18. S. C. Mukhopadhyay, IEEE Sens. J. 2015, 15, 1321.
19. How to make nanobiosensors: surface modification and characterisation of nanomaterials for biosensing applications (Meral Yuce and Hasan Kurt.)
20. D. U. Silverthorn, B. R. Johnson, W. C. Ober, C. E. Ober, A. C. Silverthorn, in Human Physiology: An Integrated Approach, 7, Pearson, 2016, 616
21. P. P. Samant, M. M. Niedzwiecki, N. Raviele, V. Tran, J. Mena-Lapaix, D. I. Walker, E. I. Felner, D. P. Jones, G. W. Miller, M. R. Prausnitz, Sci. Transl. Med. 2020, 12, eaaw0285.
22. R. Srivastava, R. D. Jayant, A. Chaudhary, M. J. McShane, J. Diabetes Sci. Technol. 2011, 5, 76.
23. J. Luan, A. Seth, R. Gupta, Z. Wang, P. Rathi, S. Cao, H. G. Derami, R. Tang, B. Xu, S. Achilefu, J. J. Morrissey, S. Singamaneni, Nat. Biomed. Eng. 2020, 4, 518.
24. M. Soler, C. S. Huertas, L. M. Lechuga, Expert Rev. Mol. Diagn. 2019, 19, 71
25. https://www.intechopen.com/chapters/57833 : [Accessed November 2022]
26. **https://www.tandfonline.com/doi/full/10.1080/15583724.2017.1405979?scroll=top&needAccess=true:** [Accessed November 2022]
27. Hydrogels and Their Role in Biosensing Applications Anna Herrmann, Rainer Haag, and Uwe Schedler
28. Exploration of biocompatible AIEgens from natural resources Yuan Gu, ab Zheng Zhao
29. J. Chao, Z. Li, J. Li, H. Peng, S. Su, Q. Li, C. Zhu, X. Zuo, S. Song, L. Wang, L. Wang, Biosens. Bioelectron. 2016, 81, 92.
30. O. Wichterle, D. LÍM, Nature 1960, 185, 117
31. Q. Tang, T. N. Plank, T. Zhu, H. Yu, Z. Ge, Q. Li, L. Li, J. T. Davis, H. Pei, ACS Appl. Mater. Interfaces 2019, 11, 19743.
32. (2020) *What causes polymer build up in emulsions?* Pilot Chemical. Available at: https://www.youtube.com/watch?v=kgN2Lx37BQI (Accessed: December 9, 2022).
33. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6432544/ : [Accessed November 2022]
34. FRET-Derived Ratiometric Fluorescent K+ Sensors Fabricated from Thermoresponsive Poly(N-isopropylacrylamide) Microgels Labeled with Crown Ether Moieties. Jun Yin et al.
35. https://www.sigmaaldrich.com/GB/en/technical-documents/technical-article/materials-science-and-engineering/polymer-synthesis/poly-n-isopropylacrylamide: [Accessed December 2022]
36. Pazos, M.D. *et al.* (2021) "Tattoo inks for optical biosensing in interstitial fluid," *Advanced Healthcare Materials*, 10(21), p. 2101238. Available at: https://doi.org/10.1002/adhm.202101238.
37. Dash, G. *et al.* (2022) "Non-laser treatment for Tattoo Removal," *Journal of Cosmetic Dermatology* [Preprint]. Available at: https://doi.org/10.1111/jocd.14819.
38. FERGUSON, J.E. *et al.* (1997) "The Q-switched neodymium: YAG laser and tattoos: A microscopic analysis of laser-tattoo interactions," *British Journal of Dermatology*, 137(3), pp. 405–410. Available at: https://doi.org/10.1046/j.1365-2133.1997.18581951.x.
39. Schreiver, I. and Luch, A. (2016) "At the Dark End of the rainbow: Data gaps in tattoo toxicology," *Archives of Toxicology*, 90(7), pp. 1763–1765. Available at: https://doi.org/10.1007/s00204-016-1740-9.
40. ANDERSON, L.A.W.R.E.N.C.E.L. *et al.* (1996) "Tattoo pigment mimicking metastatic malignant melanoma," *Dermatologic Surgery*, 22(1), pp. 92–94. Available at: https://doi.org/10.1111/j.1524-4725.1996.tb00578.x.
41. Destribats, M. *et al.* (2014) "Impact of PNIPAM microgel size on its ability to stabilize Pickering emulsions," *Langmuir*, 30(7), pp. 1768–1777. Available at: https://doi.org/10.1021/la4044396.
42. Danaei, M. *et al.* (2018) "Impact of particle size and Polydispersity Index on the clinical applications of Lipidic Nanocarrier Systems," *Pharmaceutics*, 10(2), p. 57. Available at: https://doi.org/10.3390/pharmaceutics10020057.
43. Ruscito, A. *et al.* (2020) "Microgel particles with distinct morphologies and common chemical compositions: A unified description of the responsivity to temperature and osmotic stress," *Gels*, 6(4), p. 34. Available at: https://doi.org/10.3390/gels60400

Study of the Effects of Joule Heating on Alkaline Electrolysis

Ishmail Al-Rashid and Spencer Ingall

Abstract

A Eulerian multiphase flow model, based on the multiphaseEulerFoam package available in OpenFOAM, was used to simulate a 2D alkaline electrolyser. The main goal of this study was to add a Joule Heating source term to the energy balance, and examine what effect that would have on the flow, energy, and efficiency of the system based on how much additional time was required to run simulations. A range of parameters were tested through individual simulations, ran at: temperatures of 300K, 325K and 350K and current densities of 1000Am^{-2}, 2000Am^{-2} and 3000Am^{-2}. Joule Heating was seen to cause significant deviations in the Hydrogen gas evolution along the electrolyser, bulk flow of the electrolyte, and efficiency of cells. Additionally, using the simulation timing as a proxy, it has been proven that implementation does not require a significant amount of extra computational power. Therefore, it should be implemented in future anisothermal models of alkaline electrolysers and further examined at greater current densities than those prescribed here.

1 Background

Fossil fuels accounted for 82% of the world's total energy supply in 2021 (BP 2022). Despite global corporate and political powers maintaining fossil fuel dominance in energy markets, climate change activism, research, and innovation is slowly changing the world. The energy sector has observed growth in renewable investment with the UK generating 12% more renewable energy in 2022 than the previous year (Gov.uk 2022). An exemplary area of growth is within the transport sector which has seen a massive increase in the sales of electric vehicles - expected to account for 35% of all new car sales by 2040 (MacDonald 2016). However, like most sectors undergoing transformation in energy, reliance on fossil fuels decreases incredibly gradually and is not a straightforward process. The huge growth of the EV industry means that there is a large increase in demand for lithium and cobalt (amongst many other commodities). The extraction and processing of these materials are far from renewable with the mining of rare earth metals like cobalt also being encapsulated by political and humanitarian issues.

One promising way to avoid the heavy use of rare, dangerous, and non-renewable resources is the use of water, one of earths most abundant resources, to create hydrogen fuel. Hydrogen fuel offers a solution to many of the problems outlined earlier. It is completely renewable, creates virtually no pollution, and is hugely abundant while being more energy dense than its fossil fuel alternatives. Although demand for hydrogen powered vehicles grew by 84% in 2021, this totalled at a meagre 15,500 cars sold (iea.org 2022) which shows how slow uptake and adoption can be. Currently, the largest uses for hydrogen include fertilizer production and petrol refining, both of which are crucial industrial processes. Demand for hydrogen fuel is already large and with the rising need for hydrogen as a renewable alternative in transport and other sectors, demand is projected to rise from 94Mt in 2021 to 180Mt in 2030 (iea.org 2022).

The main problem hydrogen fuel faces in becoming a reliable source of renewable energy is its production; 95% of hydrogen currently being produced is generated from natural gas and coal (Chi and Yu 2018). Water-splitting is the primary renewable method of producing hydrogen, but the amount of hydrogen produced via this method is currently not significant on an industrial scale. Nevertheless, many countries are currently investing in the research and development of "greener" hydrogen production. One example of this can be seen in 2019 when ITM Power received funding from the UK Department for Business, Energy and Industrial Strategy for a project aiming to lower the cost of electrolytic hydrogen (Gov.uk 2022).

In 2021, Kakoulaki raised various points towards the need for a better understanding of hydrogen production via water splitting. Specific emphasis is placed on the 2019 European Green Deal, in which hydrogen is considered a key input in the future energy system as a flexible energy carrier for industry and transport.(Kakoulaki et al. 2021) The validity of this statement and the role of electrolysis in the rise of hydrogen as an energy source is affirmed by the projections that the cost of producing hydrogen via electrolysis will decrease from $6.00 per kg to $2.50 by 2030, with electrolyser costs being projected to halve (Kakoulaki et al. 2021).

The barriers currently preventing industrial water-splitting from being more wide-spread is the huge associated electricity cost in delivering high yield. Currently, the efficiency of water-splitting is too low to warrant large-scale commercial use (Chi and Yu 2018). Hreiz et al states that in order to increase the efficiency of this process, we need accurate knowledge of multiphase flow within the electrolysers to make improvements on the operating conditions and apparatus (Hreiz et al. 2015). In order to expand understanding of this topic, this study will investigate the effects of Joule heating within an alkaline electrolyser on multifluid flow, as well as other relevant variables (see Section 3 Results).

Joule heating, also known as resistive heating or ohmic heating, is a process in which electrical energy is converted into heat energy. It occurs when an electric current flows through a conductor, such as a metal wire, and encounters resistance. This effect is prevalent, and in some cases taken advantage of, in various industrial electrolytic processes. For example, during the process of purifying water through electrochemical advanced oxidation, applying a current to the anode dissipates heat. This leads to an interfacial temperature that is higher than the bulk solution resulting in the interfacial temperature increasing from 25 to 70.2 degrees, which creates a rise in the bulk solution temperature by 8.6 degrees (Pei et al. 2019). Furthermore, it is observed that Joule heating contributes to the high tempera-

tures needed for solar-driven high-temperature water-splitting (Lin et al 2022).

The previous models by Zarghami et al and Hreiz et al which this paper builds on all operated isothermally(Zarghami, Deen, and Vreman 2020; Hreiz et al. 2015). Hence the purpose of this paper is to investigate the significance of the Joule heating and ultimately if it is worth accounting for in anisothermal simulation models.

2 Methodology

2.1 Model Dimensions

The 2D mesh used in this study is made up of 46,000 cells, and has a simple rectangular shape for the electrolyser, with a membrane directly in the middle - see Figure 1.

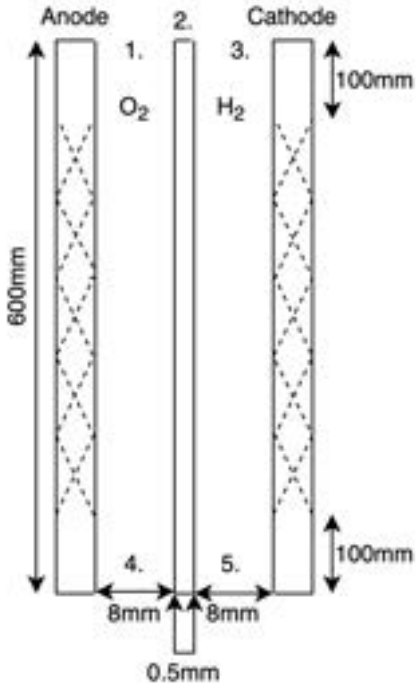


Figure 1: 2D Electrolyser Model Diagram

Table 1: Electrolyser Model Diagram Labels.

No.	Label
1.	O_2 Outlet
2.	Separator
3.	H_2 Outlet
4.	O_2 Inlet
5.	H_2 Inlet

The dotted sections denote the anode and cathode, where the first 100mm at the inlet is to allow the flow to fully develop by the time it reaches the electrodes.

2.2 Mesh Sensitivity and Boundary Conditions

For this study, a mesh sensitivity analysis was not conducted as the mesh was obtained from Zarghami et al's paper where a mesh independence test was conducted for a wide range of parameters, inclusive of those used in this study (Zarghami, Deen, and Vreman 2020).
In terms of boundary conditions, the initial velocity of the electrolyte solution was set to be 0.69ms^{-1} at the inlets (4 and 5) with a no-slip condition applied everywhere else. The temperature was initially a set value for the electrolyte, Hydrogen gas, and Oxygen gas as per the simulation choice (See Section 3 Results), and zero gradient everywhere else. The flow is subject to an atmospheric pressure drop. No-slip boundaries are implemented with the standard wall function approach (Zarghami, Deen, and Vreman 2020).

2.3 Governing Equations

In building the CFD simulation, the first point of contention was modelling the system via the Euler-Euler method or the Euler-Lagrange method. The positives and negatives of both methods are explained in detail by Taqieddin et al where the authors state that the greater accuracy of the E-L model can be attributed to the use of describing bubbles via the Langrangian approach which tracks bubbles individually using Newton's Second Law. Tracking bubbles individually would be extremely computationally expensive for larger scale simulations, such as those being undertaken in this project (Taqieddin et al. 2017). On top of this, the maximum void fraction needs to be limited to values smaller than those encountered in our simulation hence, in this study, a Eulerian multifluid flow model was created based upon the multiphaseEulerFoam model available in OpenFOAM-8. The Euler-Euler CFD model used in the simulation is based upon the following continuity and momentum equations:

$$\frac{\partial(\rho_i \alpha_i)}{\partial t} + \nabla \cdot (\rho_i \alpha_i \mathbf{U}_i) = \mathbf{S}_i \tag{1}$$

$$\begin{aligned} &\frac{\partial(\rho_i \alpha_i \mathbf{U}_i)}{\partial t} + \nabla \cdot (\rho_i \alpha_i \mathbf{U}_i \mathbf{U}_i) \\ &= -\alpha_i \nabla p + \nabla \cdot (\alpha_i \tau_i) + \rho_k \alpha_i \boldsymbol{g} + \sum \mathbf{F}_i \end{aligned} \tag{2}$$

Where ρ_i, α_i, τ_i and $\mathbf{U}_i$ denote the respective density, volume fraction, stress tensor and velocity of species i, and $\mathbf{F}_i$ denotes the interphase forces (drag, lift, virtual mass etc.) and $\mathbf{g}$ denoted the gravitational field strength. $\mathbf{S}_i$ denotes the gas phase H_2 and O_2 produced from electrochemical reaction with Faradays's Law:

$$\mathbf{S}_i = \frac{jM_i}{zF\rho_i} \tag{3}$$

Where M_i denotes the molar mass of species i, z denotes stochiometric coefficient and F is Faraday's constant. j, the current density, is given by the following equation:

$$j = (-\sigma_{eff} \nabla \phi) \cdot \boldsymbol{n} \tag{4}$$

Where σ_{eff} denotes effective conductivity and ϕ denotes the potential. The energy equation that is being solved throughout the equation is the following:

$$\frac{\partial(\alpha_i\rho_i h)}{\partial t} + \frac{\partial(\alpha_i\rho_i k_i)}{\partial t} + \nabla\cdot(\alpha_i\rho_i h_i\mathbf{U}_i) + (\alpha_i\rho_i k_i\mathbf{U}_i) = \alpha_i\frac{\partial p}{\partial t} + \nabla\cdot(\alpha_i\alpha_{eff}\nabla h_i) + h_T(T_c - T_d) + (\sigma_{eff}\nabla\phi)\cdot\nabla\phi \tag{5}$$

Where h_i is the is specific enthalpy, α_{eff} is the effective thermal diffusivity, h_T is the heat transfer coefficient between continuous and dispersed phase (a function of the Nusselt numver which is described in the chapter 2.7), T_c ad T_d are the temperatures of the dispersed and continous phases and finally where $(\sigma_{eff}\nabla\phi)\cdot\nabla\phi$ denotes the Joule heating source term.

The electrical potential is defined by the following equation:

$$\nabla\cdot(-\sigma_{eff}\nabla\phi) = 0 \tag{6}$$

And the conductivity in electrolyte and separator are given by the following respective equations:

$$\sigma_{eff} = \sigma_o(1-\alpha_{gas})^{1.5} \qquad \sigma_{eff} = \sigma\frac{\epsilon}{\tau} \tag{7}$$

Where σ_0 denotes the electrolyte conductivity, tau and ϵ are the diaphragm tortuosity and porosity respectively and α_{gas} denotes the gas volume fraction. The boundary conditions at the electrodes follow the Tafel relations found below:

$$\phi_{el} = \phi_{cell} - \eta - E_{rev}, \qquad \eta = \frac{2.303RT}{\alpha nF}log\left(\frac{j}{j_0}\right),$$
$$\phi_{el} = \phi_{cell} - \frac{2.303RT}{cnF}log\left(\frac{(-\sigma_{eff}\nabla\phi_{el}\cdot)\,\boldsymbol{n}}{j_0}\right) - E_{rev} \tag{8}$$

Where ϕ_{el} and ϕ_{cell} are the electrode and cell potential, n is number of electrons involved in the reaction, j is the current density, j_0 is exchange current density and c is charge transfer coefficient.

The flow is considered Newtonian and incompressible with the thermophysical properties of the phases being non-isothermal. The key properties - density, specific heat capacity, thermal conductivity and viscosity - were modelled using temperature-dependent 5th Order polynomials, with operating pressure assumed to be constant at 1 bar. The electrolyte is a KOH solution of concentration 1M and the considered gases are pure Hydrogen and Oxygen. The parameters which appear in Tables 3, 4, and 5 (see Appendix) require a decimal precision of at least 16 to output the correct results.

Specifically in terms of density, it was important to model the amount of water vapour present in the gaseous bubbles formed at both the anode and cathode as the additional water vapour has the potential to change the density values drastically. These models were created using experimental data with vapour pressures of water with H_2/O_2 gaseous mixtures (Kell, McLaurin, and Whalley 1989). Viscosity, thermal conductivity, and specific heat capacity were calculated using pure vapour data from NIST for Hydrogen/ Oxygen (nist.org 1997), and with extrapolation from literature for KOH (Zaytsev and Aseyev 1992).

2.4 Drag

To represent the resistance on the gas bubbles as they move through the electrolyte, a drag force was introduced which acts in the opposite direction of the bubble-liquid slip velocity. For this simulation, the Wen-Yu drag model (Wen 1966) was chosen where the dimensionless drag coefficient, C_d, is expressed as the following (for $ReynoldsNumber(Re) > 1000$):

$$C_d = 0.44\theta_g^{-2.65} \tag{9}$$

Where θ_g is the gas volume fraction.

2.5 Turbulent Dispersion

In terms of dispersed multi-phase flow, turbulence in the continuous phase causes particles in the dispersed region to be carried from areas of high to low concentration (Burns et al. 2004). It is observed that the random motion of continuous phase eddies brings about a considerable redistribution of bubbles in the axial direction. The model derived by Burns et al is used in the simulation and can be seen in Equation 10:

$$F_{td} = -\frac{3}{4}\frac{C_d}{d_b}\theta_g|\mathbf{U}_g - \mathbf{U}_l|\frac{\mu_l^{turb}}{Sc_{td}}\left(\frac{1}{\theta_g} + \frac{1}{\theta_l}\right)\nabla\alpha_g \tag{10}$$

Where d_b is the bubble diameter, μ_l^{turb} is the turbulent viscosity and Sc is the Schmidt number.

2.6 Lift

The lift force expresses the force perpendicular to its direction of motion that a bubble experiences when travelling in a shear flow. For a spherical bubble, the lift coefficient is positive and the bubble travels in the direction of decreasing liquid velocity. In the simulation, the lift coefficient is obtained through the following correlation derived by Tomiyama et al when studying the trajectories of a single bubble rising in a shear flow (Tomiyama et al. 2002).

$$\begin{cases} C_L = min\left[0.288tanh(0.121Re), f(Eo_\perp)\right] \\ \qquad\qquad Eo_\perp < 4 \\ C_L = f(Eo_\perp) \ldots\ldots\ldots\ldots\ldots\ldots 4 < Eo_\perp < 10 \\ C_L = -0.27 \ldots\ldots\ldots\ldots\ldots\ldots Eo_\perp > 10 \end{cases} \tag{11}$$

Where $Eo_\perp$ denotes the Eötvös number calculated from the maximum horizontal dimension of the bubble.

2.7 Heat Transfer

In systems where the Reynolds number does not surpass 10^4, it is proven by Buist et al that the Ranz-Marshall correlation predicts the Nusselt number, (Nu), as a function of the Reynolds number follows experimental date with a high degree of accuracy (within 2-3%) (Buist et al. 2017).

$$Nu_D = 2 + 0.6Re_D^{\frac{1}{2}}Pr^{\frac{1}{3}} \tag{12}$$

Where Pr denotes the Prandtl number. As the Reynolds number of this simulation is 7000, the correlation was deemed appropriate.

2.8 Turbulence

The turbulence within the electrolyte solution is modelled using the shear stress transport (SST) $k-\omega$ formulation (Menter 1993). The model is a Reynolds-averaged Navier-Stokes model which is mainly used for turbulent flows with low Reynolds numbers. This model was chosen over the $k-\epsilon$ model and the $k-\omega$ model for the following reasons; the $k-\epsilon$ model is known to have unreliable damping functions when applied to flows different from the calibration flows and although the $k-\omega$ model addresses this, it was still found that flow struggled to separate from the smooth surface of the body. The two-equation nature of the model means that in addition to the conservation equations, the two transport PDEs are also solved so convection and diffusion of turbulent energy are accounted for. Furthermore, in Zarghami et al's investigation of turbulence models and their similarities to simulated electrolyser's experimental data, it was found that the SST $k-\omega$ model performs best (Zarghami, Deen, and Vreman 2020) The mathematical representation can be seen in the following Equations 13, 14 and 15:
Kinematic Eddy Viscosity

$$\nu_T = \frac{a_1 k}{max\left(a_1\omega SF_2\right)} \tag{13}$$

Turbulence Kinetic Energy

$$\frac{\partial k}{\partial t} + \boldsymbol{U}_j\frac{\partial k}{\partial x_j} = \boldsymbol{P}_k - \beta^* k\omega + \frac{\partial}{\partial x_j}\left[\left(\nu + \sigma_k\nu_T\right)\frac{\partial k}{\partial x_j}\right] \tag{14}$$

Specific Dissipation Rate

$$\frac{\partial \omega}{\partial t} + \boldsymbol{U}_j\frac{\partial k}{\partial x_j} = \boldsymbol{P}_k - \beta\omega^2 + \frac{\partial}{\partial x_j}\left[\left(\nu - \sigma_\omega\nu_T\right)\frac{\partial \omega}{\partial x_j}\right] + 2(1-F_1)\sigma_{\omega 2}\frac{1}{\omega}\frac{\partial k}{\partial x_i}\frac{\partial \omega}{\partial x_i} \tag{15}$$

Where a_1, β^*, σ_k and σ_ω are modelling constants, P_k is the volumetric production rate of k, F_1 is a blending function, k is the specific turbulent kinetic energy, S is an invariant measure of the strain rate and ω is the specific turbulent dispersion rate.

2.9 Joule Heating Implementation

Implementing the Joule heating element into the simulations was a trivial task of introducing a source term in accordance with the final term in Equation 5, the energy balance.

3 Results

The main goal of this study is assessing whether Joule heating makes a significant difference to simulations of alkaline electrolysers. To that end, it begins with the evaluation of whether Joule heating has significant effects on the flow, then covers energy and efficiency. A total of 18 simulations were conducted to generate the main results. The simulations covered a range of parameters including current density (1000, 2000, and 3000Am^{-2}) and temperature (300, 325, and 350K), with Joule heating turned on and off in each case. Jon denotes the anisothermal cases with Joule Heating implementation, and Joff denotes anisothermal cases with Joule Heating off.

There are three key variables which have the potential to vary when the Joule heating element in the simulation is switched on or off as pertaining to the flow. They are; the volume fractions of the gases, the liquid temperature and the liquid velocity. In the case of energy and efficiency, this study assesses the cell voltage as it changes with current density, along with the cell efficiency as industrially defined in Equation 16. Finally, the trade-off between activating Joule heating and having a more realistic model, with computational leverage is considered with independent simulations, conducted purely for timing.

3.1 Gas Volume Fractions

The volume fractions, α, of both Hydrogen and Oxygen were plotted across the horizontal plane of the electrolyser at 0.3m, the middle of the mesh, and 0.5m, towards the top of the mesh. Along both these planes, the volume fractions of the two gases showed negligible difference (<1%) for all simulations. Both Figures 2 and 3 below show trends at 325K and 2000$\mathrm{Am}^{-}2$. These parameters were chosen at random but are representative of the trends displayed at all values of current density and temperature.

As observed in Figure 2, there is very little separation between the displayed lines as the two mole fractions remain consistently similar. However, in the same graph for Hydrogen (see Figure 3), a significant separation grows to reach a maximum at the 0.015m point where there is a difference of 0.043 equating to a 99% deviation. Although both gas fractions follow a similar downward trend, the two cases vary significantly in how Hydrogen evolves along the plane as the curve with Joule heating is much steeper.

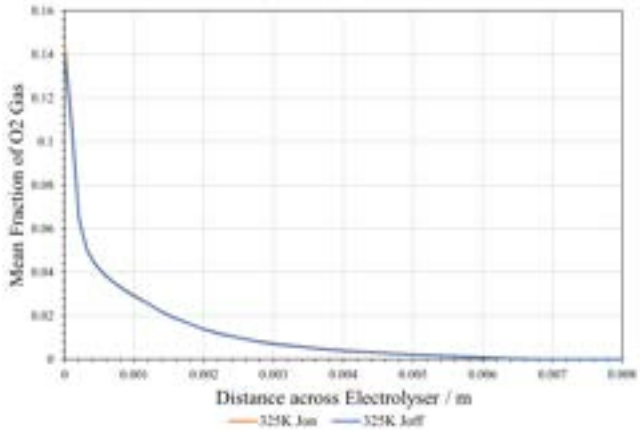


Figure 2: Graph of Oxygen volume fraction variation along the electrolyser Oxygen channel, at 2000Am^{-}2, y=0.5

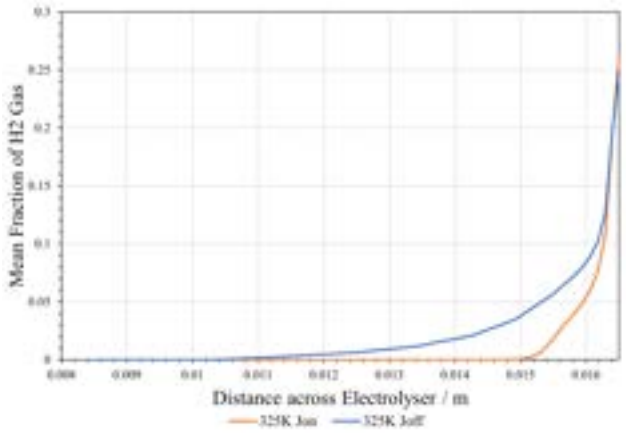


Figure 3: Graph of Oxygen volume fraction variation along the electrolyser Hydrogen channel, at 2000Am^{-}2, y=0.5

3.2 KOH Solution Temperature

While activating the Joule Heating element can cause variations in the H2 volume fraction, no such differences are observed in either the liquid temperature or velocity. It is observed that as current density increases, there are naturally greater changes in temperature - see Figure 4.

Among all the collected data, however, the maximum liquid temperature increase (at a current density of 3000Am^{-2}) was 3K, or less than 1% of the initial temperature. The Figures 4 and 5 show the case where the mean temperature difference between cases with and without Joule heating is greatest, yet is still insignificant.

3.3 KOH Solution Velocity Magnitude

A similar theme is observed for differences in the liquid velocities for lower current densities, eg 1000Am^{-2}. In Figure 6 there is no separation between the cases with and without Joule heating, across the Hydrogen channel

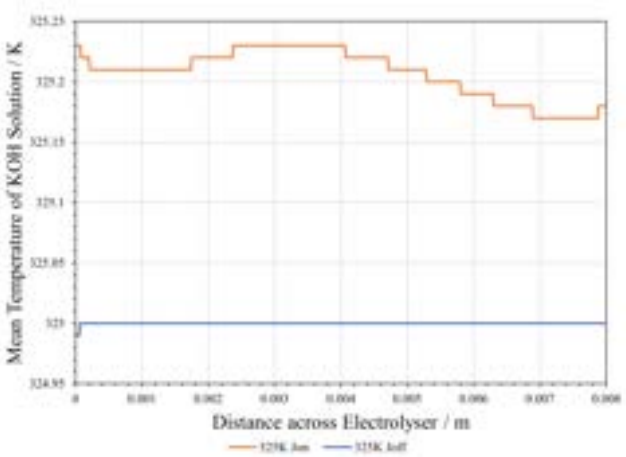


Figure 4: Graph of the KOH Solution temperature variations along the Oxygen channel of the electrolyser, at 3000Am^{-2}, y=0.5

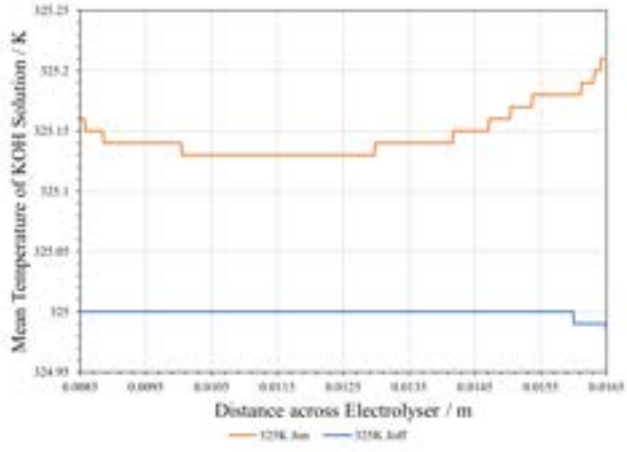


Figure 5: Graph of the KOH Solution temperature variations along the Hydrogen channel of the electrolyser, at 3000Am^{-2}, y=0.5

of the electrolyser (not plotted in this report, the same shape is observed in the Oxygen channel). Here, two lines are plotted but only one is visible.

This changes, however, when observing higher current densities, eg 3000Am^{-2}. As seen in Figure 7, there is a noticeable difference between the two mean velocity magnitudes when Joule heating is on and off. The peak of the graph without Joule heating is pushed closer to the cathodic wall of the eletrolyser where hydrogen is being produced. This peak is also higher, at 0.82ms^{-1}, while the case with Joule heating peaks at 0.8ms^{-1}.

3.4 Energy and Efficiency

Apart from the flow, Joule Heating is predicted to have an effect on the energy dynamics in the system. Cell voltage increases with increasing current density, but the trend is nonlinear and is affected by factors such as hydrogen and oxygen activation overpotentials, the concentration of the electrolyte solution, the size of

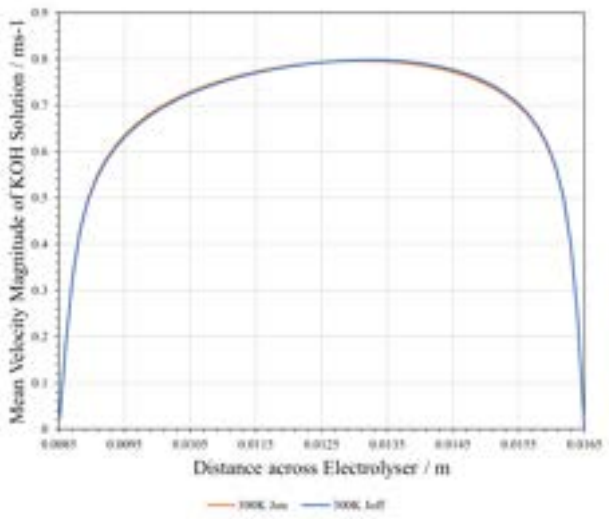


Figure 6: Graph of Water Velocity Magnitude along the electolyser Hydrogen channel at 1000Am^{-}2, y=0.5

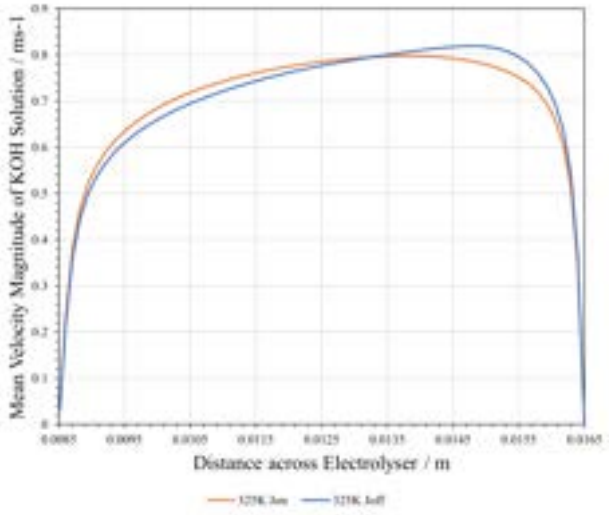


Figure 7: Graph of Water Velocity Magnitude along the electolyser Hydrogen channel at 3000Am^{-}2, y=0.5

the electrodes, and temperature of the solution. The relationship between all these factors is complex and not fully understood, but broadly, as the cell voltage increases, the hydrogen evolution reaction becomes more favorable and the hydrogen production rate at the cathode increases.

The cell efficiency is defined as follows in Equation 16:

$$Cell\ Efficiency = \frac{Standard\ Cell\ Potential}{Overall\ Cell\ Potential} \quad (16)$$

where the standard is set to 1.23V.

Equation 16 above was used to calculate efficiencies of all simulations and generate Table 2.

In addition to the data displayed in the Table above, the differences in cell efficiency are most visible in the

Table 2: Table of the absolute percentage difference in Cell Efficiency with and without Joule heating, averaged across all recorded temperatures.

Current Density / Am^{-2}	**Average Difference in Cell Efficiency / %**
1000	2.65
2000	6.60
3000	0.46

2000Am^{-2} region where the maximum change was 10% for the 325K cases.

As the current density increases, more electrons are available to drive the reaction that produces hydrogen gas, which leads to an increase in the rate of hydrogen production. However, it is important to note that there are limits to the amount of current that can be passed through an electrolysis cell without causing problems. At very high current densities, the rate of hydrogen production may begin to decrease due to a number of factors, such as increased Ohmic heating of the electrodes or an increase in activation energy required for Hydrogen and Oxygen production, thereby decreased efficiency of the electrolysis process, and result in the potential breakdown of the electrolyte solution.

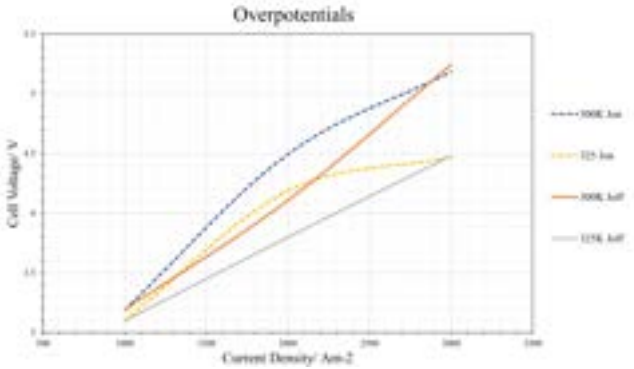


Figure 8: Graph of overpotentials, comparing the effect of Joule heating at different temperatures

Additionally, from Table 6 (see Appendix), it has been observed that the time required to run simulations with Joule heating versus without has an average difference of 2%. This suggests that the computational leverage required for the additional Joule heating source term requires is around 2% or less. At the very most, it is not significant.

4 Discussion

Firstly, the results of this study indicate that Joule heating has little effect on the peak volume fractions of gases in the simulated alkaline electrolyser. This

is consistent with the initial hypothesis that Joule heating would not significantly alter the bulk flow of gases in the system. However, the results also showed that Joule heating has the potential to have a more significant effect on the H2 volume fraction, specifically in how it evolves through the electrolyser. As seen in figure 3, there is a maximum difference of 99% at a current density of $2000\text{Am}^{-}2$ as the flow evolves through the channel.

One reason this could be the case is because the inclusion of Joule Heating affects the thermodynamics of the global reaction for the alkaline electrolyser: $H_2O(l) \rightarrow H_2(g) + \frac{1}{2}O_2(g)$. On a molar basis, Hydrogen is theoretically produced twice as much as oxygen, so if Joule Heating has indeed had an effect here, it would be more noticeable with Hydrogen than Oxygen, and that is exactly what is observed here in Figure 3 versus Figure 2.

The effects of Joule heating on velocity in the simulated electrolyser were largely unsurprising. There were some observed differences between cases with and without Joule heating, as seen in Figure 7, and these differences follow from the aforementioned considerations around Hydrogen volume fraction. As a result, at the cathode where hydrogen is produced, the difference in gas fractions causes a visible change in the shape of the bulk flow. In simulations that are at a greater scale than ours, this 5% difference in peak velocity magnitude and alteration of shape need to be well understood and accounted for in electrolyser optimisation.

When it comes to Figures 4 and 5, the temperature results are also, expectedly, insignificant. The nature of Joule Heating means that when it is included in the model, the temperature of the system will be different but only in proportion to how strong the electric field across the electrolsyer is. This is exactly what is observed in the figures. A current density of even 3000Am^{-2} does little to affect the temperature greatly.

The increase in cell voltage versus current density is known to be a non-linear relationship and in this case should have a decreasing gradient as described by the well known Butler–Volmer equation for cell voltage (Chen et al. 2017). The cases without Joule heating in Figure 8 do not exhibit this behaviour, in fact, the dependence seems to be almost linear. Hence, it is shown in Figure 8 that the cases with Joule heating more closely reflect reality. It is also clear from a graphical perspective that, again, the 2000Am^{-2} region shows the most variation in results.

In terms of execution times of various cases, it can't be said that Joule heating requires significantly more execution time or computational leverage. Table 3 (see Appendix) contains data from two cases, the first being at 1000Am^{-2} 300K where the least variance in results is expected, and 3000Am^{-2} 350K where the most variance in data is expected. This essentially covers the entire parameter space that this study has simulated, where all cases were run individually to obtain timing data. The 2% timing difference that is seen when using Joule heating is not significant enough to be attributed to Joule heating alone as there could have been random variance in the simulation physical parameter data.

5 Conclusion

In this study, the effects of Joule Heating on various physical properties in alkaline electrolysers were reported. These included the changes to: Hydrogen and Oxygen gas fractions, the bulk electrolyte velocity, the electrolyte temperature, the cell efficiency, cell potentials, and simulation execution times. While there were no significant changes to the temperature of the system, the evolution of Hydrogen gas along the horizontal plane of the electrolyser showed significant deviation. This potentially resulted in a 5% difference in peak velocity magnitudes between cases with and without Joule Heating, with the latter being displaced closer to the cathode. In terms of energy, the cell efficiency saw differences of up to 10% when applying the Joule Heating element, and the graph of overpotentials displayed that realistic behaviour was also better portrayed by cases with Joule heating when considering the Butler-Volmer equation.

Implementation of Joule Heating seems to only require a minimal amount of additional computational leverage at 2%, while it has been demonstrated that it can, on average, cause a difference in efficiency results of 6.60% for a range of temperatures at a current density of 2000Am^{-2} (see Table 2). This change is not insignificant. As a result, this study concludes that Joule Heating should be implemented in future alkaline electrolysis simulations that are anisothermal since the costs of implementation are negligible, while the resulting model will be much more physically realistic.

6 Outlook

Further research and consolidation is required to assess the results of this study. To begin with, one aspect that may require further simulation is the difference between cases in Figure 7. While it may not be directly the Joule heating that causes this change, it is attributed to the fact that the mean fraction of Hydrogen as seen in Figure 3, follows a steeper gradient with Joule heating. This brings up a key question. Given that the simulations were under galvanostatic operation, meaning that the production of hydrogen by volume should be the same for the same current density, how could there be such a large variation in results with and without Joule Heating for the 2000Am^{-2} case? The simulation-dependent as well as the physicochemical answer to this question should be investigated.

Future studies could also build on the findings here by investigating the effects of joule heating with a larger parameter space to examine greater current densities. Joule heating should have much greater effects at greater current densities but it is not well understood exactly how much effect, and when it might start to become detrimental to increase current density prior to the point of electrolyte breakdown. It would also be of importance to have more granular results in terms of temperature dependence.

7 Acknowledgements

We'd like to give a huge thanks to our primary helper Morgan Kerhouant for all the support with this study.

References

BP (2022). *Statistical Review of World Energy*. https://www.bp.com/en/global/corporate/energy-economics/statistical-review-of-world-energy.html.

Buist, KA, BJGH Backx, NG Deen, and JAM Kuipers (2017). "A combined experimental and simulation study of fluid-particle heat transfer in dense arrays of stationary particles". In: *Chemical Engineering Science* 169, pp. 310–320.

Burns, Alan D, Thomas Frank, Ian Hamill, Jun-Mei Shi, et al. (2004). "The Favre averaged drag model for turbulent dispersion in Eulerian multi-phase flows". In: *5th international conference on multiphase flow, ICMF*. Vol. 4. ICMF, pp. 1–17.

Chen, Yanan, Felipe Mojica, Guangfu Li, and Po-Ya Abel Chuang (2017). "Experimental study and analytical modeling of an alkaline water electrolysis cell". In: *International Journal of Energy Research* 41.14, pp. 2365–2373.

Chi, Jun and Hongmei Yu (2018). "Water electrolysis based on renewable energy for hydrogen production". In: *Chinese Journal of Catalysis* 39.3, pp. 390–394.

Gov.uk (2022). *Energy Trends: UK, April to June 2022*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1107502/Energy_Trends_September_2022.pdf.

Hreiz, Rainier, Lokmane Abdelouahed, Denis Fuenfschilling, and François Lapicque (2015). "Electrogenerated bubbles induced convection in narrow vertical cells: PIV measurements and Euler–Lagrange CFD simulation". In: *Chemical Engineering Science* 134, pp. 138–152.

iea.org (2022). *Global Hydrogen Review 2022*. https://www.iea.org/reports/global-hydrogen-review-2022.

Kakoulaki, Georgia, Ioannis Kougias, Nigel Taylor, Francesco Dolci, J Moya, and Arnulf Jäger-Waldau (2021). "Green hydrogen in Europe–A regional assessment: Substituting existing production with electrolysis powered by renewables". In: *Energy Conversion and Management* 228, p. 113649.

Kell, GS, GE McLaurin, and E Whalley (1989). "PVT properties of water-VII. Vapour densities of light and heavy water from 150 to 500° C". In: *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences* 425.1868, pp. 49–71.

MacDonald, Jennifer (2016). "Electric vehicles to be 35% of global new car sales by 2040". In: *Bloomberg New Energy Finance* 25, p. 4.

Menter, FLORIANR (1993). "Zonal two equation kw turbulence models for aerodynamic flows". In: *23rd fluid dynamics, plasmadynamics, and lasers conference*, p. 2906.

nist.org (1997). *Thermophysical Properties of Fluid Systems*. https://webbook.nist.gov/chemistry/fluid/.

Pei, Shuzhao, Chao Shen, Chenghu Zhang, Nanqi Ren, and Shijie You (2019). "Characterization of the interfacial joule heating effect in the electrochemical advanced oxidation process". In: *Environmental science & technology* 53.8, pp. 4406–4415.

Taqieddin, Amir, Roya Nazari, Ljiljana Rajic, and Akram Alshawabkeh (2017). "Physicochemical hydrodynamics of gas bubbles in two phase electrochemical systems". In: *Journal of The Electrochemical Society* 164.13, E448.

Tomiyama, A, GP Celata, S Hosokawa, and S Yoshida (2002). "Terminal velocity of single bubbles in surface tension force dominant regime". In: *International journal of multiphase flow* 28.9, pp. 1497–1519.

Wen, C Yu (1966). "Mechanics of fluidization". In: *Chem. Eng. Prog. Symp. Ser.* Vol. 62, pp. 100–111.

Zarghami, A, NG Deen, and AW Vreman (2020). "CFD modeling of multiphase flow in an alkaline water electrolyzer". In: *Chemical Engineering Science* 227, p. 115926.

Zaytsev, Ivan D and Georgiy G Aseyev (1992). *Properties of aqueous solutions of electrolytes*. CRC press.

Characterisation of Carbon By-products from Methane Pyrolysis in Molten Alkali Halide Salts

Tyn Suthanaruk and Korn Amnauypanit
Department of Chemical Engineering, Imperial College London, U.K.

Abstract

Methane pyrolysis in molten salts is a promising technology that can produce hydrogen with a low carbon footprint. The key challenges in methane pyrolysis are the high levelised cost of hydrogen compared to steam methane reforming and the inevitable salt loss due to the intercalation of salts in pyrolytic carbon. Therefore, the economics of this process should be enhanced through the commercialisation of carbon co-product. If methane pyrolysis in molten salts were to be widely adopted, a vast quantity of pyrolytic carbon would be generated. This makes commercialising the carbon co-product for the battery industry of particular interest as it has the demand to meet the influx in carbon supply and exploit the properties of carbon. This study explores the impact of molten alkali halide salts (LiBr-NaBr and LiCl-NaCl) on carbon morphology. Experiments were designed with the intention of lithium intercalation into the carbon co-product. This would mimic the effects of pre-lithiation and allow for the potential utilisation of pyrolytic carbon as lithium-ion battery anodes. TGA, TEM, ICP-MS, and XRD analyses revealed that lithium was intercalated into the carbon complex. All carbon samples consist of a mixture of graphitic and amorphous structures, with the carbon purity ranging between 86-88 wt%. The carbon produced with LiBr-NaBr as the salt medium has a higher degree of graphitisation, intercalated lithium-to-sodium ratio (2:1), and lithium intercalation. Thus, the carbon produced with molten LiBr-NaBr salt is a promising candidate as an anode material for lithium-ion batteries. Moving forward, testing pyrolytic carbon as an anode material for batteries would be beneficial.

Keywords: Hydrogen production, Methane pyrolysis, Molten salt, Pyrolytic carbon, Anode material, Lithium intercalation

1. Introduction and Background

The greenhouse effect is one of the most pressing environmental problems scientists must address today. This problem is caused by burning fossil fuels and the corresponding emissions of carbon dioxide due to the rise in global energy demand. New industrial-scale energy production with reduced carbon emissions is required (Acheampong, 2018). Hydrogen is a feasible alternative to fossil fuel since it has a high energy density and emits no carbon dioxide when burned. In addition, the chemical industry makes substantial use of it as a raw material.

The demand for hydrogen in 2021 is 94 Mt, and it is expected to rise to 180 Mt in 2030 (IEA, 2022). This raises a concerning issue as the vast majority of hydrogen production comes from processes that release a large amount of CO_2 emissions. 76% of the hydrogen is produced from natural gas through steam methane reforming (SMR), 22% by coal gasification, and only 2% from electrolysis (E.R. Ochu et al., 2021). Two of the three processes have a large carbon footprint and account for 98% of global hydrogen production. This fact illustrates the need for an alternative method of producing hydrogen that is cost-effective and emits minimal greenhouse gases.

Methane pyrolysis is a technology with the potential to solve this problem. Methane pyrolysis yields two moles of hydrogen per mole of methane (Eq. 1), while steam methane reforming yields four moles of hydrogen per mole of methane. Despite this drawback, methane pyrolysis has a 27% lower enthalpy of a reaction than SMR to produce one mole of hydrogen with no carbon dioxide emission (C.F. Patzschke et al., 2021). Furthermore, as global warming becomes an alarming issue and carbon taxation is enforced, methane pyrolysis may emerge as a viable alternative to the current hydrogen production technologies.

$$CH_4 \rightarrow 2H_2 + C_{(s)} \qquad (1)$$
$$\Delta H^0 = 37.4\ kJ(mol\ H_2)^{-1}$$

In economics, methane pyrolysis is a competitive process to achieve net zero emissions. In an article published in 2017, the levelised cost of hydrogen (LCOH) produced through methane pyrolysis with molten salt was calculated to be 1.76 \$ $kg^{-1}(H_2)$ (Parkinson et al., 2019b). In contrast, steam methane reforming is significantly more economical, with an LCOH of 1.26\$ $kg^{-1}(H_2)$. However, this does not account for the emissions produced by the process. To reduce CO_2 emissions, SMR should be integrated with carbon capture and storage (CCS) technology in carbon-constrained scenarios. Consequently, the LCOH for SMR with CSS would rise to 1.88\$ $kg^{-1}(H_2)$ for a 90% carbon capture rate. Therefore, the cost of manufacturing hydrogen by methane pyrolysis in a molten salt medium is comparable to the cost of producing hydrogen through SMR with CCS.

In comparison, the LCOH favours pyrolytic hydrogen more than the production of hydrogen from renewable energy, as wind energy and solar energy have LCOHs of 5.24\$ $kg^{-1}(H_2)$ and 8.87\$ $kg^{-1}(H_2)$, respectively (Parkinson et al., 2019b). Furthermore, the economy of methane pyrolysis is enhanced if the carbon produced can be commercialised (C.F. Patzschke et al., 2021). However, the problem with methane pyrolysis in molten salt is that it has yet to be at a commercial scale. Methane pyrolysis only has a technological readiness level (TRL) of 3-5, whereas SMR and SMR with CCS are 9 and 7-8,

respectively (Parkinson et al., 2019b). This demonstrates that methane pyrolysis should be improved in order to compete against current hydrogen production.

Methane pyrolysis involves the thermal decomposition of methane into its constituent (hydrogen and solid carbon). Since the 1930s, gas-phase methane pyrolysis has been used to produce carbon black (M. Voll et al., 2010). However, this process was not used to commercially produce hydrogen as other hydrogen production methods were more energy efficient, as well as the problems with the accumulation of pyrolytic carbon on the catalyst's surface, resulting in catalytic deactivation (Schneider et al., 2020). Rapid deactivation will affect process operations as catalysts would have to be replaced or regenerated, thus incurring significant downtime. Additionally, this process is uneconomical due to its low H_2 selectivity, difficulties in carbon extraction, high operating temperature (1200 °C), and high energy requirements (Holmen, Olsvik & Rokstad, 1995).

A bubble column reactor system with molten salt as the reaction medium operating at 1000 °C was developed to overcome the issues associated with methane gas pyrolysis. Since each bubble has a distinct interphase, catalyst deactivation is unlikely to occur, resulting in a more efficient catalytic reaction. In addition, methane pyrolysis with molten salt enhances heat transfer since gas-phase methane bubbles are in direct contact with molten salts (Holmen, Olsvik & Rokstad, 1995). The proposed molten salts are alkali halides that are inexpensive and environmentally safe.

To efficiently utilise carbon produced through the pyrolysis of methane, it is necessary to comprehend the main properties of carbon. Kang et al. examined the carbon morphology produced via methane pyrolysis with molten KCl-$MnCl_2$ as a medium. The results suggested that the carbon sample exhibited a low-ordered structure characterised as amorphous carbon with some graphene layers (D. Kang et al., 2019). In addition, the inevitable intercalation of salts in carbon samples resulted in the loss of expensive salts, which hindered the economics of the process (D. Kang et al., 2019). This promotes the utilisation of inexpensive salt as a reaction medium and takes advantage of the intercalated salt in carbon.

Parkinson et al. also investigated the characteristics of carbon generated from methane pyrolysis using molten alkali halide salts as a medium (B. Parkinson et al., 2021). According to the results, carbon samples contained a mixture of amorphous carbon and graphite with varying degrees of structural order. In terms of carbon properties, the performance of inexpensive salts (NaCl, NaBr, KCl and KBr) as reaction media is promising (B. Parkinson et al., 2021). Further research is necessary to establish the effects of different combinations of eutectic salt medium (LiCl, LiBr, NaBr, and NaCl) on the characteristics of pyrolytic carbon and how they might be tailored to certain carbon markets.

Due to the high global demand for hydrogen, an abundant carbon supply would be created if hydrogen was made solely from methane pyrolysis (Muradov, 2017). This implies an increase in carbon supply, making industries that consume significant amounts of carbon, such as the battery, cement, polymer, and agriculture industries, are of interest.

The battery industry, one of the most promising industries due to its vital role in decarbonising the world's energy and transportation sectors, has great potential to meet the carbon supply from methane pyrolysis. Studies have indicated that carbon produced from methane pyrolysis with molten salts has some graphitic and amorphous structure (B. Parkinson, 2020). Although this would make pyrolytic carbon produced from the process less competitive than graphite used in battery production, this disadvantage can be compensated by the intercalation of metal ions inside the salt. The interaction of metal ions could mimic the effect of pre-lithiation. Pre-lithiation is the redox reaction between high-reducing lithium compounds and the anode material, resulting in a higher lithium content in the anode material (Huang et al., 2022). Past studies have shown that pre-lithiation benefits the performance of the battery's anode. Batteries with anodes that went through pre-lithiation would have extra lithium ions in them, allowing the battery to compensate for the lithium loss from its first charge and long-term cycling (Yue et al., 2022a). Therefore, the carbon co-products with intercalated lithium from methane pyrolysis make promising anodes.

Carbon is commonly used as a battery electrode, with graphite being the ideal candidate. Carbon electrodes are distinguished by their electrical resistance and conductivity, which derive from their structure. Graphite is the most widely used anode material for lithium-ion batteries. It has excellent electrical conductivity, high crystallinity, and layered structure, making ions able to be reversibly intercalated and exfoliated. In addition to reversible ions intercalation, graphite allows lithium-ion batteries to have a specific capacity of 370 $mAhg^{-1}$ and an operating efficiency above 90% (Hou et al., 2017).

Furthermore, due to its excellent electrical conductivity, graphite can be used as an anode material for other alkaline metal ion batteries like potassium rubidium and caesium. This makes graphite an ideal electrode for making a battery. Aside from graphite, research using amorphous carbon has also provided promising results (Dresselhaus, 2002).

Amorphous carbon includes two carbon materials (soft and hard carbon). Instead of having an ordered structure like graphite, amorphous carbon comprises voids, distorted graphene sheets, and randomly distributed

graphitised microdomains. The interlayer spacing for d002 is 3.4-3.6 Å and 3.7 Å for soft and hard carbons, respectively, which is larger than that of graphite (3.35 Å). These disordered characteristics allow amorphous carbon to succeed where graphite fails, which is an anode for sodium-ion batteries. Due to its porous nature and greater interlayer spacing, amorphous carbon can accommodate large sodium ions, shorten its' diffusion path, and handle significant volume changes during the charge and discharge of sodium ion batteries. Thus, amorphous carbon can be a practical alternative to graphite as an electrode if sodium-ions batteries were needed to be produced.

This study aims to characterise pyrolytic carbon produced from methane pyrolysis in molten alkali halide salts, emphasising lithium salt. Lithium salts are of interest because lithium intercalation could enhance the characteristics of pyrolytic carbon, potentially turning it into battery electrodes due to lithium cations contamination.

2. Materials and Methods

2.1 Molten salt bubble column reactor

Material	Vendor	Description
Salts[a]	Alpha Aesar	99.9%, anhydrous
Catalyst	ExxonMobil (EMRE)	Co supported on Al_2O_3
CH_4 ,Ar	BOC Ltd.UK	99.9%
H_2	H_2 Generator	99.9%

Table 1: Chemical used for methane pyrolysis
[a] Salts refer to LiBr, NaBr, LiCl and NaCl

The eutectic mixtures were LiBr-NaBr (72:28 mol%) and LiCl-NaCl (71:29 mol%). The weight of the salts was calculated based on the salt densities, eutectic ratios, and reactor dimensions at 1000 ºC. Then, the salt mixture was combined and heated in a 200 °C oven overnight to remove excessive moisture. A third of the salt was to be evenly distributed at the top and bottom of the quartz tube reactor (250 mm x 20 mm OD / 16 mm ID Quartz tube rounded closed-end). The remainder of the dried salt was combined with the catalyst (2.5 wt% of the salt and catalyst mixture) and filled the middle section of the reactor. This was done to ensure that no catalyst particles would get stuck at the bottom of the reactor, as the catalyst is denser than the salt. The reactor head (borosilicate with two ports of 0.25" and 0.5") was sealed and secured to the reactor using silicone grease and stainless steel clamps. To provide CH_4 and Ar gas to the bottom of the reactor, a two-inlet arrangement comprising a headspace tube and an alumina injector (500 mm in length with 4 mm OD and 3 mm ID holes) was mounted to the top of the reactor head. The outlet at the top was connected to a mass spectrometer (MS) and the vent. All equipment was inspected for leaks with a bubble trap meniscus test.

To initiate the experiment, a 30 mL min^{-1} Ar flow was supplied in the reactor. The furnace was programmed to heat the reactor to 1000 °C at a rate of 5 °C minute-1. When the salt mixture had melted, the alumina injector was lowered into the molten salt until it reached 10 mm above the reactor's bottom. The inlet was switched to 30 mL min^{-1} of Ar and 5 vol% of H2 to initiate the catalyst's reduction. The catalyst reduction typically lasts 1 hour, after which the hydrogen flow rate is switched off. This duration is considered enough since the MS readings trace no water after that period. At 1000 °C, the reaction was started by switching the Ar flow (30 mL min^{-1}) to the headspace and methane flow (15 mL min^{-1}) to the bottom of the reactor. After the experiment, the furnace was set to 700 °C, and the alumina injector was lifted above the molten salt surface. While the Ar flow was on, the furnace was shut off and cooled to room temperature. The reactor was removed at room temperature for carbon collection.

2.2 Carbon recovery

Carbon was collected after the pyrolysis experiment by scraping deposits from the reactor's wall, dissolving, and sonicating the salt mixture with deionised water. A Büchner funnel was used extensively to wash and filter any salts on the carbon surface with deionised water due to the high solubility of salt in water. This procedure was repeated until the resistance of the filtrate remained constant. Then, the carbon residue on the filter paper was dried overnight and stored for further analysis.

2.3 Carbon Characterisation

A litesizer 500 was employed to measure the particle size of the carbon samples. A small amount of each carbon sample was dispersed in a tube with deionised water. The mixture was subjected to 6 minutes sonication process to achieve a more precise result. The particle size analysis will be performed on each sample until three measurements with a peak of 100% intensity are obtained. The measurement with the lowest standard deviation will then be selected as the final particle size.

Transmission electron microscopes (TEM) are microscopes that use an electron beam to provide a highly magnified image of a sample. The TEM images were obtained using JEOL 2100F machine and utilised to illustrate carbon's crystalline and amorphous structures derived from various salts. The sample was prepared by dispersing 5 mg of carbon in 2 ml of isopropanol solution. Subsequently, the material was sonicated for 5 minutes to achieve uniform dispersion. One drop of this sample was placed on a 300 carbon-copper mesh. The sample was inserted into the equipment, and measurements were taken at multiple magnification levels ranging from 0.5 μm to 20 nm.

Analysing the crystalline structure of carbon was done with the help of X-ray diffraction (XRD). The diffraction data were obtained using an X'Pert Pro (PANanalytical) with Cu-Kα radiation with λ = 0.1541nm. Measurements were done at 40 kV and 40 mA. The °2θ range was 5-90, and a step size of 0.0334 was used. Further data processing was done using Highscore software to analyse the data better.

Raman spectroscopy was conducted using a Senterra II (Bruker) with the parameters as follows; 20x magnification, a spectral range of 50 to 4260 cm^{-1} and 532 nm laser excitation. An average spectrum of over 5 samples was measured. Baseline correction was applied to the acquired spectra in the OriginPro 2023 software. The intensity and the location of peaks identified carbon structures.

The nature of the carbon and the mass fraction of the salt were examined by temperature-programmed oxidation (TPO) in a thermo-gravimetric analyser (TGA) STA 449 F5. The crucible used in the TG experiments was loaded with at least 5 mg of carbon sample before being heated under air at a temperature of 10 °C $minute^{-1}$ and a flow rate of 40 mL $minute^{-1}$ at SATP. To study the combustion temperature, carbon black and graphite were employed as references for the structures of amorphous and graphitic carbon, respectively. Following TGA analysis, the residue of a sample was collected to be further analysed for salt intercalation and carbon purity. This was accomplished through inductively coupled plasma mass spectrometry (ICP-MS).

Inductively coupled plasma mass spectrometry (ICP-MS) is an elemental analysis technique that can identify most elements on the periodic table at a low concentration. ICP-MS would be used because the residue at the end of TGA was minimal (around 1 mg). ICP-MS was utilised to detect Li^+ and Na^+ concentrations, as most of the residue should be salt. Scandium was chosen as a low-mass internal standard to evaluate system performance since it is a relatively unimportant material for analysis and is unlikely to be present in most sample types. The machine was then calibrated with standard lithium and sodium solutions at 5, 50, 100, 200, and 500 ppb concentrations. For accurate results within the calibrated range, each TGA residue was dissolved in 2% nitric acid to obtain 6 mL containing less than 500 ppb of each cation. The ratio of cations corresponds to the proportion of salt intercalated in the carbon.

3. Results and Discussion

3.1 Carbon Production and Recovery

To investigate carbon produced from methane pyrolysis, the experiment was designed to separate carbon based on its densities. The pyrolytic carbon was expected to rise to the top of the reactor, while salts and catalysts remained below. From this, carbon can be collected, as detailed in Section 2.2. In addition to carbon recovery, an eutectic mixture of salt was used to account for operational aspects. Performing methane pyrolysis with a eutectic salt mixture facilitates experiment start up and shut down by lowering the melting point of salts. Furthermore, operating with a eutectic salt mixture allows for a wider operating temperature range, which benefits industrial applications.

Upon experimenting, it was found that the carbon produced did not float to the top of the reactor, but it was either mixed with other components in the reactor or sank to the bottom of the reactor. This finding is problematic as it makes operations impractical, especially with a catalyst. Separating carbon from salt can be done by simply dissolving all the salt in water and filtering as detailed in the methods. However, separating the catalyst from carbon was not able to be done. Therefore, analysis of carbon produced with catalyst was done with catalyst contaminated in the recovered carbon.

3.2 Carbon morphology

The particle size should be evaluated to determine the potential presence of agglomerates, which may adversely affect the electrode capacity (Röder et al., 2016). The particle size of carbon samples ranges from 658 to 806 nm. The effect of sonication was only analysed on LiBr-NaBr. The results suggested that particle agglomerations were broken up by sonication, resulting in more precise results. Therefore, the remaining samples of carbon were sonicated. According to Table 2, carbon produced from LiBr-NaBr has a larger particle size than carbon produced from LiCl-NaCl. The obtained particle size was relatively small compared to the reference graphite.

Carbon Samples	Particle Size (nm)
LiBr-NaBr (No sonication)	806
LiBr-NaBr (Sonicated)	697
LiCl-NaCl (Sonicated)	658
Reference Graphite	1835

Table 2: Particle size of the carbon samples

Due to constraints, the carbon samples were only obtained from six hours of operation. The particle size of pyrolytic carbon could grow as operating time increases. As the reaction progresses, more carbon will be deposited towards the top of the reactor column due to the lower density of carbon. The deposited carbon could act as an active site which improves the rate of carbon formation (B. Parkinson, 2020).

As shown in Fig. 1, the TEM images and fast-Fourier transform (FFT) were obtained at different magnifications, ranging from 0.5 μm to 20 nm. According to Fig. 1a and 1b, the particle size of carbon samples varies from

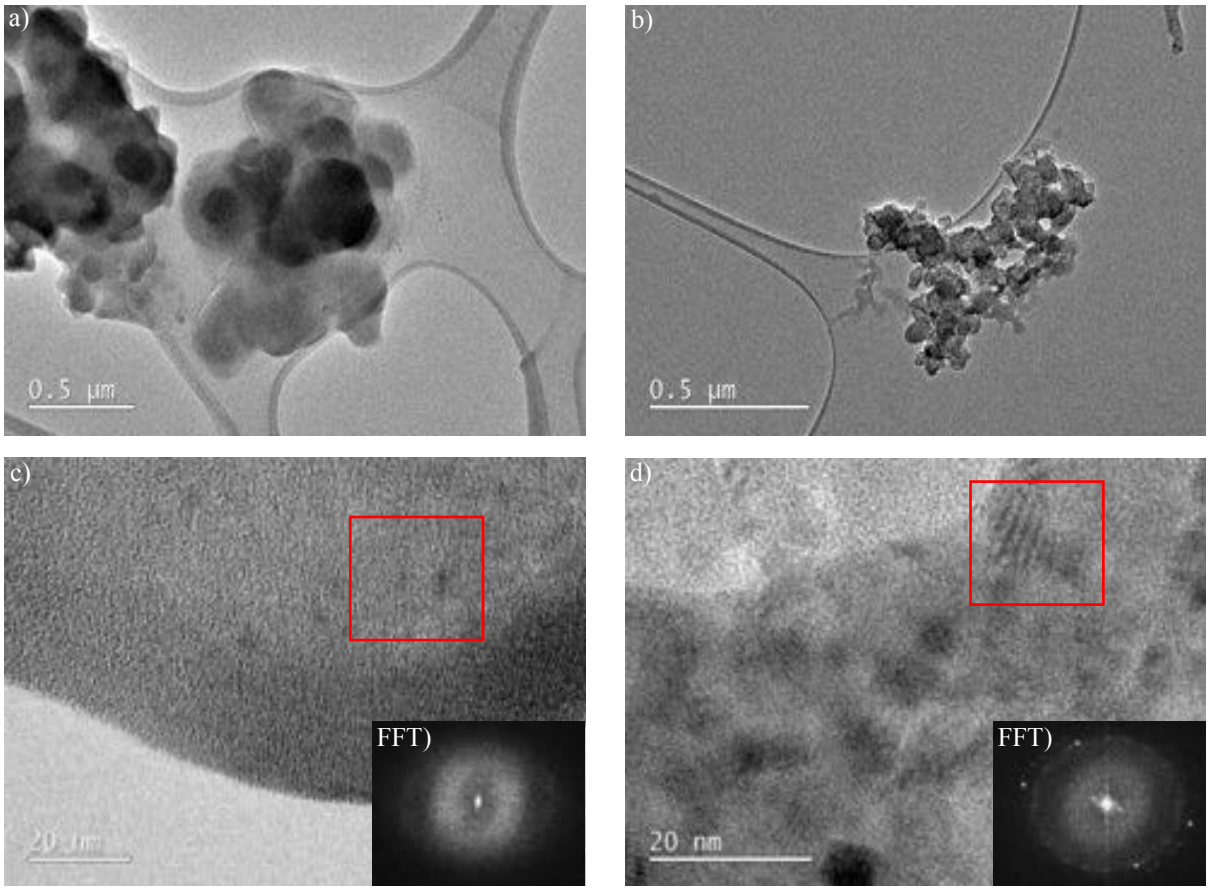


Figure 1: TEM of carbon particles retrieved from methane pyrolysis in molten salts. a) Magnification at 0.5 µm of carbon sample from LiCl-NaCl b) Magnification at 0.5 µm of carbon sample from LiBr-NaBr c) Magnification at 20 nm of carbon sample from LiCl-NaCl with FFT of the scattered electron pattern d) Magnification at 20 nm of carbon sample from LiBr-NaBr with FFT of the scattered electron pattern. The red regions indicate the pattern and where the FFT was captured.

approximately 0.5 µm to 0.8 µm which aligned with the results from the litesizer.

The TEM images of carbon samples reveal amorphous and graphitic structures, with certain locations exhibiting a high level of crystallinity. Lamella and lattice patterns were more common in the carbon sample obtained from LiBr-NaBr than LiCl-NaCl, as shown in Fig. 2c and 2d. In addition, bright diffraction spots were observed in the diffraction patterns generated from FFT from the LiBr-NaBr sample. This suggests that carbon generated from LiBr-NaBr has a higher degree of crystalline structure, resulting in higher conductivity for the battery electrode. This was further validated by the XRD and Raman analysis.

XRD analysis was done on all the carbon samples produced from the methane pyrolysis process and displayed on Fig. 2. As shown in Fig. 2, XRD analysis of the carbon sample with cobalt catalyst shows that the sample is filled with Al_2O_3 peaks. This indicates that the carbon analysed is contaminated with the catalyst. Due to this contamination, the analysis made on this sample cannot be conclusive. The presence of carbon is represented at ~26 °2θ for every sample with variation in peak size. Analysis with the HighScore software suggests that carbon produced with LiBr-NaBr as a salt medium is in the form of fullerene bromide. However, due to the software's low score for this result and the highly complex structure of fullerene bromide, it is unlikely that this is presented in the process. In addition to the low score, the peaks that correspond to carbon for all samples are broad. Thus, indicating that the carbon structure does not comprise a significant number of structural defects and a crystallite size smaller than a micrometre. (Ungár, 2004). In contrast, TEM analysis revealed that the particle size is lower than one micron, and carbon defects exist. Pyrolytic carbon is most likely to consist of amorphous and graphitic carbon.

Unlike carbon produced without a catalyst, the carbon produced with a catalyst had a much narrower peak at ~26 °2θ. This can be an indication that the catalyst could promote graphitic structure. However, due to the catalyst contamination, a conclusion cannot be drawn from this. According to the literature and the HighScore software, the peak at ~24 °2θ corresponds to lithium carbide (Missyul et al., 2017). This peak was presented in carbon

produced with LiCl-NaCl and LiBr-NaBr salt medium, with the latter having a much higher peak. This suggests that having a LiBr-NaBr salt medium promotes much more lithium intercalation than a LiCl-NaCl salt medium.

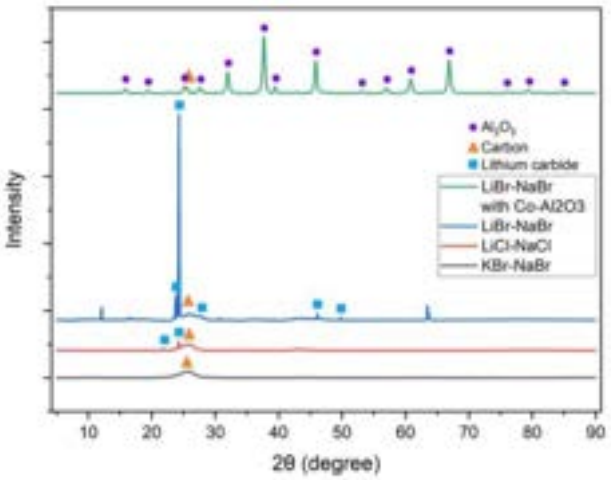


Figure 2: XRD analysis of carbon samples (the NaBr-KBr experimental data was given by the previous experiment)

As shown in Fig. 3, performing Raman analysis on the collected carbon samples reveals the presence of D and G bands (1350 cm^{-1} and 1580 cm^{-1}, respectively). D bands correspond to the structural defects of the carbon sample, which can suggest the existence of amorphous carbon in the sample. Meanwhile, G bands illustrate the existence of graphitic sp^2 structure in the sample (Pimenta et al., 2007). The ratio of D and G band peak intensity, I_D/I_G, indicates the degree of graphitisation of the sample, with I_D being the intensity of the D band and I_G being the intensity of the G band (Ferrari & Rokstad, 1995). The ideal graphitic structure will have no defects and all carbon atoms existing only in the sp2 configuration, thus making I_D/I_G = 0. As the ideal graphite experience amorphisation, its I_D/I_G will increase up to 2.0 as it gains more structural defects. At this point, the ideal graphite will be converted to nanocrystalline (NC) graphite. However, carbon atoms are still in sp^2 configurations. After further amorphisation of carbon, the I_D/I_G will keep decreasing until it reaches a ratio of 0.2, indicating that the NC graphite has turned into amorphous carbon where ~20% of atoms exist in sp^3 configuration (Ferrari & Rokstad, 1995). Another indication of carbon's crystalline structure is the G band's peak position. During amorphisation, the G band shifts from 1580 cm^{-1} (graphite) to 1600 cm^{-1} (NC graphite). Upon further amorphisation into amorphous carbon, the G band shifts to 1510 cm^{-1}. Aside from the graphitic structure, Raman analysis provides details regarding graphene layers in carbon. A second-order two-photon process creates the 2D band. The shape of this band and the I_{2D}/I_G, where I_{2D} is the intensity of the 2D band, can give details of the graphene layer. Regarding the shape of the 2D peak, a narrow peak indicates the presence of single-layer graphene. This peak results from four overlapping peaks, indicating a bilayer. Furthermore, a broadened band indicates the presence of a multilayer (Ferrari & Rokstad, 1995). As for I_{2D}/I_G, a ratio of 2-3 corresponds to a monolayer, 2> I_{2D}/I_G >1 corresponds to bilayer graphene, and an I_{2D}/I_G <1 corresponds to multilayer graphene (Van Tu Nguyen et al., 2013).

All carbon samples have an I_D/I_G between 0.85-1.12. This corresponds to either a mixture of graphite and NC graphite or NC graphite and amorphous carbon. Considering this with G bands location, it suggests that carbon produced from NaBr-LiBr salt medium and catalyst and KBr-NaBr salt medium is a mixture of NC carbon and amorphous carbon as their G bands are below 1580 cm^{-1}. However, analysis for NaBr-LiBr salt medium cannot be conclusive due to the catalyst contamination, as mentioned earlier. Analysis for experiments with NaCl-LiCl and NaBr-LiBr salt medium without catalyst is still inconclusive as their G band locations are 1586 and 1584 cm^{-1}, respectively. This could suggest that it could be a mixture of graphite and NC graphite or NC carbon and amorphous carbon. When considering this analysis in combination with TGA analysis, it could be concluded that the composition of the carbon mixture produced from NaCl-LiCl salt medium and NaBr-LiBr salt medium without catalyst are both a mixture of NC graphite and amorphous carbon.

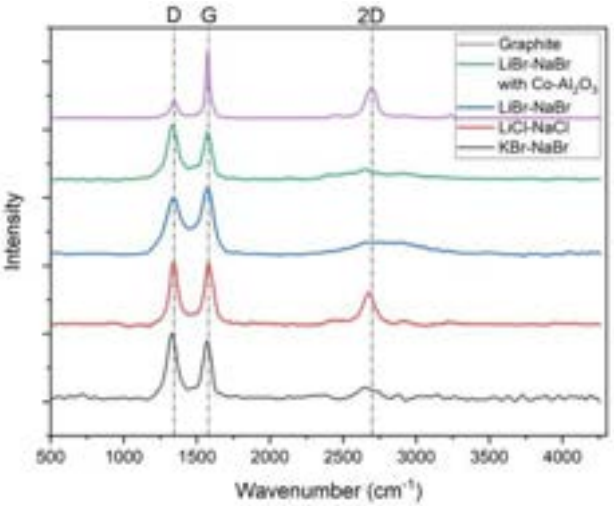


Figure 3: Raman analysis of carbon samples (the NaBr-KBr experimental data was given by the previous experiment)

Regarding the 2D band at 2700 cm^{-1}, Raman analysis on all carbon samples except for carbon produced with LiCl-NaCl salt medium have broad peaks. Thus suggesting that the carbon produced has the presence of multilayer graphene. On the other hand, carbon produced with LiCl-NaCl salt medium has a narrow peak, which indicates the presence of monolayer graphene at its 2D band. However, the relative intensity of I_{2D}/I_G is 0.55, making it inconclusive that carbon produced with LiCl-NaCl salt medium is single-layer graphene (Van Tu Nguyen et al., 2013). This narrow peak could be explained due to the carbon having a particular orientation during

Raman analysis. Furthermore, as shown in Fig.3, the I_{2D}/I_G ratio and the shape of the 2D peak of carbon produced from LiCl-NaCl salt medium have a similar shape and size to the graphite, concluding that monolayer graphene is not present in carbon produced from LiCl-NaCl salt medium. This explanation falls in line with the analysis made with TEM and TGA as both analysis suggest this carbon produced has less crystalline structure than carbon produced with LiBr-NaBr salt medium.

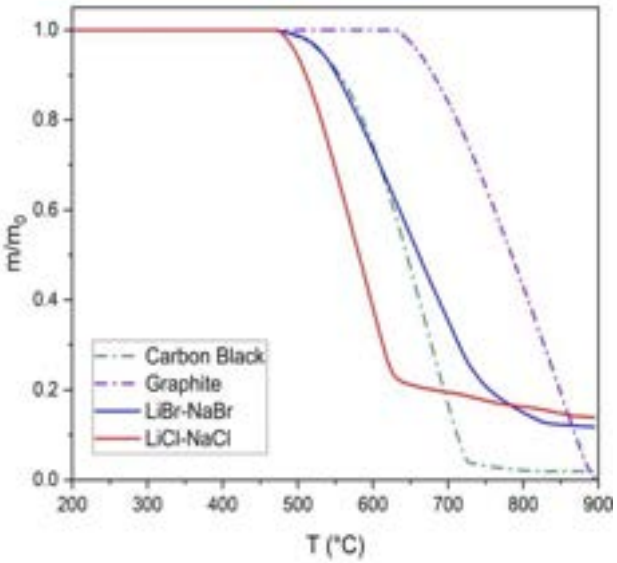


Figure 4: The TGA plot shows the remaining weight during combustion of all carbon samples and references with the adjusted contamination[b]. [a] The y-axis where m is the current mass and m0 is the initial mass). [b] The weight of the ICP-MS residue remaining after dissolving all of the salt in 2 % nitric acid was subtracted from the initial weight.

During the ICP-MS sample preparation, residue remained after dissolving all intercalated salt in 2% nitric acid. This residue was then cleaned with deionised water and completely dried in the oven at 200 °C. To account for this contamination, the residue was weighted and deducted from each TGA carbon sample's initial weight. The residue was identified as SiO_2 using XPS. This contamination results from the degradation of the internal walls of the quartz reactor.

The TGA plot in Fig. 4 illustrates that reference carbon black and graphite started to combust around 500 °C and 700 °C, respectively. Hence, carbon black has a mixture of amorphous and crystalline structures. These results for reference carbon were consistent with the study that suggested the combustion of amorphous carbon begins around 400 °C, while graphitic carbon starts at 700 °C (Devrim & Albostan, 2016). Fig. 4 shows curves with one distinct weight loss starting around 500 °C in all carbon samples without a catalyst. TGA plots will often reveal two distinct weight losses due to the combustion of amorphous and graphitic carbon. At a high combustion rate of 10 °C $minute^{-1}$, there might not be sufficient time for all amorphous carbon to burn off and reach a plateau before the graphitic carbon starts to combust (Devrim & Albostan, 2016).

Fig. 4 indicates that both carbon samples without catalyst began to combust approximately at the same temperature as the carbon black reference but at a varying rate corresponding to different degrees of order structure. The carbon produced by the LiBr-NaBr salt has a less steep slope than reference carbon black and carbon made by the LiCl-NaCl. This may imply that the carbon derived from LiBr-NaBr salt has a higher crystalline structure, resulting in a higher melting point and higher electrode conductivity (Devrim & Albostan, 2016). This result is consistent with the XRD and ICP-MS results, which showed that carbon from LiBr-NaBr salt contained a high peak at ~24 °2θ, indicating intercalated lithium graphite or lithium carbide. Furthermore, carbon formed by the LiCl-NaCl salt could be more amorphous, resulting in a faster combustion rate than carbon black. All of the carbon samples started to burn before the reference graphite, indicating that none of the carbon samples was as crystalline as the graphite reference.

Parkinson et al. concluded that there was a strong correlation ($R^2 = 0.93$) between the internuclear spacing of molten salts (NaCl, NaBr, KCl and KBr) to the amount of intercalated salt in the carbon (B. Parkinson et al., 2021). NaCl has the smallest internuclear separation of the examined samples (B. Parkinson et al., 2021). Therefore, the carbon derived from the NaCl had the lowest intercalated salt and, thus, the highest carbon purity of 90.1 wt% (B. Parkinson et al., 2021).

Carbon samples	Carbon Purity (wt%)[a]
LiBr-NaBr	88 ±5%[b]
LiCl-NaCl	86 ±5%

Table 3: Weight percentage of the carbon samples after TGA
[a] The carbon purity was calculated by subtracting 100% by the total weight remaining (%) at the end of TGA. [b] The uncertainty of ± 5% caused by the calibration process in the TGA equipment.

At the end of the TGA, the carbon purity was determined by the total weight percentage loss during the combustion. The carbon purity produced by LiBr-NaBr and LiCl-NaCl were 88 wt% and 86 wt%, respectively. However, the latter results were expected to be higher than the carbon purity of carbon produced from pure NaCl (90.1 wt%) due to the smaller internuclear spacing of Li-Cl (B. Parkinson et al., 2021). In addition, the carbon purity derived from the larger internuclear spacing salt (LiBr-NaBr) was higher than that from, the smaller internuclear spacing salt (LiCl-NaCl). This does not align with the relationship proposed by Parkinson et al. The deviation from the expected trend could have arisen from the uncertainty due to the calibration of the TGA instrument. After validating the reference carbon in TGA and discussed with the lab technician, the uncertainty was concluded to be ± 5%. In addition, the relationship between internuclear distance and carbon purity may not

hold at LiCl-NaCl spacing. ICP-MS further investigated the salt residues to determine the ratio of intercalated salt in the TGA residue.

As demonstrated in Table 4, all cation values fell within the calibrated range of less than 500 ppb. The ratio of lithium-ion to sodium ion (Li:Na) reveals the proportion of intercalated salt in carbon samples. The Li: Na ratio of carbon generated from LiBr-NaBr was around 2:1. This was favourable, as lithium-ion batteries would benefit from a higher lithium content (Shellikeri et al., 2017). However, the Li: Na ratio of carbon generated from LiCl-NaCl was approximately 1:26. Furthermore, the high lithium intercalated ratio was consistent with the XRD data, which showed a high peak at ~24 °2θ that might be intercalated lithium graphite or lithium carbide.

Carbon samples	Li^+ (ppb)	Na^+ (ppb)
LiBr-NaBr	434.0	206.0
LiCl-NaCl	16.8	436.0

Table 4: Concentrations[a] of Li^+ and Na^+ in carbon samples
[a] This concentration does not represent the total concentration of salt in the whole carbon sample but it can be used to approximate the ratio of the intercalated salt in the carbon sample

4. Conclusions

With a demand of 94 Mt of hydrogen in 2021 and an expected rise to 180 Mt of hydrogen by 2030, hydrogen production becomes a concerning issue. SMR and coal gasification are responsible for 98% of hydrogen production. This makes hydrogen a significant contributor to CO_2 emissions. In contrast, methane pyrolysis in molten salts produces hydrogen with minimal CO_2 emissions and has the potential to create valuable carbon byproducts. However, due to the low technological readiness of this process, it is not employed commercially. In addition, a large amount of pyrolytic carbon would be generated if methane pyrolysis were widely implemented. This makes commercialising pyrolytic carbon to the battery industry of particular interest as it has the demand to handle the carbon produced. Prior studies on methane pyrolysis indicated salt loss due to intercalation into the carbon product, hindering economic viability. This study aims to characterise carbon produced from methane pyrolysis in a molten lithium salt mixture to enhance the economics of this process, as pyrolytic carbon with lithium intercalated is suitable anodes for lithium-ion batteries.

Performing methane pyrolysis in molten salts (LiBr-NaBr and LiCl-NaCl) has provided promising results for carbon commercialisation. All samples of pyrolytic carbon produced from the process exhibit a mixture of graphitic and amorphous structures. TGA, TEM, ICP-MS, and XRD analysis suggested that lithium was intercalated into the carbon complex. The carbon derived from LiBr-NaBr salts contained a higher degree of graphitisation, lithium-to-sodium ratio and high degree of lithium intercalation. This indicates that it could be a suitable anode material for lithium-ion batteries

5. Outlook

This study provides an analysis of the primary characteristics of pyrolytic carbon that are associated with the performance of the Li-ion anode. The following step would be to benchmark the capabilities of pyrolytic carbon as the anode of Li-ion batteries, given that the features of pyrolytic carbon produced were promising for Li-ion batteries. Each carbon sample will be used to generate the anode, which will then be tested in a coin cell. The analysis would be focused on how pyrolytic carbon produced from different salt mixtures and catalysts will affect the capacity of the coin cell.

In this investigation, not all analytical procedures were utilised perfectly. The TGA results provided in Section 3.2 had some trends that were not aligned with the literature. The discrepancy from the expected trend could be due to the high combustion rate. Therefore, in future experiments, the combustion rate should be lowered. Additionally, CHNS elemental analysis should be conducted on carbon samples to obtain and validate carbon purity.

Acknowledgements

The authors would like to express gratitude to Dimitrios Nikolis for advising throughout the project. The authors would like to thank Patricia Carry and Kaho Cheung from the analytical laboratory in the Chemical Engineering Department at Imperial College London for their assistance with equipment training.

6. References

Shellikeri, A., Watson V., Adams D., Kalu E.E., Read J.A., Jow T.R., Zheng J.S., Zheng J.P., 2017. Investigation of Pre-lithiation in Graphite and Hard-Carbon Anodes Using Different Lithium Source Structures. *Journal of The Electrochemical Society*, 164(14).

Acheampong, A. O. (2018) Economic growth, CO2 emissions and energy consumption: What causes what and where? *Energy Economics*. 74 677-692. 10.1016/j.eneco.2018.07.022.

B. Parkinson, C.F. Patzschke, D. Nikolis, S. Raman, D. Dankworth, K. Hellgardt, Methane pyrolysis in monovalent alkali halide salts: kinetics and pyrolytic carbon properties, *Int. J. Hydrogen. Energ.* 46 (9) (2021)

C.F. Patzschke, B. Parkinson, J.J. Willis, P. Nandi, A.M. Love, S. Raman, K. Hellgardt, Co-Mn catalysts for H2 production via methane pyrolysis in molten salts, *Chem. Eng. J*. 414 (2021)

D. Kang, N. Rahimi, M.J. Gordon, H. Metiu, E.W. McFarland, Catalytic methane pyrolysis in molten MnCl2-KCl, *Appl. Catal. B Environ.* 254 (2019) 659–666

Devrim, Y. ı & Albostan, A. (2016) Graphene-Supported Platinum Catalyst-Based Membrane Electrode Assembly for PEM Fuel Cell. *Journal of Electronic Materials*. 45 10.1007/s11664-016-4703-2.

Dresselhaus, M. S. & Dresselhaus, G. (2002) Intercalation compounds of graphite. *Advances in Physics*. 51 (1), 1-186. 10.1080/00018730110113644.

E.R. Ochu, S. Braverman, G. Smith, J. Friedmann, Hydrogen Fact Sheet: Production of Low Carbon Hydrogen, *Center of Global Energy Policy*, June 2021. https://www.energypolicy.columbia.edu/sites/default/files/pictures/HydrogenProduction_CGEP_FactSheet_052621.pdf

Ferrari, A. & Robertson, J. (2000) Interpretation of Raman spectra of disordered and amorphous carbon. *Physical Review B - Condensed Matter and Materials Physics*. 61 (20), 14095-14107. 10.1103/PhysRevB.61.14095.

Holmen, A., Olsvik, O. & Rokstad, O. A. (1995) Pyrolysis of natural gas: chemistry and process concepts. *Fuel Processing Technology*. 42 (2), 249-267. 10.1016/0378-3820(94)00109-7.

Hou, H., Qiu, X., Wei, W., Zhang, Y. & Ji, X. (2017) Carbon Anode Materials for Advanced Sodium-Ion Batteries. *Advanced Energy Materials*. 7 (24), 1602898. 10.1002/aenm.201602898.

Huang, Z., Deng, Z., Zhong, Y., Xu, M., Li, S., Liu, X., Zhou, Y., Huang, K., Shen, Y. & Huang, Y. (2022) Progress and challenges of prelithiation technology for lithium-ion battery. *Carbon Energy*. 4 (6), 1107-1132. 10.1002/cey2.256.

IEA (2022), Hydrogen, IEA, Paris https://www.iea.org/reports/hydrogen, License: CC BY 4.0

M. Voll, P. Kleinschmit, Carbon, 6. Carbon Black, in Ullmann‘s Encyclopedia of Industrial Chemistry, *Wiley-VCH,* Weinheim 2010.

Missyul, A., Bolshakov, I. & Shpanchenko, R. (2017) XRD study of phase transformations in lithiated graphite anodes by Rietveld method. *Powder Diffraction*. 32 (S1), S56-S62. 10.1017/S0885715617000458.

Muradov, N. (2017) Low to near-zero CO2 production of hydrogen from fossil fuels: Status and perspectives. *International Journal of Hydrogen Energy*. 42 (20), 14058-14088. 10.1016/j.ijhydene.2017.04.101.

Parkinson, B. J. (2020). Methane pyrolysis in molten salt environments Ph. D. Thesis. *Imperial College London*. Available at: https://doi.org/10.25560/96506 (Accessed: 12 November 2022).

Parkinson, B., Balcombe, P., Speirs, J. F., Hawkes, A. D. & Hellgardt, K. (2019b) Levelized cost of CO2 mitigation from hydrogen production routes. *Energy & Environmental Science*. 12 (1), 19-40. 10.1039/C8EE02079E.

Pimenta, M. A., Dresselhaus, G., Dresselhaus, M. S., Cançado, L. G., Jorio, A. & Saito, R. (2007) Studying disorder in graphite-based systems by Raman spectroscopy. *Physical Chemistry Chemical Physics*. 9 (11), 1276-1290. 10.1039/B613962K.

Röder, F., Sonntag, S., Schröder, D. & Krewer, U. (2016) Simulating the Impact of Particle Size Distribution on the Performance of Graphite Electrodes in Lithium-Ion Batteries. *Energy Technology*. 4 (12), 1588-1597. 10.1002/ente.201600232.

Schneider, S., Bajohr, S., Graf, F. & Kolb, T. (2020) State of the Art of Hydrogen Production via Pyrolysis of Natural Gas. *ChemBioEng Reviews*. 7 (5), 150-158. 10.1002/cben.202000014.

Shellikeri, A., Watson V., Adams D., Kalu E.E., Read J.A., Jow T.R., Zheng J.S., Zheng J.P., 2017. Investigation of Pre-lithiation in Graphite and Hard-Carbon Anodes Using Different Lithium Source Structures. *Journal of The Electrochemical Society*, 164(14).

Ungár, T. (2004) Microstructural parameters from X-ray diffraction peak broadening. *Scripta Materialia*. 51 (8), 777-781. 10.1016/j.scriptamat.2004.05.007.

Van Tu Nguyen, Huu Doan Le, Van Chuc Nguyen, Thi Thanh Tam Ngo, Dinh Quang Le, Xuan Nghia Nguyen and Ngoc Minh Phan, 2013. Synthesis of multi-layer graphene films on copper tape by atmospheric pressure chemical vapor deposition method. *Advances in Natural Sciences: Nanoscience and Nanotechnology*, 4(3).

Yue, X., Yao, Y., Zhang, J., Yang, S., Li, Z., Yan, C. & Zhang, Q. (2022a) Unblocked Electron Channels Enable Efficient Contact Prelithiation for Lithium-Ion Batteries. *Advanced Materials*. 34 (15), 2110337. 10.1002/adma.202110337.

Benchmarking the performance of the SAFT-g Mie approach in the prediction of solubility of pharma compounds

Khalid Osman and Mihnea Stefan
Department of Chemical Engineering, Imperial College London, U.K.

Abstract In this work, we intend to investigate the performance of a fully computational method to predict the pK_a value of two API of interest, ibuprofen and procaine. Using these pK_a values along with the predicted values of the activity coefficient of these API, obtained using a cutting-edge equation of state SAFT-γ-Mie, we try to predict the pH dependent solubility profiles of the API of interest and assess the performance of these models by comparing them to experimental data. We approach this problem by implementing a quantum-mechanical method to calculate the values of the pK_a which as opposed to the original research by Ho and Ertem will further be linearly regressed to correct for the systematic error given by the quantum-mechanical method. Using this linear regression technique, we managed to increase the accuracy of the predicted values of the pK_a of the API of interest threefold compared to the original research. In terms of pH-solubility profiles, the predicted solubility tends to be higher than the experimentally determined one, moreover in the case for ibuprofen an unusual curvature is observed in the predicted profile which has no physical meaning and is most likely due to numerical issues generated by the choice of buffering agent.

1.Introduction

The poor solubility of drugs is a common problem that can lead to low bioavailability and limited efficacy. In the gastrointestinal (GI) tract, where drugs are absorbed, the pH can vary greatly (Prescott, 1974). The solubility of ionisable drugs is dependent on the pH of the surrounding medium and the drug's pK_a value (Shoghi et al., 2013). Therefore, obtaining the pH-solubility profile of a new drug is essential for predicting its bioavailability. While experiments are the most reliable way to determine pK_a, they can be inconvenient and costly at the early stages of drug design for a new API (active pharmaceutical ingredient). In addition, conducting experiments on molecules that may be discarded at later stages of development can lead to a loss of time, money, and resources. In contrast, a fully computational approach offers a more desirable alternative for obtaining pH-solubility information about potential API at an early stage. This method can quickly identify and eliminate poor candidates, allowing experimental resources to be focused on the most promising API.

Obtaining the pK_a of a new ionisable API is crucial before it's pH-solubility profile can be produced. Previous work has been conducted by Ho and Ertem, where they used quantum-mechanical methods to predict pK_a values of different molecules (Ho and Ertem, 2016). In particular the thermodynamic cycle method was used and pK_a was found to be within 3.8-6.2 pH points of experimental values. In determining these pK_a values, a variety of solvation models were used and optimised at different levels of theory.

Ho and Ertem attribute the discrepancy between the predicted pK_a and experimental pK_a values, to a systematic error. This has also been identified by researchers in the MSE group at Imperial College London. In this paper, we aim to address this discrepancy and improve the pK_a prediction through systematic error elimination achieved via a simple linear regression model.

The focus of this research is on two ionisable API of interest, ibuprofen and procaine. Ibuprofen is a non-steroidal anti-inflammatory drug which is widely used and available over the counter (Drugbank, 2005a). Procaine is a local anaesthetic drug which is used for peripheral and spinal nerve block (Drugbank, 2005b). Since these drugs are common, pH-solubility experimental data will be easier to acquire as they have been well studied, this allows for a good assessment into the effectiveness of this proposed computational method.

The aim of this research is twofold. The first aim is to predict the values of the pK_a of these API using a two-stage procedure, which consists of a preliminary prediction via a quantum-mechanical approach followed by refinement of the preliminary predictions using a linear regression model. The second aim is to then use the predicted values of pK_a to obtain the pH-solubility profile of the API of interest and assess their performance against experimental data.

The remainder of this paper is laid out as follows. Firstly, we will outline the two-stage method used to obtain the predicted values of pK_a for both API of interest. We then outline the method used to obtain pH-solubility profiles for the API of interest and assess their performance against experimental data. Next, we will present the results of our investigation along with some analysis. Finally, we will summarise the conclusions of the research and discuss the implications of our findings

2.Background and Methods

The general purpose of this work is to study the performance of a fully computation approach to predict the pK_a values and the solubility profile at varying pH values for two ionisable active pharmaceutical ingredients (API) of interest, ibuprofen and procaine. To predict the pK_a values for the API of interest, the thermodynamic cycle

method (Ho and Ertem, 2016) was chosen for this work. To predict the solubility profile of an ionisable API one must know both the pK_a and activity coefficient value for the API. The pK_a will be obtained using the method specified above and will be refined using a linear-regression model, and the activity coefficient will be predicted using a cutting-edge equation of state SAFT-γ-Mie.

2.1 pKa Prediction:

In general, the acidity constant of the API or of its conjugated acid, if the API is a base is used, K_a. This constant quantifies the chemical dissociation of the API in water and is defined according to Equation (2), this definition is used here to be consistent with the original research made by (Wehbe, 2022). The chemical dissociation of the API of interest in water is described using Equation (1), using the general notation for the dissociating acid as HA and the ionised form of the acid A^-.

$$HA + H_2O \rightarrow A^- + H_3O^+ \quad (1)$$

$$K^m_{a,\,HA} = \left(\frac{m_{A^-}\, m_{H_3O^+}}{m_{HA}}\right)\left(\frac{\tilde{\gamma}_{m,\,A^-}\tilde{\gamma}_{m,\,H_3O^+}}{\tilde{\gamma}_{m,\,HA}}\right) \quad (2)$$

In Equation (2), $K^m_{a,\,HA}$ represents the molality dissociation constant of the API, where HA represent the API or its conjugated acid if the API is a base. In the same equation m_i and $\tilde{\gamma}_{m,i}$ represent the molality and the asymmetric molal activity coefficient of compound i. In general, however, it is not the dissociation constant that is mostly used, but its negative logarithm in the base of 10, the pK_a, this is defined in Equation (4) below.

$$pK_a = -\log_{10}(K^m_{a,\,HA}) \quad (4)$$

In our study we decided to use a quantum-mechanical based thermodynamic cycle method plus a systematic error elimination method to predict the value of the pK_a of the API of interest. The preliminary estimate of this value is calculated using Equation (5) and the thermodynamic cycle presented in Figure 1.

$$pK_a = \frac{\Delta G^*_{aq}}{RT\ln(10)} \quad (5)$$

To do this, we follow the approach set up in Ho and Ertem (Ho and Ertem, 2016). We use a thermocycle approach, Figure 1, in conjunction with the continuum solvation model to calculate the energies of solvation and liquid phase geometry optimizations. The reason we chose to use the thermocycle approach instead of the "direct method" defined by Ho and Ertem is the fact that the thermocycle method gives us a greater flexibility in selecting the level of theory for different calculations, allowing us to reach a better match between the model used and the level of theory. This means that for the gas phase calculations we select a high level of theory – namely G3MP2 - to assure good accuracy and for the solvation energy calculations we select the level of theory that is consistent with the parametrization scheme of the solvation model in Ho and Ertem. The solvation model we used to calculate the solvation energies and optimise geometries was SMD-M062X with a basis set 6-31+G(d).

Below, Figure 1, we present the general thermodynamical cycle used to perform all the calculations in our work. The cycle represents the dissociation reaction of a general acid HA, Equation (6). In order to calculate the value of the pK_a of HA one should use Equation (5). As it can be seen from it, all that is needed to use Equation (5) is the change in the standard Gibbs free energy of the dissociation reaction in the liquid phase, $\Delta G^*_{(aq)}$. According to the thermocycle method in the original work (Ho and Ertem 2016) this change in energy should be obtained from Equation (7).

$$HA + H_2O \rightarrow A^- + H_3O^+ \quad (6)$$

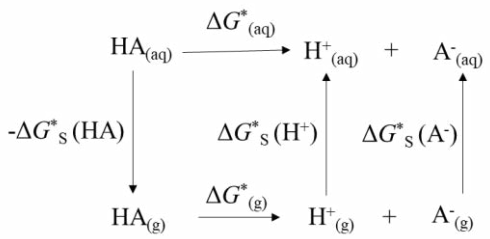


Figure 1. Generic thermodynamic cycle reproduced from Ho and Ertem (Ho and Ertem, 2016).

$$\Delta G^*_{(aq)} = \Delta G^*_{(g)} - \Delta G^*_S(HA) + \Delta G^*_S(H^+) + \Delta G^*_S(A^-) \quad (7)$$

In Equation (7), $\Delta G^*_{(g)}$ stands for the change in standard Gibbs free energy of the dissociation reaction in gas phase and $\Delta G^*_S(HA)$, $\Delta G^*_S(H^+)$ and $\Delta G^*_S(A^-)$ represent the solvation energies for HA, H^+ and A^-. All of these are calculated using the quantum-mechanical approach described above and in (Ho and Ertem, 2016). All the calculations were performed in the software Gaussian 16. The value of pK_a obtained from this method represents the preliminary estimation of it. We call it preliminary estimation as it has been proven by Ho and Ertem and by the work of other members in the MSE group at Imperial College London that this quantum-mechanical based technique gives a systematic error in the predicted value of the pK_a. In our work we intend to correct this systematic error by applying a scaling factor and a shifting

term that can be derived from a linear-regression model, as it has been proved by the work of members in our group (MSE Group, Imperial).

2.1.1 Linear-Regression Model:

Once the preliminary estimates for the API of interest are calculated, they can be refined using a linear-regression model. To do this, the linear model must be based on experimental values for the pK_a of small molecules. The molecules chosen for the model must be similar in structure to the target API for which the preliminary pK_a value must be corrected. It is natural then to expect that each of the API of interest will have its own separate linear model and if the API is diprotic then two linear model should be built for that API. In this work the API of interest are ibuprofen and procaine. Ibuprofen is an acidic monoprotic API, therefore only one linear model should be built for it. On the other hand, procaine is a diprotic basic API, hence two linear models will have to be built for it. Regardless of the API, the linear model is built in the same manner. For each API - or more accurately for each acidic/basic site, a set of molecules must be compiled, ones that are similar in structure to the acidic/basic site that is desired to be studied. These sets of molecules must be well known and experimental data for their value of the pK_a must exist and must be collected. All these sets are inputted into the quantum mechanical method chosen for this study and their preliminary estimations for the values of the pK_a are calculated. Then the preliminary estimated values for the pK_a are linearly regressed against their experimental counterparts, hence obtaining the linear model. The model takes as input the preliminary estimate for the pK_a value of the API, and outputs the refined value. These refined values are considered to be the final values for the pK_a of the API of interest in this work. All of the results, the linear models and their accuracy, and the refined values for the pK_a of the API of interest are presented in the Results and Discussion section.

2.2 Activity Coefficient Prediction:

The values for the activity coefficients for ibuprofen and procaine at different values of the pH of the solution were calculated using a highly accurate, cutting-edge equation of state SAFT-γ-Mie. This equation of state was deemed to be very accurate in predicting the values of the activity coefficients of various compounds in different solutions, as shown in the work done by (Wehbe, 2022). SAFT-γ-Mie is an equation of state based on statistical-mechanics perturbation theory. Its main function is to calculate the total Helmholtz free energy of the system of interest (Papaioannou et al., 2014). After this is obtained, it can be manipulated using standard thermodynamic relations to get the value of the activity coefficient (Wehbe, 2022). The Helmholtz free energy calculated by SAFT-γ-Mie is defined in Equation (7) below, in the same way Wehbe defined it in her work (Wehbe, 2022)

$$A = A^{Ideal} + A^{Mono} + A^{Chain} + A^{Assoc} + A^{Ion} + A^{Born} \quad (8)$$

In Equation (8), A stands for total Helmholtz free energy, A^{Ideal} represents the ideal contribution to the total Helmholtz free energy and the rest of terms stand for the monomeric, chain, association, ion and Born contributions to the total free energy of the system, respectively (Wehbe, 2022). As it is illustrated in Equation (8), the total Helmholtz free energy of the system is modelled as the sum of different terms, each term representing the contribution to the total free energy of different factors. This follows the underlying perturbation theory approach (Papaioannou et al., 2014). The activity coefficient is defined in the same way Wehbe defined it in her work (Wehbe, 2022) Equation (9).

$$\gamma_{HA}(T,P,x^L) = \frac{\hat{\varphi}_{HA}(T,\ P,\ \mathbf{x^L})}{\varphi^*_{HA}(T,\ P)} \quad (9)$$

$$\tilde{\gamma_i} = \frac{\gamma_i}{\gamma_i^\infty} \quad (9a)$$

$$\tilde{\gamma_{m,i}} = x_{H_2O}\tilde{\gamma_i} \quad (9b)$$

In Equation (9), γ_{HA} represents the symmetric activity coefficient of HA at given temperature T, pressure P and liquid phase composition vector x^L. Moreover, $\hat{\varphi}_{HA}$ represents the fugacity coefficient of HA in the solution at the given temperature pressure and composition, and φ^*_{HA} represent the fugacity coefficient of pure HA at the same temperature and pressure. In Equation (9a), $\tilde{\gamma_i}$ stands for the asymmetric molar coefficient of component i and γ_i^∞ stands for the symmetric activity coefficient at infinite dilution of component i. Equation (9b) is used to obtain the asymmetric molal activity coefficient of component i, $\tilde{\gamma}_{m,i}$ which will be used extensively in this work. Coming back to Equation (9), to obtain the fugacity coefficient $\hat{\varphi}_{HA}$ use Equation (10) below. In this equation $\mu_i^{Res}(T, P, x)$ stands for the residual chemical potential of component i (which can of course be our general acid HA) at temperature T, pressure P and composition x, and R represents the universal gas constant.

$$\ln \hat{\varphi}_i(T, P, x) = \frac{\mu_i^{Res}(T, P, \boldsymbol{x})}{RT} \quad (10)$$

To obtain the residual chemical potential in Equation (10), Equation (11) is used. In this equation A^{Res} stands for the residual Helmholtz free energy at the given temperature T, volume V and composition x. N_i represents the number of moles of component i in the mixture and Z is the compressibility factor.

$$\mu_i^{Res}(T, P, \boldsymbol{x}) = \left.\frac{\partial A^{Res}(T, V, \boldsymbol{x})}{\partial N_i}\right|_{T, V, N_{j\neq i}} - RT \ln Z(T, P, \boldsymbol{x}) \quad (11)$$

To obtain the residual Helmholtz free energy in Equation (11), the compressibility factor Z, and the pressure P of the system use Equations (12), (13), (14), (15), respectively.

$$A^{Res} = A - A^{Ideal} \quad (12)$$

$$Z = Pv_P/(RT) \quad (13)$$

$$v_P = V_P/N \quad (14)$$

$$P = -\left.\frac{\partial A(T, V, x)}{\partial V}\right|_{T, N} \quad (15)$$

In Equation (14) V_P represents the volume at the given pressure P, and N represents the total number of moles and hence v_P stands for the molar volume at the given pressure. For in-depth analysis of these equations check the work done by (Wehbe, 2022). It must also be noted that the bolded variables are vectors, and the normally written ones are scalars. To be able to use SAFT-γ-Mie one must first create the molecular model of the molecule of interest. This equation of state treats the chemical compounds as fused chains of spherical segments; this set of segments, represents the molecular model of the compound and these segments represent the different chemical groups that form the molecule of interest (Papaioannou et al., 2014). The molecular models for ibuprofen, procaine and their ionised forms were taken from another work (Wehbe, 2022) and we reproduce them in the Results and Discussion section, in the respective Case Studies.

Finally, to use SAFT-γ-Mie, one must calculate all the parameters required for the groups used in the molecular models of the molecules of interest. In the case of ibuprofen and procaine all the required parameters for all the groups of interest were calculated in another work by (Wehbe, 2022). We take these parameters from their work; for more details about how these parameters were calculated check the work done by (Wehbe, 2022). Using the molecular models for ibuprofen, procaine and their ionised forms, and the parameters calculated by Wehbe for their groups, we implement the SAFT-γ-Mie equation of state to predict the value of the activity coefficients of the molecules of interest in solution at different values of pH.

2.3 Solubility Profile Prediction:

At this stage we have the means to calculate the value for both the pK_a and the activity coefficient for the API of interest, ibuprofen and procaine. Hence, we can proceed to calculate the solubility of these API at different values of the pH of the solution. To do this, one must solve Equations (16) to 21 simultaneously, these have been taken from (Wehbe, 2022), while specifying the value of the pH of the solution of interest.

$$\ln x_{HA}^{L}(T, P) = \frac{\Delta h_{HA}^{fus}(T_{HA}^{fus}, P)}{R}\left(\frac{1}{T_{HA}^{fus}} - \frac{1}{T}\right) + \frac{1}{RT}\int_T^{T_{HA}^{fus}} \Delta c_{p, HA}(T', P)dT' - \frac{1}{R}\int_T^{T_{HA}^{fus}} \frac{\Delta c_{p, HA}(T', P)}{T'}dT' - \ln \gamma_{HA}(T, P, \mathbf{x}^{L}) \quad (16)$$

$$K_{a, HA}^{m} = \left(\frac{m_{A^-} m_{H_3O^+}}{m_{HA}}\right)\left(\frac{\tilde{\gamma}_{m, A^-}\tilde{\gamma}_{m, H_3O^+}}{\tilde{\gamma}_{m, HA}}\right) \quad (17)$$

$$K_w = \left(m_{H_3O^+} m_{OH^-}\right)\left(\tilde{\gamma}_{m, H_3O^+}\tilde{\gamma}_{m, OH^-}\right) \quad (18)$$

$$\sum_{i=1}^{N} q_i m_i = 0 \quad (19)$$

$$pH = -\log_{10}\left(a_{H_3O^+}\right) \quad (20)$$

$$\sum_{i=1}^{N} x_i \quad (21)$$

Equation (16) represents the solid-liquid equilibrium between the solid API and its solubilized neutral form. Equation (17) represents the chemical dissociation of the API. Equation (18) represents the chemical dissociation of water. Equation (19) represents the electroneutrality equation of the solution. Equation (20) represents the definition of pH and Equation (21) represents the mass conservation equation - the sum of the molar fractions of all components in a mixture must be equal to unity.

Once all of these equations are solved one can retrieve the value of the solubility of the API in the solution at that specified pH. The solubility of the API was defined, Equation (22), in the same way as (Wehbe, 2022) did in her work. It must be noted that Equation (16) will be modified for procaine, that is the integration terms will be taken out.

$$S_{API} = \rho M_w (m_{HA} + m_{A^-}) \quad (22)$$

In Equation (22), ρ stands for the density of the solution, M_w for the molecular mass of water and m_{HA} stands for the molality of HA and m_{A^-} for that of A^-.

It must be noted that in this work the solubility analysis was performed only until pH_{max}, a quantity defined by (Wehbe, 2022) in her work as the value of pH at which the salt of the ionisable API starts to precipitate. To calculate pH_{max} one must solve Equations (16) to (21) and Equation (23) simultaneously. Below, we present Equation (23), which represents the solubility product equation for the salt of the API. It must be noted the solubility product equation used here is not the same as the one used by Wehbe.

$$K_{sp} = \left(m_{A^-}\, m_{B^+} \right) \left(\tilde{\gamma}_{m,\,A^-} \tilde{\gamma}_{m,\,B^+} \right) \qquad (23)$$

Now pH_{max} can be calculated. Using the values of pK_a, activity coefficient and pH_{max} one can now obtain the solubility profiles for the API of interest. These profiles are presented in the Results and Discussion section together with all the results obtained from this work.

3. Results and Discussion

3.1 Case Study 1: Ibuprofen

3.1.1 Thermodynamic Cycle:

Here we present the particular form of the thermodynamic cycle used to predict the pK_a value for ibuprofen. This cycle represents the dissociation rection of ibuprofen.

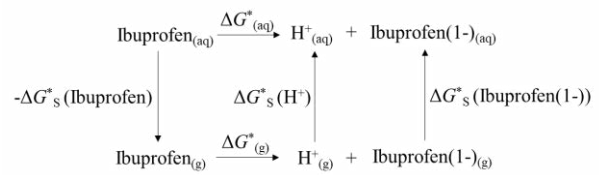


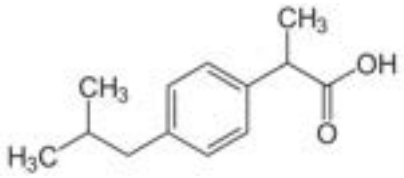
Figure 2. The thermodynamic cycle for ibuprofen.

3.1.2. Linear-Regression Model:

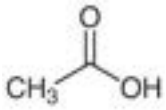
For ibuprofen, the small molecules chosen to make the linear regression were taken to be as similar as possible in structure to ibuprofen. The data source for the experimental pK_a for the small molecules were determined using a pH-metric method and were taken from (Haynes, Lide and Bruno, 2014). This was done since the pH-metric method is one of the most common methods of pH determination and this allowed for a larger selection of potential molecules for the regression, however there were still difficulties in obtaining pK_a values for some molecules which had large similarity to ibuprofen.

The chosen molecules for ibuprofen were:
Acetic acid, propanoic acid, phenylacetic acid, 2-phenylpropanoic acid, 4-methylbenzoic acid , 4-tert-butylbenzoic acid. However, phenylacetic acid and 4-tert-butylbenzoic acid were found to be outliers and removed from the final model. The structures for the final regression model are presented along with the structure for ibuprofen in the figure below

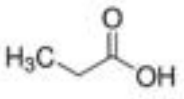
a)



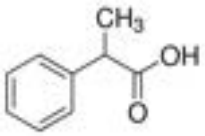
b)



c)



d)



e)

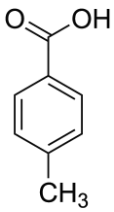


Figure 3. Structures for the small molecules used in the linear regression along with the structure for ibuprofen. a) ibuprofen b) acetic acid c) propanoic acid d) 2-phenylpropanoic acid f) 4-methylbenzoic acid

The linear regression model created using the small molecules shows excellent correlation between the Gaussian pK_a and the experimental pK_a, this is due to the systematic error present in the Gaussian calculations. These models are shown in Figure 4 along with the R^2 value and the visualisation. While these models could be constructed using any combination of molecules, by using molecules which have similar chemical groups to ibuprofen we can have pK_a values which are in close proximity to each other. This allows a linear model to be easily created. The high correlation could also be due to the systematic error being specific to different chemical groups, which was also a suspicion that Ho and Ertem had.

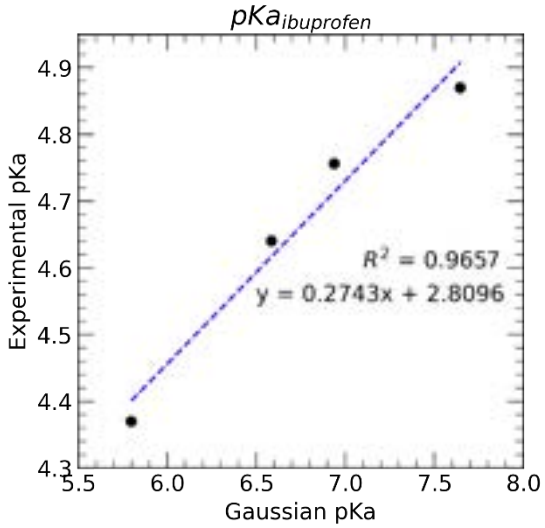


Figure 4. Linear regression model used to correct the preliminary Gaussian-predicted pK_a value for ibuprofen. The blue dashed line represents the model and the R^2 value is given along with the model equation. The variable y is the experimental value and x is the Gaussian pK_a. To correct the preliminary Gaussian-predicted pK_a value it is input as x into the model equation to generate y; the corrected Gaussian-predicted pK_a value. Each of the black data points corresponds to a different molecule used to construct the linear regression model. By increasing values of Gaussian pK_a these are:
4-methylbenzoic acid, 2-phenylpropanoic acid, acetic acid, propanoic acid.

3.1.3. pKa prediction

Using the linear regression model created in the prior section, the preliminary Gaussian-predicted pK_a value for ibuprofen was corrected. This has been reported in Table 1 along with the literature values for the pK_a.

Ho and Ertem (Ho and Ertem, 2016) reported that the general precision of predicted pK_a values for carboxylic acids using the SMD-M062X solvation model and thermodynamic cycle method is within 2.0 pH points of the experimental values. Using the same model and same thermodynamic cycle method, by applying a linear regression it can be seen that our precision has increased to be within 0.05 pH points of the mean literature pK_a value for ibuprofen, illustrating the large increase in precision that has been achieved by applying the systematic error correction.

Literature pK_a			QM calculated pK_a
4.51	4.42	4.85	4.54

Table 1. Literature values for the pK_a of the ibuprofen along with the corrected Gaussian-predicted pK_a values at T=298.15K P=1atm. The literature pK_a values were taken from (Domańska et al., 2009). These literature values were determined using the pH-metric method as it was the most common method for determining pK_a experimentally.

3.1.4. Activity Coefficient prediction:

The molecular models used to model ibuprofen and its ionised form were taken from another work along with all the required parameters for the groups of interest, this work is (Wehbe, 2022). These models are presented in Figure 5, below.

Figure 5. The molecular model used to model ibuprofen, image taken from (Wehbe, 2022). The molecular groups used to model ibuprofen were: 3 x CH_3, 4 x aCH, 1 x aCCH, 1 x CCH_2, 1 x CH, 1 x COOH. The ionised form of ibuprofen was modelled using the same groups as the molecular ibuprofen, the only difference is that the COOH group in molecular ibuprofen was replaced by the COO^- group in the ionised form.

Using these molecular models and the parameters for all the required groups - all taken from (Wehbe, 2022) - one can now implement the SAFT-γ-Mie equation of state to predict the value of the activity coefficient of ibuprofen in solution.

3.1.5. Solubility Prediction:

Using the corrected quantum-mechanical pK_a value and the activity coefficient, the pH-solubility profile was produced for ibuprofen. This was done by following the methodology outlined in section 2.2. pH_{max} was also calculated as outlined in section 2.2 and using a pK_a of 4.54 the pH_{max} was calculated to be 6.84 for ibuprofen. The pH-solubility profile was plotted against experimental data for pH-solubility, which was determined via saturation shake flask and pH-metric methods. A reasonably good fit is observed qualitatively.

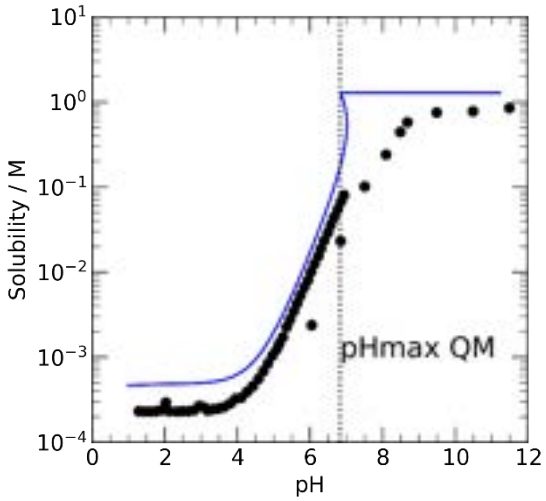


Figure 6. pH-solubility profile for ibuprofen against experimental data taken from (Avdeef, Berger and Brownell, 2000). Blue curve represents the QM pH-solubility profile. Black dots represent the experimental data, obtained using pH-metric and SSF method. Dotted vertical line is plotted at pH_{max} to outline the sections before and after pH_{max}. Plotted at conditions T = 298.15K, P= 1 atm.

An overprediction is observed in the predicted solubility, this is most probably due to the calculated pK_a not being in complete agreement with the literature pK_a value. Our investigation was only concerned with the pH-solubility behaviour prior to pH_{max}, however a strange curvature was seen in the profile beyond pH_{max}. This was also observed by other members in the MSE group and is suspected to be due to numerical error. Due to time constraints in this investigation, this deviation was explored however remains unresolved. This deviation is a point for improvement which can be explored in the future. Perhaps a better agreement with experimental data can be achieved at pH values beyond pH_{max}.

3.2 Case Study 2: Procaine

3.2.1. Thermodynamic Cycle:

Here we present the particular forms of the thermodynamic cycles used to predict the two pK_a values for procaine. In Figure 7, the thermocycle used to predict the pK_a value for the first dissociation reaction for procaine is shown, and in Figure 8 the thermocycle used for the second dissociation reaction for procaine.

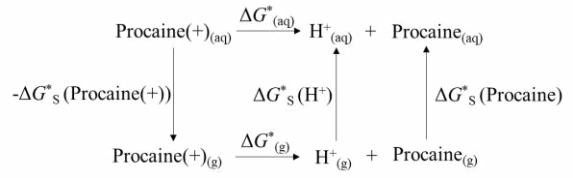


Figure 7. The thermodynamic cycle used to predict the value of the pK_a of the first dissociation reaction for procaine.

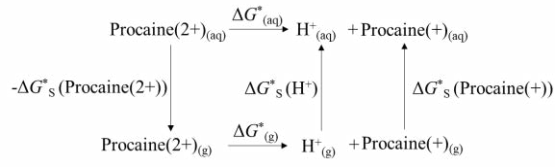


Figure 8. The thermodynamic cycle used to predict the value of the pK_a of the second dissociation reaction for procaine.

The two dissociation reactions for procaine are presented below in Equation (24) and (25).

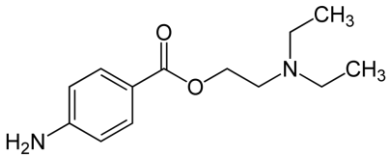
$$\text{Procaine}(+)_{(aq)} \rightarrow H^{+}_{(aq)} + \text{Procaine}_{(aq)} \quad (24)$$

$$\text{Procaine}(2+)_{(aq)} \rightarrow H^{+}_{(aq)} + \text{Procaine}(+)_{(aq)} \quad (25)$$

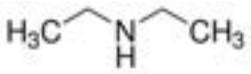
3.2.2. Linear-Regression Models:

Using all of the equations presented in the Section 2.1 one can reproduce the initial estimate of the pK_a value for the API of interest that we obtained using a quantum-mechanical approach. These predicted values of the pK_a of procaine have been refined using a linear-regression model meant to correct the systematic error given by the original quantum-mechanical method. For procaine, which is a diprotic API, two linear models will need to be developed, one for the aliphatic amino basic group and another for the aromatic amino basic group. For the aliphatic basic site of procaine the following molecules were selected to build the linear refinement model: diethylamine, dimethylamine and ethylamine. For the aromatic amino basic site, the following molecules were considered fit to be included in the linear model: 4-nitroaniline, ethyl 4-aminobenzoate and aniline. These structures, along with the structure for procaine have been shown in Figure 9 below. Comparing the structures of these aromatic and aliphatic groups with the structure of the molecules selected for the sets to be linearly regressed, a great similarity should be observed between the target amino group and its respective set of selected molecules.

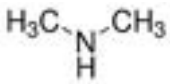
a)



b)

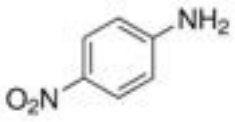


c)

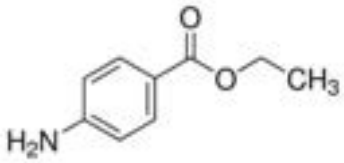


d)

e)



f)



g)

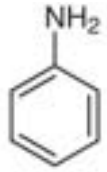


Figure 9. Chemical Structure of a) Procaine. The small molecules used in the linear regression are shown as well. For the aliphatic basic site, the molecules used were: b) diethylamine c) dimethylamine d) ethylamine For the aromatic basic site, the molecules used were: e) 4-nitroaniline f) ethyl 4-aminobenzoate g) aniline

Using these two sets of molecules the linear models have been developed, they are presented in Figure 10 and 11.

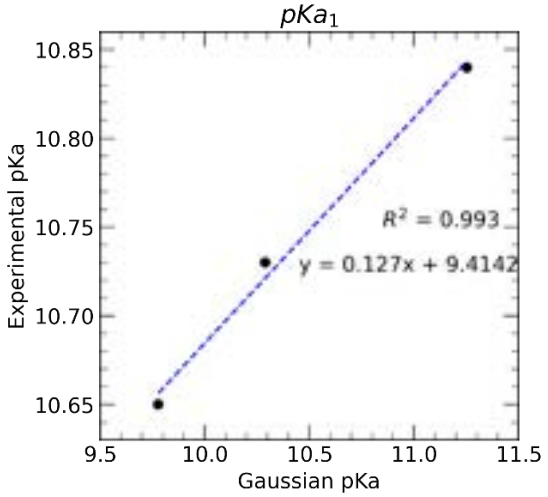


Figure 10. Linear-Regression Model for the aliphatic amino site of procaine. It can be seen from the $R^2 = 0.993$ as well as from the graph itself that this is a good fit. The molecules used in the final fit were: diethylamine, dimethylamine and ethylamine.

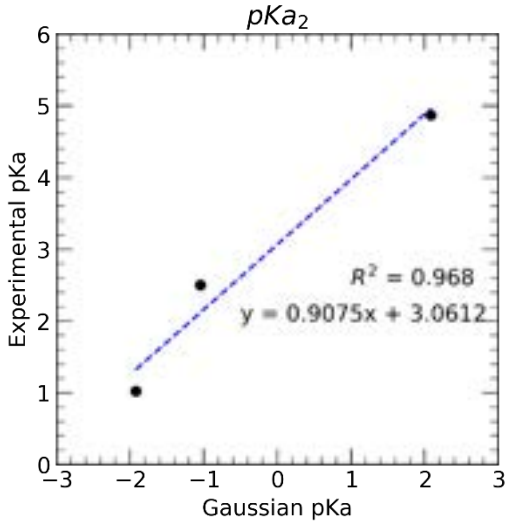


Figure 11. Linear-Regression Model for the aromatic amino site of procaine. It can be seen from the $R^2 = 0.968$ as well as from the graph itself that this is a good fit. The molecules used in the final fit were: 4-nitroaniline, ethyl 4-aminobenzoate and aniline.

3.2.3. pKa Prediction:

Using these linear models, we have the intention to remove the systematic error given by the quantum-mechanical method and hence increase the precision of the predicted values of the pK_a for the API of interest. So, for the aliphatic site of procaine a pK_a of 10.62 is obtained, using the linear model presented in Figure 10. When compared to the experimental value obtained by (Cairns, 2012, p.70) it doesn't seem to be a good agreement between the two. However, if we are to compare to other quantum-mechanical calculated pK_a value for the aliphatic site of procaine we would discover that the precision of the method presented in this work is threefold better than the general prediction presented by other studies. For, instance Ho and Ertem (Ho and Ertem, 2016) report a general precision of the original method of 4.6 pK_a points, whereas in our work we obtain a precision of 1.6 pK_a points at our lowest performance.

Moving to the aromatic amino basic site of procaine, using the proposed linear model in Figure 11 the predicted value for this pK_a was found to be 2.04. When compared to the experimental value of 2.5 (Cairns, 2012, p.70) a very good agreement can be seen. On top of this good agreement, there is also the knowledge that the general precision of the original method was found to be 4.6 pK_a points (Ho and Ertem, 2016). In this case our precision is of 0.46 pK_a points, tenfold better.

3.2.4. Activity Coefficient Prediction:

The molecular models used to describe procaine and its ionised forms were taken from another work, (Wehbe, 2022) and we just reproduce them here in Figure 12 to 14, below. All the required parameters for all the groups of interest used in these models were calculated by (Wehbe, 2022) and we retrieve them from there.

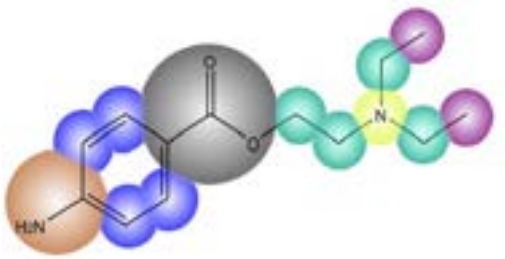


Figure 12. The molecular model used to describe procaine, image taken from (Wehbe, 2022). The procaine molecule was modelled using the following groups: 2 x CH_3, 4 x CH_2, 1 x N, 4 x aCH, 1 x aCCOO and 1 x $aCNH_2$.

Figure 13. The molecular model used to describe procaine mono-cation, image taken from (Wehbe, 2022). To describe procaine monocation the same groups were used as to describe molecular procaine, the only difference is that the N group in molecular procaine was replaced by the N^+ group in procaine monocation.

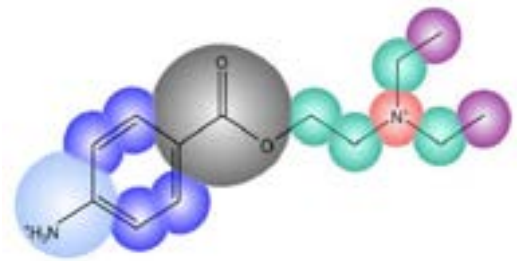


Figure 14. The molecular model used to describe procaine dication, image taken from (Wehbe, 2022). The groups used to describe procaine dication are the same as those used to describe procaine monocation, the only difference is that the $aCNH_2$ group in procaine monocation was replaced by the $aCNH_3^+$ group in procaine dication.

Using the molecular models presented above in Figure 12 to 14 and the parameters retrieved from (Wehbe, 2022) one can now implement the SAFT-γ-Mie equation of state to calculate the value of the activity coefficient for procaine in solution.

3.2.5. pH-Solubility Profile Prediction:

Following the method described in the Background and Method section of this work together with the particularities for procaine (linear-regression models, molecular models and the fact that procaine is a diprotic base hence two chemical dissociation reaction may occur and the pK_a value will be calculated for both of these reactions. etc.) one can now predict the solubility of procaine in solution at a given value of the pH, and from here the pH-solubility profile for procaine can be obtained. This profile was calculated only to pH_{max} (Wehbe, 2022), which was found for procaine to be 7.83. This value is not in good agreement with the one calculated by Wehbe for procaine, 6.21, judging that pK_a is measured in logarithmic scale, hence an error of 1.62 in logarithmic scale, represents in linear scale an error of over one order of magnitude. All the work done in this research was performed according to the method described by Wehbe. The only difference between her work and ours is the value used for the pK_a of procaine. In the work by Wehbe this value was taken from literature whereas in our work we calculated it using a fully computational quantum-mechanical method. Hence, the discrepancies between the work done by Wehbe and this work can be attributed to the predicted value of the pK_a not being identical to the experimental one used by Wehbe.

In Figure 15 we present our calculated pH-solubility profile for procaine. Its performance would have ideally been compared to a set of experimental data, however the data available for procaine is not sufficient to draw a conclusive analysis, hence we decide to compare our profile to the one obtained by Wehbe. The procaine pH-solubility profile of Wehbe has been shown to agree very well with existing experimental data available for procaine and we decide to use this profile as our standard, considering it to be experimental level data.

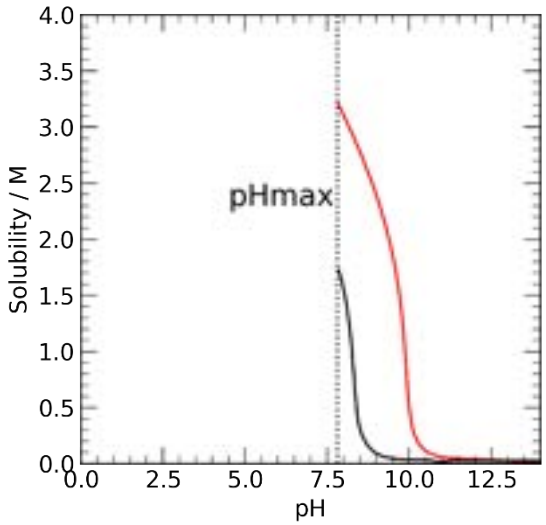


Figure 15. The pH-solubility profile for procaine calculated using the pK_a values obtained from the thermodynamic cycle method and the activity coefficient value obtained from SAFT-γ-Mie (red curve) compared to the pH-solubility profile obtained by Wehbe (black curve). It can be seen that the two profiles agree qualitatively, they have the same shape, but disagree from a quantitative point of view, they do not superimpose. We attribute this disagreement to the calculated values for pK_a of procaine not being identical to the experimental ones used by Wehbe. Comparison is performed at conditions of T=298.15K and P = 1atm.

As it can be seen from Figure 15, the pH dependent solubility profile for procaine developed in this work captures the qualitative essence of this property by having the same shape as the profile developed by Wehbe which we consider to be experimental level data. Hence from a qualitative point of view our model performs well, it manages to predict the steep increase in the solubility of procaine at the point where ionization starts to take place and play an important role in the overall solubility of procaine. On the other hand, quantitatively the performance of the solubility model developed in this work decreases drastically compared to the standard we chose. The pH-

solubility model for procaine developed in this work agrees quantitatively with the one developed by Wehbe only in the pH region 11-14 and disagrees all throughout the rest of the pH spectrum. A shift in the curve can be seen in Figure 15, where the original model by Wehbe (Wehbe, 2022) predicts that the solubility of procaine starts to increase at a lower pH value than the one predicted by the model developed in this work. As stated before, all calculations performed in this work respect the method presented in the original work in (Wehbe, 2022) the only difference is that the two pK_a values for procaine were calculated, in our work, using a fully computational method, whereas in the original work (Wehbe, 2022) they were taken from literature. Thus, we attribute this inaccuracy of our model to the values of the pK_a of procaine not being identical to the experimental ones used by Wehbe.

4. Conclusions

Based on our investigations, we can conclude that the linear regression models have improved the accuracy of the predicted values for the pK_a of the API of interest, as compared to the values obtained by Ho and Ertem. We have obtained a threefold increase in accuracy of the predicted values of the pK_a. In terms of solubility predictions, we manage to obtain a fairly good qualitative agreement with the experimental data for both API of interest, however when the point of view is switched to a quantitative one, it can be easily concluded that the performance of the models developed in this work is quite poor in predicting the experimental data for the solubility of the API of interest. Moreover, there should be further investigations into improving the curvature issue observed in the ibuprofen pH-solubility profile. We suspect this issue could be due to the choice of the buffering agent used to control the pH value of the solution. Investigations should be performed with a variety of basic buffers to see whether this curvature issue can be addressed. To improve the general predictions of the solubility for the API of interest one should endeavor to increase the accuracy of the predicted values for the pK_a. This could be done by implementing a stronger correction model than the simple linear one. It could be argued that only one predictor or one feature, namely the preliminary prediction for the pK_a of the API of interest, is not enough to describe nor correct for the systematic error given by the quantum-mechanical method used in this work and the original research. One would be interested in pursuing a non-linear model that takes into account, along with the preliminary prediction of the pKa of the API, of interest some information regarding their chemical structure, such as aromaticity etc.

References

Cairns, D. (2012). *Essentials of pharmaceutical chemistry*. London ; Chicago: Pharmaceutical Press, p.70.

Domańska, U., Pobudkowska, A., Pelczarska, A. and Gierycz, P. (2009). pK_a and Solubility of Drugs in Water, Ethanol, and 1-Octanol. *The Journal of Physical Chemistry B*, 113(26), pp.8941–8947. doi:10.1021/jp900468w.

Drugbank (2005a). *Ibuprofen*. [online] go.drugbank.com. Available at: https://go.drugbank.com/drugs/DB01050.

Drugbank (2005b). *Procaine*. [online] go.drugbank.com. Available at: https://go.drugbank.com/drugs/DB00721

Haynes, W.M., Lide, D.R. and Bruno, T.J. (2014). *CRC handbook of chemistry and physics a ready-reference book of chemical and physical data*. Boca Raton, Fla Crc Press.

Papaioannou, V., Lafitte, T., Avendaño, C., Adjiman, C.S., Jackson, G., Müller, E.A. and Galindo, A. (2014). Group contribution methodology based on the statistical associating fluid theory for heteronuclear molecules formed from Mie segments. *The Journal of Chemical Physics*, 140(5), p.054107. doi:10.1063/1.4851455.

Prescott, L.F. (1974). Gastrointestinal Absorption of Drugs. *Medical Clinics of North America*, 58(5), pp.907–916. doi:10.1016/s0025-7125(16)32088-0.

Serajuddin, A.T.M. (2007). Salt formation to improve drug solubility. *Advanced drug delivery reviews*, [online] 59(7), pp.603–16. doi:10.1016/j.addr.2007.05.010.

Shoghi, E., Fuguet, E., Bosch, E. and Ràfols, C. (2013). Solubility–pH profiles of some acidic, basic and amphoteric drugs. *European Journal of Pharmaceutical Sciences*, 48(1-2), pp.291–300. doi:10.1016/j.ejps.2012.10.028.

Wehbe, M (2022), 'Development of a predictive group-contribution platform for the phase behaviour and the effect of pH on the solubility of pharmaceuticals and excipients', PhD thesis, Imperial College London, London

Imperial College of London

Department of Chemical Engineering

Training in immersive and non-immersive environments: A comparative case study with VR & EEG

Authors: Adrien Lienau, George Spencer, Nitesh Bhatia and Omar K. Matar

Abstract:

The high cost associated with training employees is often mitigated by using non-immersive videos. VR has recently immerged as an alternative, but it remains to be seen if it is objectively more effective. EEG technology will therefore be used to bridge the brain-computer barrier (BCI) and objectively assess the mental performance during an assessment of two groups: Group A having received immersive training and Group B, non-immersive training. A combination of linear power spectral density ratios and machine learning models was used to quantify the levels of restfulness, mindfulness, engagement, alertness, and workload of the candidates. The values and the trends observed were compared to the commonly used subjective Nasa TLX survey to justify the use of EEG. It was found that immersive training allowed for higher concentration and focus than non-immersive training, but it proved to be a less restful experience for the user which could cause mental fatigue and negatively affect alertness through continued exposure. Although this was partly shown through the Nasa TLX survey, the results of the EEG were more consistent from one index to the other and from one participant to the other, as shown by the lower coefficient of variation. Further research is required to assess the effect of mental workload in an immersive environment on mental fatigue.

Keywords: Electroencephalography (EEG), Virtual Reality (VR), Brain Computer Interface (BCI), Interactive Training

1. Introduction

Skill training is a significant expense in the operation of any complex system with UK employers paying an average of £1,530 to train an employee in 2020 [1]. Transforming a novice into a skilled personnel is cost, time and resource intensive often requiring large-scale training infrastructure. Many industries such as medicine, manufacturing, education, maintenance and safety, therefore, try to mitigate these issues by utilizing video-based training, which is generally made by subject matter experts [2] [3] [4]. Interactive training has emerged as an alternative to these tutorials as individuals can train at their own pace by repeating simulations to improve their skills, practice often dangerous situations with less risk and learn from anywhere geographically. This switch to interactive training is expected to be accelerated by the recent advancements in immersive technologies such as virtual reality (VR).

Electroencephalographic (EEG) technology can be used to bridge the brain-computer interface (BCI) and assess the effectiveness of such VR training [5]. EEG indices such as power spectral densities and their ratios are commonly used to characterise the response and attitude of individuals performing various tasks, but little research has been conducted in applying these methods to VR training [5]. This type of immersive training is usually only assessed through self-reporting instruments such as the Nasa TLX survey or the IEQ which are rarely objective. Another limitation of these surveys is that the users are expected to gauge and remember their feelings throughout the whole process at the end.

This paper aims to assess the effectiveness of VR training by direct comparison with non-immersive training through EEG with two groups:

1. Group A - Immersive: Subjects trained in immersive and interactive VR training, which we consider close to real-world training.
2. Group B - Non-Immersive: Subjects trained in non-immersive and non-interactive video-based training, for which we opted for a first-person point-of-view recording of VR training.

2. Background

2.1 Immersive and Non-Immersive Training

The effectiveness of immersive compared to non-immersive training has been quantified before by comparing the performance of their subjects on an assessment after having either read a textbook, watched a video, or trained in VR [6]. However, the behaviour of the candidates in this study was only reported subjectively by making use of nine adapted emotion scales. Various meta-analyses' of the effectiveness of immersive training has been performed over the years, especially in the field of education, which quantify the learning outcomes of students in different year groups taught in an immersive environment [7]. While the results also showed immersive training to be more effective than non-immersive, these studies looked directly at academic achievements and did not characterise the learning through the use of EEG technology. The limitations of immersive training have also been explored, however, to show that physical training is still more effective by comparing the performance of candidates who used a VR or a physical wooden puzzle [8].

2.2 Performance Measurements with EEG

A systematic review of the application of EEG technology to effectively bridge the BCI showed that power spectral density (PSD) indices can be calculated to measure performance during various tasks ranging from flying a plane to arithmetic tests [5]. In literature, different EEG spectral powers have been associated with separate aspects of mental performance. A high alpha power (8-13 Hz), for example, has been linked to lower alertness but can also be a sign of cognitive fatigue [9] [10]. A decrease in mental awareness results in an increase in alpha power, and conversely a drop in alpha levels has been linked to the difficulty of a task increasing [9]. Beta power (13-30 Hz), on the other hand, has been associated with short-term memory alongside a change in working memory [11] [12] [13] [14] [15]. A higher beta power is a sign of a higher mental workload and of an increase in concentration levels [16] [17]. Finally, theta power (4-8 Hz) has been shown to rise with the difficulty of a task especially if it requires continued concentration which also leads to lower levels of alertness [18] [19] [9]. These frequency bands are the foundation of indices/ratios that enable further evaluation of mental performance.

The use of multiple indices to quantify overall mental performance is necessary due to its diverse nature [20]. Given theta increases with concentration and alpha increases with lower alertness, a power spectral density ratio suggested in literature is theta/alpha which can be used to assess mental workload [21]. Moreover, fatigue has been linked to both a decrease in theta and an increase in alpha powers, which both impact mental workload scores [22] [23]. This ratio will be used alongside two other ratios: engagement, defined as the ratio of beta-to-alpha power and alertness, defined as the inverse of alpha power [24]. Given that beta increases with higher concentration and alpha decreases with increasing difficulty, engagement reflects the ability to visually process and synthesise information [25]. Furthermore, from the definition of alertness, we expect to see an increase in the alertness value when the alpha band decreases. The procedures to analyze the data laid out by Freeman and inspired by Pope et al. will also be used to calculate all the PSDs [24].

As suggested by the systematic review of the application of EEG technology, two additional non-linear indices were implemented from the Brainflow library in python which utilize a machine-learning model trained on all five EEG power bands [5].

2.3 EEG within VR

EEG has also specifically been combined with VR to either create an adaptive environment such as a game that adjusts to the mental workload of the player to maximize engagement or to characterise behaviour in an immersive setting [26]. EEG was, for example, used to measure the cognitive load of people playing a game similar to Tetris in VR [27]. EEG indices have also been used to evaluate the effect that playing a VR game in a first-person, or third-person perspective has on engagement [28].

The research that is most closely related to this paper is one that investigates the technical feasibility of adaptive training in VR using alpha-based indices [29]. As mentioned earlier, our paper will use a combination of PSD ratios because of the diverse nature of mental workload. Although this research explores the feasibility of VR training, it only uses EEG as feedback to adapt the training and not as a performance assessment tool and therefore also does not make the comparison with non-immersive training.

3. Methods

3.1 Experiment Protocol

To compare the effectiveness of VR and non-immersive training, two groups of four were formed, with Group A receiving VR training and Group B receiving non-immersive training. The eight candidates were recruited through a survey in which they had to supply their personal contact details, age, and gender, and indicate whether they wore glasses. The age range of the candidates was 18 to 23 years old, and all were male. As part of the survey, the participants were also required to measure their head circumference with provided measuring tapes as well as complete a mental math test. The candidates were selected if their head circumference lay between 51 cm and 61 cm, as these were the only sizes the Ant Neuro Eego medium and large EEG caps were able to accommodate, and if they did not wear glasses so that they could comfortably wear the VR headset, a Pico Neo 2 Eye (All-in-one android-based VR headset). Furthermore, given this research contained human subjects, we also collected ethical approval from the department and written consent from all the subjects. All data from the study was also anomalysed using a subject identification number.

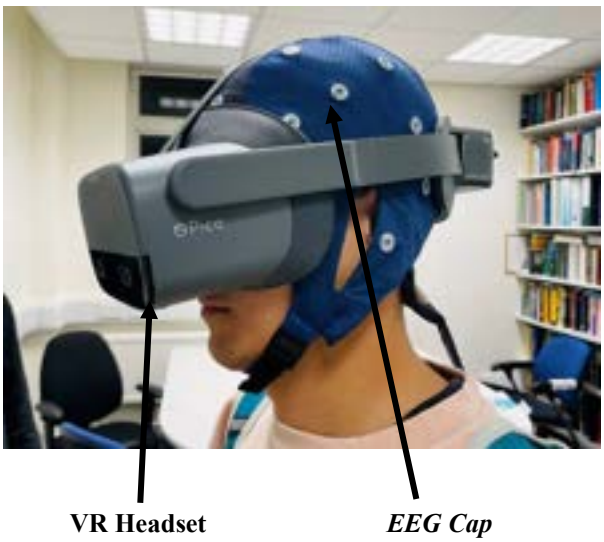


Figure 1: Subject wearing the VR Headset and EEG Cap

During the experiments, the EEG caps were fitted onto the heads of the participants (**Figure 1**) by aligning all the electrodes with the nine scalp sites of the 10-20 system considered (F3, F4, C3, C4, P3, P4, Cz, Pz and Fz) [30][Appendix 9.3]. The central region of the scalp was chosen to prevent interference with the VR headset and help reduce noise. A conductive gel was added onto each electrode to bridge the scalp-electrode barrier and the EEG cap was adjusted until the impedance at each electrode was below 60 Hz. Each site was referred to Cpz, grounded at Fpz and the recorded sampling rate was of 500 Hz. The signals were notch filtered at 50Hz to remove UK-specific electrical interference during data acquisition [31].

The VR environment was then calibrated to the floor to ensure that all the elements of the test environment were at the correct height. Additionally, the fit of the headset was adjusted using a built-in application.

3.2 Experimental Task

To be able to normalize the EEG readings of the training and assessment, the candidates were first asked to complete a 30-second relaxation exercise with their eyes closed to serve as a base level, and a math test until they reached the median score of the survey results, to serve as the high level. The candidates were all given a pre-training task to get familiar with the buttons on the controllers and learn how to grab objects in VR. The training for group A candidates then consisted of performing the experimental task with audio commands and Group B candidates watched a recording of someone performing the experimental task.

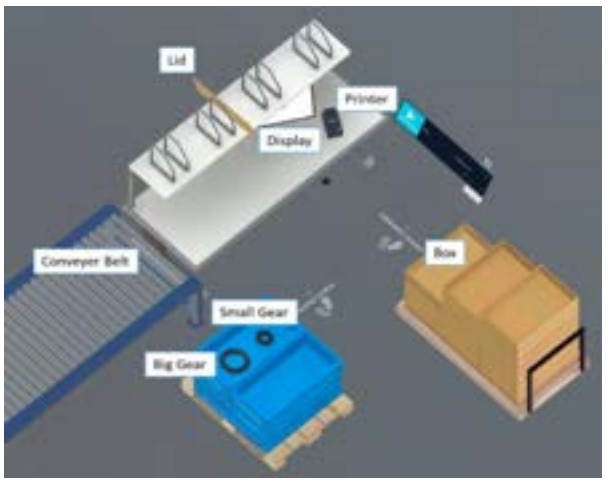


Figure 2: VR Experimental Task [33]

As shown in **Figure 2**, the experimental task was a packing exercise in which the user is required to place an empty cardboard box on a work surface, attach a label on the side of the box, place two gears in the box in a specific order, close the box with a lid and send it off to the next processing step by placing it on a conveyer belt.

Both groups were then assessed on the same experimental task in VR without any audio commands but if the participants could not remember the steps, they had the possibility to refer to a textbox indicating the next step. Participants were given as much time as they needed.

3.3 Data Description

The EEG activity of all the participants was measured during the relaxation exercise, math test, training, and assessment at the nine locations on their scalp. Each participant also completed a NASA TLX survey to collect a subjective assessment of their performance solely during the assessment [Appendix 9.1]. The participants were first asked to rate out of a 100 their perceived workload with six sub-scales (**Table 1**):

Table 1: Nasa TLX Definitions for a given Index

Index	Definition
Mental	How mentally demanding was the task?
Physical	How physically demanding was the task?
Temporal	How hurried or rushed was the pace during the task?
Performance	How successful were you in accomplishing the task you were asked to do?
Effort	How hard did you have to work to accomplish your level of performance?
Frustration	How insecure, discouraged, irritated, stressed, annoyed were you?

The reported scores from the scales were then weighed using results from paired choices between the same six indices. The survey data for groups A and B were analysed separately. The average completion time of the assessment was also measured for both groups.

3.4 Analysis Methods

The data collected with the EEG caps was first filtered for noise using the ANT Neuro software. A Fast Fourier Transform was then performed on data from each electrode for each participant to compute power estimates. These were split into five frequency bands (**Table 2**):

Table 2: Frequency Intervals for a given Band

Band	Frequency interval (Hz)
Delta	1-4
Theta	4-8
Alpha	8-13
Beta	13-30
Gamma	30-50

The power spectral density was then estimated using Welch's method on two-second windows without considering the first and the last chunk to account for start-up and shutdown [32]. Mindfulness and restfulness indices were calculated using the Brainflow library in python with all five frequency bands for each two-second window. Engagement, alertness, and workload were calculated for each two-second chunk using the following equations:

$$\text{Engagement Index} = \frac{\text{Beta}}{\text{Alpha}} \quad (1)$$

$$\text{Workload Index} = \frac{\text{Theta}}{\text{Alpha}} \quad (2)$$

$$\text{Alertness Index} = \frac{1}{\text{Alpha}} \quad (3)$$

This analysis was performed in python and the script can be found in the appendix [Appendix 9.2]. Each index's minimum and maximum values in all the two-second windows during the relaxation and mathematics exercises were used as low and high limits for each participant. The average values of each index for each participant were then normalized using those limits. The coefficient of variation (CV) was also calculated for each index, for each group, both for the EEG data and the Nasa TLX survey to compare the precision and consistency of the data.

4. Results

4.1 EEG Data

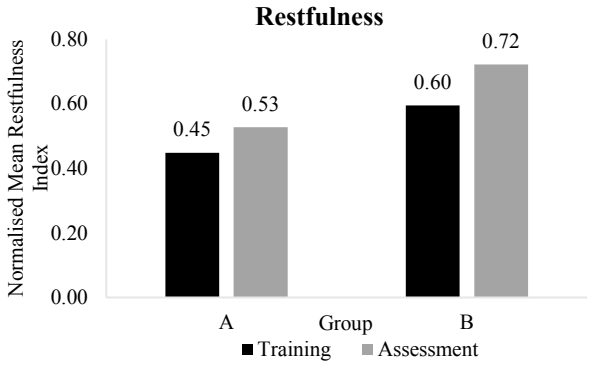


Figure 3: Calculated Restfulness Index for Immersive (Group A) and Non-Immersive (Group B) Training

Firstly, from **Figure 3**, we can observe that candidates who followed non-immersive training had higher restfulness. Secondly, we can see a 17.7% and 21.3% increase in restfulness from the training to the assessment for Group A and Group B, respectively. As can be seen from **Table 3**, the training values for Group B have the highest CV at 66% which is still considered low variance as it lies below 100%.

Looking at **Figure 4**, we can see that Group A has a mindfulness score for training higher than Group B. Furthermore, looking into the change between training and assessment, Group A and B had a 14.4% drop and 31.3% decrease respectively in mindfulness. Group B's training data, with a CV of 97% (**Table 3**), had the highest variance but again, given it is below 100%, can be considered low.

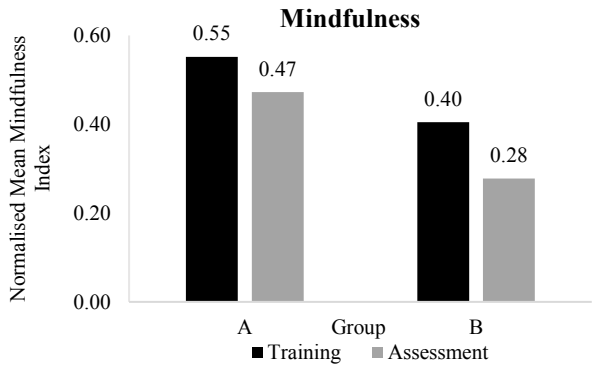


Figure 4: Calculated Mindfulness Index for Immersive (Group A) and Non-Immersive (Group B) Training

As seen in **Figure 5**, the engagement index followed the same trend for both Group A (-4.4%) and B (-40.2%) experiencing a drop from the training to the assessment. Looking into the CV values from **Table 3**, the values for both training and assessment for Group A are close or at 100% possibly indicating that these values have high variance and therefore lower reliability.

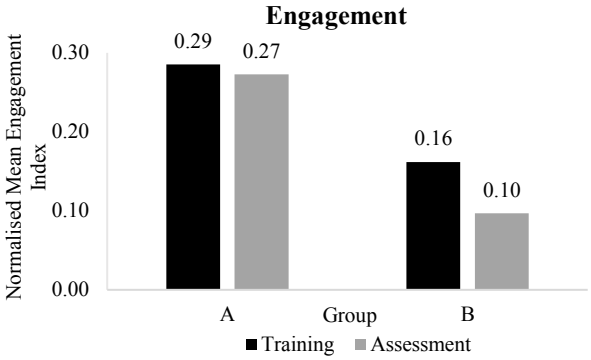


Figure 5: Calculated Engagement Index for Immersive (Group A) and Non-Immersive Group B) Training

Looking at **Figure 6**, Alertness was quite unique as the changes from the training to the assessment in both groups were opposite. Group A saw a decrease of 13.1% and Group B gave an increase of 35.4%. This resulted in a very small difference between the A and B assessment ratios of 0.02. Given that no CV value was greater than 51%, all alertness ratios can be considered to have low variation and therefore high precision.

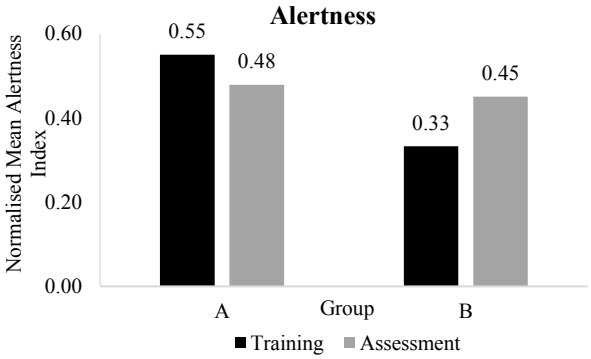


Figure 6: Calculated Alertness Index for Immersive (Group A) and Non-Immersive (Group B) Training

Looking into the workload from **Figure 7**, we can observe the same decreasing trend from training to assessment as seen in prior data between Group A and B. Regarding the precision, the CV values for training and assessment for Group A were 94% and 97% respectively. Since these values still lie below 100%, they can be considered low variance.

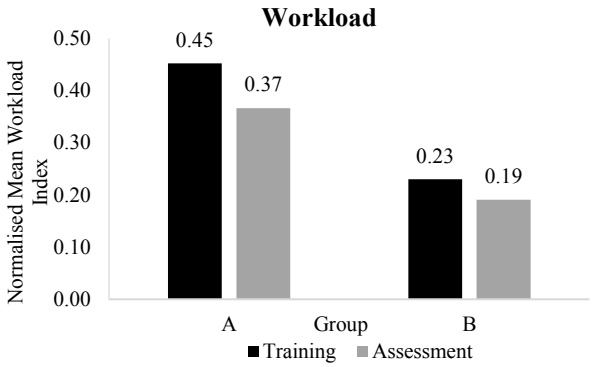


Figure 7: Calculated Workload Index for Immersive (Group A) and Non-Immersive (Group B) Training

4.2 Average Assessment Completion Time

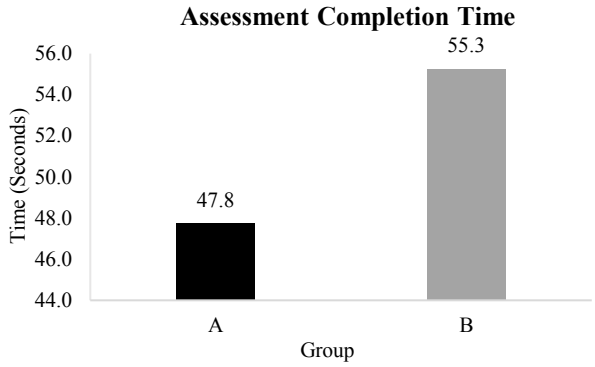


Figure 8: Average Assessment Completion Time for Immersive (Group A) and Non-Immersive (Group B) Training

As can be seen from the average completion time in **Figure 8**, the immersive group was able to complete the assessment 15.7% faster than the non-immersive.

4.3 NASA TLX Survey

As can be seen from **Figure 9**, Groups A and B reported very different scores for most of the indices. The largest difference was found to be for physical demand with Group A's average reported score 10 times higher than Group B's. The candidates from Group B also reported an 81.0% lower mental effort score than group B. The average reported temporal and performance indices were 73.0% higher and 9.8% lower for Group A than for Group B, respectively. These were the closest scores of any metric indicating that on a subjective level, candidates did not feel differently between training mediums as heavily. The percentage difference between Group A and B for the Effort Index was 51.2% and Group B reported being 6.2 times more frustrated than Group B. Given the relatively large differences between the individual metrics, it was quite surprising to get very similar overall scores (**Figure 10**) with the weighted averages for Group A being 12.0% lower than for Group B.

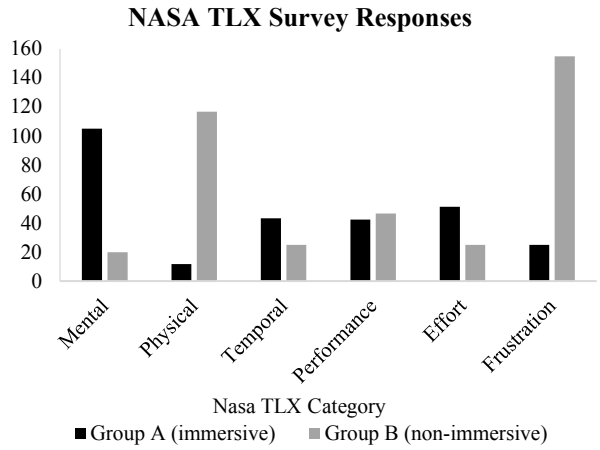


Figure 9: Weighted Nasa TLX Survey Responses for the assessment from both Group A and Group B

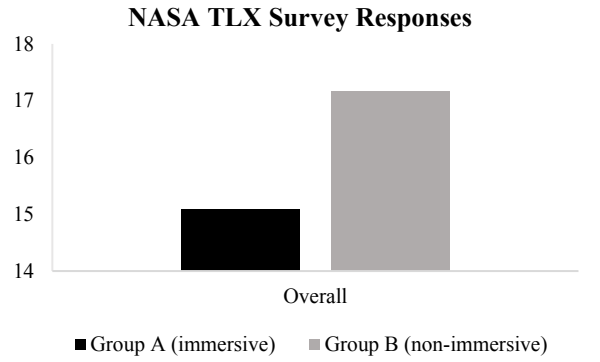


Figure 10: Combined Weighted Nasa TLX Survey Responses for the assessment from both Group A and Group

It should also be noted that the reported scores in the Nasa TLX survey for both groups had very high coefficients of variation: 1.01 for group A and 0.94 for group B (**Figure 10**).

4.4 Measurement Error

Furthermore, measurement error was considered throughout the experiment. Firstly, we used wet electrode EEG caps to minimise EEG measurement errors due to signal quality and variation in skull geometry. We also created a custom montage [Appendix 9.3] for each subject when carrying out the tasks which helped ensure connection between the EEG cap and the amplifier was stable. Moreover, the recording software (ANT Neuro Eego) provided online noise and signal filtering to help ensure reliability of the data. This data, for all four tasks, was also recorded in a single trial having an approximate five-minute duration. To minimise subject-specific errors due to noise, we further normalised the metrics using EEG data from the first two tasks.

Table 3: Coefficients of Variation for each of the five Indices Measured for both Groups

	Group A					
	Training			Assessment		
Index	St. Dev	Mean	CV	St. Dev	Mean	CV
Mindfulness	0.15	0.55	27%	0.047	0.47	10%
Restfulness	0.15	0.44	34%	0.047	0.53	9%
Engagement	0.26	0.29	93%	0.28	0.27	100%
Alertness	0.28	0.55	51%	0.17	0.48	35%
Workload	0.42	0.45	94%	0.36	0.37	97%
	Group B					
	Training			Assessment		
Index	St. Dev	Mean	CV	St. Dev	Mean	CV
Mindfulness	0.39	0.40	97%	0.13	0.28	46%
Restfulness	0.39	0.60	66%	0.13	0.72	18%
Engagement	0.06	0.16	37%	0.04	0.10	45%
Alertness	0.078	0.33	23%	0.20	0.45	44%
Workload	0.035	0.23	15%	0.11	0.19	56%

5. Discussion

5.1 EEG Data

When comparing the data from both groups, two aspects are of interest: the first is the values of the indices during the training and the second is the change between training and assessment. What is of particular interest is the shift from training to assessment. If different for both groups, it would indicate this change is caused by the medium (immersive versus non-immersive) through which the content is taught and not the content itself. Firstly, it can be observed that participants who received immersive training were less relaxed than their counterparts in Group B from the lower calculated restfulness score (**Figure 3**). A possible explanation for this lower restfulness is that the stress of completing the activity oneself in the VR environment and the performance anxiety associated with the task were greater than the stress caused by simply watching someone else do it, possibly highlighting some drawbacks associated with training in VR. Fortunately, this metric increased for both groups when moving onto the assessment, but the Group A participants remained less relaxed. As the trend is the same in both cases, it is more likely caused by the content of the exercise and could simply be explained by the increased confidence that both groups had with the task at hand which shows that the training was effective in delivering the material in both cases. However, given that very little, if any, research has been carried out on the link between restfulness and immersive versus non-immersive training, more work needs to be carried out to fully justify these explanations.

Participants from Group A were also much more concentrated on the task during the training than participants from Group B as seen from the higher mindfulness (**Figure 4**) and engagement (**Figure 5**) scores recorded. Being immersed in a 3D environment and having to continuously perform tasks to complete the training could logically increase the given indices and decrease the chance of a trainee losing focus. Knowing that higher engagement scores are linked with increased concentration and alertness levels, helps confirm that immersive training is more effective than non-immersive training [16] [25]. However, higher engagement is also linked with increasing difficulty of a given task showing that subjects possibly found the VR training to be more challenging than watching the video [25]. This is also shown through the increased workload index for training in VR (**Figure 7**). More research, therefore, needs to be carried out to see the difference in workload between the two groups when a more complex and mentally demanding task is performed to see if the consequence of this effect on a subject over a sustained period leads to other outcomes not considered. Furthermore, as the mindfulness index is based on a novel machine-learning model, there is not sufficient literature to fully understand it. A possible way to quantify the weight of each of the five power bands in the machine-learning model would be to perform a sensitivity analysis. Finally, going from training to assessment, concentration seems to decrease in both cases which is therefore likely caused by the content and not the medium, although both metrics decreased more heavily for subjects in Group B. This can be logically understood as the act of learning requiring more focus than the act of completing an assessment. The respective percentage drops of 4.4% between Group A's training and assessment compared to the 40.2% decrease from Group B's training to assessment illustrate that training in a VR environment can improve concentration when the task is eventually assessed.

Group A was also much more alert than Group B during the training (**Figure 6**) which confirms that the subjects remained much more aware of the material being taught [9]. However, as discussed before, alertness can also be an indication of mental fatigue which might explain the trend from training to assessment for both groups [10]. Group A became less alert whilst Group B's alertness index increased to almost match the level of Group A. Given the higher workload of Group A during training and assessment (**Figure 7**), the VR may have challenged the participants too much and negatively affected mental fatigue. While this shows alertness is more heavily linked to the environment in which candidates perform a task than to the content of the training, it is concerning that even a simple task caused a drop in alertness from training to

assessment. Given this decrease was relatively small, more research would need to be carried out to measure the effect of more complex tasks in a VR environment on fatigue and alertness.

5.2 NASA TLX

The higher reported mental and effort scores in the survey for candidates who received immersive training are consistent with the high mental workload index measured. The slightly higher temporal pressure felt by Group A could also explain why they completed the assessment eight seconds faster on average (**Figure 8**), but all the other metrics seem to give a poor representation of the assessment. The high frustration for the non-immersive group is contradictory to the EEG data as this should have impacted the restfulness score. Moreover, the performance index is also higher for Group B even though almost all candidates had to refer to the written instructions. These differences are most likely caused by the subjectivity of this survey which is shown by the high coefficient of variation for both Groups A and B. This shows how much more consistent and precise the EEG data is compared to the Nasa TLX. To confirm these results, a higher number of participants should be selected.

6. Conclusion

In conclusion, EEG technology enabled us to bridge the BCI and objectively show through both the linear engagement metric and mindfulness machine learning model that immersive training permits higher concentration and focus than non-immersive training. However, immersive training also proved to be a less restful experience for the user which could cause mental fatigue and negatively affect alertness through continued exposure. This adverse effect could be caused by the higher workload required to complete the task in the VR environment. Although this was partly shown through the Nasa TLX survey, the results of the EEG were more consistent from one index to the other and from one participant to the other, as shown by the lower coefficient of variation. These results differ from previous studies as the superiority of immersive learning is put into question which might stem from our use of a combination of linear and non-linear ratios. The use of multiple indices to get a more complete understanding of the effect of immersive training is only one of the steps required to compare training in immersive and non-immersive environments. An optimum between mental fatigue and high mental workload which increases engagement could be found by conducting a range of experimental tasks with a spread of difficulties and duration. With more time, a greater number of candidates with more age and gender diversity should also be used to increase confidence in the data. Additionally, more sensitive EEG caps would enable more accurate readings in more precise areas of the brain, which with more knowledge of the different cortexes could lead to a more comprehensive answer to our research question.

7. Acknowledgements

We would like to thank Prof. Omar K. Matar and Dr. Nitesh Bhatia for their continued support throughout this project, as well as the Department of Chemical Engineering.

It is also a privilege to be invited to present our findings as a full paper at the Human Computer Interaction 2023 International (HCII2023) Conference.

8. References

[1] F. Hardcastle, "Baltic Apprenticeships," 2021. [Online].

[2] B. Dencker, H.-J. Balzer, W. Theuerkauf and Schweres, "Using a production-integrated video learning system in the assembly sector of the car manufaturing industry," *International Journal of Industrial Ergonomics, Volume 23, Issue 5-6,* pp. 525-537, March 1999.

[3] C. McCarthy and R. Uppot, "Advances in Virtual and Augmented Reality—Exploring the Role in Health-care Education," *Journal of Radiology Nursing,* vol. 38, no. 2, pp. 104-105, 2019.

[4] S. Leblanc, "Analysis of Video-based training approaches and professional development," *Contemporary issues in Technology and Teacher Education,* vol. 18, no. 1, 2018.

[5] L. Ismail and W. Karwowski, "Applications of EEG indices for the quantification of human cognitive performance: A systematic review and bibliometric analysis," 2020.

[6] A. Von Mühlenen and D. Allcoat, "Learning in virtual reality: Effects on performance, emotion and engagement," *Research in Leaning Technology,* vol. 26, 2018.

[7] R. Villena-Taranilla, S. Tirado-Olivares, R. Cozar-Gutierrez and J. Gonzalez-Calero, "Effects of virtual reality on learning outcomes in K-6 education: A meta-analysis," *Educational Research Review,* vol. 35, 2022.

[8] P. Carlson, A. Peters, S. Gilbert, J. Vance and A. Luse, "Virtual Training: Learning Transfer of Assembly Tasks," *IEEE Transactions on Visualization and Computer Graphics,* vol. 21, no. 6, pp. 770-782, 2015.

[9] A. Kamzanova, A. Kustubayeva and G. Matthews, "Use of EEG workload indices for diagnostic monitoring of vigilance decrement," *The Journal of the Human Factors and Ergonomics Society,* vol. 56, no. 6, 2014.

[10] G. Borghini, G. Vecchiato, J. Toppi, L. Astolfi, A. Maglione, R. Isabella, C. Caltogirone, W. Kong, D. Wei, Z. Zhou, L. Polidori, S. Vitiello and F. Babiloni, "Assessment of mental fatigue during car driving by using high resolution EEG activity and neurophysiologic indices," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2012.

[11] M. MacLean, K. Arnell and K. Cote, "Resting EEG in alpha and beta bands predicts individual differences in attentional blink magnitude," *Brain and Cognition,* vol. 78, no. 3, 2012.

[12] M. Mazher, A. Aziz, A. Malik and H. Amin, "An EEG-Based Cognitive Load Assessment in Multimedia Learning Using Feature Extraction and Partial Directed Coherence," *IEEE Access,* vol. 5, 2020.

[13] C. Tallon-Baudry, A. Kreiter and O. Bertrand, "Sustained and transient oscillatory responses in the gamma and beta bands in a visual short-term memory task in humans," *Cambridge University Press,* vol. 16, no. 3, 1999.

[14] S. Palva, S. Kulashekhar, M. Hamalainen and M. Palva, "Localization of Cortical Phase and Amplitude Dynamics during Visual Working Memory Encoding and Retention," *Journal of Neuroscience,* vol. 31, no. 13, 2011.

[15] B. Spitzer and S. Haegens, "Beyond the Status Quo: A Role for Beta Oscillations in Endogenous Content (Re)Activation," *eNeuro,* vol. 4, no. 4, 2017.

[16] S. Coelli, R. Sclocco, R. Barbieri, G. Reni, C. Zucca and A. M. Bianchi, "EEG-based index for engagement level monitoring during sustained attention," in *37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2015.

[17] I. Kakkos and G. Dimitrakopoulos, "Mental Workload Drives Different Reorganizations of Functional Cortical Connectivity Between 2D and 3D Simulated Flight Experiments," *IEEE Transactions on Neural Systems and Rehabilitation Engineering,* vol. 27, no. 9, 2019.

[18] P. Antonenko, F. Paas, R. Grabner and T. van Gog, "Using Electroencephalography to Measure Cognitive Load," *Educational Psychology review,* vol. 22, 2010.

[19] A. Gevins and M. Smith, "Neurophysiological measures of cognitive workload during human-computer interaction," *Theoretical Issues in Ergonomics Science,* vol. 4, no. 1-2, 2010.

[20] R. F. Rojas, J. Fidock, M. Barlow, D. Essam, K. Kathryn, A. Sreenatha, G. Matthew and A. Hussein, "Electroencephalographic Workload Indicators During Teleoperation of an Unmanned Aerial Vehicle Shepherding a Swarm of Unmanned Ground Vehicles in Contested Environments," 2020.

[21] B. Raufi and L. Longo, "An Evaluation of the EEG Alpha-to-Theta and Theta-to-Alpha Band Ratios as Indexes of Mental Workload," *Frontiers in Neuroinformatics,* vol. 16, 2020.

[22] I. Kathner, S. Wriessnegger, G. Muller-Putz, A. Kubler and S. Halder, "Effects of mental workload and fatigue on the P300, alpha and theta band power during operation of an ERP (P300) brain–computer interface," *Biologival Psychology,* vol. 102, 2014.

[23] J. Xie, G. Xu, J. Wang, M. Li, C. Han and Y. Jia, "Effects of Mental Load and Fatigue on Steady-State Evoked Potential Based Brain Computer Interface Tasks: A Comparison of Periodic Flickering and Motion-Reversal Based Visual Attention," 2016.

[24] F. Freeman, P. Mikulka, L. Prinzel and M. Scerbo, "Evaluation of an adaptive automation system using three EEG indices with a visual tracking task," *Biological Psychology,* vol. 50, no. 1, 1999.

[25] C. Berka, D. Levendowski, M. Lumicao, Alan Yau, G. Davis, V. Zivkovic, R. Olmstead, P. Tremoulet and P. Craven, "EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks," *Aviation Space Environmental Medicine,* 2007.

[26] M. Aksoy, C. Ufodiama, A. Bateson, S. Martin and A. Asghar, "A comparative experimental study of visual brain event-related potentials to a working memory task: virtual reality head-

mounted display versus a desktop computer screen," *Experimental Brain Research,* vol. 239, 2021.

[27] E. Redlinger and C. Shao, "Comparing brain activity in virtual and non-virtual environments: A VR & EEG study," *Measurement: Sensors,* vol. 18, 2021.

[28] D. Monteiro, H.-N. Liang, A. Abel, N. Bahaei and R. d. C. Monteiro, "Evaluating Engagement of Virtual Reality Games Based on First and Third Person Perspective Using EEG and Subjective Metrics," in *2018 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, 2018.

[29] A. Dey, A. Chatburn and M. Billinghurst, "Exploration of an EEG-Based Cognitively Adaptive Training System in Virtual Reality," in *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 2019.

[30] M. Nuwera, G. Comi, R. Emerson, A. Fulgsang-Fredeiksen, J.-M. Guerit, H. Hinrichs, A. Ikeda and P. Rappelsburger, "IFCN standards for digital recording of clinical EEG," *Electroencephalography and Clinical Neurophysiology,* vol. 106, no. 3, 1998.

[31] W. Gibbs, "Electricity Supply in the UK: A chronology," 1987.

[32] P. Welch, "The use of fast fourier transform for the estimation of power spectra," *IEEE transctions on audio and electroacoustics,* vol. 15, no. 2, 1967.

[33] N. Bhatia, "CET-VR: A cognitive education and training framework using virtual reality," 2022.

Optimisation of Solar Energy Use Within Communal Buildings

Rohit Thota & Mukund Murali
Department of Chemical Engineering, Imperial College London, U.K.

Abstract

Adequate energy management systems and distribution algorithms are essential within buildings and communities that share solar power, especially so during recent years where carbon emission reduction is a priority and countries around the world are facing an energy crisis. This paper addresses this problem by developing a novel optimisation model, which uses a multi-objective function formulated and solved within python. This determines which energy management system (EMS) strategy and distribution algorithm to implement to make the best use of generated PV within a building. This model is used with the Flamsteed Estate and factors in the savings residents will experience, the fraction of PV generated that is unused, and the equalised annual cost when making a decision. It is also used for cases outside of the Flamsteed Estate where the solar panel capacity has not yet been determined in order to find the optimum configuration of panels along with EMS strategies and distribution algorithms. For the Flamsteed Estate, Peer-to-Peer (P2P) trading and 80kWh of battery storage capacity with equal distribution of PV between residents results in a reduction in the building's electricity costs of 66%. Future projects on buildings of a similar size should use the same distribution algorithm with 100kW of solar panel capacity, P2P trading and 90kWh of battery storage capacity which will reduce annual electricity costs by 72%.

Keywords: *Solar PV, Energy Sharing, Energy Management System (EMS), Multi-Objective Function (MOF)*

Introduction

Building operations contribute more than 9.9 Gt of carbon emissions annually, accounting for 27% of all energy related CO_2 emissions (IEA, 2022). Decarbonisation of these sectors through the reduction of the dependence on fossil fuels is now a priority, with the Paris agreement stating buildings must be net-zero by 2050 (World Resources Institute, 2019). This pathway to a low-carbon future can be achieved with smart energy management systems in combination with on or off-site renewable energy.

Solar power and PV technology has seen the largest fall in cost of any electricity technology over the last decade with an 82% decline from 2010-2019. (IRENA, 2020) PV technologies are expected to provide 25% of the global electricity requirement by 2050 with 40% of this energy coming from PV panels integrated within buildings (Masson et al, 2019). This fall in price has led to a year-on-year increase in the rate of solar panel adoption. By 2030, the UK's solar panel capacity is expected to triple according to the IEA (IEA, 2022), hence, it is vital that adequate energy management systems (EMS) are implemented to optimise the use of household electricity. Soaring electricity prices further emphasise the need for such energy management systems allowing for resident's financial, as well as environmental concerns to be alleviated.

Through the implementation of such smart algorithms and EMS strategies, the interaction between the communal building with equipped PV and the grid should be minimised by ensuring the building load at a given time matches with the production of PV. On-site production of PV also allows residents to act as prosumers, this being an individual who both consumes and produces power. This allows for energy sharing to occur between residents through dynamic pricing and allocation according to supply-demand ratios, ultimately providing a framework which potentially improves economic performance known as Peer-to-Peer (P2P) which will be discussed further in this paper. Therefore, the study on how this solar energy use should be optimised between residents will be carried out. By better consuming PV produced onsite with EMS and battery storage, there is scope to realise considerable monetary savings and reduce dependency on the grid whilst lowering greenhouse gas emissions.

Background

The increase in PV adoption in recent years has led a number of researchers to conduct studies into optimal allocation strategies. Due to occupant's role as prosumers, they can now play an active role within the energy market and this ability to change their roles between buyers and sellers allows them to partake in an internal energy market between grid-connected peers. This new, collaborative network where the users of the grid can self-organise and trade renewable energy directly with each other without an intermediary is known as Peer-to-Peer (P2P) trading – an energy sharing model that has potential to provide significant savings. The financial benefits of such a model has been explored and it was discovered that the energy bills of participating households would see a reduction between 15.1% and 23.6% amounting to an average of £2.92 per customer over a six-month period. Additional benefits of this included balancing of the energy grid and the mitigation of transmission losses due to reduced congestion within the distribution network (Klein et al, 2019). The use of a P2P model has been further investigated by exploring internal pricing procedures based on hourly and daily forecasts in demand and electricity market prices

which ultimately saves costs for prosumers when compared to trading with the utility grid (Liu et al, 2017). In some cases where the use of battery systems can be incorporated alongside P2P, energy costs for a community were able to be reduced by 30% (Long et al, 2018).

Allocation strategies where a uniform amount of the generated renewable energy is distributed between residents is shown to encourage responsible usage and ease of interlinkage with P2P mechanisms (Syed et al, 2020). Stackelberg game approaches have also been used in which pricing and allocation is determined using an hour-ahead model. Deviations between actual energy consumption and scheduled energy consumption are then reflected within the final bill which enhances the utility of the prosumer while improving the profit for the operator (Erol et al, 2022). Similarly, it has been found by using a model predictive control which accounts for the weather forecast and changes in price in the electricity market, linear programs can provide monetary individual savings between 5.4% and 7.7% (Vand et al, 2021). Optimisation models have been developed for the operation of production and storage technologies using a 15-minute balance, however, it is suggested that smaller time intervals should be explored to provide for more granular data which will be explored in this paper (Savolainen et al, 2022). These methods, whilst improving the utility of on-site generated energy, did not present an effective way of combining the respective allocation strategies with the use of agent-based models and storage capabilities. Research also focussed on solely the benefit to operator or individual and not both parties, a crucial element to our case study where social benefit is vital.

Aim and Motivation

This paper aims to identify the best combination of EMS strategy and PV distribution algorithm for a twelve-dwelling building in Flamsteed Estate within the Royal Borough of Greenwich. Currently, the PV generated by the rooftop PV panels (Capacity = 72.5kW) is used to power the communal areas in the building, with the excess PV generated sold to the grid. Even with the most generous export tariff being £0.075/kwh (Gridcog, 2022), the council does not earn a significant revenue from doing so. The excess PV generated could instead be distributed amongst the residents of the building, who are mostly low-income families. Doing so would be beneficial not only from a social aspect, but also from an environmental standpoint as it would reduce the building's dependence on grid electricity, whose share of renewable energy is only ~40% (GOVUK, 2022), which aligns strongly with the Paris Agreement's requirement for buildings to be net-zero by 2050 and the council's own sustainability targets of reaching net-zero carbon emissions by 2030 (RBG, 2021).

This paper also generalises the model developed to identify the optimal solar panel capacity that should be installed and pairs it with the optimal EMS strategy and distribution algorithm combination for similarly sized communal buildings.

The optimal strategy will be the scenario which achieves the best mix of the following objectives:

1. Maximise collective savings per year of the residents/Minimise the collective costs per year of the residents.
2. Minimise unused electricity generated by the panels.
3. Minimise the equalised annual cost (EAC) of the investment.

Method

A: Scenarios explored

Scenario	Description
1.Equal percentage cost savings distribution with solar panels	At each 1-minute interval, PV generated is distributed proportionally to the dwelling's electricity demand as a ratio of the total building's electricity demand. Demand deficits are met by grid imports whereas surplus after allocation is exported to the grid.
2.Equal amount distribution with solar panels	At each 1-minute interval, PV generated is distributed equally amongst the residents. Demand deficits are met by grid imports whereas surplus after allocation is exported to the grid.
3.Equal percentage cost savings distribution with solar panels and Peer-to-Peer trading	At each 1-minute interval, PV generated is distributed as per scenario 1. Dwellings allocated a surplus can 'sell' their surplus to dwellings in deficit at the export to grid price. Earnings from/purchase of P2P trading is proportional to the dwelling's contribution to/demand of the total excess PV allocated. Post P2P, demand deficits are met by grid imports whereas

	surplus after allocation is exported to the grid.
4.Equal amount distribution with solar panels and Peer-to-Peer trading	At each 1-minute interval, PV generated is distributed as per scenario 2. Dwellings allocated a surplus can 'sell' their surplus to dwellings in deficit at the export to grid price. Earnings from/purchase of P2P trading is divided equally amongst the 'sellers/buyers'. Post P2P, demand deficits are met by grid imports whereas surplus after allocation is exported to the grid.
5.Equal percentage cost savings distribution with solar panels and Li-ion battery	PV generated is distributed as per scenario 1. Excess PV generated is stored in the battery during hours of PV generation and distributed when no PV is generated. Dwellings 'purchase' the battery power as per demand at the grid export price from the council. Demand deficits are met by grid imports whereas surplus after allocation is exported to the grid.
6.Equal amount distribution with solar panels and Li-ion battery	PV generated is distributed as per scenario 2. Excess PV generated is stored in the battery during hours of PV generation and distributed when no PV is generated. Dwellings 'purchase' the battery power as per demand at the grid export price. Demand deficits are met by grid imports and surplus after allocation is exported to the grid.
7.Equal percentage cost savings distribution with solar panels, Peer-to-Peer trading and Li-ion battery	Post P2P surplus is stored in the battery which discharges when there is no PV generation. Distribution algorithm follows the same rules as other equal percentage savings models.
8.Equal amount savings distribution with solar panels, Peer-to-Peer trading and Li-ion battery	Post P2P surplus is stored in the battery which discharges when there is no PV generation. Distribution algorithm follows the same rules as other equal amount models.

B: Collection of dwelling demand data and solar panel output by minute

Occupants have a diverse range of energy consumption patterns which is modelled using an open-sourced thermal-electrical demand model (McKenna, Thomson and Barton, 2015) that integrates three models – domestic occupancy (Richardson et al., 2008), domestic lighting demand (Richardson et al., 2009) and domestic electricity use (Richardson et al., 2010) – designed for low-voltage network analysis for houses in India and the UK. This model was made publicly available by the *Centre for Renewable Energy Systems Technology (CREST)*. The dwelling parameters i.e. number of residents, occupancy at each time and appliance distribution were selected stochastically and per-minute electricity data was collected for three representative weeks (One week from January, one week from July and one week from September - Winter, Summer and Autumn in the year. Figure 1 below shows the trend between power demand by the building over a day. The trend over all the days in the year follows a similar pattern.

Real-world weather data for the representative weeks is also used to predict patterns of PV production and account for the seasonality and intermittency of solar power. Per-hour PV generation data was obtained from an open source platform (Pfenninger et al, 2016) and an equal distribution of PV generation was assumed for each minute of every hour.

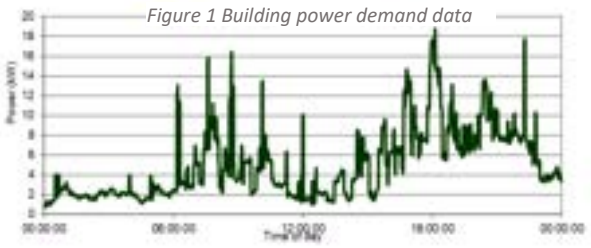


Figure 1 Building power demand data

C:Modelling of EMS Strategies

The models developed to simulate the different strategies use Mixed Integer Linear Programming (MILP). They have been formulated and solved on Python using a graphical method.

Multi-Objective Function

$$\min_{P,B_{max}} \sum_{t=1}^{525600} \sum_{d=1}^{12} \left[\left(w_1 \frac{c_{t,d}}{\sum_{d=1}^{12} bc_{t,d}} \right) + \left(w_2 \frac{\sum_{d=1}^{12} ex_{t,d}}{PV_{out}} \right) + \left(w_3 \frac{eac}{\sum_{d=1}^{12} Sav_{t,d}} \right) \right] \quad (1)$$

$$c_{t,d} = \left(Imp_{t,d} \times p_{imp} \right) + \left(E_{bat-A_{t,d}} \times p_{bat_{t,d}} \right) + \left(\left(E_{P2P-A_{t,d}} - E_{P2P-C_{t,d}} \right) \times p_{P2P} \right) \quad (2)$$

$$bc_{t,d} = \left(dem_{t,d} \times p_{imp} \right) \quad (3)$$

$$exp_{t,d} = \begin{cases} \left\{ \left(E_{PV_{t,d}} - dem_{t,d} - E_{P2P-C_{t,d}} \right) \right. & for\ scenarios\ w/o\ battery \\ \left(\sum_{d=1}^{n(D)} E_{bat-C_{t,d}} - B_{max} \right) & for\ scenarios\ w/\ battery \end{cases} \quad (4)$$

$$imp_{t,d} = \left(dem_{t,d} - E_{PV_{t,d}} - E_{bat-A} - E_{P2P-A_{t,d}} \right) \quad (5)$$

$$EAC = EAC_{panel} + EAC_{bat} \quad (6)$$

$$AF_{panel} = \frac{\left[1 - \frac{1}{(1+cc)^{L_{panel}}} \right]}{cc} \quad (7)$$

$$AF_{battery} = \frac{\left[1 - \frac{1}{(1+cc)^{L_{battery}}} \right]}{cc} \quad (8)$$

$$EAC_{panel} = \left(\frac{P_{cost}}{AF_{Panel}} \right) + \left(\frac{PM_{cost}}{n(P)} \right) \quad (9)$$

$$EAC_{battery} = \left(\frac{B_{cost}}{AF_{battery}} \right) \quad (10)$$

Equation (1) is a weighted sum of electricity costs incurred by the building as a fraction of the cost incurred in the base case scenario, total electricity generated by the solar panels exported to the grid as a fraction of the electricity generated by the building and equalised annual cost of the investment as a fraction of the total savings experienced by the residents of the building. w_1, w_2 and w_3 are weights assigned to each of the objectives in the cost function and are proportional to the priority of each of the objectives. The function is heavily weighted towards minimising the residents' electricity costs as social benefit is the greatest priority for the council and building emissions from grid imports are consequently reduced. The aim of minimising this function to strike the ideal balance between the three objectives listed. The score obtained from this function will be the primary comparison metric used to determine the best scenario to be adopted for the current building and future projects.

The decision variables across all scenarios are the solar panel system's capacity in kW. For the strategies involving the use of a battery, the maximum battery capacity in kWh is another decision variable considered. The Solar panel capacity for the Flamsteed Estate specific case is set at 72.5kW as this is the capacity of the solar panels that is currently in place. These two variables do not directly appear in the objective function however are primarily responsible for the values of the variables that are present in it.

Distribution methods

Equal percentage savings Panel PV Allocation method

$$E_{PV\,d,t} = \frac{dem_{d,t}}{\sum_{d=1}^{12} dem_{d,t}} \times PV_{out} \quad (11)$$

Equal amount Panel PV Allocation method

$$E_{PV\,d,t} = \frac{PV_{out}}{n(D)} \quad (12)$$

Equal percentage savings Peer2Peer Allocation method

$$E_{P2P-A\,d,t} = \sum_{d=1}^{12} [(E_{PV})_{d,t} - (dem)_{d,t}] \times \frac{dem_{d,t}}{\sum_{d=1}^{12} dem_{d,t}} \quad (13)$$

Equal percentage savings Peer2Peer Contribution method

$$E_{P2P-C\,d,t} = \frac{[(E_{PV})_{d,t} - (dem)_{d,t}]}{\sum_{d=1}^{12} [(E_{PV})_{d,t} - (dem)_{d,t}]} \times E_{P2P-A\,d,t} \quad (14)$$

Equal Amount Peer2Peer Allocation method

$$E_{P2P-A\,d,t} = \frac{\sum_{d=1}^{12} [(E_{PV})_{d,t} - (dem)_{d,t}]}{n(D)_{underallocated}} \quad (15)$$

Equal Amount Peer2Peer Contribution method

$$E_{P2P-C\,d,t} = = \frac{E_{P2P-A\,d,t}}{n(D)_{overallocated}} \quad (16)$$

Battery Energy Allocation method (Panels only and Peer2Peer)

$$\sum_{d=1}^{n(D)} E_{Bat-A\,d,t} = \begin{cases} \sum_{d=1}^{n(D)} dem_{d,t} \ if \left(PV_{out_t} = 0 \right) and \left[B_t - \sum_{d=1}^{n(D)} dem_{d,t} \geq 0.2 \times B_{max} \right] \\ 0.2 \times B_{max} \ if \left(PV_{out_t} = 0 \right) and \left[B_t - \sum_{d=1}^{n(D)} dem_{d,t} < 0.2 \times B_{max} \right] \\ 0 \ if \ Building\ PV\ Output_t > 0 \end{cases} \quad (17)$$

Constraints

Demand Balance Constraint

This constraint is in place to ensure that all the dwellings' electricity demand is met at all time intervals.

$$dem_{d,t} = E_{PV\,d,t} + \left(E_{P2P-A\,d,t} - E_{P2P-C\,d,t} \right) + (E_{bat-A\,d,t} - E_{bat-c_{d,t}}) + Imp_{d,t} \quad (18)$$

Energy Balance Constraint

This constraint is in place to ensure the model satisfies the law of conservation of energy.

$$PV_{out} = \sum_{d=1}^{12}(E_{PV_{d,t}} + exp_{d,t}) \quad (19)$$

Battery State of Charge Constraint

The operating Depth of Discharge (DoD) of the battery in this model was assumed to be 80%.

$$0.2 \leq SOC \leq 1.0 \quad (20)$$

$$0.2 \times B_{max} \leq B_{cap_t} \leq B_{max} \quad (21)$$

Initial Conditions

Battery Level Initial Condition

$$B_{t=0} = B_{max} \quad (22)$$

Results and Discussion

Solar Panels Only

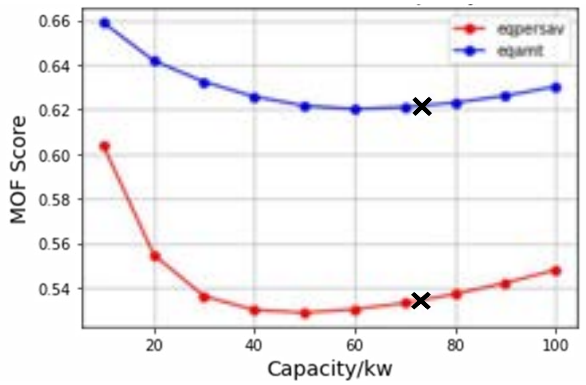


Figure 2 MOF Score against Solar Capacity for a General Case. Note the 'X' marks the Flamsteed Case

Allocation		Flamsteed	General
Eq. % Cost Savings	Panels Capacity/kW	72.5	50
	MOF Score	0.534	0.530
Eq. Amt Distribution	Panels Capacity/kW	72.5	60
	MOF Score	0.621	0.619
Eq. % Cost Savings	EAC/£	5760	3970
Eq. Amt Distribution	EAC/£	5760	4760

Table 1 Allocation method for Flamsteed and General Case showing Panel Capacity, MOF Score and EAC

The equal percentage cost savings distribution algorithm results in a lower MOF score across all solar panel capacities trialled in this study, suggesting that it is the better of the two distribution algorithms for this EMS strategy. Additionally, the model suggests that the ideal panel capacity that minimises the MOF score to 0.530, is 50kW for the equal percentage cost savings distribution algorithm. The ideal capacity is 60kW for the equal amount cost distribution algorithm which minimises the MOF score to 0.619. These are both lower than the 72.5kW capacity that is installed on the building being modelled within the Flamsteed Case.

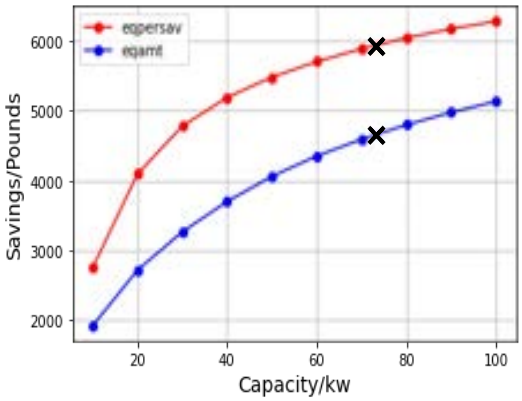


Figure 3 Savings against Solar Capacity for a General case. Note the 'X' marks the Flamsteed Case

Figure 3 above shows that the total savings experienced by the building over the year increases logarithmically with solar panel capacity. Across all solar panel capacities trialled, the equal percentage cost savings distribution algorithm results in higher total annual building savings, and as a result, lower total electricity costs for the residents in the building. Consequently, it will also result in a lower equalised annual cost as a fraction of savings as the equalised annual cost is independent of the distribution algorithm for the same solar capacity.

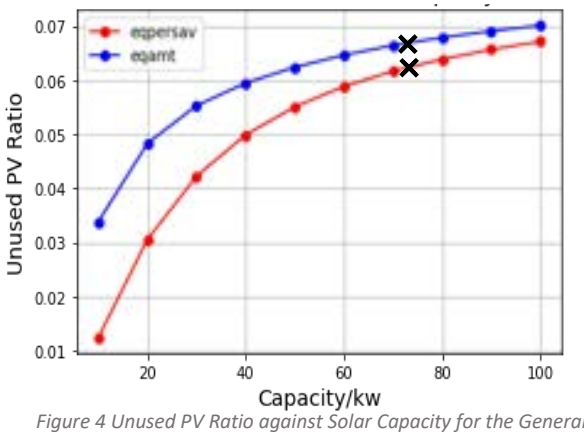


Figure 4 Unused PV Ratio against Solar Capacity for the General Case. Note the 'X' marks the Flamsteed Case

Figure 4 above shows that the ratio of electricity generated that goes unused by the building i.e., exported back to the grid, also increases logarithmically with solar panel capacity. Across all solar panel capacities trialled, the equal percentage cost savings distribution algorithm results in a lower fraction of generated electricity that goes unused.

The logarithmic nature of the increase of both building annual savings and unused PV ratio in the capacities tested means that rate at which they

increase with panel capacity i.e., the slope of the graphs, decreases with time. This means that the increase in savings slows down with an increasing solar panel capacity and there will be a capacity above which the increase in savings experienced does not outweigh the penalty of the rise in the equalised annual cost, which increases linearly with solar panel capacity as seen in figure 5 below. This linear trend is also observed for the battery EAC in the scenarios including a battery. The multi-objective function accounts for this with the equalised annual cost as a fraction of savings term and thereby concludes that a 50kW panel for the equal percentage cost savings distribution method and a 60kw panel for the equal amount distribution method best meets the objectives of this report.

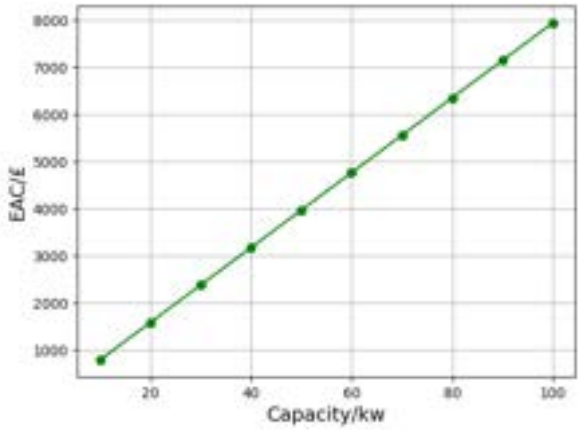


Figure 5 EAC against Solar Capacity

The reason why the equal percentage savings distribution algorithm performs better in the above two metrics is due to the solar power generated being distributed proportionately to the dwelling's demand. This results in the solar power generated being used more efficiently as opposed to the equal amount distribution algorithm which sees smaller households, which demand less electricity on average, being heavily overallocated and larger households, which demand more electricity on average, being heavily under allocated. This mismatch in allocation results in a greater fraction of the PV being exported in the equal amount distribution algorithm, reflected by the trend in figure 4, leading to a greater amount of electricity having to be imported from the grid, which leads to greater costs and lower savings from the residents, reflected by the trend in figure 3**.**

Solar Panels with Battery Storage

The equal percentage cost savings distribution algorithm results in a lower MOF score across all battery capacities trialled for the Flamsteed Estate specific case (72.5kW Solar Panel Capacity in place) and across all solar panel and battery capacity combinations trialled for the generalised case in this study, suggesting that it is the better of the two distribution techniques for this

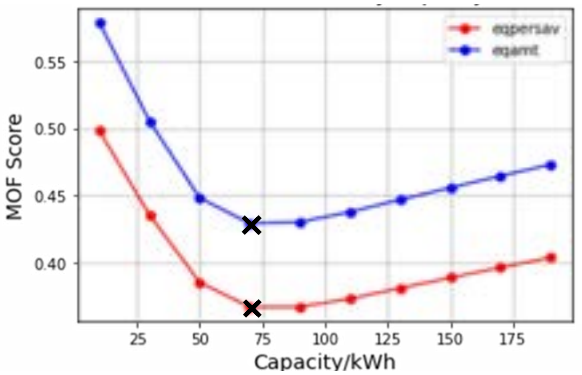


Figure 6 MOF Score against Battery Capacity for the Flamsteed Case. Note the 'X' marks the minimum point

Allocation		Flamsteed	General
Eq. % Cost Savings	Panels Capacity (kW)	72.5	100
	Battery Capacity (kWh)	70	90
	MOF Score	0.367	0.358
Eq. Amt Distribution	Capacity (kW)	72.5	100
	Battery Capacity (kWh)	70	90
	MOF Score	0.429	0.387
Eq. % Cost Savings	EAC	10100	13500
Eq. Amt Distribution	EAC	10100	13500

Table 2 Allocation method for Flamsteed and General Case showing Panel & Battery Capacity, MOF Score and EAC

EMS strategy. Additionally, the model suggests that the ideal panel and battery capacity that minimises the MOF score to 0.358 and 0.387 is 100kW and 90kWh respectively for both distribution algorithms.

Figure 7 below shows that the total savings experienced by the building over the year increases logarithmically, just like in the solar panels only EMS strategy, with battery capacity for the Flamsteed Estate specific case. A similar trend can be expected for every solar panel capacity trialled.

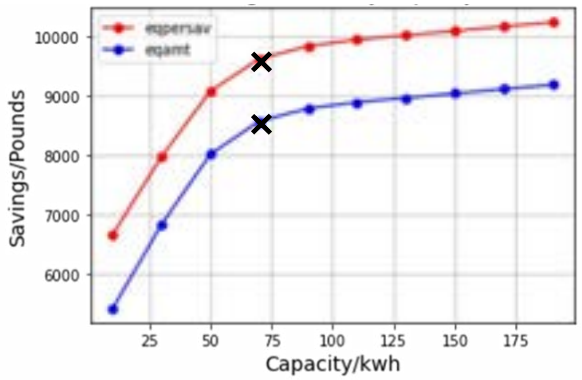


Figure 7 Savings against Battery Capacity for the Flamsteed Case. Note the 'X' marks the ideal case for Flamsteed

Across all battery and panel capacity and combinations trialled, the equal percentage cost savings distribution algorithm results in higher total annual building savings, and as a result, lower total electricity costs for the residents in the building.

Consequently, it will also result in a lower equalised annual cost as a fraction of savings as the equalised annual costs is independent of the distribution algorithm for the same solar capacity as per equation 9.

Figure 8 below shows that the ratio of electricity generated that goes unused by the building i.e., exported back to the grid, also decreases with battery capacity for the Flamsteed Estate specific case due to fewer exports to the grid. A similar trend can be expected for every solar panel capacity trialled. Across all combinations of solar panel and battery capacities trialled, the equal percentage cost savings distribution algorithm results in a lower fraction of generated electricity that goes unused.

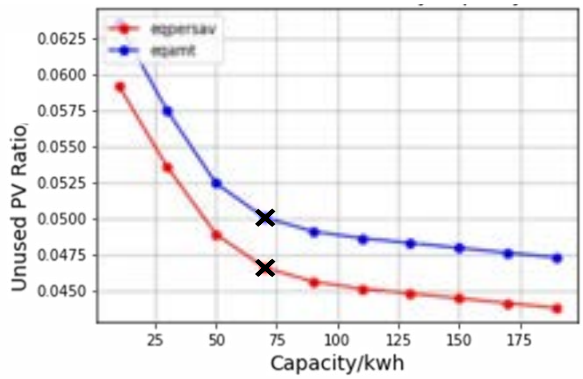


Figure 8 Unused PV Ratio against Battery Capacity. Note the 'X' marks the ideal case for Flamsteed

The logarithmic nature of the increase of building annual savings and decrease in unused PV ratio in the capacities tested means that rate at which the benefit experienced increases with battery capacity i.e., the absolute value of the slope of the graphs decreases with capacity. This means that the rate of increase in savings and decrease in the fraction of unused PV slows down with increasing the capacity and there will be a capacity above which these benefits experienced do not outweigh the penalty of the rise in the equalised annual cost, which increases linearly with solar panel and battery capacity. The multi-objective function accounts for this with the equalised annual cost as a fraction of savings term and concludes that a 100kw panel with a 90kWh battery for both distribution algorithms best meet the objectives of this report.

The equal percentage cost savings algorithm performs better across both metrics in figures 7 and 8 for the same reasons as that of the solar panel only model. The battery charges and discharges in the same manner in both distribution algorithms, hence the same explanation applies here too. The only difference being the overall savings experienced by the residents which is greater, and the fraction of PV generated getting exported which is lower. This is because a portion of what would otherwise be exported is stored in the battery and redistributed amongst the dwellings when there is no PV generation and within the confines of the battery's operating constraints. This further reduces the dwelling's reliance on expensive grid electricity.

Solar Panels with Peer-to-Peer trading

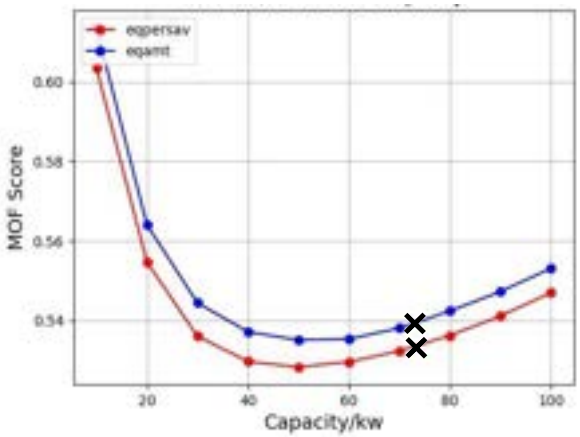


Figure 9 MOF Score against Solar Capacity for the General Case. Note the 'X' marks the Flamsteed Case

The MOF scores for both distribution algorithms are very similar to each other in figure 9, with the equal percentage cost savings algorithm having a very slight advantage across all capacities. However, considering the various sources of errors and the assumptions made when building the model, which are highlighted in the error analysis section, the difference between the MOF scores is not high enough to conclusively decide that this is the best algorithm for this EMS strategy. The final decision should then be made considering the ease of implementation of the two distribution algorithms and the fairness of distribution. In any case, the benefits observed from either distribution algorithm selection will be very similar, meaning the differences between each option will be negligible.

Allocation		Flamsteed	General
Eq. % Cost Savings	Panels Capacity (kW)	72.5	50
	MOF Score	0.533	0.520
Eq. Amt Distribution	Panels Capacity (kW)	72.5	50
	MOF Score	0.529	0.525
Eq. % Cost Savings	EAC	5756	3970
Eq. Amt Distribution	EAC	5760	3970

Table 3 Allocation method for Flamsteed and General Case showing Panel Capacity, MOF Score and EAC

Additionally, the model suggests that the ideal panel capacity that minimises the MOF score is 50kW for both distribution algorithms, which is lower than the 72.5kW capacity that is installed on the building being modelled.

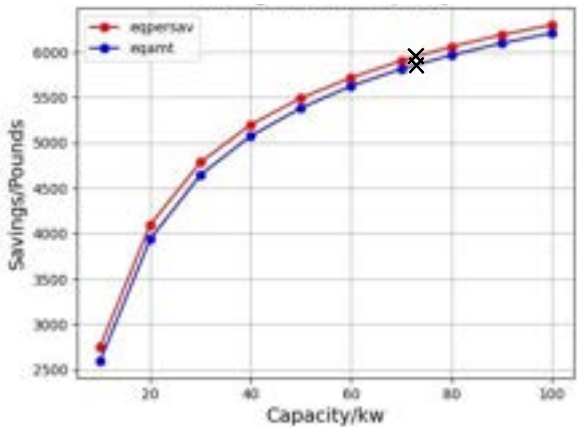


Figure 10 Savings against Solar Capacity. Note the 'X' marks the Flamsteed Case

Figure 10 above shows that the total savings experienced by the building over the year increases logarithmically with solar panel capacity. Across all solar panel capacities trialled, both algorithms result in similar annual building savings, with the difference being too low to conclusively determine that one distribution algorithm outperforms the other. Consequently, the equalised annual cost as a fraction of savings will be very similar for the same solar capacity.

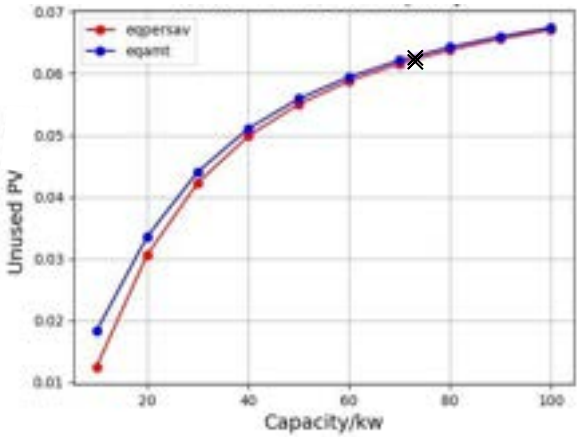


Figure 11 Unused PV ratio against Solar Capacity for the General Case. Note the 'X' marks the Flamsteed Case

Figure 11 above shows that the ratio of electricity generated that goes unused by the building i.e. exported back to the grid also increases logarithmically with solar panel capacity. Again, the difference between the two distribution algorithms is very minute.

The logarithmic nature of the increase of both building annual savings and unused PV ratio in the capacities tested means that rate at which they increase with panel capacity i.e., the slope of the graphs, decreases with time. This means that the increase in savings slows down with increasing the solar panel capacity and there will be a capacity above which the increase in savings experienced does not outweigh the penalty of the rise in the equalised annual cost, which increases linearly with solar panel capacity as seen in figure 5. The multi-objective function accounts for this with the equalised annual cost as a fraction of savings term and thereby concludes that a 50kw panel for both distribution algorithms best meets the objectives of this report.

When comparing this EMS strategy to the solar panel only strategy, we see that there is a very small improvement when peer-to-peer trading is added for the equal percentage cost savings algorithm while the equal amount distribution algorithm sees significant improvements in all three metrics of the multi-objective function. This is because the problem of over-allocation and under-allocation that weakens the equal amount distribution algorithm in the solar panel only strategy is corrected in this strategy. The over-allocated households sell the excess electricity that was allocated to the under-allocated households at a price that is lower than the grid import price. Not only does this reduce the fraction of electricity generated that gets exported, but it has a great impact on the savings of the residents as they earn a revenue for using less electricity than what was allocated to them/are able to meet their deficit at a much lower price than before.

Solar Panels with Battery and Peer-to-Peer trading

At lower battery capacities, we see in figure 12 the two distribution algorithms have almost identical MOF scores for the Flamsteed Estate specific case (72.5kW Solar Panel Capacity in place). From a battery capacity of 60kWh onwards, we see the equal amount distribution model has consistently

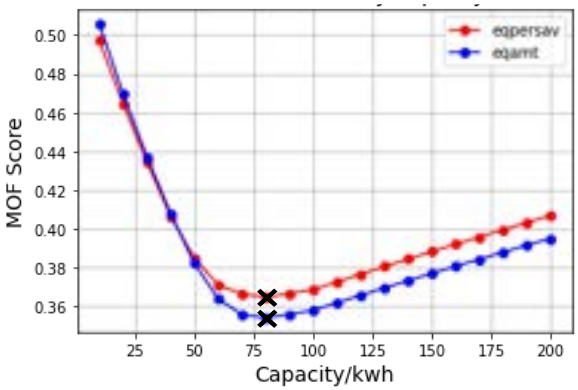


Figure 12 MOF Score against Battery Capacity for Flamsteed Case. Note the 'X' marks the lowest MOF score for the Flamsteed Case.

lower MOF scores, albeit by a relatively small margin. We expect this trend to follow for all the solar panel sizes trialled in this study.

Allocation		Flamsteed	General
Eq. % Cost Savings	Panels Capacity (kW)	72.5	100
	Battery Capacity (kWh)	80	90
	MOF Score	0.363	0.352
Eq. Amt Distribution	Panels Capacity (kW)	72.5	100
	Battery Capacity (kWh)	80	90
	MOF Score	0.358	0.317
Eq. % Cost Savings	EAC	10700	13500
Eq. Amt Distribution	EAC	10700	13500

Table 4 Allocation method for Flamsteed and General Case showing Panel & Battery Capacity, MOF Score and EAC

The model suggests that the ideal panel and battery capacity that minimises the MOF score is 100kW and 90kWh respectively for both algorithms.

This EMS strategy combines the benefits achieved from all the strategies previously mentioned and therefore will result in greater annual savings for the building as a whole, a lower fraction of unused electricity that was generated by the solar panels and a lower equalised annual cost to annual building savings ratio for the reasons mentioned in the previous strategies. These combine to give this strategy a lower MOF score than the other strategies for both of the distribution algorithms.

When comparing the MOF scores for all scenarios, we see that for the Flamsteed Case using an equal amount distribution algorithm with P2P trading and a battery of 80kWh capacity has the lowest score of 0.358. This allows total building energy costs to be reduced by 66% and meets all three objectives the best. The Flamsteed Estate should therefore adopt this scenario for their building.

For the general case, an equal amount distribution algorithm is used with 100kW of solar capacity, 90kWh of battery capacity and P2P trading resulting in the lowest MOF score of 0.317. Future projects of similar sized communal buildings should adopt this approach.

Error Analysis:

Within this report several assumptions have been made which can have an impact on our final findings. Firstly, in order to get PV generated per-minute, hourly electricity generation data was obtained from a 2019 open-source database and divided by 60 as minute-by-minute data is not available. Given the intermittency of solar generation and changes within our climate, these results may not accurately represent the true PV generation at the time of implementation. There is also an assumption that the solar panels and battery work without loss in performance and degradation which can affect the efficiency of the model in future years after implementation.

The crest model used to obtain per-minute dwelling data is modelled using data from 2015. Electricity consumption patterns may have changed since then which can potentially affect the what the ideal scenario may be.

Conclusion

The work presented in this paper proposes an optimum EMS strategy and distribution algorithm for solar energy generated in a communal building and buildings of similar sizes. This was done by developing a multi-objective function which made use of real-world weather data in addition to dwelling electricity consumption data from the CREST model. This model was tested on a building in Flamsteed Estate which had 12 dwellings within it. The optimal scenario deduced the total annual electricity costs for residents within the building can be reduced by 66% given that the building has 72.5kW of solar panel capacity and there is a battery with 80kWh capacity. For future projects for communal buildings of a similar size, resident's annual electricity costs could be reduced by 72% given 100kW of solar panel capacity and a battery with 90kWh capacity.

Future Work

To further increase the scope of this project, the model should be expanded to deduce the optimal scenario for buildings of different sizes and types. This would also allow other areas to be explored outside of social housing where energy management systems will be necessary such as hotels with on-site PV. The multi-objective function could be adapted with different weightings to prioritise various areas such as social and economic benefit. In some cases, a fixed tariff may not be used, hence exploring these cases with a variable tariff is also recommended as it may affect the optimal scenario given the price of electricity at a certain time. With reduction in carbon emissions also being a consideration in this project, a quantitative measure of carbon emissions reduced as a result of the implementations of these scenarios can also investigated. Finally, any embodied emissions in the creation of solar panels and batteries has also not been considered within this project as they often involve mining rare earth metals and other energy intensive materials (Kilgore, 2022) which can be done so for future projects.

Acknowledgements

We are very thankful for Dr. Edward J O'Dwyer and Max Bird for their continuous support and insight that was provided to us throughout the project.

Nomenclature

P	Solar Panel System Capacity / kW
P_{cost}	Solar Panel System Installation Cost / £
PM_{cost}	Solar Panel System maintenance cost / £
$n(P)$	Number of 325W Solar Panels
EAC	Equalised Annual Cost / £
AF	Annuity Factor
cc	Cost of Capital = 0.04
$L_{battery}$	Battery Lifetime = 10 years
L_{panel}	Solar Panels Lifetime = 25 years
B_{max}	Maximum Battery Capacity / kWh
B_{cost}	Battery installation Cost / £
B_t	Battery Capacity at a specified time interval / kWh
t	1 min time intervals; $t \in T$
T	$T = [1,2,\ldots,525600]$
$n(T)$	Number of 1 min time intervals in a year
d	Dwelling Index; $d \in D$
D	$D = [1,2,\ldots,12]$
$n(D)$	Number of dwellings in the building
dem	Dwelling Electricity demand / kWh
E_{imp}	Electricity Imported from grid / kWh
E_{PV}	Electricity generated by solar panels allocated / kWh
$E_{P2P\text{-}A}$	Electricity allocated from Peer-to-Peer trading / kWh
$E_{P2P\text{-}C}$	Electricity contributed towards Peer-to-Peer trading / kWh
$E_{bat\text{-}A}$	Electricity purchased battery / kWh
$E_{bat\text{-}C}$	Electricity contributed to battery / kWh
c	Cost of electricity in the scenario tested / £
bc	Base case electricity costs / £
PV_{out}	Total Electricity generated by solar PV panels on building rooftop / kWh
EAC	Equalised annual cost of the investment in scenario implemented / £
Sav	Savings experienced on electricity costs relative to base case cost of electricity in the scenario tested / £
Imp	Electricity Imported from the Grid / kWh
exp	Electricity Exported to Grid / kWh
$Bat. Purchase$	Electricity Purchased from the Battery / £
p_{imp}	Grid import price / £. Taken to be £0.34/kWh
p_{exp}	Grid export price / £. Taken to be £0.075/kWh
p_{bat}	Battery electricity purchase price / £. Set at £0.075/kWh
p_{p2p}	Peer-to-Peer trading price / £. Set at £0.075/kWh

References

McKenna, E. and Thomson, M. (2016) "High-resolution stochastic integrated thermal–electrical domestic demand model," Applied Energy, 165, pp. 445–461. Available at: https://doi.org/10.1016/j.apenergy.2015.12.089.

Delmastro, C. (2022) Buildings – analysis, IEA. Available at:https://www.iea.org/reports/buildings (Accessed: December 10, 2022).

World Resources Institute Research (2019) Zero carbon buildings are possible where you might least expect them. Available at: https://www.wri.org/news/release-new-research-shows-zero-carbon-buildings-are-possible-where-you-might-least-expect (Accessed: December 10, 2022).

IRENA (2020), Renewable Power Generation Costs in 2019, International Renewable Energy Agency, Abu Dhabi, pp. 15,16

Masson, G. et al. (2019) "A snapshot of global PV markets - the latest survey results on PV markets and policies from the IEA PVPS programme in 2018," 2019 IEEE 46th Photovoltaic Specialists Conference (PVSC) [Preprint]. Available at: https://doi.org/10.1109/pvsc40753.2019.8981142.

IEA (2022), Renewables 2022, IEA, Paris https://www.iea.org/reports/renewables-2022 , License: CC BY 4.0

Pires Klein, L. et al. (2019) "A novel peer-to-peer energy sharing business model for the Portuguese Energy Market," Energies, 13(1), p. 125. Available at: https://doi.org/10.3390/en13010125.

Liu, N. et al. (2017) "Energy-sharing model with price-based demand response for microgrids of peer-to-peer prosumers," IEEE Transactions on Power Systems, 32(5), pp. 3569–3583. Available at: https://doi.org/10.1109/tpwrs.2017.2649558.

Long, C. et al. (2018) "Peer-to-peer energy sharing through a two-stage aggregated battery control in a community microgrid," Applied Energy, 226, pp. 261–276. Available at: https://doi.org/10.1016/j.apenergy.2018.05.097.

Syed, M.M., Morrison, G.M. and Darbyshire, J. (2020) "Energy Allocation Strategies for common property load connected to shared solar and battery storage systems in Strata Apartments," Energies, 13(22), p. 6137. Available at: https://doi.org/10.3390/en13226137.

Erol, Ö. and Başaran Filik, Ü. (2022) "A Stackelberg game approach for energy sharing management of a microgrid providing flexibility to entities," Applied Energy, 316, p. 118944. Available at: https://doi.org/10.1016/j.apenergy.2022.118944.

Vand, B. et al. (2021) "Optimal Management of Energy Sharing in a community of buildings using a model predictive control," Energy Conversion and Management, 239, p. 114178. Available at: https://doi.org/10.1016/j.enconman.2021.114178.

Savolainen, R. and Lahdelma, R. (2022) "Optimization of renewable energy for buildings with energy storages and 15-minute power balance," Energy, 243, p. 123046. Available at: https://doi.org/10.1016/j.energy.2021.123046.

Solar export tariffs in GB - Gridcognition (2022) Solar Export Tariffs in GB. Available at: https://gridcognition.com/solar-export-tariffs-in-gb/ (Accessed: December 11, 2022).

GOV.UK 2022. Energy Trends UK, April to June 2022. [online] Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1107502/Energy_Trends_September_2022.pdf (Accessed: December 11, 2022)

Royal Borough of Greenwich (no date) Carbon neutral plan, Carbon neutral plan | Royal Borough of Greenwich. Available at: https://www.royalgreenwich.gov.uk/carbonneutralplan#:~:text=The%20threat%20to%20our%20climate,zero%20carbon%20emissions%20by%202030. (Accessed: December 12, 2022).

Pfenninger, S. and Staffell, I. (2016) "Long-term patterns of European PV output using 30 years of validated hourly reanalysis and Satellite Data," Energy, 114, pp. 1251–1265. Available at: https://doi.org/10.1016/j.energy.2016.08.060.

Kilgore, G. (2022) Carbon Footprint: Solar panel manufacturing in 1 simple explanation, 8 Billion Trees: Carbon Offset Projects & Ecological Footprint Calculators. Available at: https://8billiontrees.com/solar-panels/carbon-footprint-solar-panel/ (Accessed: December 13, 2022)

Hasan Ahmed and Muhammed Imdad Kadir
Department of Chemical Engineering, Imperial College London, U.K.

Abstract

As energy demands rise in developing countries, it is becoming increasingly important to transition to renewable sources of energy in order to meet the demands in a sustainable manner. Lao PDR is aware of this and has introduced energy policies to commit to increasing the share of renewable energy in the country's total energy supply. With agriculture being the main sector in the Laotian economy, there is potential to produce biomass that can be used as feedstock to generate energy. This study presents the energy generation potential of cassava, sugarcane, and maize as energy crops by mapping the crops onto suitable land in Lao PDR using QGIS. Several conversion technologies were analysed, and granular-level comparisons were made to identify optimum energy crop choice. In addition, the energy potential from the residues of currently grown rice, cassava, sugarcane, and maize crops were investigated. Future impacts of climate change scenarios on the yield of potential energy crops were also considered. The results of the study found that a distribution of cassava and sugarcane generated the highest potential energy of 66.5 billion kWh, and an additional 20 billion kWh can be produced from existing residue. Maps of these results were created, which can be used when assessing the implementation of biomass energy systems in Lao PDR. As this study was primarily focused on energy potential, economic and other considerations must be made to identify the optimum strategy for the country.

1 Introduction

Global energy supply is currently dependent on fossil fuels, accounting for 82% of primary energy use in 2021 [1]. Fossil fuels are finite resources which produce greenhouse gas emissions, causing global climate change. Therefore, it is crucial that countries focus on the development of renewable energy sources. Developing countries such as Lao PDR are experiencing significant increases in energy demand due to economic and social developments. Lao PDR has been extending its electricity grid to bring electricity to more remote areas, increasing energy demand. Total energy consumption in Lao PDR is expected to increase at an average of 4.7% per year over 2015 to 2040 [2]. Lao PDR has introduced energy policies, including reducing its fossil fuel consumption and increasing the share of renewable energy in total energy supply by 30% in 2030 [2] and achieving electrification of the whole country. This places importance for Lao PDR to consider renewable energy alternatives such as biomass energy in order to reach the country's goals.

Biomass energy is energy generated from biological material, such as plants. It currently accounts for around 12% of the world's final energy demand [3]. Biomass is a renewable energy source that produces significantly less emissions than fossil fuels and would reduce the country's dependence on importing fossil fuels. Other factors when considering using biomass is that it provides social and economic benefits especially in rural areas. Biomass feedstock from residues of crops is abundant in rural areas of Lao PDR; if bioenergy systems are introduced, this can create employment in rural areas and reduce the waste produced.

Lao PDR has an opportunity to utilise its current agricultural production and land to implement biomass energy production as a component of its energy strategy. This report aims to address the following three main goals using a map-based approach:

1. Lao PDR's potential for bioenergy production through existing agricultural residues;
2. Lao PDR' potential for bioenergy production through growing energy crops, considering optimum crop growth, the conversion processes available, food security, land preservation and other factors.
3. Investigate effects of climate change on future energy crop yields to ensure future feasibility.

By meeting these goals, this report aims to perform a national-scale analysis of Lao PDR to provide insights for the biomass energy strategy. The maps generated can also provide a foundation for future work in this field.

2 Background

2.1 Literature Review

Bioenergy potential from existing crop residue can be estimated by utilising crop production data found in FAOSTAT and performing calculations. A study published in 2018 did this approach for Lao PDR [4]. This however does not provide geo-spatial data of the production of these crops. This study aimed to use a map-based approach as it provides the opportunity in future work to assess location-dependent factors that are necessary when considering the implementation of bioenergy production. These factors include spatial awareness of existing grid lines, transportation links, costs, energy use, and others.

GIS-based approaches have been used previously for the evaluation of bioenergy potential through energy crop growth. In 1980, one of the earliest applications of this approach, woody biomass production potential in southeast United States was analysed by Ranney and Cushman [5]. County-level maps were produced to show the potential areas of biomass supply from woody crops by analysing land availability, soil conditions, and woody crop productivity. More recent studies have used more sophisticated GIS-based approaches. Miscanthus

production in England for the purpose of bioenergy potential was analysed by taking into account both spatial supply and demand relationships [6]. Factors such as the spatial energy use in different regions and the cost of transporting biomass feedstock was analysed to determine efficient usage of the distributed feedstock. To identify areas suitable for Miscanthus production, the approach excluded areas that did not meet the criteria for production such as unsuitable land types, natural habitats, water bodies, urban areas and others.

GIS-based approaches have also been used for the evaluation of bioenergy potential through agricultural residue and waste. Beccali et al. performed analysis of bioenergy potential in Sicily, Italy, by using GIS to identify the potential areas for residue collection from the pruning of olive groves, vineyards, and other agricultural crops [7]. The database used included data on land cover, land use, regional cartography, climatic data, and other factors. With this GIS-map generated, potential areas of residues from chosen crops were identified. With the areas highlighted, yield coefficients of the various crops were assumed and used in calculating the theoretical potential productivity. The bioenergy potential in Uganda was estimated using a GIS-based approach by Barata [8]. In the study, land available for the growth of energy crops was mapped, accounting for food security by excluding land already being used for crop growth. Selected crops were then evaluated by splitting into crop parts and calculating bioenergy potential using different conversion processes. Maps were then generated of individual crop parts with the highest energy potentials. Limitations of the study include generating maps on an individual crops basis and not looking at the crops holistically to identify differences in regions.

This report aims to build on the foundations of prior GIS-based approaches but include additional results that are critical in understanding to aid in the implementation of biomass energy. This includes evaluating and comparing energy potentials of different crops in the same area to identify optimum crop choice. Various climate change scenarios were also analysed to take into consideration the effects on crop production.

2.2 Workflow

The production of biomass has been addressed from two sources. The first is growth of energy crops, which are crops grown for the purpose of energy production. The second is agricultural residues, the waste materials from existing food crop production.

2.2.1 Energy Crops

Energy crops are plants grown specifically for the purpose of producing bioenergy which has three forms: direct combustion of biomass, biogas, and biofuels. The Lao Institute for Renewable Energy has determined jatropha curcas to be the most likely commercial bio-fuel crop in Lao PDR [9]. Additionally, saccharose- and starch-producing crops are effective energy crops as they can also be used for ethanol production, with sugarcane, cassava, and maize being the most promising in Lao PDR [10]. As these crops are already being cultivated in Lao PDR for energy purposes, they are deemed suitable for further investigation.

Cassava is a resilient woody shrub that is easily cultivated in Lao PDR. Sugarcane is a perennial grass and has high potential to provide bioenergy, as seen in Brazil's electricity mix where sugarcane mills are the fourth most important electricity suppliers, providing more than 22,5000 GWh in 2019 [11]. Maize is a cereal grain and staple crop. It can grow in hot and dry conditions, enabling it to be grown in land that is not suitable for other crops. Jatropha curcas is valued due to its resilience. It can grow in many different types of soil, thriving in arid conditions and poor-quality land. This makes it suitable to survive dry seasons in Lao PDR.

2.2.2 Agricultural Residue

Lao PDR has a large agriculture sector where, on estimate, 80% of the total population is engaged in farming [4]. This production of crops leads to large quantities of agricultural residues being generated. Often the residues are left in order to improve soil fertility. However, they can also be used for energy generation as feedstock. Lao PDR has an opportunity generate biomass energy from current residues. This report analysed the crop residue potential from rice, cassava, sugar cane, and maize. These selected crops cover a majority of the crops grown in Lao PDR, being the four highest crops in production quantity in Lao PDR [12], and with 72% of total cultivated area dedicated to rice [13].

2.2.3 Conversion processes

The majority of the energy consumption in Lao PDR comes from residential use, where about 51% of all power is consumed [10]. Wood fuel, fuelwood and charcoal accounted for 69% of the average energy use in rural areas where it is mainly used for cooking and heating [10]. Therefore, renewable energy alternatives for Lao PDR's primary energy use were studied. Hence, direct combustion of biomass and biogas were examined as they can be used for cooking and heating. Despite the fact that cassava, sugarcane, and maize can also be used for ethanol production for biofuels, this was not further investigated as focus was placed on bioenergy that can be utilised for Lao PDR' main energy needs. Biofuel processing plants are non-existent in Lao PDR and so the feasibility of biofuel production would be hard to assess. Biodiesel production from jatropha curcas was not further investigated for the same reason. However, the GIS work performed, including the maps generated, can still be used as a basis for further investigation in these alternative fuels in future work.

When choosing suitable conversion processes for energy generation, emphasis was placed on mature, proven processes that are commercially available or in the early commercial stage. Previously implemented conversion processes in Lao PDR were also considered due to the country's familiarity with them. According to International Finance Corporation (IFC), combustion, gasification and anaerobic digestion are conversion processes most suitable [14].

- **Combustion**: Electricity is produced by direct combustion when biomass is burned to produce high-pressure steam which then allows turbines to rotate. This drives a generator which produces energy.
- **Gasification**: Solid biomass material is exposed to high temperatures with very little oxygen present, to produce synthesis gas (syngas). The gas can then be burned to produce energy.
- **Anaerobic digestion**: Organic waste material such as animal manure or human sewage can be collected in tanks called digesters. In these oxygen-free tanks, the waste is decomposed by anaerobic bacteria, which produce methane and other by-products. This forms a renewable natural gas that can be purified and used to generate electricity.

Different conversion processes require engines suitable for the process and therefore multiple engines were considered. Conversion efficiency and practical factors were considered for engine choice used in final energy calculations.

- **Combustion**: The two main conversion processes considered were Stirling engines and Externally Fired Gas Turbines (EFGT). Both Stirling Engines and EFGT have a rather low electrical efficiency of 20-25% but both engines can be operated using all biomass feedstocks. The Stirling engine was selected over the EFGT as it would be able to cope with fast load changes better. [15]
- **Gasification**: Microturbines can directly run-on biogas and are suitable due to their low maintenance requirements and high efficiency of around 35%. As it is of importance to provide energy to remote areas where maintenance and operation area easy, microturbines were the most suitable choice [15].
- **Anaerobic Digestion**: Due to similar reasons mentioned above, microturbines have been selected as the most suitable engine for the process.

2.2.4 Yield data

The Global Agro-Ecological Zones (GAEZ) modelling framework [16], developed in cooperation of The Food and Agriculture Organization of the United Nations (FAO) and the International Institute for Applied Systems Analysis (IIASA), has generated a spatial database for the cultivation potentials of around 50 crops. GAEZ v4 was used extensively in this study to determine yield potentials of crops when calculating bioenergy potentials. Two main data sources from GAEZ v4 were used:

- Actual Yield and Production: providing downscaled historical yields of crops;
- Agro-climatic Potential Yield: providing potential yields of crops under different input and management assumptions for historical, current and future climate.

2.2.4.1 Actual Yield and Production

Aggregate data regarding agricultural production exists at a national level, but this does not provide data at a finer resolution which is needed to assess crop yield potential across Lao PDR. GAEZ v4 uses a "downscaling" method in order to convert the national-scale data into individual spatial unit data. This is performed using optimization principles to combine spatial data from several factors that affect the distribution of crop production. These factors include: land characteristics, land cover, soil type, terrain slopes, climate and others.

2.2.4.2 Agro-climatic Potential Yield

The potential crop yield is calculated in two parts. First, calculating the biomass and yield potential and second, applying adjustment factors to account for agro-climatic restraints.

The biomass and yield potential is calculated by Land Utilization Types (LUTs). Each location (grid-cell) has a specific LUT which comprises technical specifications for crop production within the given socioeconomic setting. A LUT will consist of attributes such as water supply type, type of main produce, cultivation practices, and other agronomic data. The temperature and radiation regime of the grid-cell is then combined with the LUT and the specified crop's characteristics such as growth cycle information, photosynthesis rates, respiration rates, sensitivity to heat etc. This allows a potential yield to be calculated. Each grid-cell yield can be calculated for different water source conditions (rain-fed or irrigated) and different input levels (low, intermediate, and high). Under rain-fed conditions, water requirements of each LUT and crop are identified, and a water-stress and water-deficit yield reduction factor is applied where necessary. A low-level input uses conditions where production relies on the use of 'labour intensive techniques, and no application of plant nutrients, no use of chemicals for pest and disease control and minimum conservation measures'. A high-level input uses conditions where production is 'fully mechanized where possible with low labour intensity and uses optimum applications of nutrients and chemical pest, disease and weed control'. An intermediate level input was not used in this study.

After the yield potential is calculated by LUTs, individual agro-climatic constraint factors are combined and applied, including pests, diseases, weeds, effect on farming operations and others. Yield adjustment factors also include crop-specific responses to CO2 concentrations, which is crucial when evaluating climate scenarios. An example equation is provided in Appendix A for how the factors is calculated.

2.2.5 Climate Scenarios

As a goal of this report is to provide a foundation for Lao PDR in developing a sustainable plan for renewable energy, it is important to understand the impacts of different climate scenarios on bioenergy production.

Scenarios have been run using various Representative Concentration Pathways (RCP), which provide scenarios of the emissions trajectory and resultant radiative forcing projections. RCP 2.6, known as 'Low emissions', would have CO_2 emissions stay at today's level until 2020, then decline and go to zero in 2100. RCP 4.5. known as 'Intermediate emissions' and the most probable baseline scenario, have CO_2 emissions increase slightly before declining starts

around 2040. RCP 8.5, known as 'High emissions' and the basis for the worst-case scenario, have emissions continue to rise throughout the century.

3 Energy Potential of Energy Crops & Residues

3.1 Method

A step-by-step methodology was established to determine the total potential energy that could be generated from biomass use in Lao PDR. Background research on Lao PDR provided a foundation for creating the workflow, which considered land distribution, conversion processes, and suitable energy crops.

All calculations and mapping work were carried out on Quantum Geographical Information System (QGIS), and all relevant shapefiles containing this work has been saved to be used for future work in this field.

3.1.1 Mapping suitable land for energy crop growth

The first step was to determine the area of land suitable for growing energy crops. Initially, a map of Lao PDR was imported into QGIS. 'Land use' and 'Forest classification' outlined in 'Forestry Strategy to the year 2020 of Lao PDR' [17] were then used to identify regions of Lao PDR to remove from the map and have been defined below:

- **Production Forests** are considered areas used in regularly providing forest products such as timber on a sustainable basis to help social and economic development requirements and for people's livelihoods.
- **Conservation Forests** are regions classified for the purpose of protecting and conserving animal and plant species, natural habitats, and various other entities of historical, cultural, tourism, environmental, educational, or scientific value.
- **Protection Forests:** are regions classified for the protection of watershed areas and prevention of soil erosion. They also include areas of forestland with national security significance, areas for protecting against natural disaster and areas for protection of the environment.

Land in Lao PDR that is already being used for existing crop production were also identified and removed in order to account for food security. To create a map of Lao PDR without protected regions or crop land cover, the AEZ shapefiles for forest regions and existing crop production areas were downloaded and processed using QGIS [18]. The shapefile layers were then overlaid on the map of Lao PDR and geoprocessing tools were used to remove areas of overlap, resulting in the final map representing suitable land for energy crop use.

3.1.2 Calculating production of energy crop

The next step involved collecting yield data corresponding to the areas on the created map in order to calculate the total production of each crop to be used as biomass feedstock. The yield data for the respective crops was downloaded from the 'Actual Yield and Production' theme in GAEZ as outlined in section 2.2.5.2. This data was in a raster file format, which is a rectangular array of values known as pixels, and therefore converted into a vector shapefile using the 'raster pixel to polygon' processing tool to be compatible with other layers in QGIS. The yield data had a resolution of 5 arc minutes, providing granular information on a 9km-by-9km scale for Lao PDR. Geoprocessing tools were then used to merge the yield data with the corresponding areas on the map, converting the map into small 9km-by-9km grids of the potential production of each energy crop.

3.1.3 Calculating energy potential

Crop production for each grid cell was then converted to energy potential. The purpose of considering different conversion processes available was to compare the maximum potential energy generated using the production data and calorific value for each energy crop and by accounting for types of engines and engine efficiency. The following equation was used to derive energy potential and has been adapted for each conversion processes [19]:

$$E = (CV/C1) \times C2 \times DM \quad (1)$$

Where CV is the calorific value (MJkg^{-1}), $C1$ is the coefficient to transform MJ unit to kWh (1kWh = 3.6MJ), $C2$ is the efficiency of the engine, DM is the dry matter of the crop (g/ha), C3 is syngas efficiency to convert biomass to gas, E is the energy potential (kWh/ha).

Combustion:

$$E_{Combustion} = DM \times CV \times C2 \times C1 \quad (2)$$

Gasification:

$$E_{Gasification} = DM \times CV \times C3 \times C2 \times C1 \quad (3)$$

Anaerobic Digestion:

$$E_{AD} = DM \times Biogas\ Yield \times CV \times C2 \times C1 \quad (4)$$

1m^3 of Biogas Yield = 22MJ/kg

3.1.4 Energy Crops

Cassava, sugarcane, and maize were chosen as the energy crops to analyse, as explained in Section 2.2.1. Jatropha curcas was not included in this section as historical yield data was not available due to jatropha curcas cultivation being negligible in Lao PDR in the past. Jatropha curcas was analysed in Section 3. Energy crops investigated in this report were split into different parts of the plant. This approach ensured that the suitable parts of each crop would be used dependant on the conversion processes and remaining parts could be used for other purposes, such as food or ethanol production. The breakdown of each crop has been included below and a more detailed summary has been included in the Appendix B. All relevant data required was collected from literature.

- Cassava has been separated into stalks, leaves, and husks

- Maize has been separated into corncob, stover and trash
- Sugarcane has been separated into bagasse, trash, straw, and leaves

These values were then applied to the equations in Section 3.2 to output total potential energy for each energy crop and conversion processes.

3.1.5 Grid-by-grid Comparison

To determine the distribution of the different crops across Lao PDR in order to maximise potential energy production, analysis was carried out to compare energy potentials of different crops in each grid cell location. To perform this, the individual maps generated for each crop were combined into one layer. Once combined, each grid was evaluated using an expression to output the crop that generates the most energy.

3.1.6 Calculating energy potential of residues

Rice, cassava, sugarcane, and maize were selected to analyse the potential energy production from residues, as explained in Section 2.2.2. To calculate the energy potential, land crop cover of Lao PDR was collected using data from AEZ [18] and generated into QGIS. As the data source was agglomerated, data filtering work was carried out in order to produce isolated maps of the specific crop. Yield data was then 'merged' as explained in Section 3.1.2, producing a map of the specified crop production across Lao PDR.

To calculate the potential energy associated with the specified crop, each crop was broken down into its residue parts. The total potential energy of each crop was then calculated using the formula [20]:

$$E_i = \sum_j P_i \times RPR_{i,j} \times LHV_{i,j} \quad (5)$$

Where i is the crop and j is the residue part of the selected crop. P is the mass crop production, RPR is the residue to crop ratio, and LHV is the lower heating value (MJ/kg). Breakdown of each crop can be found in Appendix B.1.3

3.2 Results

3.2.1 Energy potential of residues

Crop	Residue part	Energy of part (kWh)	Total Energy (kWh)
Rice	Husk	3.27E+09	7.28E+09
	Straw	4.01E+09	
Cassava	Stalk	2.82E+09	5.79E+09
	Roots	2.97E+09	
Sugarcane	Bagasse	8.89E+08	2.03E+09
	Top & trash	1.14E+09	
Maize	Husk	1.70E+08	4.93E+09
	Stalk	4.54E+09	
	Cob	2.22E+08	

Table 1 Potential energy generation using residues from existing crop land cover in Lao PDR.

Table 1 displays the potential energy that can be generated using residue from current crop land cover in Lao PDR. A substantial amount of energy can be generated from existing agriculture already in place, with a total combined potential of 20 billion kWh, which is promising for Lao PDR.

3.2.2 Energy potential of energy crops

For the base case, all forest area including unprotected forest regions were removed from the map in addition to crop land cover.

Crop	Combustion (kWh)	Gasification (kWh)	AD (kWh)
	Without Forests		
Cassava	**6.52E+10**	4.51E+10	6.45E+10
Maize	**2.12E+10**	1.23E+10	1.85E+10
Sugarcane	**5.14E+10**	3.12E+10	1.90E+10
	With Forests		
Cassava	**1.31E+11**	9.04e+10	1.29E+11
Maize	**4.14E+10**	2.40E+10	3.62E+10
Sugarcane	**1.02E+11**	6.21E+10	3.78E+10

Table 2 Comparison of potential energy generated for different energy crops using various conversion processes

Table 2 shows the total potential energy calculated for each energy crop when using different conversion processes. The figures reported are the total energy potentials for all of Lao PDR and a more detailed breakdown of the figures for each province can be found in the Appendix C.1.1. From these figures, it is clear that energy potential is highly affected by conversion processes type. Moreover, cassava can generate the highest total potential energy under current climate conditions and agriculture input level when using combustion as the conversion processes. The energy generated from cassava and sugarcane using combustion were both significant. Grid-by-grid comparison (as explained in Section 3.1.5) generated the map shown in Figure 1 which showcases the outcome of crop distribution by prioritising energy potential. From the map, it was apparent that whilst cassava would be the predominant crop used for energy generation, there were certain provinces that favoured sugarcane production. Energy potential generated from maize crop is significantly lower compared to the other crops as the total production of the crop is not as large. From these results, it is evident the yield of the crop plays a large factor in the total energy production as the heating value of individual parts of the crop are similar.

3.2.3 Expanding suitable land area

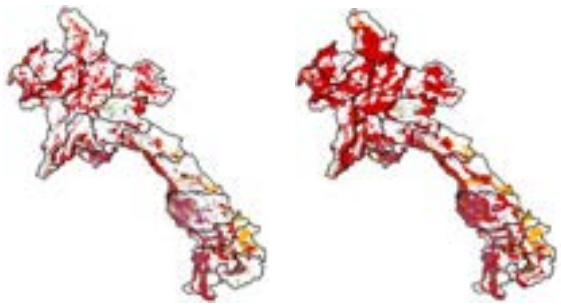


Figure 1. Mapping of energy crops and crop residues in Lao PDR using all forest areas (right) and without unprotected forest area (left). Red regions represent the regions recommended to grow cassava;

yellow regions represent the regions to grow sugarcane; purple regions represent regions using residue.

Figure 1 shows a side-by-side visual comparison, mapping the energy crops using forest regions not protected by law (right) and without (left). A significant increase in energy crop cover can be seen, mainly in the northern provinces such as Louangnamtha and Oudomxay, when accounting for forest areas which is expected as about 20% of Lao PDR is covered in forest areas not protected. The increase in area of land that can be utilised is translated into a greater energy potential value as seen in table 2. However, as mentioned before it is unrealistic to remove all forest area for the sole purpose of growing energy crops and therefore a more detailed breakdown of energy potentials is provided on a provincial scale in C.1.1 to help individual provinces if they wish to remove some of the forest cover.

3.3 Discussion

3.3.1 Agricultural Residues

One of the project aims was to identify the bioenergy potential through existing agricultural residues. Calculations show by using residue from existing crops, about 20 billion kWh can be generated which equates to about 20% of the total energy that can be generated from biomass in this report. With current demand levels in Lao PDR, using only residues for energy generation exceeds the current annual consumption of 5.47 billion kWh [21] in the country.

3.3.2 Energy Crops

3.3.2.1 Crop Selection

Table 2 shows that, for maximising energy production, cassava using anaerobic digestion outperforms the studied crops and conversion processes, making cassava a promising energy crop to start growing in Lao PDR, especially in the northern provinces. The driving reason behind cassava's energy potential is that cassava is able to generate higher yields and thus higher energy potentials. An explanation for this is that the climate is favourable to cassava in these regions. According to AEZ [18], for the periods when yield data was collected (2010), precipitation levels were on average below 1500mm throughout the northern provinces and some south-western provinces. According to [22], optimum precipitation levels for cassava are 1000 to 2000mm per year which explains why it outperforms maize and sugarcane yields in regions of low precipitation. In contrast, in areas of high precipitation such as south-eastern provinces, where over 2200mm were recorded in 2010, sugarcane outperforms maize and cassava.

The results found that northern provinces of Lao PDR tend to achieve higher yields for all the energy crops compared to other provinces. This can be more easily seen in Appendix C.1.1 where energy crop maps for each individual crop is provided. This can be explained by more suitable agro-climatic conditions in these regions. When evaluating energy crop growth, these northern provinces should be prioritised where greater yields lead to greater energy potential.

The results show that it important to consider growing a variety of energy crops in Lao PDR. It was found that using one energy crop across all of Lao PDR did not maximise energy potential, but a mixed distribution of crops was optimum as showed by Figure 1. Whilst about 65 billion kWh could be generated from cassava alone, by replacing some regions where higher sugarcane yields are achievable, an additional 1.3 billion kWh of energy can be produced.

3.3.2.2 Conversion Processes Selection

The results show that the energy potential generated by the crops are dependent on the conversion process used. Table 2 shows that for all the energy crops, combustion is able to produce the highest energy potential. However, anaerobic digestion is able to almost match the production of energy for cassava by generating 64.5 billion kWh. This can be explained as selection of a preferred conversion processes is complex and requires consideration of the type of biomass and moisture content in the fuel. For biomass with moisture content above 65%, the calorific value becomes too low for combustion, making a biogas plant the more suitable option [14]. Moisture content in cassava crop can range from 62.5-75.4% [23] and therefore, the energy generated from anaerobic digestion almost matches combustion. Combustion outperforms anaerobic digestion in this instance due to a larger proportion of the feedstock being used in the process. In contrast, as moisture content in maize (14-22%) [24], and sugarcane (45-55%) [25], are lower than the suggested 65% moisture content, combustion conversion processes was the ideal conversion processes.

As the energy potential generated is still significant in each of the processes, factors other than just energy potential should be examined. As most of Lao PDR's population are in rural areas, anaerobic digestion and gasification is more suitable as they are lower maintenance and are ideal for small-scale energy generation. This makes anaerobic digestion of cassava potentially the more attractive combination to use in Lao PDR.

3.3.2.3 Expansion of suitable land area

Table 2 shows that energy potential is almost doubled when expanding the land cover used to grow energy crops to include forest areas in Lao PDR that are not protected. With the help of this data, decision-makers can assess whether it is beneficial to use forest area for the purpose of bioenergy production. It is important to account for proper use of forest areas and implications on wildlife. Therefore, more detailed energy potential data has been provided on a provincial level in the Appendix C.1.1, to aide in the evaluation.

4 Climate Scenarios

4.1 Method

In this section, the aim of the investigation was to understand the impacts of climate change on the predicted yields of selected energy crops by comparing it with historical averages which formed a baseline. This would provide useful insights to the suitability of

growing energy crops in the future for Lao PDR. The workflow in this section was adopted from the methodology outlined in Section 3.1, using QGIS to identify suitable land for energy crop growth, with key differences being the theme used to collect data and focusing mainly on predicted yields to draw conclusions rather than maximum potential energy. Therefore, the area used for growing energy crops in our climate scenarios were equivalent to our base case. As mentioned in Section 2.2.4, there were multiple themes created by GAEZ for different classes of data. In this section, the 'Agro-Climatic Potential Yield' theme was used to calculate potential yields by applying adjustment factors to account for different agro-climatic scenarios.

This theme also provided simulations using historical data between 1981-2010, which would be used as a baseline for comparing against future estimated yields. A limitation, however, was that the climate pathways were only modelled with a high input level which, as mentioned in Section 2.2.4, is a scenario where crop production is fully mechanised and uses optimum farming techniques. To maintain consistency, the baseline data was adjusted by factoring in a high input level scenario. This allowed for a comprehensive comparison between the data generated from different climate pathways and the baseline scenario. The impacts of using a high input level will be explored in the discussion. Conversion processes were not applied as improvements in technology and efficiency are unknown for the future. Therefore, the energy potential calculated for each crop is the maximum thermal energy that can be generated and has been included in Appendix C.1.3.

Historical data for cassava, maize, sugarcane and jatropha curcas from 1981 to 2010 were obtained using the CRUTS32 data source provided by GAEZ. For future climate data, the IPSL-CMIP5 climate model was used to generate potential yields for different representative concentration pathways (RCP2.6, RCP4.5, RCP8.5) for the periods 2011-2040 and 2041-2070. Maize was replaced by jatropha curcas as previous results suggest that the performance of the crop is not promising. Jatropha curcas could only be assessed in the climate pathway work due to the availability of crop data provided in this theme.

4.2 Results

Crop	Baseline Yield (Kg/ha)	% Difference 2011-2040		
		RCP 2.6	RCP 4.5	RCP 8.5
Cassava	29395	-3.5%	-2.9%	-4.0%
Sugarcane	104456	-16.6%	-5.5%	-16.6%
Jatropha	3524	-2.0%	-2.9%	-2.6%

Table 3 Predicted average yields generated for different energy crops under different climate pathways between 2011-2040

Crop	Baseline Yield (Kg/ha)	% Difference 2041-2070		
		RCP 2.6	RCP 4.5	RCP 8.5
Cassava	29395	-3.7%	-4.4%	-9.3%
Sugarcane	104456	-16.3%	-15.6%	-29.1%
Jatropha	3524	-2.4%	-3.4%	-7.7%

Table 4 Predicted average yields generated for different energy crops under different climate pathways between 2041-2071

The performances of the energy crops were assessed by the difference in yields generated under each climate pathway, relative to the baseline as the simulations were modelled with the same control variables. Tables 3 and 4 show how the average potential yields vary under different climate pathways for each energy crop, between 2011-2040 and 2041-2070. Potential energy generated from these yields can be located in Appendix C.1.3.

Table 3 shows that between 2011-2040, for all future climate scenarios, the predicted average yield is generally expected to decline and the variation in decline are energy crop specific. Jatropha curcas appears to be the most resilient crop under different climate scenarios with the lowest changes of 2%, 2.9% and 2.6% for RCP 2.6, 4.5 and 8.5 respectively. Cassava is also predicted to maintain decent crop yields with a maximum decline of 4% under RCP 8.5. Sugarcane yields are expected to decline the most between 2011-2040 under all climate scenarios with expected declines of -5.5% under RCP4.5 and -16.6% under RCP 2.6 and 8.5. As seen in table 4, between 2041-2070 the decline in predicted crop yields area expected to increase. This is the case for all energy crops for all climate scenarios except sugarcane under RCP 2.6 which slightly increases predicted yields from a -16.6% decline to -16.3%. RCP 8.5 climate pathways cause the largest yield declines between 2041-2070 of -9.3%, -29.1% and -7.7% for cassava, sugarcane and jatropha curcas respectively. Jatropha is predicted to have the lowest yield decline across all climate scenarios once again.

4.3 Discussion

4.3.1 Impacts of high input scenario

Tables 3 and 4 and table 13 in Appendix C.1.3 shows the changes in potential yields and energy potential for different climate pathways relative to the baseline when operating with a high input level, which is promising for the prospect of growing energy crops in the future for Lao PDR. However, due to the high input level assumption, the absolute values can be misleading and therefore, an emphasis is placed on the relative difference to the baseline data. This is because modelling for expected yields using only high input levels does not necessarily reflect the overall agriculture industry/landscape in Lao PDR This can be seen by the difference in energy potentials generated using data from 2010 operating with current input levels used to create our base case and data from historical averages operating with a high input level. Refer to table 2 in section 3.2.2 and table 12 in appendix C.1.3. The

difference in values show that Lao PDR can generate a significant amount of extra energy through adopting a high input level but, currently this is not the case. The value that can be extracted from this work is that under all climate scenarios in the future, the difference in crop yields relative to historical averages are reasonable and therefore, expected yields can be utilised as a source of energy for Lao PDR.

4.3.2 Impacts of climate pathways

From tables 3 and 4, it is evident that the impacts of different climate pathways vary between the first half of the century and the second half. Between 2011-2040, there is less variance in yield declines for the different RCP scenarios compared to 2041-2070, where the range in yield declines are larger. For example, between 2011-2040, cassava crop yield decline ranges from -3.5% to -4% whereas between 2041-2070, the yield decline ranges from -3.7% to -9.3%. In addition, RCP 8.5 is expected to cause the biggest change in yields between 2011-2040 and 2041-2070 compared to all other climate scenarios.

Similar yield declines for all climate pathways between 2011-2040 can be attributed to the fact that until 2050, the expected outcomes between RCPs ae relatively small. This is because climate change systems respond slowly to changes in greenhouse gas (GHG) emissions [26]. After 2050, more importance must be given to the different RCPs as they represent very different scenarios by accounting for rate of warming, and bigger changes to water temperature and precipitation levels. Whilst RCP 2.6 and 4.5 models are less distinguished, with increasing temperatures and GHG emissions expected to slow down, RCP 8.5 leads to greater temperature increases and increased GHG emissions [27]. The findings in tables 3 and 4 support the aim of using crops for energy generation as predicted yields can still generate a substantial portion of energy consumed in Lao PDR annually. However, it is important to be wary of climate scenario RCP 8.5 where there is a much more significant decline in crops, as if global climates area exacerbated further, the impacts on yields could be much worse

4.3.3 Yield Responses

As seen in table 3 and 4, the percentage decrease in predicted yields for cassava and jatropha curcas are lower than sugarcane. This suggests that energy crops have different sensitivity levels to changes in climate which affects predicted yields. The data supports that cassava and jatropha curcas are more resilient to changes in climate which make them more suitable for use in Lao PDR as they will generate relatively consistent yields for all climate pathways. The response of the energy crops can be explained by the following reasons:

- Yield response to water
- Yield response to CO2 levels

Yield reductions are related to changing rainfall patterns, evaporative demand, and reduced availability of water [28]. K_y values have been derived by FAO to quantify the link between production and water use by a crop. A K_y value greater than 1 suggests that the crop response is very sensitive to water deficit with proportional larger yield reductions when water use is reduced because of stress [29]. The impacts of reduced availability of water caused by climate change are reflected in the decreasing yields observed in tables 3 and 4. According to FAO, sugarcane has a K_y factor of 1.2 which is higher than jatropha curcas and cassava and explains why a relatively greater decrease in predicted yields are seen.

According to IPCC, CO2 levels are expected to increase under all climate scenarios [30]. The 'fertilisation' effect on crop yields caused by increasing CO_2 in the atmosphere has been accounted for by GAEZ [16]. Crop species respond differently to changes in CO_2 levels depending on physiological characteristics. The empirical correction factor, f_{CO2}, captures the yield responses of five broad crop groups. Cassava and jatropha curcas have been classified as group 2 crops, whilst sugarcane has been classified as a group 3 crop. Group 3 crops have consistently higher increment factors for all CO2 levels compared to group 2. By accounting for yield response to water deficit and CO2 levels, results in tables 3 and 4 seem reasonable. Cassava and jatropha curcas seem more resilient to climate change compared to sugarcane due to the positive response to increasing CO2 levels and the lower sensitivity to water deficit. When deciding the best energy crops to utilise, it is important to recognise the ability to grow under changing climates alongside the yields that can produced. By factoring in all these indicators, cassava, jatropha curcas and sugarcane all seem viable options.

5 Conclusions

This report achieved the aims of the project by using a map-based approach to understand the energy generation potential of suitable energy crops and agricultural residues in Lao PDR. The results obtained found that a combination of cassava and sugarcane across Lao PDR generated the highest potential energy of 66.5 billion kWh. This involved using combustion as the conversion process, however, anaerobic digestion was found to perform similarly and may be more suitable for use in rural areas in Lao PDR due to lower maintenance requirements and more suited for small-scale generation. Results also found that an additional 20 billion kWh of energy can be produced from existing residue. Additionally, results from climate pathway analysis showed that decline in crop yields are expected under all climate scenarios but are more significant between 2041-2070. However, the results show cassava and jatropha curcas are more resilient to the changing climate.

Geo-spatial data in the form of maps have been produced, to allow spatial evaluation when considering how to implement the biomass energy potential that has been found in this study for Lao PDR. These maps have been made accessible to provide a foundation for future work in this field. In deciding which energy crops to grow and conversion processes to choose, whilst this

report has evaluated energy production potential, it is also important to consider the economic factors that will determine what strategy is more suitable for Lao PDR. Some factors that should be included in a future economic analysis are:

- Cost to produce crop: Seeds, fertilizers, plant protection products, cost of capital goods such as equipment & machinery, labour, land costs and others. FAO have produced the *Handbook on Agricultural Cost of Production Statistics* [31] which outlines a workflow for calculating cost of production of crops, especially focused for developing countries.
- Cost required for energy production: transportation of feedstock, processing of feedstock, initial investment & operational costs of conversion processes, implementation costs with existing grid lines and others.

The geo-spatial nature of the data provided in this study can provide a basis for assessing these location-dependent economic factors.

The crops evaluated in this report were the most promising, but not exhaustive. Other potential energy crops, such as soybean and oil palm, can be evaluated using the same methodology. The conversion processes evaluated also focused on mature conversion processes that have been used in Lao PDR previously. However, biomass conversion processes in the research and development stage and demonstration stage, namely pyrolysis and torrefaction, could be looked into. Pyrolysis involves heating biomass with high temperatures in the absence of oxygen to produce solid charcoal, liquid pyrolysis oil, and a product gas. Torrefaction is a mild form of pyrolysis where biomass is heated in the absence of oxygen to produce char to use for bioenergy production.

Although biofuel production was not assessed as explained in Section 2.2.3, the spatial data of crop production in this report can be used to calculate the energy production of the bioethanol production from cassava, sugar cane and maize. Factors to consider when using the selected crops for biofuel production are outlined in Table 2.1 of the Appendix in [10].

Finally, to ensure future viability, a sensitivity analysis should be performed on the production of the studied crops as production is variable and vulnerable to significant changes. Although climate scenarios were considered, other factors could impact the production in the future such as diseases. Future improvements in agricultural techniques, water availability, irrigation and efficiency of biomass conversion processes would also be important to consider.

6 References

[1] '2022 - Statistical Review of World Energy 2022.pdf'. Accessed: Dec. 13, 2022. [Online]. Available: https://www.bp.com/content/dam/bp/business-sites/en/global/corporate/pdfs/energy-economics/statistical-review/bp-stats-review-2022-full-report.pdf

[2] Lao PDR Department of Energy Policy and Planning, Ministry of Energy and Mines, 'Lao PDR Energy Outlook 2020', p. 94, 2020.

[3] nternational Renewable Energy Agency, 'Bioenergy for the energy transition: Ensuring sustainability and overcoming barriers', p. 128, 2022.

[4] B. Vongvisith *et al.*, 'Agricultural waste resources and biogas energy potential in rural areas of Lao PDR', *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, vol. 40, no. 19, pp. 2334–2341, Oct. 2018, doi: 10.1080/15567036.2018.1488017.

[5] J. W. Ranney and J. H. Cushman, 'Regional evaluation of woody biomass production for fuels in the southeast', Oak Ridge National Lab., TN (USA), CONF-791072-2, Jan. 1979. Accessed: Dec. 06, 2022. [Online]. Available: https://www.osti.gov/biblio/5428722

[6] A. Thomas, A. Bond, and K. Hiscock, 'A GIS based assessment of bioenergy potential in England within existing energy systems', *Biomass and Bioenergy*, vol. 55, pp. 107–121, Aug. 2013, doi: 10.1016/j.biombioe.2013.01.010.

[7] M. Beccali, P. Columba, V. D'Alberti, and V. Franzitta, 'Assessment of bioenergy potential in Sicily: A GIS-based support methodology', *Biomass and Bioenergy*, vol. 33, no. 1, pp. 79–87, Jan. 2009, doi: 10.1016/j.biombioe.2008.04.019.

[8] E. R. Barata, 'A GIS approach to estimate the bioenergy potential in Uganda', p. 98, 2017.

[9] R. Gaillard, 'Biofuel Assessment Study in Lao PDR Inception Report', 2009, Accessed: Nov. 27, 2022. [Online]. Available: https://policy.asiapacificenergy.org/sites/default/files/Biofuel%20Assessment%20Study%20in%20Lao%20PDR%20-%20Inception%20Report.pdf

[10] Greater Mekong Subregion Economic Development Initiative, 'Status and Potential for the Development of Biofuels and Rural Renewable Energy: The Lao People's Democratic Republic', p. 38, 2009.

[11] 'Bioelectricity', *SugarCane*. https://www.sugarcane.org/sugarcane-products/bioelectricity/ (accessed Dec. 05, 2022).

[12] 'FAOSTAT'. https://www.fao.org/faostat/en/#data/QCL (accessed Sep. 25, 2022).

[13] 'Laos at a glance | FAO in Laos | Food and Agriculture Organization of the United Nations'. https://www.fao.org/laos/fao-in-laos/laos-at-a-glance/en/ (accessed Dec. 09, 2022).

[14] 'BioMass_report_06+2017.pdf'. Accessed: Dec. 11, 2022. [Online]. Available: https://www.ifc.org/wps/wcm/connect/fb976e15-abb8-4ecf-8bf3-

8551315dee42/BioMass_report_06+2017.pdf?MOD=AJPERES&CVID=lPHGOaN

[15] M. Loeser and M. Redfern, 'MICRO-SCALE BIOMASS GENERATION PLANT TECHNOLOGY: STAND_ALONE DESIGNS FOR REMOTE CUSTOMERS', [Online]. Available: https://purehost.bath.ac.uk/ws/files/402852/Paper_-_Micro-scale_Biomass_Generation_Plant_Technology_-_Stand-alone_Designs_for_Remote_Customers.pdf

[16] *Global agro-ecological zone V4 – Model documentation*. FAO, 2021. doi: 10.4060/cb4744en.

[17] Ministry of Agriculture and Forestry, *FORESTRY STRATEGY TO THE YEAR 2020 OF THE LAO PDR*. 2005. Accessed: Dec. 04, 2022. [Online]. Available: https://data.opendevelopmentmekong.net/dataset/2020-1/resource/0ef2b050-13dd-4890-9cbf-21d2092085fe/view/1f777cb7-afa8-49f1-8812-f6c4ec20a034

[18] 'LRIMS-DALAM'. https://lrims-dalam.net/?thematic=aez (accessed Dec. 10, 2022).

[19] R. Ambrosio, V. Pauletti, G. Barth, F. P. Povh, D. A. da Silva, and H. Blum, 'Energy potential of residual maize biomass at different spacings and nitrogen doses', *Ciênc. agrotec.*, vol. 41, no. 6, pp. 626–633, Dec. 2017, doi: 10.1590/1413-70542017416009017.

[20] M. A. Gonzalez-Salazar *et al.*, 'Methodology for estimating biomass energy potential and its application to Colombia', *Applied Energy*, vol. 136, pp. 781–796, Dec. 2014, doi: 10.1016/j.apenergy.2014.07.004.

[21] 'Energy consumption in Laos', *Worlddata.info*. https://www.worlddata.info/asia/laos/energy-consumption.php (accessed Dec. 11, 2022).

[22] 'Cassava Project'. https://www.uq.edu.au/_School_Science_Lessons/CasProj.html#:~:text=Annual%20rainfall%20should%20be%20greater,than%202500%20mm%20per%20year (accessed Dec. 09, 2022).

[23] P. Pornpraipech, M. Khusakul, R. Singklin, P. Sarabhorn, and C. Areeprasert, 'Effect of temperature and shape on drying performance of cassava chips', *Agriculture and Natural Resources*, vol. 51, no. 5, pp. 402–409, Oct. 2017, doi: 10.1016/j.anres.2017.12.004.

[24] Z. Weinberg, Y. Yan, S. Finkelman, G. Ashbell, and S. Navarro, 'The effect of moisture level on high-moisture maize (Zea mays L.) under hermetic storage conditions - in vitro studies', *Journal of Stored Products Research*, vol. 44, pp. 136–144, Dec. 2008, doi: 10.1016/j.jspr.2007.08.006.

[25] 'Analysis of Sugarcane Bagasse, Cellulose Content of Sugarcane Bagasse, Lignin Content of Sugarcane Bagasse'. https://www.celignis.com/feedstock.php?value=13 (accessed Dec. 11, 2022).

[26] F. Gyllensvärd, 'What do different RCPs mean?', *Climate Information*, Apr. 20, 2020. https://climateinformation.org/data-variables/what-do-different-rcps-mean/ (accessed Dec. 11, 2022).

[27] '15-117-NCCARFINFOGRAPHICS-01-UPLOADED-WEB(27Feb).pdf'. Accessed: Dec. 11, 2022. [Online]. Available: https://coastadapt.com.au/sites/default/files/infographics/15-117-NCCARFINFOGRAPHICS-01-UPLOADED-WEB%2827Feb%29.pdf

[28] J. I. L. Morison and R. B. Matthews, 'Agriculture and Forestry Climate Change Impacts Summary Report, Living With Environmental Change'. Living With Environmental Change (LWEC) Network, 2016. [Online]. Available: https://www.ukri.org/wp-content/uploads/2021/12/131221-NERC-LWEC-AgricultureForestryClimateChangeImpacts-ReportCard2016-English.pdf

[29] P. Steduto and Food and Agriculture Organization of the United Nations, Eds., *Crop yield response to water*. Rome: Food and Agriculture Organization of the United Nations, 2012.

[30] N. Nakićenović and Intergovernmental Panel on Climate Change, Eds., *Special report on emissions scenarios: a special report of Working Group III of the Intergovernmental Panel on Climate Change*. Cambridge ; New York: Cambridge University Press, 2000.

[31] 'HANDBOOK ON Agricultural Cost of Production Statistics.pdf'. Accessed: Dec. 10, 2022. [Online]. Available: https://www.fao.org/3/ca6411en/ca6411en.pdf

Techno-Economic and Environmental Assessment of Ethylene Electrosynthesis from Carbon Dioxide

Xu, Yuzhe and Leung, Chiyuen
Department of Chemical Engineering, Imperial College London, U.K.

Abstract: Global concern about greenhouse gas emission rose from last century. Researchers have developed numerous methods to reduce the impact of from chemical processes to the environment. Current technology enables people to build nano-level of catalyst structures for desired conversion of previously non-reactive components. Based on former study of electrolysis of carbon dioxide in alkaline solution, we have managed to accomplish a model of the process alongside with its life cycle assessment result to observe how the process could help to the environmental improvement and its corresponding economic price. Aspen Plus V11 is used to model the process, consisting of three different separation units to obtain four main sales product of CO_2-H_2O electrical reduction: ethylene gas, hydrogen gas, pure ethanol, and pure acetic acid. The process is dedicated to reducing the emission of carbon dioxide; hence it is not its duty to pose a positive economic payback. In the results section, a negative economic potential is presented, yet the contribution to net-zero and carbon-neutral target of global politics of the process is obvious and apparent. Furthermore, current constraints of this report and the discussion of it is also available in the conclusion.

Keywords: Process Simulation, CO_2 Electrical Reduction, CO_2 Absorption, Ethanol Distillation, Acetic Acid Extraction, CO_2 Life cycle assessment

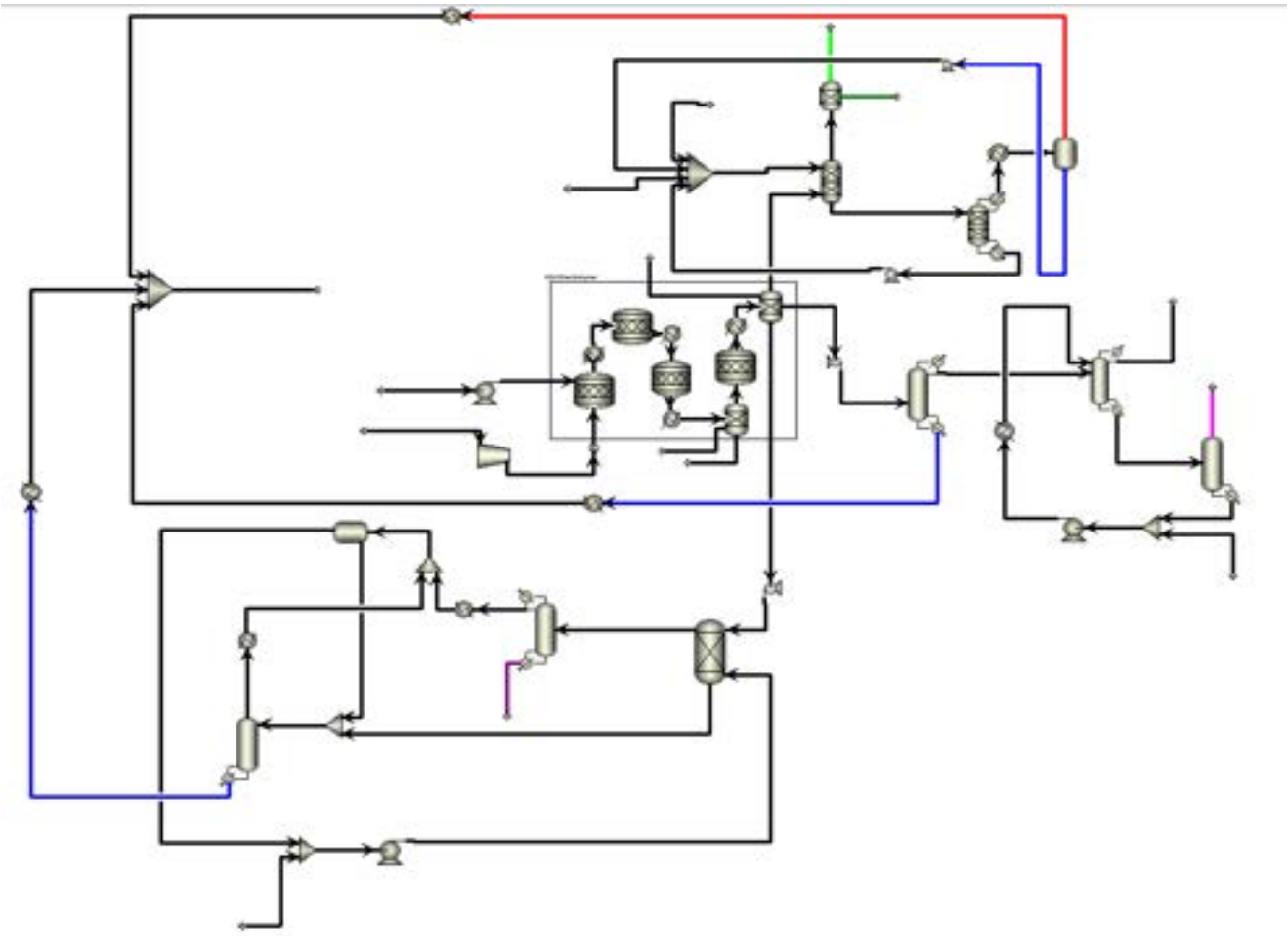


Figure. Overall Flowsheet of the Process

1. Introduction

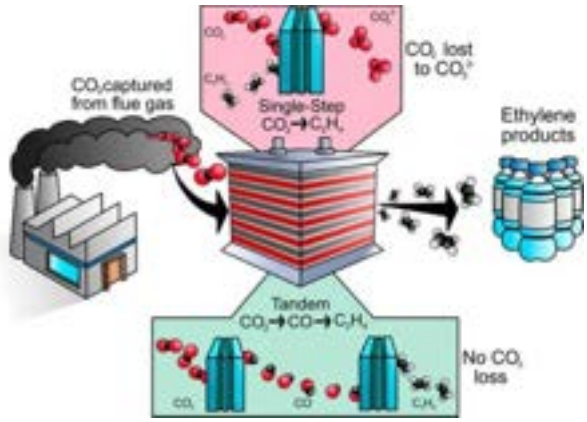


Figure 0. Process Framework

Increased global greenhouse gas emissions have been a growing concern for mankind as the accumulation of those gases in the atmosphere could lead to devastating consequences such as severe heatwaves, melting of glaciers and arctic ice caps, loss of habitats for animals etc. In December 2015, 195 countries reached a historical climate agreement that to keep the rise of global temperature below 2°C to avoid dangerous climate change. Many governments then announced the gradual phase-out of fossil fuels and switch to renewable energy as means to achieve the target [1].

Carbon dioxide (CO_2) is the most significant greenhouse gas, accounting for about three-quarters of total greenhouse gas emissions [2]. CO_2 is produced by a variety of sources, including the burning of fossil fuels, as well as by certain industrial processes and agriculture for example the production of livestock and the use of fertilizers. One of the most common methods to consume CO_2 is through the process of photosynthesis, in which plants and other photosynthetic organisms use light energy to convert CO_2 into glucose and other organic compounds. This process acts as the basis of the Earth's food chain and is essential for sustaining life on the planet.

In recent years, carbon capture and utilization (CCU) has gained attention as a potential solution to reduce CO_2 emissions and mitigate climate change. This approach involves capturing CO_2 from industrial emissions or the atmosphere and using it as a starting material for the synthesis of other chemicals, particularly ethylene as it has a large market size and many applications, for instance, acts as a starting material in the production of synthetic rubber and plastics. This technology enables CO_2 to be used as a valuable resource rather than simply being released into the atmosphere. However, CCU also has some barriers to implementation and limitations. For example, the process of retrieving CO_2 and purification of desired products could be energy-intensive and expensive, therefore the chemicals produced via this process may not be price competitive with those derived from fossil fuels. Additionally, there are concerns about the environmental impacts of using CO_2 as a feedstock for chemical reactions since it could be difficult to dispose process waste. Despite these challenges, CCU remains an active area of research and development, and it holds promise as a potential solution for reducing CO_2 emissions and mitigating climate change. In the case that the carbon tax imposed by governments rises when the world has a more urgent energy crisis, the cost of the process may be less than the cost of emissive carbon dioxide to the atmosphere. Hence, the process would perform a positive effect on a local factory and shall be developed a better scheme as soon as possible.

2. Background

This report explores the possible routes of ethylene electrosynthesis from carbon dioxide. Similar research on carbon dioxide electrolysis has already been started by pioneers [3]. Previous studies developed routes relatively effective to produce ethylene and several by-products. The main characteristics of the process include gas diffusion electrodes and copper-based catalyst, and a full combination of separation techniques. The process could be operated by two different routines, the direct CO_2 to ethylene process (CO_2P) and the two-step CO_2/CO to ethylene process (COP); the latter performs a better economic potential [4].

In this report, simulations of both routes are performed, a detailed result discussion is established for the CO_2P process, and the corresponding life cycle assessment is produced. Based on the previous art of work [4], a better separation technique is introduced, and the detail of gas absorption is developed using Aspen Plus V11 to determine the necessary parameters of the block units. Furthermore, a complete economic analysis is performed by the software to give a direct view of the process' s total cost and revenue.

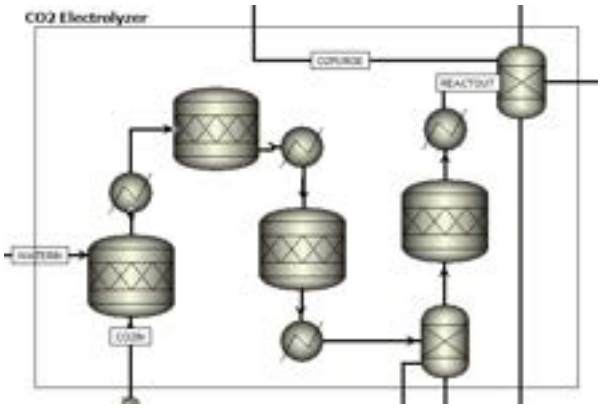


Figure 1. CO_2 Electrolysis Reactor Group

Block Name	Number of Stages	Feed Stage	Distillate Rate (kmol/hour)	Reboiler Duty (MW)	Solvent Feed Stage
Absorber	**9**	**9**	**N/A**	**N/A**	**1**
Stripper-CO_2	**20**	**2**	**1453**	**133**	**N/A**
ODC	**25**	**16**	**28.4**	**1.65**	**N/A**
ADC-Ethanol	**33**	**12**	**22.65**	**0.67**	**2**
Stripper-Glycerol	**5**	**2**	**5.8**	**0.32**	**N/A**
Liquid-Liquid Extractor	**15**	**1**	**N/A**	**N/A**	**15**
ADC-Acetic Acid	**50**	**18**	**354.45**	**8.25**	**N/A**
Stripper-Ethyl Acetate	**10**	**1**	**9**	**N/A**	**N/A**

Table 1. parameter of gas separation columns for single step process

The electrolyser used to reduce CO_2 consists of 3-compartment and is operated at a high-pressure condition (10 bar) with a total current density of 5000 A/m2 to achieve better conversion [4]. Unfortunately, Aspen Plus V11 does not include a model block as an electrolyser, hence a replacement series of the stoichiometric reactor (RSTOIC) and numerical separator (SEP) blocks using electrolysis model ENRTL-RK are performed to resemble the electrolysis operator, shown in Figure 1. Carbon dioxide is set to react with water and produce four main products: ethylene, ethanol, acetic acid, and hydrogen, discarded dividedly by the electrolyser along with unreacted carbon dioxide and water. This report then presents the enhanced separation technique to gain high purity products in the section Process Simulation; Methodology and how much they influence the environment in the section Life Cycle Assessment. The process involves several chemicals which recycle as entrainer or solvent inside separation units; with the products and the reactants, a life cycle of all participants is in turn performed as a cradle-to-gate level estimation of the process' s impact to the environment.

3. Methodology

3.1 Process Feed

The feed to the reactors is assumed to be purified carbon dioxide from a plant' s emission, since a non-pure syngas would disturb electrolysis and evades selectivity of carbon dioxide. The relative Faraday Efficiencies of the products are ethylene (50%), acetic acid (20%), ethanol (20%), and hydrogen (10%); the reaction conversion is set to 50%, the unreacted carbon dioxide would be recycled back to the electrolyser. The selection of main techniques to purify carbon dioxide from emission, absorption, adsorption, and membrane permeation, depends on the composition of syngas of the existing chemical flowsheet. This report does not discuss the pre-reaction procedures but may offer a case of absorption. The yield of product i is calculated by Eqn.1

$$Yield_i = \frac{CD_{total} \times t \times A}{n_i F} \times FE_i\% \quad \text{(Eqn.1)}$$

which CD, t, A, n, F, FE% stand for current density, time, area of electrode, stoichiometric moles of transferred electron during reaction, Faraday constant, and Faraday Efficiency, respectively. The area of electrodes can be calculated by Eqn.2

$$A = \frac{N_{CO_2}}{\frac{CD_{total}}{F} \sum -v_i \left(\frac{FE_i}{n_i}\right)} \quad \text{(Eqn.2)}$$

which N_{CO_2} and v_i are number of moles of reacted carbon dioxide and stoichiometric number of carbon dioxide as reactant corresponding to each product. In this case, 455 kmol/hour of carbon dioxide is supplied with 1910 kmol/hour of water to Aspen electrolyser; 50% of the carbon dioxide is converted for CO_2P method, 75% for COP method. Three main product flows are obtained from the outlet of compartments.

3.2 Gas Separation Unit

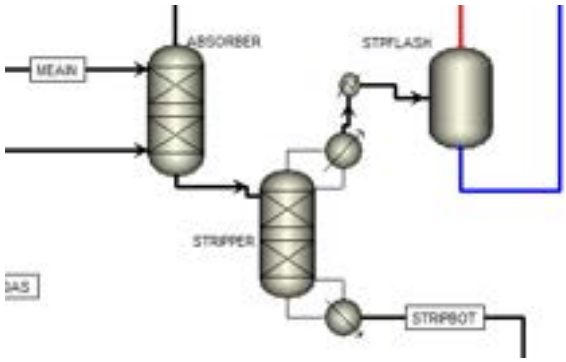


Figure 2. Absorber-Stripper-Flash Column Gas Separation Sequence

The main product of the reaction, ethylene, is the target product in this section, and the gas separation unit also takes responsibility to recycle unreacted carbon dioxide back to the electrolyser. Coming out from the electrolyser unit, the syngas stream, with a total flowrate of $352.6\,\mathrm{kmol/hour}$, is composed of 19.4%mol hydrogen, 16.1%mol ethylene, and 64.5%mol carbon dioxide.

To remove carbon dioxide, a RADFRAC block as absorber is imposed, fed with $5500\,kmol/hour$ stream of monoethanolamine (MEA) with a concentration of 20%mol as the reactive solvent of carbon dioxide, based on thermal dynamic model ENRTL-RK and its following system properties of Henry' s constants of water and MEA. Furthermore, few more properties are user-defined; Henry' s constants for water-CO_2 and CO_2-MEA are obtained from [5,6] respectively. The Arrhenius equation parameters of kinetic and activated energy are obtained from [7]. The manually imposed properties serve the basis of equilibrium and kinetic coefficients of matters in packed absorber and stripper units. Once all set up, ethylene and hydrogen leave the absorber column from top as renewed syngas, and carbon dioxide is absorbed by MEA which leaves from the bottom as carbonate solution. The leaving syngas would then be sent to adsorption unit using activated carbon as the gas bed to filter product ethylene. The bottom solution of absorber is sent to stripper unit for carbon dioxide stripping, and MEA is recycled back to absorber. When carbon dioxide leaves the stripper from top alongside with water vapour, a flash unit operated at 10 bar and 80℃, is in turn to separate the two components, recycling carbon dioxide back to the purification unit before electrolyser. Table 1 presents the optimal parameters of the absorber and stripper column units used in Aspen Plus V11. The gas separation unit section is totally operated at a pressure of 10 bar to match the condition of former electrolysis for a more convenient recycle process. Scaling calculation for the columns is performed manually; take the packed absorber column as an example, Eqn.3 estimates the diameter of the column and Eqn.4 estimates the height of it.

$$A_C = \frac{V_w}{K\sqrt{\rho_V}}\sqrt{\frac{F_P}{g\rho_L}}\left(\frac{\mu_L}{\mu_0}\right)^{0.05} \quad \text{(Eqn.3)}$$

$$H = 1.3 \times \frac{V_w}{k_{OG,a}A_C}\ln\left(\frac{y_B}{y_T}\right) \quad \text{(Eqn.4)}$$

which V_w is the mass flowrate of gas, K is the capacity factor at 80% flood, ρ_V is the gas density, F_P is the packing factor, g is gravitational constant, ρ_L is solvent density, μ_L is solvent kinematic viscosity, μ_0 is constant of 1×10^{-6}, $k_{OG,a}$ is the overall mass transfer coefficient (obtained from [8]), and y_B, y_T are mole fractions of target product in feed and in product respectively. The calculation matches well with the simulation process, absorber column with diameter of 0.9m and height of 25m performs a good hydraulic result and operates below flooding condition. Such scaling calculation will not present in later column sections; the calculation is self-operated by Aspen Plus V11.

3.3 Ethanol Distillation Unit

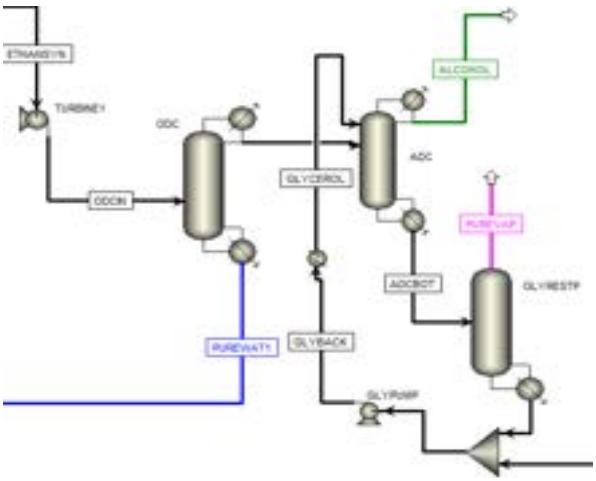


Figure 3. Ethanol Distillation-Stripping Column Separation Sequence

This section is done using the thermal dynamic model of UNIQUAC in Aspen Plus V11. Once left from the electrolyser, the high-pressure stream is first depressurized to 1 bar through a hydraulic turbine, which also could be considered a contribution to energy recycle. The depressurized ethanol alongside with water is sent to the first distillation column, RADFRAC block in Aspen, for rough distillation. Since there exists an azeotrope for the 2-component system of ethanol and water at 78℃, the target of the column is to conduct an ethanol stream consisting of 80%mol ethanol from 10%wt, the feed condition. In the first distillation column, the ethanol-rich stream leaves at the top, flowing to the next azeotropic distillation column; the pure water stream leaves at the bottom, serves as a heating resource for the factory since the exit temperature reaches 99.9℃.

In the second column, glycerol is used as an intermediate solvent to form a solution with water. Purified ethanol of 99.95%mol comes out from the top of the column as distillate, with a flowrate of $22.65\,kmol/hour$; the remaining solution flows to the next stripping column to separate glycerol solvent from water, creating a recycle stream for glycerol. In the stripping column, water is evaporated as steam from the top, which could be used to thrust power generation through a turbine or serve as heat mediate. The >99.999%mol pure glycerol recycles back to the

azeotropic distillation column with a flowrate of 20 $kmol/hour$. The overall ethanol distillation entity operates at constant pressure, 1 bar, and the corresponding column specifications are presented in Table 1.

3.4 Acetic Acid Extraction Unit

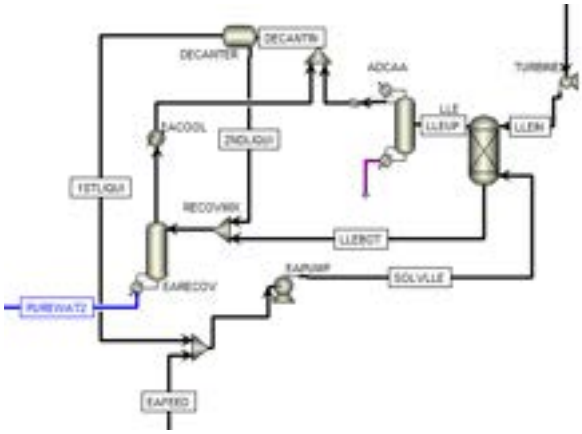


Figure 4. Acetic Acid Extraction-Distillation Column Separation Sequence

This section is done by using thermal dynamic model of UNIQUAC in Aspen Plus V11. The third stream left from 3-compartment electrolyser consists of 20%wt acetic acid and water. The total flowrate of the solution is 488.63 $kmol/hour$ with the pressure from electrolyser outlet of 10 bar. Resembling the procedure dealing with ethanol stream, a depressurisation turbine is amounted to release the extra potential of the liquid stream. After depressurisation, the solution is sent to a liquid-liquid extractor, using model EXTRACT in Aspen for simulation. An extractor solvent is imposed to the column, composed of 86%mol ethyl acetate, to carry acetic acid out from water. The extract solution leaving from the top of column then enters an azeotropic distillation column, again simulated using model RADFRAC in Aspen, to obtain a pure acetic acid bottom stream. The product stream from the distillation column has a flowrate of 34.1 $kmol/hour$, and the purification reaches 99.9%wt acetic acid.

Furthermore, the liquid distillate leaving from top of the column goes to a decanter to separate ethyl acetate from water, giving two streams with different compositions: one ethyl-acetate-rich stream, one water-rich stream. The organic rich stream is recycled directly back to liquid-liquid extractor, and the water-rich stream goes to a similar stripper used in ethanol distillation unit, together with the raffinate stream, which is also water-rich solution with mere composition of ethyl acetate. Finally, as the remaining ethyl acetate in water-rich stream is stripped and leaves the column from top, it is sent to mix with the outcome of azeotropic distillate and re-joins the separation of ethyl acetate from water to recycle the extractor solvent in a maximized utility. At the bottom of the stripping column, pure water stream is produced and can be served as a heat intermediate or recycling reactant. The mentioned columns, specification parameters are produced in Table 1.

3.5 Life Cycle Assessment

Figure 5. Representation of the relations between the impact categories midpoint and the areas of production [19]

Table 2. Economic Report Generated by Aspen Plus V11, component price from website [10, 11, 12, 13] Process assumed to run for 20 years

CO_2 Feed Cost	**69986.9**	**$/hour**
Acetic Acid (99.9%) Sales	**2450.6**	**$/hour**
Ethanol (99.9%) Sales	**1450.0**	**$/hour**
Ethylene Sales	**7620.4**	**$/hour**
Hydrogen Sales	**6809.7**	**$/hour**
Oxygen Sales	**2191.8**	**$/hour**
Total Project Capital Cost	**4.21E+07**	**$/Year**
Total Operating Cost	**7.40E+08**	**$/Year**
Total Raw Materials Cost	**6.14E+08**	**$/Year**
Total Utilities Cost	**6.83E+07**	**$/Year**
Total Product Sales	**1.80E+08**	**$/Year**

The goals of the life cycle assessments (LCAs) are determined based on the scope of this project and literature [18] as follows: To compare and analyse the environmental impact of the simulation model with a conventional ethylene production line. To identify the contributions to the environmental impact of the production of ethylene by the CO_2P process. LCA was performed with the platform openLCA 1.11.0. Ecoinvent was selected as the database and the methods pack was loaded with openLCA LCIA_pack. To obtain a thorough understanding of the environmental impact, the calculations were performed using ReCiPe2016 Endpoint (H) methodology as it aggregates the information into three endpoint categories namely human health, ecosystems, and resource scarcity. The results were normalised with the year 2010 as a reference for impact comparison. The performance of the CO_2P and reference process systems was standardised using 1kg of ethylene produced as the functional unit. Also, a cradle-to-gate system boundary was implemented to satisfy the aim. However, it was difficult to determine the carbon footprint of the inlet stream in the CO_2P process as it was captured from the flue gas which has various compositions and may require pre-treatment depending on the type of industrial processes. Therefore, the CO_2 inlet in the LCA calculation was assumed to be obtained in the air. Whereas other materials input for production were assumed to be obtained from global markets.

4. Results and Discussion

4.1 Techno-Economic Assessment

Aspen Plus V11 is equipped with a full-scale economic analyser. This section is based on the results of simulation streams and blocks, including the price of import of the reactant, solvent, and chemical equipment and the sales product revenue. Table 2. presents the economic status of the process. It is worthy to notice that even though carbon dioxide stream is from an existing flowsheet process, the cost of purification still causes an industrial expenditure of producing pure carbon dioxide, and it takes up to 3 million USD per hour [9]. Water is assumed to be supplied with zero expenditure, and the cost of pump is calculated with other equipment of the process. Product sales price of acetic acid, ethanol, ethylene, and hydrogen are obtained from Made-in-China suppliers; respective cost/revenue of the component is presented in Table 2. Noted that the electrolysis does not solely consume carbon dioxide; it consumes water as well. In the electrolyser, approximately $1000\ kmol/hour$ of water is reduced into corresponding hydrogen and oxygen output. The two basic gases could be recycled as clean energy of the factory and have their values presented in the table, assuming they are worth the market price.

From investment analysis, the cost of this process way exceeds the revenue it brings to the producer. Due to the mechanism of the process in modern industry, petroproducts are alongside with the creation of carbon dioxide as an unwanted by-product, the research is based on CO_2 reduction. Therefore, profiting is not the main purpose. In this report, all carbon dioxide emission from the existing plant participated the electrolysis reaction. The whole process is a decontaminating fitting. It could be useful to factories which locate in strict environmental legislation nations, but currently, this process prohibits a relatively high expenses to major production lines in the world. In the future, however, with rising concern of global warming and the spread of net-zero ideology, this technique could then serve its uses.

Beyond electrical reduction, there are numerous research dedicated to decrease the emission of carbon dioxide. In 2017, a research team in Chinese Academy of Science has explored a reaction pathway to convert carbon dioxide into fuel-qualified gasoline [15]. Recently, a pilot plan was built in Dalian, 2022, which successfully tested the theory with a production rate of 1000 tonne/year of gasoline [16].

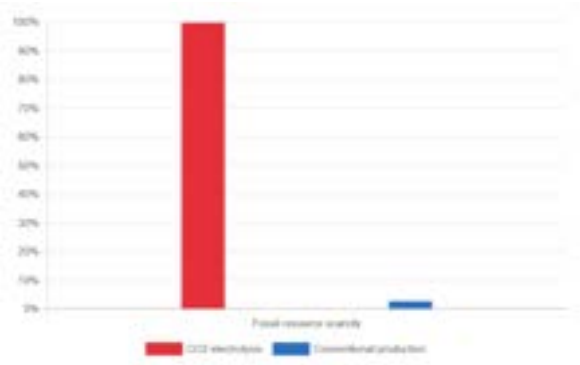


Figure 6. Comparison of the process impact on fossil fuel scarcity

Table 3. Normalized results for CO_2P and conventional production method

Impact category/Process	CO2 Electrolysis	Conventional Process
Resource Scarcity	9.98E+08	2.67E+07
Human Health	9.74E-03	8.18E-05
Ecosystems	6.31E-07	6.21E-09

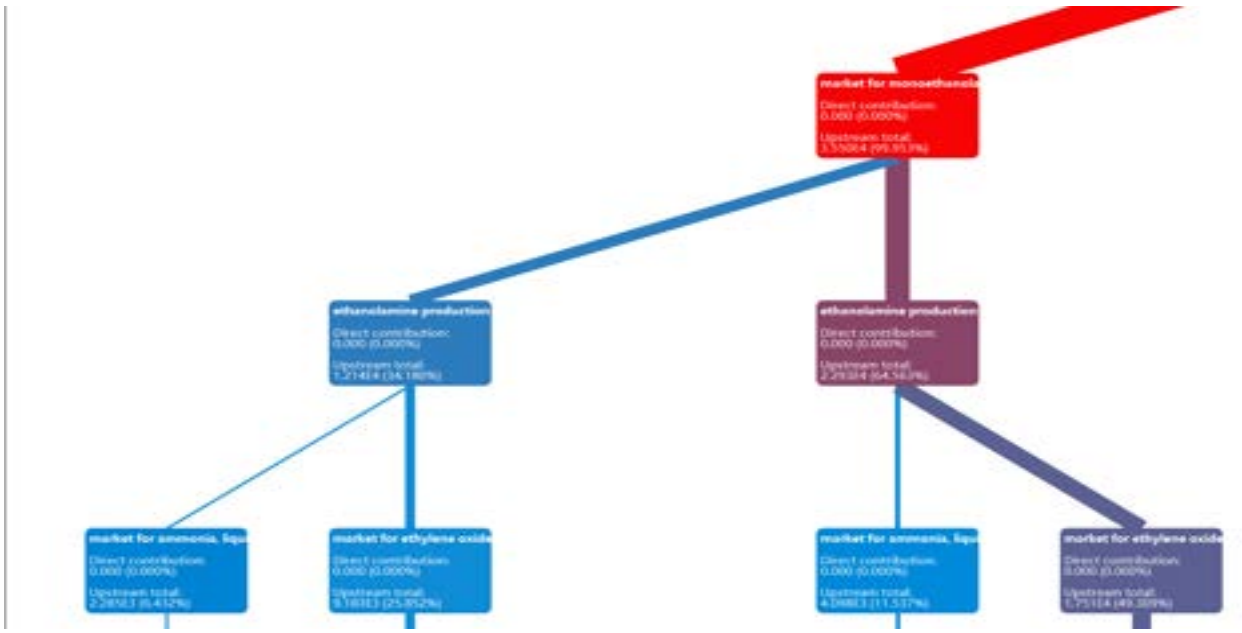


Figure 7. Sanky diagram of the CO_2P process on fossil resource scarcity

4.2 Environmental Assessment

The results were analysed in the categories mentioned above and compared to an existing ethylene production derived from fossil carbon sources. Table 3 illustrates the normalised result for both processes.

According to the impact values, to produce ethylene in either way, the impact exhibited on resource availability is dominating relative to the other two categories. Following a further breakdown of this category, the exploitation of fossil and mineral resources contributes 75% and 25% respectively. The comparison in figure 5 revealed that the CO2 electrolysis process was approximately 100 times harmful to the fossil resource than the conventional process. This result is not desirable as it is indicating that, from environmental perspective, the process being not implemented to factories as means to convert CO2 from the fume gas to other valuable products without further optimisation.

A sanky diagram of the CO2P process was produced to get a better understanding of the reason for the observed large value in fossil resource scarcity. A cut off 3.613% was selected to produce a general shape of the flowsheet. According to the colour scale of figure xx, the production of monoethanolamine (MEA) accounted for over 99.9% impact. It was noticeable that the flowsheet of the upstream process block involved the production of ethylene oxide which was mainly formed from the oxidation of ethylene that extracted from fossil carbon source. Since the absorption process consumed a large amount of MEA, the environmental impact on resource availability outweighed the reference process to a large extent. Possible ways of optimisation include the tuning of separation column parameters to reduce the MEA input without affecting the purity of desired product and the implementation of substitute solvent with less overall damage to environment.

5. Conclusion

Overall, this report presents a possible solution to an existing environmental issue. However, it is still not at an optimal status to conduct a practical pilot plant building. There still exists a blank of adsorption modelling in Gas Separation Unit. This requires a further simulation using Aspen Adsorption as the basis theoretical program to proceed. The adsorption simulation consists of Langmuir equation for accurate adsorbing behaviour to extract ethylene from hydrogen. Current majority uses activated carbon to purify hydrogen [17], yet there are numerous choices of which depend on the economic and efficiency consideration. Additionally, the cost of capital may be further reduced since this report does not imply any heat integration analysis yet. Current assumption is to build heat exchangers between the reboilers and condensers of different columns, which the least amount of 0.9-megawatt energy usage could be saved, contributing approximately 8% of current energy consumption, yet still, the cost to purify carbon dioxide before electrolysis is the main part of expenditure in the process. Furthermore, current international relationship may present a negative impact to global chemical industry, especially in Europe and United Kingdom. It is the time which carbon dioxide emission is not considered as a main issue to enterprises, hence the electric reduction

process of CO_2 may have a more sufficient time to develop better reaction mechanism.

6. Supplementary Information

Simulation Model Available at here*. Calculator spread sheet available here. Aspen economic analysis report available here***. This report is based on the previous research of CO_2 electric reduction by Bert de Mot et al., for electrolyser detail please visit** 10.1021/acs.iecr.1c03592.

References

1. WWF-UK, International work on climate change, [Online] Available at: https://www.wwf.org.uk/what-we-do/projects/international-work-climate-change [Accessed 12 December 2022].

2. Centre for climate and energy solutions, Global Emissions, [Online] Available at: https://www.c2es.org/content/international-emissions/ [Accessed 12 December 2022].

3. Hori, Y. Electrochemical CO2 Reduction on Metal Electrodes. In Modern Aspects of Electrochemistry; Vayenas, C. G., White, R. E., Gamboa-Aldeco, M. E., Eds.; Springer: New York, NY, 2008; Vol. 42; pp 89−189..

4. Electroreduction of CO_2/CO to C_2 Products: Process Modeling, Downstream Separation, System Integration, and Economic Analysis. Mahinder Ramdin, Bert De Mot, Andrew R. T. Morrison, Tom Breugelmans, Leo J. P. van den Broeke, J. P. Martin Trusler, Ruud Kortlever, Wiebren de Jong, Othonas A. Moultos, Penny Xiao, Paul A. Webley, and Thijs J. H. Vlugt, *Industrial & Engineering Chemistry Research* **2021** *60* (49), 17862-17880. DOI: 10.1021/acs.iecr.1c03592

5. S. Takenouchi and G.C. Kennedy, " The Binary System H2O–CO2 at High Temperatures and Pressures" . Am. J. Sci., 262, 1055–1074 (1964)

6. Y. W. Wang, S. Xu, F. D. Otto, A. E. Mather, " Solubility of N2O in Alkanolamines and in mixed solvents" , Chem. Eng. J. 48, 31-40 (1992)

7. H. Hikita, S. Asai, H. Ishikawa, M. Honda, " The Kinetics of Reactions of Carbon Dioxide with Monoethanolamine, Diethanolamine, and Triethanolamine by a Rapid Mixing Method" , Chem. Eng. J., 13, 7-12 (1977)

8. Bravo, J.L., Rocha, J.A., Fair, J.R., 1985. Mass-transfer in Gauze packings. Hydrocarb. Process. 64, 91–95.

9. Qingdao Guida Special Gas Co., Ltd., [Online] Available at: 99.999% 40L Carbon Dioxide - China Carbon Dioxide Cylinder and Carbon Dioxide (made-in-china.com) [Accessed 13 December 2022].

10. Chengdu Taiyu Industrial Gases Co. [Online] Available at: Chinese Supplier 200-815-3 C2h4 Price Ethene - China Ethylene Oxide Gas and Sterilant Mixture Gas (made-in-china.com) [Accessed 12 December 2022].

11. HaiNan chuangyi chemical Co., Ltd. [Online] Available at: Textile Dyeing Use Glacial Acetic Acid 99.85% - China Acetic Acid Glacial Importers and Market Price of Glacial Acetic Acid (made-in-china.com) [Accessed 12 December 2022].

12. Qingdao Eurasia Chemical Technology Development Co., Ltd. [Online] Available at: Hot Sale 99% Purity Ethanol CAS 64-17-5 - China Alcohol and Industrial Use Ethanol (made-in-china.com) [Accessed 14 December 2022].

13. Qingdao Guida Special Gas Co., Ltd. [Online] Available at: 50L 99.999% High Purity Hydrogen H2 Gas Cylinder - China Hydrogen and 50L Gas Cylinder (made-in-china.com) [Accessed 11 December 2022].

14. Chongqing Tonghui Gas Co., Ltd. [Online] Available at: High Pressure 47L Argon/Nitrogen/Oxygen Industrial Gas Cylinder Gas - China 47L Argon and Nitrogen Gas (made-in-china.com) [Accessed 13 December 2022].

15. Wei, J., Ge, Q., Yao, R. *et al.* Erratum: Directly converting CO_2 into a gasoline fuel. *Nat Commun* **8**, 16170 (2017). https://www.nature.com/articles/ncomms16170

16. Carbon dioxide can be turned into gasoline? Scientists have an affirmative answer. [Online] Available at: http://www.xinhuanet.com/politics/2017-06/23/c_1121195021.htm [Accessed 12 December 2022].

17. M.R. Rahimpour, M. Ghaemi, S.M. Jokar, O. Dehghani, M. Jafari, S. Amiri, S. Raeissi, The enhancement of hydrogen recovery in PSA unit of domestic petrochemical plant, Chemical Engineering Journal, Volume 226, 2013, Pages 444-459, ISSN 1385-8947,

https://doi.org/10.1016/j.cej.2013.04.029.

18. Ana Somoza-Tornos, Omar J. Guerra,2 Allison M. Crow,1,2 Wilson A. Smith,1,2 and Bri-Mathias Hodge, Process modelling, techno-economic assessment, and life cycle assessment of the electrochemical reduction of CO2: a review. iScience 24, 102813, July 23, 2021

19. Source: Huijbregts MAJ et al.(2017) Department of Environmental Science, Radbound University Nijmegen.

Graph actor-critic for automated flowsheet synthesis

Samuel Andersson*, Jacob Whyte*
Department of Chemical Engineering, Imperial College London, U.K.
*Contributed equally

Abstract

Reinforcement learning is a promising approach for process flowsheet generation. This paper presents a novel reinforcement learning agent architecture called graph actor-critic (GAC) for graph building tasks such as process flowsheet synthesis. GAC is based on graph convolutional networks and allows for local perception and operation on the flowsheet graph. The proposed GAC is validated in two case studies: one in a simple graph building game and another in a simple chemical process environment for the synthesis of para-xylene. The results demonstrate the potential of deep reinforcement learning in graph building tasks such as chemical process design.

Keywords: Reinforcement Learning; graph convolutional networks; flowsheet synthesis; machine learning; graph actor-critic

1 Introduction

Process synthesis is the art of designing chemical processes that transform raw materials into desired products. It involves a combination of disciplines within chemical engineering, including process design, simulation, and optimization. Currently, process synthesis is mostly carried out by human experts who use their knowledge and experience to create process flowsheets that meet product specifications and constraints such as safety and environmental regulations.[12] However, this process is time-consuming and often involves trial and error, making it difficult to design optimal processes quickly. As demand for sustainable and circular processes increases, there is a growing need for more efficient and effective methods of process synthesis.[10] Artificial intelligence (AI) and other emerging technologies offer promising possibilities for process design and could help the chemical industry transform and become more sustainable.[3] The design of process flowsheets is a complex, time-consuming, and expensive task, that requires expertise, creativity, and intuition. It also involves multiple objectives, constraints, and uncertainties, that need to be balanced, coordinated, and optimized. The current design methods are limited by their heuristics, their assumptions, and their approximations. Therefore, there is a need for a new design method that is more flexible, more accurate, and more efficient, that can handle the complexity, the diversity, and the dynamics of the process design. Reinforcement learning (RL), a sub-field of machine learning (ML), is a promising approach to process flowsheet generation because it allows the design of chemical processes and their flowsheets to be learned and optimized through interaction with a process simulation tool. This allows the RL agent to explore the space of possible flowsheet designs and evaluate their performance in a simulated environment without requiring a priori knowledge or expert guidance. Overall, the characteristics of RL make it well-suited for the task of process flowsheet generation and can enable the automation and optimization of chemical processes in a data-driven and efficient manner.

1.1 Previous approaches and state-of-the-art

Recently, RL has shown its potential to tackle complex decision-making problems by outperforming experts in Chess and GO[14], and has also shown success in process control.[11,7,13] RL allows for the exploration outside the user-defined structure and find alternative solutions without heuristics or prior knowledge of the problem. RL shows great potential in handling large open-ended problems and recent research has shown its applicability to process synthesis, itself a creative design task.[5,9,4,6,15,8] Midgley[8] demonstrates the implementation of RL for designing and optimising a distillation column sequence for non-azeotropic mixtures. He created a simple process simulator environment in which a soft actor-critic[2] RL agent learns to design distillation trains; the agent first decides whether to add a new column and subsequently selects the column's operating conditions. Göttle et al.[4] models flowsheet synthesis as a game of two competing players, who in turn attempt to create more profitable flowsheets than the last. This problem formulation enabled them to reuse DeepMind's AlphaZero monte-carlo tree search RL algorithm[1]. They have since enhanced their work by structuring the agent's actions to consist of several hierarchy levels and using convolutional neural networks (CNNs) to perceive the process state as represented by large flowsheet matrices, thus improving their approach in terms of scalability to large and more complex flow sheeting problems.[6] They successfully demonstrated the usability of their framework for

fully automated synthesis of the *ethyl tert-butyl ether* process. Until recently, a major gap in the literature on RL for process synthesis was in the state representation of flowsheets. In the past, flowsheets were represented in matrices containing features such as unit design specifications, stream data, and topological information. Stops et al.[15] recognised that these matrices are limited when passed through CNNs because they are designed to operate on Euclidean data and do not consider geometric features. Graph convolutional networks (GCNs) are a generalized version of CNNs that can handle varying numbers of node connections and unordered nodes. They can produce useful feature representations and leverage structural information in networks, allowing the topology to be part of the network's input. By using graph neural networks (GNNs), the RL agent can learn to represent and manipulate the flowsheet graph in a more expressive and efficient way, and it can use this knowledge to explore the space of possible flowsheet designs and evaluate their performance. Stops et al.[15] designed an agent using a combination of GCNs and hierarchically structured multilayer perceptrons (MLPs) and demonstrates its ability to design economically viable flowsheets for a case study involving equilibrium reactions and azeotropic distillation.

1.2 Contribution

In this paper, we propose a novel RL agent architecture for graph building tasks such as process flowsheet synthesis, called graph actor-critic (GAC). Our approach is based on GCNs, which enable the learning agent to represent and reason about complex process flowsheets using graph structures. Previous actor-critic approaches, such as the one proposed by Stops et al.[15], use flowsheet fingerprint vectors as global representations of the entire flowsheet graph in a latent space. The GAC approach allows the actor and critic to operate at a local level on the flowsheet graph, making it possible to distinguish between different nodes and their potential actions. Refer to section 2.4 for details about the GAC's architecture. We argue that state representation is a crucial factor for efficient and effective RL. By carefully designing the state representation, we can help the learning algorithm to focus on the most relevant and useful information, and to ignore irrelevant or noisy information that may hinder learning. To this end, we propose a new state representation that is described in detail in section 2.3. We believe that our proposed state representation has several advantages over existing approaches, and that it can improve performance and reliability.

2 Reinforcement learning for process synthesis

RL focuses on sequential decision making and enables an agent to learn in an interactive environment by trial and error, using feedback from its own actions and experiences. In RL, the agent observes the state of the environment, selects an action according to its policy, and receives a reward or penalty from the environment based on its action. The agent then updates its policy to maximize its expected cumulative reward. A RL problem can be expressed as a Markov decision process (MDP), meaning the current action of the agent depends only on the current state, and not on the history of the past states and actions. An MDP consists of two entities: the agent and the environment. The agent observes the state of the environment and selects an action according to its policy, which is a mapping from states to actions. The environment then transitions to a new state based on the selected action and provides feedback to the agent in the form of a reward. This sequence continues until a terminal state is reached, resulting in a series of state-action-reward tuples {S, A, R}. The agent attempts to learn a behaviour that maximizes its cumulative reward by updating its policy incrementally.

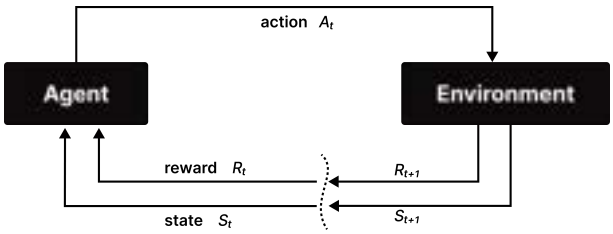


Figure 1: Basic depiction of a Markov Decision Process.

2.1 Graph Game

The Graph Game is an environment designed to test and evaluate the performance of the GAC RL agent. The Graph Game environment has several advantages over a process simulation environment, such as interpretability, computational efficiency, and controllability. The Graph Game environment allows us to gain insight into how both the actor and the critic make decisions, and to modify the agent's hyperparameters before testing on a process environment. In order for an agent to succeed at the game, it must gain an understanding of the directed structure of the graph in order to maximise the game's non-linear reward structure.

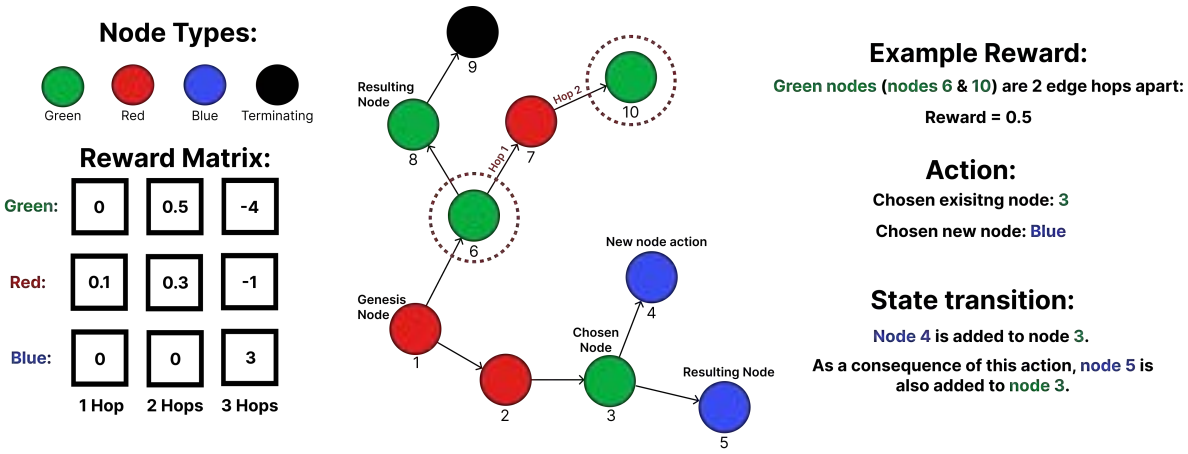


Figure 2: Example illustration of the Graph Game depicting an action, resulting state transition and example reward calculation.

2.1.1 Agent-environment interaction for the Graph Game

The Graph Game consists of four types of nodes: Red, Blue, Green, and Black (a terminating node). The game is initialized with a single genesis node, and the graph is created in an iterative manner. The agent observes the graph and chooses an action, which consists of selecting a node type and an existing node in the graph to create an edge from. The chosen node type is then added to the graph, and the graph transitions to the next state. The process is repeated until either the maximum number of nodes is reached, or there are no available nodes left in the graph to add to. A node is considered unavailable if it is a terminating node, or if it has reached a given maximum number of connections. Some actions result in additional resultant nodes being added to the graph, for example, if the agent chooses to add a blue node to an existing green node, a resulting blue node is added to the green node, as to increase the complexity of the game and to more resemble a process environment.

Rewards in the Graph Game are calculated based on the number of edge hops between nodes of the same colour. For example, red nodes directly connected to each other receive a reward of *0.1*, while green nodes two edge hops apart receive a reward of *0.5*. An edge hop is the act of crossing an edge from one node to another. The cumulative reward of the entire graph is computed at every transition and is used as feedback for the agent. Figure 2 illustrates the Graph Game with an example action and resulting state transition, as well as an illustration of how rewards are calculated.

2.2 Process environment

The purpose of the process environment is to demonstrate that a GAC agent can be effective in a process design context. The environment consists of reactor units and separation units. Reactors are modelled as gas phase isothermal plug flow reactors (PFRs) with a heterogeneous catalyst. Four reactions take place in the reactor, both parallel and consecutive, and the reaction rates are a function of the mass of the catalyst. Separators are modelled as distillation columns (DCs), and several correlations are used to model their operational conditions, physical dimensions, and cost. The feed steam consists of primarily methanol and toluene, with para-xylene being the most profitable product. The less valuable side products produced by side reaction were also sold if the minimum purity was met.

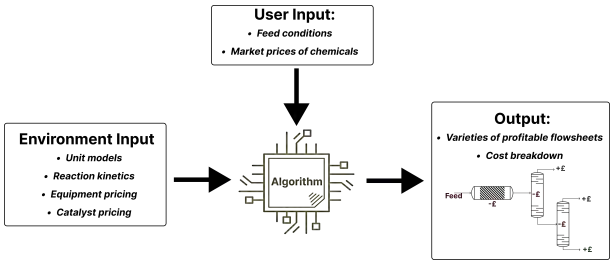


Figure 3: Objective of the GAC algorithm in a process design context.

2.2.1 Agent-environment interaction for the process environment

The process graph is generated iteratively, like in the Graph Game. However, the state and action space in the process environment are more complex. The agent must make both discrete and continuous actions. The available discrete actions are to add a PFR, a DC, or a terminating stream. Continuous actions associated with the PFR include its physical dimensions, such as length and cross-sectional area, as well as the mass of catalyst. Cost correlations are used to calculate the annualized cost of the specified PFR unit based on its dimensions and the amount of catalyst used. Actions associated with the DC include selecting a component as the light-key, and light-key split, which are used to calculate the internal column dimensions, number of stages, and operating conditions. An annualized cost is calculated from the process unit's respective design variables. Positive rewards are given for selling components in terminating streams if their purity is above 90 mol%. Process flowsheet synthesis can be formulated as a MDP as depicted in figure 4. The goal of the agent is to produce varieties of profitable flowsheets, with their associated cost breakdowns given the initial

feed conditions and market price of chemicals, as seen in figure 3

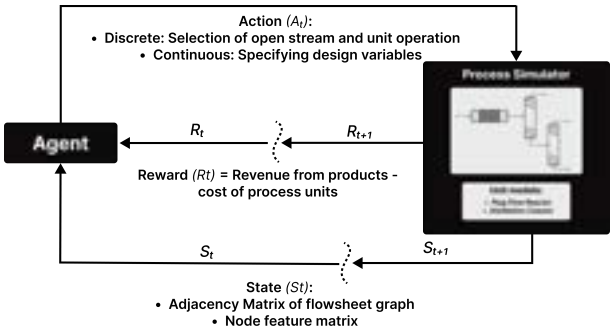


Figure 4: Markov Decision Process in a flowsheet synthesis context.

2.3 State representation

Stops et al.[15] represents process units as nodes, and streams as edges with edge features. Although streams are represented as edges in process flowsheets for human engineers, this is not necessarily the best representation for a GNN. In this paper, streams were also represented as nodes, which we believe has numerous advantages. Representing process units as nodes and streams as edges provides a more compact representation of the process flowsheet, which may make the graph easier to process and reduce computational complexity. However, this representation may not capture the full complexity of the process flowsheet because it does not explicitly show the interactions between process units and streams. On the other hand, representing both process units and streams as nodes allows for a more detailed representation of the process flowsheet because it explicitly shows the interactions between process units and streams in the graph structure. This may provide the agent with more information and enable it to make better decisions. Additionally, when streams are represented as nodes then there is no need to add unspecified process units upon node classification as was necessary in Stops et al.[15] This would be problematic in GAC where every legal node is considered, meaning that unspecified process units would be created for every open stream upon every forward pass.

2.4 Agent architecture

We devised a novel RL agent architecture which we believe is especially suited to flowsheet generation. GAC is a RL method that uses GNNs to learn and optimize policies for process flowsheet synthesis tasks. In GAC, the RL agent takes actions on the flowsheet graph by proposing operations and design variables for each node in the graph, and it uses GNNs to evaluate the expected reward of the proposed actions. The GAC agent can learn from the feedback of the environment, such as the energy and cost efficiency of the generated flowsheets, and it can adaptively update its policies to improve the performance over time. GAC has several advantages over actor-critic for process flowsheet synthesis tasks. For example, GAC can provide more fine-grained and flexible control over the action selection process, as it can evaluate and select actions for every node in the flowsheet graph, rather than only for the whole graph. This can enable the GAC agent to explore the space of possible flowsheets more efficiently and to learn more complex and dynamic policies. In addition, GAC can decouple the actor and the critic in the action evaluation process, which can improve the accuracy and stability of the learning process, and it can avoid bias and overfitting in the policy optimization. GCNs are a type of GNN that have been widely used and studied for a variety of RL tasks, including process flowsheet synthesis. We chose to use GCNs as they can effectively capture the local and global structural patterns in the flowsheet graph, and they can be trained efficiently using gradient-based optimization algorithms. Overall, GAC can be a more effective and efficient method for process flowsheet synthesis tasks than traditional actor-critic. By using GNNs to evaluate the expected reward of the proposed actions, GAC can enable the RL agent to learn complex and dynamic policies that can optimize the performance of the generated flowsheets.

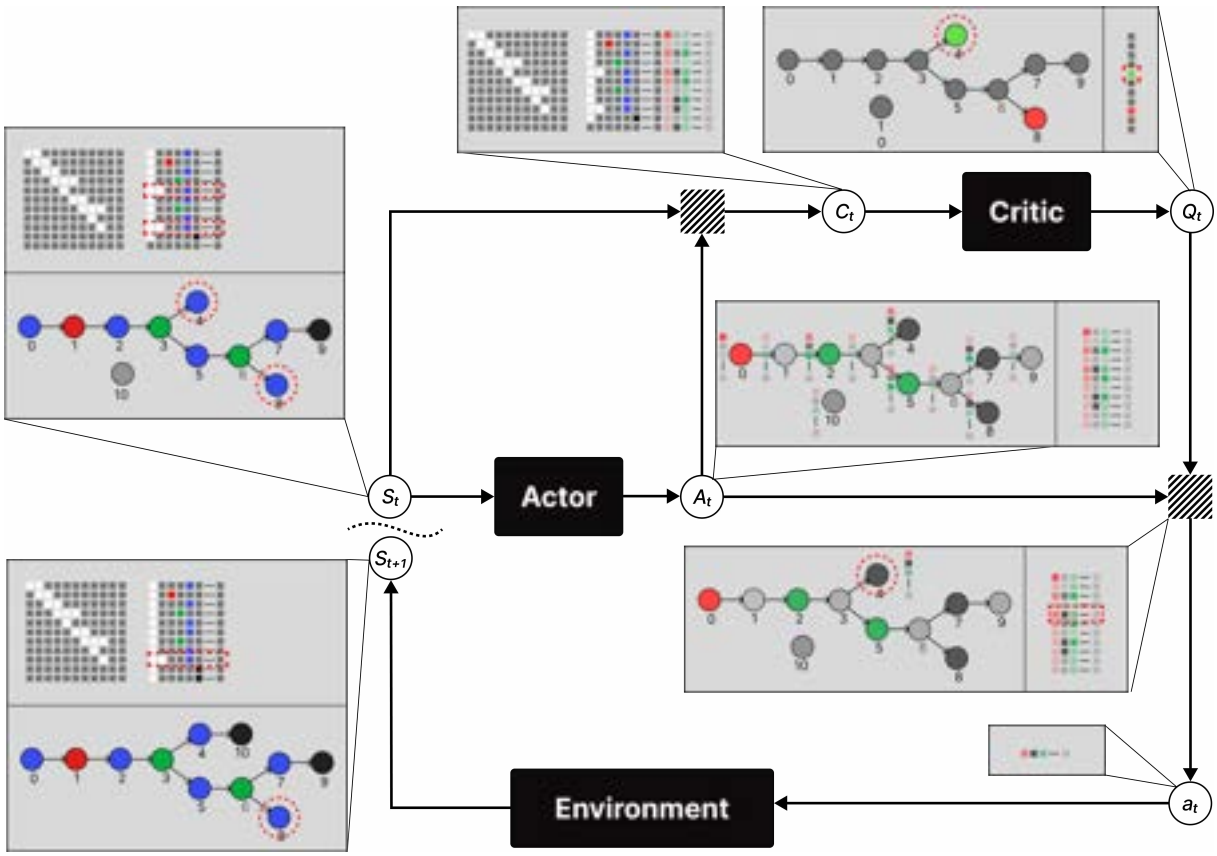


Figure 5: Graph actor-critic architecture

2.5 Training and Implementation

In this paper, we present a simple exploration strategy for a reinforcement learning agent applied to the chemical flowsheet design problem. Our approach involves running the agent for 1000 games between episodes to explore the design space and collect experience in the form of transitions. A transition is a tuple consisting of the current state, the next state, and the associated reward. The set of transitions resulting from one game is referred to as a trajectory. Transitions are sorted and stored in replay buffers, which are a type of data structure. During training, batches of transitions are sampled from the replay buffers and used to update the agent. The method of sampling and batching of transitions is critical to the efficiency of training, and is a central problem in RL for which sophisticated strategies such as curriculum learning are a hot topic of research. In this paper, we employ simple strategies for this purpose.

2.5.1 Method

First, a completely random agent is run for 1000 games to collect basic information about the environment. The agent is trained on random batches of transitions, which allows it to gain a basic understanding of the process environment. From this point on, the agent is run for 1000 games with a depth-based exploration and the experience is stored and sorted into a separate set of policy buffers. The exploration is a function of both depth and episode number. In the first episode, the agent has an exploratory phase lasting a few steps in the process design, after which it mostly follows its policy greedily. In each subsequent episode, the zone of exploration shifts a constant number of steps deeper in the design. Ultimately, the choice of how to batch samples of transitions from the replay buffers was a significant factor in performance. See section 3.2.2 for a comparison between random batch transition sampling and batch trajectory sampling.

2.5.2 Balancing Exploration and Exploitation

In the context of reinforcement learning, it is crucial to strike a balance between exploitation and exploration. The agent should exploit its previous experience by avoiding poor trajectories and designs, while also exploring variations of successful trajectories and unexplored design space. This balance is particularly difficult in process synthesis, where a minor alteration in a design trajectory may result in significantly different profitability. In order to address this challenge, we propose an exploration strategy that is suitable for the chemical flowsheet design problem. Our strategy allows the agent to first gain a basic understanding of the process environment through random exploration of the design space. Subsequent depth-based exploration allows the agent to focus on specific areas of the design space, exploring deeper in a targeted manner as the training episodes progress. The use of a normal distribution for the exploration function allows the agent to explore a range of depths within each episode,

balancing exploration and exploitation to improve performance over time.

2.5.3 G-DDPG Update

The paper introduces G-DDPG (Graph Deep Deterministic Policy Gradients), an adaptation of the Deep Deterministic Policy Gradients (DDPG) algorithm for use with Graph Neural Networks (GNNs). Unlike traditional DDPG, where the critic network produces a single scalar value as the predicted Q-value, GNNs produce a vector of q-values, one for each node in the graph. To update the GNN critic, a target Q-value must be specified for every node-action pair. However, only one action is taken in the environment with only one corresponding reward received. This raises the question of how to determine the target Q-values for the remaining node-action pairs. To address this issue, we propose G-DDPG, which adapts the DDPG algorithm for use with GNN actor-critic frameworks, where both the actor and critic networks are GNNs. This allows for the evaluation of a proposed action for every node in the graph, as well as the corresponding expected Q-value for every node-action pair. In G-DDPG, the target Q-value vector is computed in the following way: for the chosen action, the corresponding target Q-value in the target Q-value vector is updated using the true reward of the next state, as in DDPG. The remaining target q-values are set to the legal predicted q-values of the current state, with illegal q-values being nullified. This approach allows the critic to learn from a single example without altering its predictions for other actions.

3 Results and Discussions

3.1 Results from graph game

Figure 6 shows the change in performance of the GAC agent as it is trained in the Graph Game. The distributions of rewards are plotted for 1000 game simulations in a graph with a maximum size of 50 nodes. The performance of GAC's noisy policy is compared with that of a random agent, which takes completely random actions at every time step. GAC performs worse than the random agent at the beginning of training but shows an outperforming reward distribution after the third training episode. It converges to its optimal policy and the distribution remains relatively constant after the seventh episode of training. This shows it can reliably outperform a baseline random actor for an open ended, sequential decision-making problem with a nonlinear reward structure. The most useful results from the graph game can bee seen in figure 7, the training dashboard. The dashboard is allows for gaining insight to how both the actor and critic are thinking, and how their decision making evolves with each training episode. The three graphs in the top right are called inspection graphs, they illustrate every possible sequence of nodes for a three edge-hop neighbourhood. Halos around the nodes represent how valuable the critic thinks it is to act on each node. A red halo indicates a negative Q-value, while a green halo indicates a positive Q-value. The size of the halo is indicative of the magnitude of the Q-value. The plots on the left give insight on the critic and show the Q-values of adding different node types to the end of different node sequences. The plots under the inspection graphs indicate the actions the actor is likely to propose given certain node sequences. After five episodes of training, the GAC agent develops a good understanding of which node sequences are worth while pursuing, as well as which actions to take to maximise cumulative reward. Furthermore, we know the agent has essentially solved, or figured out the optimal policy of the game as the inspection graphs are representative of how to fully maximise the provided reward structure.

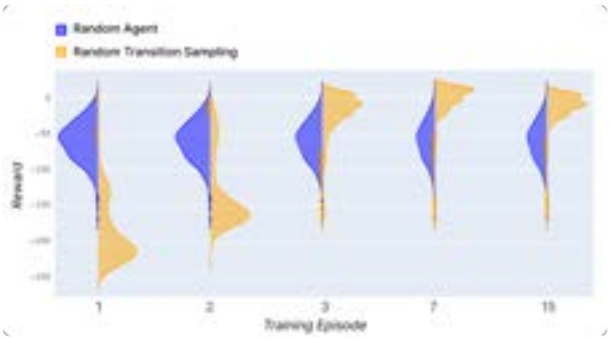


Figure 6: Comparative distribution of reward over 1000 simulation on a graph with maximum of 50 nodes. Violin plots comparing GAC agent performance with a random agent over several training episodes.

Figure 7: Training Dashboard for Graph Game after 5 episodes of training.

3.2 Results from process environment

3.2.1 Benchmarking against random agent

Figure 8 shows distributions of annual profit for 1000 simulations in the process environment. An agent with a completely random policy is shown in blue and is used to benchmark the performance of our agent trained with random transition batch updates, shown in yellow. The agent following the random policy never creates a flowsheet with annual profit above $10 million, with most of its flowsheets resulting in a negative profit. The trained agent exhibits a bimodal distribution, with one its mode being centred around a similar annual profit of that seen in the random agent. The other mode is centred around a much higher annual profit of $15 million, with some flowsheets producing annual profits as high as $22 million. The bimodal distribution is a consequence of the agent following a noisy policy, which means it's not always making what it thinks is the most optimal decision at every step. A noisy policy can allow the agent to explore a broader range of possibilities and consider a wider variety of options. This can potentially lead to the discovery of new and improved flowsheet designs that may not have been found with a deterministic policy. Flowsheet synthesis is particularly sensitive to exploratory actions policy because it is a complex and dynamic problem that involves many variables and decision points. As a result, many flowsheets produced by the agent result in a low annual profit, as its exploratory actions on that particular simulation did not pay off. An example of such actions could be not creating a reactor early in the process, thus not producing the valuable product para-xylene until too late in the process. Alternatively, the agent could have made a very promising flowsheet, but then decided not to create product streams, thus not selling the valuable streams it produced. These results indicate that GAC is able to outperform an agent following a random policy for flowsheet generation. It is worth noting that the economic potential of the process is $40.05 million in annual profit, assuming perfect separation and reactor conversion. This potential value is determined by the value of the product and the cost of the raw materials, without considering the costs of the necessary process units. The agent using GAC was able to produce flowsheets with over half the profit of the economic potential, which is a promising result. This suggests that there may not be much room for improvement in the agent's performance.

3.2.2 Comparing trajectory and random transition updates

Figure 9 compares the annual profit distributions of two agents trained with different methods: random batch transition updates (yellow) and trajectory batch updates (blue). Both agents follow a noisy policy and exhibit two main modes in their distributions, one around $0 million in annual profit and the other with a much higher annual profit. Upon closer inspection of the flowsheets that yielded high profits, the blue distribution shows that the agent trained with trajectory updates generally outperforms the other. Most of the flowsheets it creates have annual profits above $20 million, while the majority of those created by the agent trained on random transitions have annual profits of around $14 million. Furthermore, the trajectory trained agent converged on its optimal policy almost three times faster than the agent trained on random transitions. This suggests that the trajectory training

method allows the agent to learn the best actions more quickly and effectively, due to its exploration being a function of depth. Additionally, the blue distribution is less broad than the yellow one, indicating that the trajectory trained agent is able to consistently achieve higher annual profits with fewer fluctuations. This suggests that the trajectory training method leads to more stable and reliable performance.

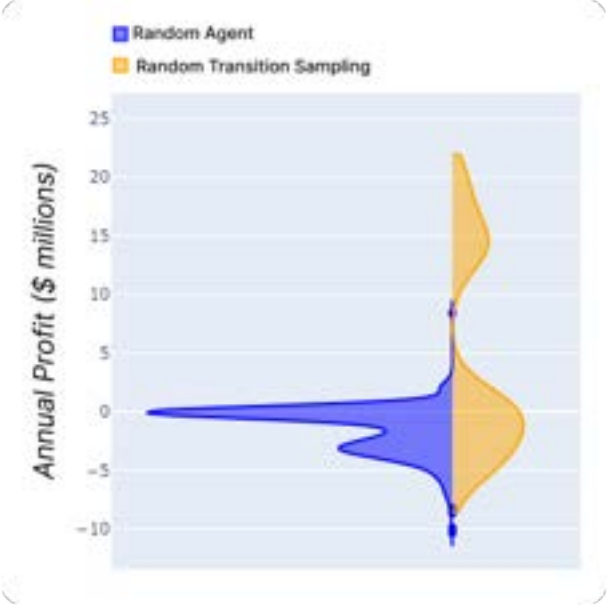


Figure 8: Distribution of annual profit over 1000 simulation comparing an agent with a random policy and an agent trained with random transition sampling.

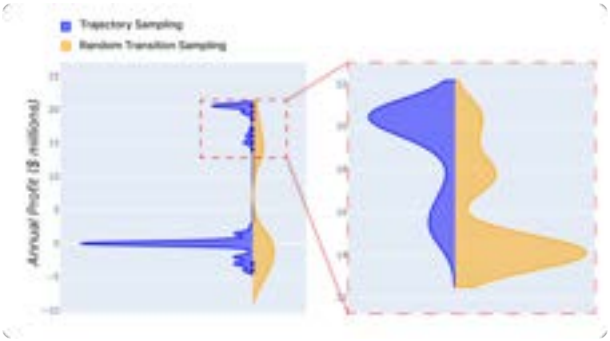


Figure 9: Distribution of annual profit over 1000 simulations comparing an agent trained with random transition sampling and an agent trained with trajectory sampling.

3.3 Future Work

3.3.1 Broadening the scope and applicability in industry

One major limitation is that the model currently only considers a single objective, such as maximizing profit or minimizing emissions. In the real world, process synthesis often involves multiple objectives, such as maximizing profit while minimizing emissions and ensuring safety. Therefore, future work on GAC should consider developing methods for handling multiple objectives in the process synthesis problem. Additionally, the current implementation does not consider constraints on the process, such as maximum energy consumption. Incorporating such constraints into the algorithm would make it more realistic and applicable to real-world process synthesis problems.

The current implementation is limited to a specific type of process synthesis problem, namely the synthesis of continuous processes. However, many industrial processes are batch processes, which require different modelling approaches. Therefore, future work should consider extending it to handle batch processes as well. One potential way to broaden the scope of application is with the use of a commercial process simulation software such as ASPEN. It is designed to model a wide range of processes from simple distillation columns to complex multistage processes with heat and mass transfer, reaction kinetics, and other phenomena. This would allow for the creation of a more realistic and accurate representation of a process compared to a simplified model.

3.3.2 Implementation of hierarchical reinforcement learning

Göttle et al.[6] and Stops et al.[15] have shown the application of hierarchical reinforcement learning (HRL) is very promising in the space of process synthesis. HRL can be applied to process flowsheet synthesis by decomposing the design process into multiple levels or hierarchies of decision making. At the high-level, a RL agent could be trained to learn a policy for selecting the overall structure of the flowsheet that is most likely to meet the specified performance criteria. This policy could be based on the characteristics of the process and the flowsheet design, such as the chemical reactions involved, the properties of the process inputs and outputs, and the constraints on the flowsheet design. The agent could learn to take actions that maximize the reward signal, such as minimizing the cost or environmental impact of the flowsheet, and it could use this knowledge to select the most promising flowsheet structures. At the low-level, the RL agent could be trained to learn a policy for selecting the detailed operating conditions within the flowsheet. This policy could be based on the characteristics of the individual process units, such as the kinetics of the chemical reactions, the heat and mass transfer rates, and the energy and material balances. The agent could learn to maximize a reward such as the yield of the process or minimizing the energy consumption of a unit operation. One of the key benefits of HRL is improved learning efficiency. By dividing a complex task into smaller subtasks, HRL enables the learning algorithm to focus on learning each subtask separately, which can improve learning efficiency compared to trying to learn the overall task all at once. This can reduce the amount of computational resource required to learn a given task, allowing you to train your model more quickly. Another benefit of HRL is better generalization. By breaking a complex task into subtasks, HRL enables

the learning algorithm to learn more generalizable skills that can be applied to a wider range of tasks and environments. This can help improve the adaptability and robustness of your model, allowing it to perform well on a greater variety of process flowsheet generation problems. In addition to these benefits, HRL can also increase modularity and reusability. By learning and reusing individual subtasks across different tasks and environments, HRL improves modularity and makes it easier to reuse components of your learning model in different contexts. This can save time and effort compared to starting from scratch each time you want to solve a new process flowsheet synthesis problem.

3.3.3 Other Applications

The GAC framework introduced in this paper could potentially be applied to other areas relevant to chemical process design, such as retro-synthesis for the discovery of alternative synthetic routes. Retro-synthesis is a problem in chemical engineering that involves predicting the sequence of reactions needed to synthesize a given target molecule from a set of starting materials. One approach to solving the retro-synthesis problem is to use machine learning algorithms to automatically search through the space of possible reaction pathways and identify the most promising ones. This can be achieved by representing molecules and reactions as graphs, with atoms and bonds as nodes and edges, respectively, and using graph neural networks to learn the structural and chemical properties of these graphs. By defining a reward function that rewards the agent for finding pathways that are feasible, efficient, and cost-effective, the agent can be trained to predict the optimal pathway for synthesizing a given molecule.

Overall, the GAC framework has the potential to be applied to a wide range of problems in chemical engineering, making it a valuable tool for chemical engineers in the future.

3.3.4 Use of Attention

Another type of GNN that may be suitable for the proposed GAC agent is graph attention networks (GATs). GATs are a variant of GCNs that use attention mechanisms to weight and combine the features of the flowsheet graph nodes and edges, which can enable the agent to focus on the most relevant and informative parts of the graph One of the potential advantages of using GATs for process synthesis is their interpretability. This is because the attention mechanisms in GATs can provide insights into the decision-making process of the GAT. For example, the node- and edge-level attention can provide information about the most relevant operations and equipment, and the most relevant interactions between them. This can help to explain the decisions made by the GAT, and to understand the reasons behind them.

4 Conclusion

In this paper, we present a novel approach to the task of chemical process synthesis by combining graph neural networks and deep reinforcement learning. Our proposed framework, known as the Graph Actor-Critic (GAC), utilizes both GNNs for the actor and critic, and the critic plays a key role in the decision making process by simultaneously considering and weighing multiple actions at each step in the flowsheet design.

To enable the GNN critic to learn from a single example without affecting its predictions for other actions, we propose a new version of the Deep Deterministic Policy Gradient algorithm called Graph Deep Deterministic Policy Gradients (GDDPG). We test GAC in two case studies. The first study validates its ability to learn a robust policy within a non-linear reward structure based on graph topology and node features using a simple graph building environment known as Graph Game. The second study applies GAC to the more complex task of chemical process design within a simple process environment, incorporating reactors, distillation columns, and product streams as discrete actions, and process unit specifications as continuous actions for the synthesis of para-xylene.

In order to effectively explore the vast design space, we employ a depth-based exploration strategy. Despite operating in a hybrid action space that is not well-suited for a DDPG algorithm and the actor lacking a hierarchical structure, the GAC agent successfully learns to generate profitable and reasonable process designs. Furthermore, we observe that updating in batches of trajectories leads to significant performance improvements compared to random samples of transitions.

Overall, our results demonstrate the potential of deep reinforcement learning in graph building tasks such as chemical process design. We argue that there is no fundamental reason why hierarchical RL could not be extended from a process flowsheet down to a P&ID or even a full CAD design. This could greatly impact the chemical engineering industry by allowing for the detailed automated design of complex chemical processes.

However, we also identify several avenues for future research and development. In particular, we argue that further innovation is needed in the state representation to enable local and global process perception. In the current work, a major limitation is that the directed graphs only support upstream perception of the process. In order for an agent to effectively utilize recycles, it is critical that it incorporates downstream information in its policy. Thus, new state representations and GNN architectures should be explored to this end.

In the context of an increasingly circular chemical industry, this kind of model could be used to explore alternatives to valorize waste streams. As chemical processes become more complex and environmentally sustainable, the ability to automatically design profitable and sustainable processes will become increasingly important. Our GAC framework represents an important step towards this goal, and we believe it has the potential to be a valuable tool for chemical engineers in the future.

References

[1] David Silver et. al. "Mastering the game of Go with deep neural networks and tree search". In: *Nature* 529.7587 (Jan. 2016), pp. 484–489. DOI: 10.1038/nature16961. URL: https://doi.org/10.1038%5C%2Fnature16961.

[2] Haarnoja et. al. *Soft Actor-Critic Algorithms and Applications*. 2018. DOI: 10.48550/ARXIV.1812.05905. URL: https://arxiv.org/abs/1812.05905.

[3] Peter Fantke et al. "Transition to sustainable chemistry through digitalization". In: *Chem* 7.11 (Nov. 2021), pp. 2866–2882. DOI: 10.1016/j.chempr.2021.09.012. URL: https://doi.org/10.1016%5C%2Fj.chempr.2021.09.012.

[4] Quirin Göttl, Dominik G. Grimm, and Jakob Burger. "Automated synthesis of steady-state continuous processes using reinforcement learning". In: *Frontiers of Chemical Science and Engineering* 16.2 (May 2021), pp. 288–302. DOI: 10.1007/s11705-021-2055-9. URL: https://doi.org/10.1007%5C%2Fs11705-021-2055-9.

[5] Quirin Göttl, Dominik G. Grimm, and Jakob Burger. "Using Reinforcement Learning in a Game-like Setup for Automated Process Synthesis without Prior Process Knowledge". In: *Computer Aided Chemical Engineering*. Elsevier, 2022, pp. 1555–1560. DOI: 10.1016/b978-0-323-85159-6.50259-1. URL: https://doi.org/10.1016%5C%2Fb978-0-323-85159-6.50259-1.

[6] Quirin Göttl et al. "Automated Flowsheet Synthesis Using Hierarchical Reinforcement Learning: Proof of Concept". In: *Chemie Ingenieur Technik* 93.12 (Aug. 2021), pp. 2010–2018. DOI: 10.1002/cite.202100086. URL: https://doi.org/10.1002%5C%2Fcite.202100086.

[7] Zhong-Ping Jiang, Tao Bian, and Weinan Gao. "Learning-Based Control: A Tutorial and Some Recent Results". In: *Foundations and Trends® in Systems and Control* 8.3 (2020), pp. 176–284. DOI: 10.1561/2600000023. URL: https://doi.org/10.1561%5C%2F2600000023.

[8] Stephan C. P. A. van Kalmthout, Laurence I. Midgley, and Meik B. Franke. *Synthesis of separation processes with reinforcement learning*. 2022. DOI: 10.48550/ARXIV.2211.04327. URL: https://arxiv.org/abs/2211.04327.

[9] Ahmad Khan and Alexei Lapkin. "Searching for optimal process routes: A reinforcement learning approach". In: *Computers &amp Chemical Engineering* 141 (Oct. 2020), p. 107027. DOI: 10.1016/j.compchemeng.2020.107027. URL: https://doi.org/10.1016%5C%2Fj.compchemeng.2020.107027.

[10] S.Venkata Mohan and Ranaprathap Katakojwala. "The circular chemistry conceptual framework: A way forward to sustainability in industry 4.0". In: *Current Opinion in Green and Sustainable Chemistry* 28 (Apr. 2021), p. 100434. DOI: 10.1016/j.cogsc.2020.100434. URL: https://doi.org/10.1016%5C%2Fj.cogsc.2020.100434.

[11] M. Mowbray et al. "Safe chance constrained reinforcement learning for batch process control". In: *Computers &amp Chemical Engineering* 157 (Jan. 2022), p. 107630. DOI: 10.1016/j.compchemeng.2021.107630. URL: https://doi.org/10.1016%5C%2Fj.compchemeng.2021.107630.

[12] Naonori Nishida, George Stephanopoulos, and A. W. Westerberg. "A review of process synthesis". In: *AIChE Journal* 27.3 (May 1981), pp. 321–351. DOI: 10.1002/aic.690270302. URL: https://doi.org/10.1002%5C%2Faic.690270302.

[13] Joohyun Shin et al. "Reinforcement Learning – Overview of recent progress and implications for process control". In: *Computers &amp Chemical Engineering* 127 (Aug. 2019), pp. 282–294. DOI: 10.1016/j.compchemeng.2019.05.029. URL: https://doi.org/10.1016%5C%2Fj.compchemeng.2019.05.029.

[14] David Silver et al. "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play". In: *Science* 362.6419 (Dec. 2018), pp. 1140–1144. DOI: 10.1126/science.aar6404. URL: https://doi.org/10.1126%5C%2Fscience.aar6404.

[15] Laura Stops et al. *Flowsheet synthesis through hierarchical reinforcement learning and graph neural networks*. 2022. DOI: 10.48550/ARXIV.2207.12051. URL: https://arxiv.org/abs/2207.12051.

Technoeconomic Assessment on Hf-Beta Zeolite-Catalysed Glucose-Fructose Isomerisation

Noor Mellina Abu Kasim and Marsya Maisarah Ahmad Sharifuddin
Department of Chemical Engineering, Imperial College London, U.K.

Abstract The transformation of non-consumable second-generation biomass to high value platform chemicals such as 5-hydroxymethylfurfural (5-HMF) via glucose-fructose isomerisation (GI) contributes to the collective effort of shifting energy dependency away from the non-renewable fossil fuels, thus allows a step further into building a more sustainable future. Large-scale GI currently relies on the catalytic activity of the enzyme xylose isomerase as seen in the production of high-fructose corn syrup (HFCS). However, since the biological catalysts present some drawbacks such as the need to operate at narrow pH and temperature windows, studies revolved around the search of a chemo-catalyst that can perform better than the enzymes have become more popular. Hafnium-beta zeolites (Hf-Beta) are one of the many Lewis acid catalysts being studied for this subject and it has been recently proven to perform well in continuous operations of GI. To give an economical point of view on the Hf-Beta catalysed GI reaction and provide context to their economic values, a technoeconomic assessment was conducted on this process in which optimisation procedures and economic evaluations were carried out and compared alongside the biological system. This was done through modelling a simulation on Aspen Plus V11 based off kinetic data from experimental results found in literature. The results of the model showed that the optimum temperature to operate the model at was 140ºC with 30 wt% water in solvent for a specified glucose conversion of 50%, which gave fructose yield and selectivity of 45.6% and 91.2% respectively. The total fructose production cost amounted to a value of $1.68 per kilogram of fructose and was compared against fructose price in commercialised HFCS-42 and HFCS-55 where it was concluded that further refining on the cost comparison method should be worth pursuing.

Introduction & Background

As the world slowly transitions into greener and more sustainable methods of supplying energy and chemical production, dependency on non-renewable fossil fuels is being reduced by introducing environmentally friendly feedstock substitutes such as biomass[1]. Biomass can be any organic material or waste that contains chemical building blocks (i.e. carbon and hydrogen) which can be used to generate bioenergy for application as fuels and power production[2]. Although both fossil fuels and biomass originates from living organisms, the main contrast between the two is that the former is unable to re-absorb the carbon it emits whereas the latter has the ability to do so and contributes to the exchange occurring in the carbon cycle[3]. This is simply because the growth of biomass themselves remove carbon dioxide from the atmosphere[4] as they come from recently living organisms unlike fossil fuels which releases carbon that has been concealed away from thousands of centuries ago[3]. As fossil fuels currently supply around 80%[5] of the world's energy while contributing over 75%[6] and almost 90%[6] of greenhouse gases and carbon dioxide emissions respectively, this transition is a very much important step towards achieving world sustainability goals.

Lignocellulosic biomass is biomass rich in cellulose, hemicellulose, and lignin, and is a form of second-generation biomass feedstock which are not suitable for consumption[7]. It is a promising source of bioenergy as they are a highly abundant and renewable natural resource on Earth[7]. The conversion of lignocellulosic biomass to platform chemicals such as furans and 5-hydroxymethylfurfural (5-HMF) are of high interest not only in the scope of biofuel production, but also because they can be readily upgraded into molecules with high potential for generating fuel-derived or polymer-derived products[8]. This said "conversion" involves the critical step of transforming the simple carbohydrate glucose, which is a constituent of the complex carbohydrate cellulose, into fructose through an isomerisation process. Instead of the viable route of directly converting glucose into these platform chemicals, it is more challenging to do so than starting with fructose[9] and studies have shown that better yields are observed when the latter route is taken[10].

Glucose to fructose isomerisation (GI) can occur in hot water at high pressure even without catalytic activity, however, this only happens as a side reaction and is unable to achieve a high selectivity of fructose[11]. This explains the reliance of GI reactions on catalytic activity for large-scale fructose production. Lobry de Bruyn and Alberda van Ekenstein were the first to report the reciprocal interconversion of carbohydrate isomers in 1895[12]. The equilibrium between glucose, mannose, and fructose was found to be possible through the formation of enediolic intermediates in alkali solutions[12]. Further exploration on their interconversions were then carried out extensively which mainly focused on the biocatalytic approach via enzymes and mineral acid/base catalysis[13].

In 1957, Marshall and Kooi discovered that the enzyme isolated from *Pseudomonas hydrophilia* in the presence of xylose and arsenate exhibited GI activity and exploited this feature to produce high-fructose corn syrup (HFCS)[14]. It was then emerged into an industrial scale production in 1997 by Clinton Corn Processing Co. in the US[14]. Until today, large-scale industrial HFCS production utilises xylose isomerase (EC 5.3.1.5)[13] to catalyse GI in aqueous phase. However, major drawbacks are inherent in the bio-catalytic process which is mainly centered around the activity and stability of the enzyme throughout the reaction[13]. The narrow operating pH (7.0 – 9.0)[15] and temperature range (55ºC – 60ºC) due to irreversible inactivation of the enzyme at higher temperatures limits the equilibrium glucose conversion at 50% and increases the risk of microbial infection[14], thus implying the necessity of expensive chromatographic enrichment to achieve higher fructose concentration in HFCS[16]. Moreover, the sensitivity of the enzyme towards impurities such as

heavy metals require pre-reaction purification steps, further reducing the process economic efficiency[17]. One way to address these flaws is through immobilisation of the enzyme to enable continuous HFCS[18] production and widen its stability range[15]. It was recently reported that silica/chitosan microspheres immobilisation increases the operating pH range to 5.8 – 8.0 and temperature range to 40ºC – 80ºC[15]. Even though massive studies around the biocatalytic process were carried out to further improve the system[18], isomerisation via chemo-catalysis is seen to be more attractive as it is more operationally versatile thus allowing process intensification.

In chemo-catalysis, both Bronsted bases and Lewis acids can be used as catalysts for the GI reaction. In the case of Bronsted bases, reports have shown that this catalysis resulted in low fructose yield and was only able to reach a high selectivity at low glucose conversion due to the instability of monosaccharides in strong alkaline media[19,20]. This explains why Lewis acids are typically preferred but it is important to note that there have been interesting developments on the Bronsted base catalyst such as the high performance achieved by using hydrotalcites catalysts in GI with ethanol solvent[21].

Zeolites have drawn significant attention in the catalysis field on account of its highly crystalline structure and tuneable composition[22]. Numerous reports have revealed the effectiveness of Sn-Beta, a three-dimensional zeolite beta containing isolated Lewis acidic tin (Sn) sites, in catalysing GI[23]. Recent work comparing the performance of different Lewis acidic silicates namely Sn, zirconium (Zr), titanium (Ti), and hafnium (Hf) under continuous operation of GI has demonstrated Hf to exhibit the highest stability after a brief induction period (subsequently eliminated by methanol pre-treatment), surpassing Sn which lost 40% of its activity upon 113 hours on stream[23]. At high glucose conversion of 66.2%, high fructose selectivity remained with zero loss of carbon balance, indicating the absence of competitive side reactions which are present in the case of Sn[23]. The high fructose selectivity was even maintained at a high operation temperature of 140ºC which deemed Hf-Beta to be the first ever catalyst to achieve this in GI catalysis[23].

Before allowing process scale-ups, preliminary economic evaluations are necessary to gauge whether or not a process system is worth upgrading. As limited literature was available on the economics related to the Hf-Beta catalysis, a technoeconomic assessment was important to identify specific aspects of the process that would be profitable and thus allow suggestions of more specific optimisation procedures to be made. This paper aims to conduct a technoeconomic assessment on the Hf-Beta catalysed GI reaction through optimisation of the process and obtain its economic feasibility with respect to existing bio-catalysed GI reaction implied on an industrial HFCS production.

Methods

Setting up the simulation model

Aspen Plus V11 was used to build a model simulating the continuous GI reaction over Hf-Beta catalyst based on published experimental data in Angew. Chem. Int. Ed. 2020, 59, 20017-20023 which involved a continuous isomerisation of glucose at varied contact times, reacting at two different temperatures. The process flowsheet in Aspen started from a feed stream of glucose dissolved in methanol solvent passing through a pump into a plug-flow reactor (RPlug) producing an outlet stream. The reactor was set to operate isothermally at a defined temperature and at a pump outlet pressure of 20 bar.

The thermodynamic model applied was the Non-random Two-liquid (NRTL) method due to its proven agreement with highly non-ideal systems and extensive use in the chemical industry, which also achieved high consistency when tested against experimental data in a study involving sugar-alcohol-water system[24].

The overall GI reaction was modelled considering the following two reaction pathways,

$$Glucose \underset{k_2}{\overset{k_1}{\rightleftharpoons}} Fructose \quad (1)$$

$$Glucose \underset{k_4}{\overset{k_3}{\rightleftharpoons}} Mannose \quad (2)$$

where k_i represents the rate constants of reactions i. The process modelling involved some assumptions as listed below:

1. The backward reaction of mannose to glucose ($i = 4$) is negligible
2. All other reactions follow the first-order kinetics with respect to the concentration of its reactants
3. The catalyst effect is inherent in the kinetic parameters and not simulated explicitly in the model
4. Pressure drop across the reactor is negligible
5. The catalyst deactivation rate is negligible up to 800 hours of operation[23]
6. Mass transfer limitation inside the reactor is negligible

Deducing the kinetic parameters (i.e. pre-exponential factor and activation energy) of the reactions required the expression of the experimental data into Arrhenius plots using the following plug-flow reactor design equation and reaction rate equation:

$$r_{Glu} = -\frac{dn_{Glu}}{dV_R} \quad (3)$$

$$r_{Glu} = \frac{dC_{Glu}}{dt} = k_1 C_{Glu} - k_2 C_{Fru} + k_3 C_{Glu} \quad (4)$$

where r_{Glu} is the rate of disappearance of glucose, n_{Glu} is the molar flowrate of glucose, V_R is the reactor volume, C_{Glu}, and C_{Fru} are the concentrations of the compounds glucose and fructose respectively.

Further manipulations to equation (3) and (4) led to the following three equations which can be solved to obtain k_i values for the Arrhenius plot:

$$\frac{X_{Glu}}{X_{Glu,eq}} = 1 - e^{-(k_1+k_3+k_2)\tau} \quad (5)$$

$$k_1 = \frac{Y_{Fru}}{X_{Glu}}(k_1 + k_3) \quad (6)$$

$$K_{eq} = \frac{C_{Fru}}{C_{Glu}} = \frac{k_1+k_3}{k_2} \quad (7)$$

where $X_{Glu,eq}$ is the equilibrium glucose conversion, τ is the contact time, Y_{Fru} is the fructose yield and K_{eq} is the equilibrium constant. It is worth noting that the constants at equilibrium point were approximated using the data point at the highest contact time as available from the experiment.

Validating the simulation model

Before proceeding with further steps of optimising the GI process and obtaining its related costs using the model, validation plots comparing the model to the experimental data were generated by running the model at the exact same operating conditions (1 wt% glucose in methanol, 110ºC) of the experiment. The metrics that were used to validate the model included glucose conversion, fructose yield and selectivity, and mannose yield.

Optimisation of process parameters

In line with the objective of optimising the Hf-Beta catalysed GI reaction, a set of performance parameters were used as a measure to compare between the manipulated process parameters. Glucose conversion, selectivity of fructose, production costs, and catalyst productivity were regarded as performance parameters whereas operating temperature and water content were the desired process parameters for optimisation. Each simulation run ensured the volumetric feed flow rate was fixed at 1.5 mL min^{-1} and the reactor diameter at 0.41 cm, to match the scale of the experimental work.

1. Temperature effects

The effects of different temperatures on the performance parameters were observed by implementing the sensitivity analysis tool onto the model. The reactor length was set as the manipulated variable as it directly affects the contact time, and this feature was activated for separate runs at temperatures of 110ºC, 120ºC, 130ºC, and 140ºC each. This temperature range was chosen as the absence of Maillard browning of sugars was observed during the experimental runs up to this temperature[23].

2. Water fraction effects

Previous simulations with methanol as solvent operated at only 1 wt% of glucose in the feed stream owing to the limited solubility of glucose in methanol. This limitation however can be mitigated by the addition of water into the solvent to increase its solubility. A solubility data of glucose in a methanol-water mixture was extrapolated from Figure 9 of van Putten, R.-J. et al. (2014) to aid this evaluation. As the presence of water in a zeolite-catalysed reaction is known to reduce the catalyst activity due to the leaching of the active sites[26], further alterations to the reaction kinetics must be made to account for the reduced catalyst activity. This was done by fitting a decreasing exponential function on SI Figure S9 of Angew. Chem. Int. Ed. 2020, 59, 20017-20023 to deduce a scale factor that relates the decrease in catalyst productivity with respect to water fraction. This scale factor was then used to alter the pre-exponential factor of all reaction kinetics at each water fraction as an approximate to 'correct' for the reduced rate of reaction due to the reduction of catalyst activity in the presence of water based on equation 8:

$$R_{fru|\,x_{water}>0} = scale\ factor \times R_{fru|\,x_{water}=0} \quad (8)$$

where $R_{fru|\,x_{water}>0}$ is the rate of fructose production in the presence of water, and $R_{fru|\,x_{water}=0}$ is the rate of fructose production in absence of water.

Though in reality, the presence of water may alter the kinetics in a different manner, this was taken to be the best available approach in accounting for the effect of water on catalyst activity. To validate this approach, the simulation output using the 'corrected' kinetics was compared against the experimental data of adding 5% water[23] and the model was found to be within a fair 10% deviation from the data. Upon this, a range of 0 wt% to 80 wt% water was evaluated at each temperature where glucose was fed at its maximum solubility at each water fraction, and the glucose conversion was fixed at 50% using the design specification feature on Aspen Plus. The model is simulated in a steady-state setting, with the lowered rate of reaction to account for the reduction in catalyst productivity due to its deactivation.

Costs evaluation

The estimate of the production cost of dry basis fructose was calculated only taking into account the cost of raw materials, catalyst, and utilities associated with operating the reaction at a laboratory scale. A basis of 8,000 hours of operation per year was used to reflect reactor downtime during catalyst regeneration, where the catalyst is regenerated after 800 hours of operation and replaced once a year. Every fresh/regenerated catalyst will undergo 20 hours of pre-activation with methanol solvent at operating temperature to eliminate the induction period of the catalyst[23]. At 140ºC, this is not required as no induction period was observed at this temperature[23]. This pre-activation period of 20 hours was reflected in the cost calculation where no glucose is fed, and no fructose is produced during these hours.

The cost of the catalyst was assumed to be dominated by the cost of materials needed to synthesise the catalyst, hence the associated utilities required for synthesising and regenerating the catalyst were neglected. As listed in SI of Angew. Chem. Int. Ed. 2020, 59, 20017-20023, the cost of each material was taken from SigmaAldrich and converted into USD using a currency rate of 1 GBP = USD 1.19.

The costs of glucose and methanol were taken from Alibaba, while the cost of water was taken from a vendor[27]. Since methanol is highly volatile, it was assumed that 95% of methanol was recovered and recycled back into the reactor, at a negligible operating cost.

Utilities composed mainly the electricity required to run the pump and the heater for the reactor. The pump duty and the reactor duty required were both taken from Aspen Plus results. The electricity cost per kwh was taken from the average electricity price for industrial consumers in the US from the year 2021[28]. The unit price for each cost component is summarised as follows:

Table 1. Unit price of each cost component

Cost component	Unit price	Unit
Hf-Beta Catalyst	9.30	$/g
Glucose	99.00	$/kmol
Methanol	3.50	$/kmol
Water	0.0013	$/kg
Electricity	0.0726	$/kwh

Results and Discussion

Model output validation

Figure 1 A-D presents the comparison between the experimental and simulated data of glucose conversion, fructose yield, mannose yield, and fructose selectivity against the duration of which the solution mixture is in contact with the catalyst whilst residing inside the reactor ("contact time"), respectively. The shaded area in these plots represents the boundaries calculated with respect to the experimental data to evaluate the closeness between them.

As seen in Figure 1, good agreement was achieved for glucose conversion, fructose yield, and fructose selectivity. However, for mannose yield, the model deviates considerably far from experimental values. This may be due to the error in experimental data, having over 100% carbon balance of the product stream, which may suggest an over measurement of mannose product. As mannose yield is not the main focus of this project, and the fructose selectivity predicted by the model is still within tolerable range at higher conversions, the model is deemed valid for further optimisation on fructose production process.

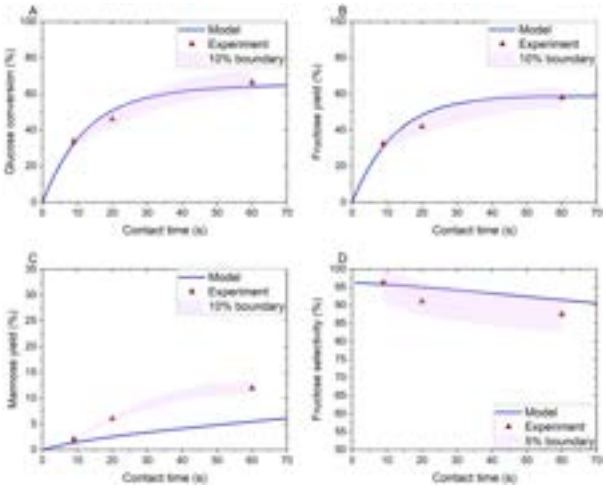


Figure 1. Plots of comparison between model results and experimental data with boundaries as a validation method. General reaction conditions: 1 wt% glucose in methanol, 110ºC. A) glucose conversion against contact time; B) fructose yield against contact time; C) mannose yield against contact time; D) fructose selectivity against contact time.

Temperature effects

Figure 2 depicts the effect of temperature on glucose conversion at different contact times. Up until 50% glucose conversion, a consistent trend can be seen across the different temperatures where the contact times required to achieve the same glucose conversion decreases exponentially as temperature increases. This is in line with the increased rate of reaction due to increased kinetic energy as temperature increases.

Beyond this point, the conversion is expected to plateau at a faster rate and at increasing conversion as the temperature increases owing to the fact that the reaction proceeded endothermically[29]. However, the model failed to attain an equilibrium as can be seen from the non-plateau lines in Figure 2**Figure** . This can be presumed so due to the limited experimental data points that the model is based on, which were absent of data points near reaction thermodynamic equilibrium, thus, the approximate equilibrium concentrations were far off the true values. Other than that, this lack of model prediction at higher conversion was also due to neglecting the backward mannose reaction to glucose in the kinetic model hence the reaction proceeded without being bounded by a finite equilibrium point, thus, increasing the glucose conversion further beyond the supposed equilibrium. With this model limitation, further investigation proceeded at 50% glucose conversion, keeping it comparable to the existing bio-catalysis, and ensuring the validity of subsequent results.

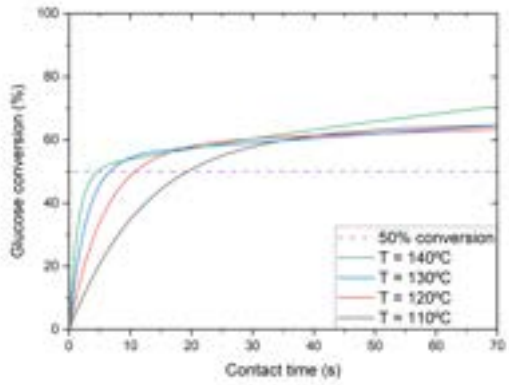


Figure 2. Effects of temperatures on contact times required to achieve a variation of fixed glucose conversion.

The temperature effect on fructose production was further evaluated at a standardised glucose conversion of 50% by inputting the correlated contact times at different temperatures through manipulation of reactor length. Figure 3 depicts the carbon mol% of the product stream for every temperature. It can be observed that a consistent trend of decreasing fructose selectivity with increasing temperature was achieved. This is because the GI reaction involved two competing parallel reactions, hence increasing the temperature would increase both rates of productions. Therefore, the concentration of mannose in the product stream became more and more significant as the temperature increased, growing from around 3% at 110ºC to around 5% at 140ºC. This thus lowered the fructose selectivity that was being achieved from 95% to 91% as temperature increases. As the drop in selectivity is undeniably marginal, deducing the optimum temperature based on fructose selectivity alone is not fair. Further evaluation of temperature effects on other design aspect must therefore be taken.

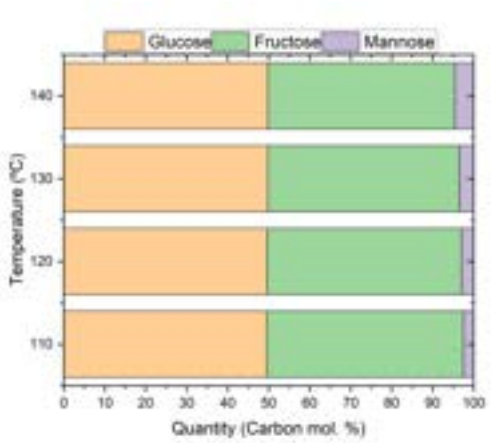


Figure 3. Effects of temperatures on fructose selectivity at a fixed glucose conversion of 50%.

The influence of temperature on reactor size and ultimately fructose production costs was studied to rationalise the temperature selection. While keeping the glucose conversion at the same fixed value of 50% as before, the results of simulations showed that the reactor volume decreased with increasing temperature (Figure 4A). The reactor volume at 140ºC only came to about 0.1 mL whereas at 110ºC, it went up to around 0.5 mL, which is 5 times bigger. This can be explained by the requirement of a larger contact time value hence a larger reactor length for the lower temperature operations to reach the same level of glucose being converted at the higher temperature operations.

The fructose production costs for each operating temperature were evaluated based on simulation results and broken down into four different categories (utilities, solvent, feedstock, and catalyst) as visualised in Figure 4B. The utility cost rose with temperature at approximately 11% – 13% for every 10ºC and came to a maximum value of $2.096 per kilogram of fructose. This observation is consistent with the fact that a larger heat duty is required to heat up reactors set to operate at higher temperatures, hence more cost.

The cost of methanol solvent and glucose feedstock gave similar trends and the maximum differences across the four temperatures were $0.031 kg_{fru}^{-1} and $0.058 kg_{fru}^{-1} respectively. Without normalising to the amount of fructose being produced, the cost of methanol solvent ($3.125 $year^{-1}$) was the same for all temperatures as the inlet volumetric flowrate was kept constant at a value to allow contact time and reactor length variation for the purpose of achieving a fixed conversion.

Similar to solvent cost, the cost of feedstock also depended on the inlet volumetric flowrate thus its value per unit time was consistent at $3.090 $year^{-1}$ across the different temperatures except at 140ºC. This can be explained by the absence of pre-activation period at this temperature that allows more glucose to be fed for the same normalised time unit which gave a value of $3.160 $year^{-1}$ instead.

The catalyst cost, however, showed an opposite trend from the rest of the cost categories where the increase in temperature was accompanied by a decrease in its own value. When compared with the plot of reactor volume against temperature (Figure 4A), the trends can be seen to agree with each other. This was expected as the mass of catalyst per reactor volume ("bed density") was kept constant. The mass of catalyst and equivalently the costs it is associated with should display a linear relationship as proven.

The overall production cost of fructose was discovered to attain the highest value of $5.592 kg_{fru}^{-1} at the lowest temperature of 110ºC with catalyst cost dominating around 33% of the said total. The total production cost at the other operating temperatures was governed by the utility costs instead of the catalyst cost, and the lowest production rate cost was achieved at 130ºC with an amount of $4.894 kg_{fru}^{-1}. However, this was slightly over 1% smaller than $4.947 kg_{fru}^{-1} that was obtained when the reactor was operated at 140ºC. This small difference implies that the additional heating being put into the system was compensated by the relatively small volume of the reactor, therefore operating at a high temperature can be beneficial, nevertheless.

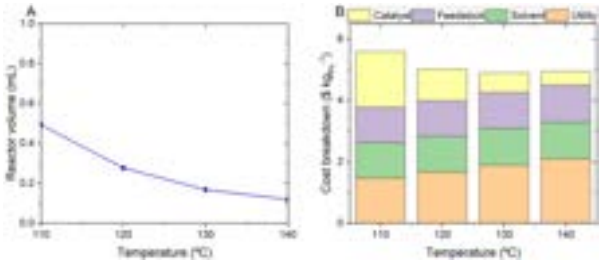


Figure 4. Effects of temperature on reactor sizing and fructose production costs, at a fixed glucose conversion of 50%. A) Reactor volume as a function of temperature; B) Breakdown of costs for fructose production as a function of temperature.

Water fraction effects

Using the solubility data mentioned earlier in the method section, the increase in glucose loading in the feed stream with water content can be seen in Figure 5A. To alter the kinetics of the reaction, the scale factor deduced from two experimental data in Angew. Chem. Int. Ed. 2020, 59, 20017-20023 can be found in Figure 5B. From this graph, a general decrease in the trend of the Hf-Beta catalyst productivity can be seen. This is contradictory to a published paper which studied the activity of Sn-Beta of the same isomerisation reaction[30]. From the paper, an optimum window of catalyst productivity was found in between 0 - 10 wt%, which then followed by a decreasing trend as more water was added. Although this seemed to be the case with Sn-Beta, the water effect with Hf-Beta is still unknown as the two species exhibit different intrinsic properties which is beyond the scope of this project. Therefore, a general decrease in catalyst productivity as water content increases was deemed as a valid assumption for Hf-Beta, based on the two experimental points[23] which showed a decrease in productivity when 5% of water is added.

From these graphs, the simulation model operating at the maximum glucose loading, and 'corrected' kinetics at each water fraction were repeated for each temperature, and the catalyst productivity against water content was analysed as shown in Figure 6. As can be seen from the Figure 6, a consistent optimum point of

water content was obtained across the different temperatures investigated. This trend can be explained by two competing effects which dictate the productivity of the catalyst. The first one is the positive effect of increasing water content in the solvent where it increases the solubility of the glucose. More glucose can therefore be fed into the reactor, increasing the concentration of glucose in the feed stream. As GI is taken as a first-order reaction, increasing the glucose concentration increases the rate of reaction which directly relates to increasing the fructose production. Conversely, increasing the water content impedes the reaction. This is due to the leaching effect of water[26] on the zeolite crystals as mentioned earlier which caused irreversible catalyst deactivation, hence reducing the catalyst productivity.

From the result, a 30% water content was found to be the optimum point as it correlated to the highest catalyst productivity. This means that before this point, the effect of increasing glucose concentration outweighs the degrading effect of water on the catalysis. Hence, increasing the water content will directly increase the productivity of the catalyst. Beyond this point, the deactivation effect of water on the catalyst influenced the productivity by great amount, causing the productivity to drop despite the increase in glucose loading.

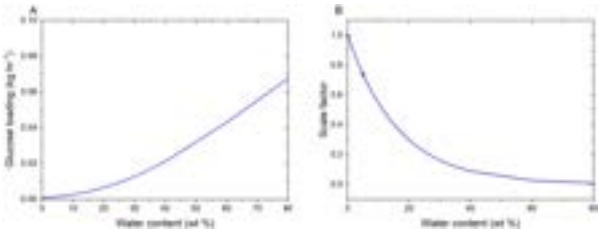


Figure 5. A) Glucose loading against water content in methanol-water solvent; B) Scale factor for 'corrected' kinetics against water content

To see how this impacts cost, further investigation was carried out to figure out the effect of water percent variation on cost of fructose production. Same approach was taken as before apart from the catalyst cost where it was considered to be replaced nine times per year after 800 hours of operation to account for irreversible deactivation caused by water. From Figure 6B, as the water content increases, the total production cost for each temperature decreases until it reached a minimum, and then increases. The minimum cost was found near the 30% optimum point which suggested that the total production cost is reflective of the catalyst productivity where the higher catalyst productivity relates to the lower total cost. The initial decreasing trend can be explained by the faster increase in fructose production relative to increase in reactor volume required to achieve the defined conversion. In other words, the increase in reactor volume which relates directly to increase in utility, catalyst, and feed cost were compensated by the larger increase in fructose production per hour, which then translates into a decreased in total production cost. The opposite is true for the points beyond the minimum, where the increase in fructose production is no longer large enough to reduce the effect of increasing reactor volume. This is due to the drop of catalyst productivity as discussed earlier. Therefore, based on these two graphs, a water content of 30% was deemed to be the best operating parameter for GI.

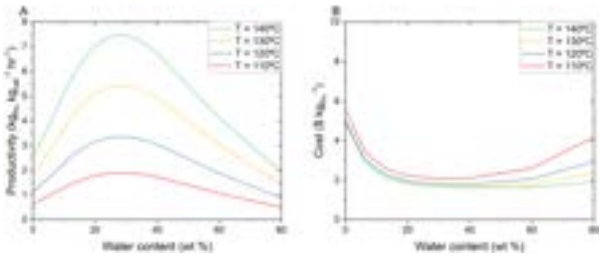


Figure 6. Effects of water content in methanol-water solvent on catalyst productivity and fructose production cost at fixed glucose conversion of 50%. A) Catalyst productivity as a function of water percent; B) fructose production cost as a function of water percent.

Cost breakdown at optimum operating parameters

The final model was evaluated with inputs of the optimised conditions (140ºC, 30% water in solvent) concluded earlier for 50% glucose conversion. The results of this model gave fructose yield and selectivity values of 45.6% and 91.2% respectively. To put this into context, this yield value surpassed the 42% yield observed with the biological system[23] which provides an optimistic future in the scope of utilising Hf-Beta for large-scale GI reactions. It can be observed from Figure 7 that the total cost of production was dominated by the cost of glucose feed (72%), followed by utilities (15.7%) and catalyst (9.2%).

The high glucose cost contribution suggests its role as one of the cost drivers and highlights the critical importance of minimising glucose waste in GI reactions. This can be achieved by maximising glucose conversion through reactor operation at equilibrium conversion of higher temperatures.

One other possible factor leading to this observation is the process model setup that did not consider recovery of unreacted glucose feed after the reaction was completed, since it was beyond the scope of this project. From literature, it was discovered that the current sugar separation technique involves an expensive simulated bed chromatography unit operation[31]. A study on an alternative method of separation known as simultaneous isomerisation reactive extraction (SIRE)[32] has been found for the biocatalytic process. In the same way, separation methods for the chemo-catalytic process should be explored as it not only eliminates the need of expensive separation unit, but also overcomes equilibrium limitations of the GI reaction and becomes another to compete against the biological system.

The final reasoning that was considered to justify the large cost contribution by glucose was the usage of highly purified glucose cost as the feedstock price. In the real-world case of utilising second generation biomass for platform chemicals production, the initial step before GI that biomass must go through is the saccharification process to break its complex carbohydrate compound into the simple carbohydrate glucose. This process does not achieve a high purity as that of glucose used for this costing therefore this was an overestimation.

The utility cost takes the second place after glucose feedstock as the highest contributor towards the whole

fructose cost production. This suggests the necessity of applying energy integration in the overall process of converting biomass to platform chemicals on the larger scale.

Next in line is the catalyst cost which accounts for 9.2% of the total fructose production cost. Although not as much as glucose contribution, this percentage is quite a significant value and can increase once the aforementioned steps towards reducing glucose cost contributions are realised. It thus highlights the importance of picking the right catalyst by studying their respective catalytic activity and regeneration methods to give the best performance in GI.

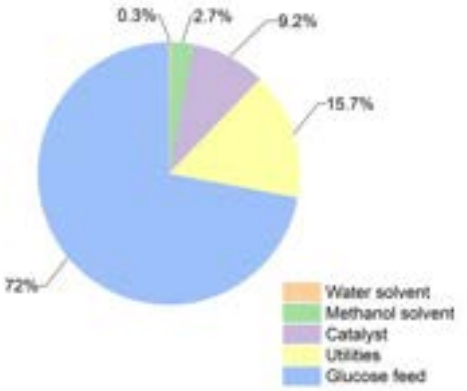


Figure 7. Cost breakdown of fructose production at optimum conditions (140ºC, 30% water in solvent) for Hf-Beta catalysed GI.

Cost comparison to biological catalyst

Looking further into the economical aspect of this process, comparisons against the cost breakdown of the biological system would be favourable as it could directly conclude if the heterogeneous system has potential in beating the biological system. Ideally, this comparison should be done using processed or experimental data of the same scale which considers the same process operation with similar goals. However, due to time limitations and mismatch of data available in a sense that only industrial scale ones were obtainable for the biological system, it was impossible to realise this idea. Regardless, a comparison against the available data was deemed necessary to put the result of $1.680 kg_{fru}^{-1} for total fructose production cost into context and its results are presented in the form of a bar chart on Figure8.

As the HFCS manufacturing process demonstrates bulk formation of fructose from glucose through the bio-catalysed GI reaction, it was selected to represent the biological system where prices of fructose on a dry weight basis in commercialised HFCS-42 and HFCS-55 obtained from US wholesale spot price[33] were utilised. With the commercialised HFCS having went through a high-level degree of optimisation for its millions of tonnes per annum[34] production, it was expected that this comparison would be in favour of the biological system.

It should be appreciated, nonetheless, that the differences between the Hf-Beta bar and the two commercialised HFCS bars on Figure 8 can be considered medium-sized as they are within similar orders of magnitude with a minimum percentage difference of 42.6%. Here, it is important to mention the caveats related to comparing data of incompatible scale such as inconsistent inclusion of additional or reduced costs from heat integration, labour, maintenance, and downstream process of separation. Generating lab-scale data for the biological process or carrying out an extensive data search that can be used to set up a similar model to the one built for Hf-Beta would be highly beneficial for a fairer comparison to be made. However, this was not possible within the time limitations that was imposed upon this project.

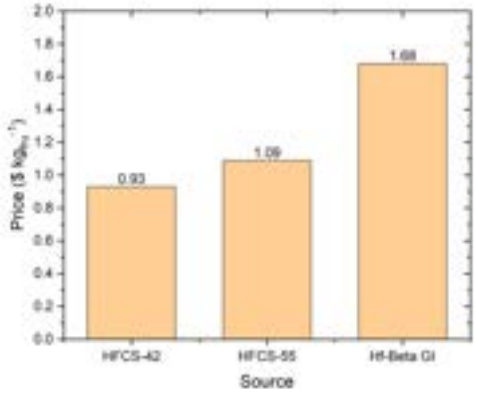


Figure 8. Price of fructose (dry weight basis) from different sources of production

Conclusions

The discovery of biomass as a non-renewable energy source has opened a door to a greener and more sustainable planet Earth. Biomass are a promising alternative feedstock to fossil resources as they can be integrated into high-value platform chemicals like 5-HMF. This involves a critical step of isomerisation from glucose to fructose which is currently being achieved at large scale using the biological catalyst xylose isomerase. Although effective, this process suffers from several drawbacks which include requirement of strictly controlled operating conditions which inspires the pursuit of a heterogeneous catalyst to challenge the position of enzymes in the industrial scene status quo.

From literature, it was reported that Hf-Beta exhibited good performance in GI catalysis and was therefore used to represent heterogeneous catalyst for the purpose of gauging its economic feasibility. A technoeconomic analysis was conducted on Hf-Beta zeolite catalysed GI reaction by first setting up an Aspen simulation that was modelled and validated using kinetic parameters obtained from the available experimental data. The model was then used to optimise the GI reaction and it was found that higher temperatures favour the productivity and economics, while the addition of water only improved the economics. However, this was limited to a temperature range of 110ºC to 140ºC and in the case of water addition, an optimum point existed. From the aforementioned observations, it was concluded that the optimum conditions to operate the reaction specified at 50% glucose conversion were 140ºC and 30:70 mass ratio of water to methanol solvent. Operating at these conditions

attained a fructose yield of 45.6% and selectivity of 91.2%.

On the economics side, it was observed that the glucose feedstock dominated the fructose production cost which amounted to $1.68 per kilogram of fructose. To put this into context, this value was compared to fructose price in commercialised HFCS as a representative of the biological process. Although it was expected that the economy of scale would favour the highly optimised HFCS fructose production cost, it should be made clear that exploration into the economic comparison with lab-scale biological data is worth pursuing as the differences among the values compared were of similar orders of magnitude.

Outlook

The Aspen model developed in this paper could be used as a supporting tool in labs to predict reactor outputs at low conversions. However, limitations to the model must be addressed, where firstly, it did not produce satisfactory result near reaction equilibrium, suggesting future work to be done on improving the model. In order to achieve this, more experimental data which allows the reaction to operate at equilibrium conversion is needed to obtain more accurate kinetics.

The rough assumption of exponential decrease of Hf-Beta productivity with water content must also be validated in labs as only two points of data were available during the analysis. This relationship is pivotal in obtaining the optimum operating water content when operated at maximum glucose solubility as this optimum point is highly influenced by the extent of water effect on catalyst productivity.

On top of that, as the model predicted a 30% water content to be the optimum, further evaluation on cost-benefit of this amount of water must be taken into account when considering the downstream cost related to wastewater treatment required as considerably large amount of water is involved in the system as opposed to the initial consideration of methanol alone as the solvent.

Lastly, it is worthwhile to carry out lab-scale GI over the enzyme catalyst, xylose isomerase in order to obtain enough data for subsequent process modelling of a comparable scale to the current model developed. This is beneficial to aid the identification of limits to the operating parameters in terms of profitability when compared to the biocatalyst system.

Acknowledgements

The authors would like to express their gratitude towards the whole Hammond group, especially Laurie Overtoom for the overwhelming support and guidance throughout the duration of the project.

References

[1] Souzanchi, Sadra. (2016) "Catalytic Conversion of Fructose, Glucose and Industrial Grade Sugar Syrups to 5-Hydroxymethylfurfural, A Platform for Fuels and Chemicals" Electronic Thesis and Dissertation Repository. 4070. Available at: https://ir.lib.uwo.ca/etd/4070 (Accessed: December 15, 2022).

[2] Biomass Energy Basics (no date) NREL.gov. Available at: https://www.nrel.gov/research/re-biomass.html (Accessed: December 15, 2022).

[3] Biomass Energy (no date) National Geographic Society. Available at: https://education.nationalgeographic.org/resource/biomass-energy (Accessed: December 15, 2022).

[4] A burning issue: Biomass is the biggest source of renewable energy consumed in the UK (no date) A burning issue: biomass is the biggest source of renewable energy consumed in the UK - Office for National Statistics. Available at: https://www.ons.gov.uk/economy/environmentalaccounts/articles/aburningissuebiomassisthebiggestsourceofrenewableenergyconsumedintheuk/2019-08-30#:~:text=It%20is%20considered%20a%20renewable,the%20soil%2C%20plants%20or%20trees (Accessed: December 15, 2022).

[5] Environmental and Energy Study Institute (EESI) (no date) Fossil fuels, EESI. Available at: https://www.eesi.org/topics/fossil-fuels/description#:~:text=Fossil%20fuels%E2%80%94including%20coal%2C%20oil,percent%20of%20the%20world%27s%20energy. (Accessed: December 15, 2022).

[6] Causes and effects of climate change (no date) United Nations. United Nations. Available at: https://www.un.org/en/climatechange/science/causes-effects-climate-change#:~:text=Fossil%20fuels%20%E2%80%93%20coal%2C%20oil%20and,of%20all%20carbon%20dioxide%20emissions. (Accessed: December 14, 2022).

[7] Lignocellulosic biomass (no date) Lignocellulosic Biomass - an overview | ScienceDirect Topics. Available at: https://www.sciencedirect.com/topics/engineering/lignocellulosic-biomass (Accessed: December 14, 2022).

[8] van Putten, R.-J. et al. (2013) "Hydroxymethylfurfural, a versatile platform chemical made from renewable resources," Chemical Reviews, 113(3), pp. 1499–1597. Available at: https://doi.org/10.1021/cr300182k.

[9] Zhou, C. et al. (2017) "Conversion of glucose into 5-hydroxymethylfurfural in different solvents and catalysts: Reaction kinetics and mechanism," Egyptian Journal of Petroleum, 26(2), pp. 477–487. Available at: https://doi.org/10.1016/j.ejpe.2016.07.005.

[10] Gogar, R. L. (2020). "Economic Production of Furans from Lignocellulosic Sugars [Doctoral dissertation, University of Toledo". OhioLINK Electronic Theses and Dissertations Center. Available at: http://rave.ohiolink.edu/etdc/view?acc_num=toledo1595977336480846. (Accessed: December 15, 2022).

[11] Steinbach, D. et al. (2020) "Isomerization of glucose to fructose in hydrolysates from lignocellulosic biomass using hydrotalcite," Processes, 8(6), p. 644. Available at: https://doi.org/10.3390/pr8060644.

[12] Wang, Z. (2010). "Lobry de Bruyn-Alberda van Ekenstein Transformation. In Comprehensive Organic Name Reactions and Reagents," Z. Wang (Ed.). Available at: https://doi.org/10.1002/9780470638859.conrr396

[13] Li, H. et al. (2017) "Glucose isomerization by enzymes and chemo-catalysts: Status and current advances," ACS Catalysis, 7(4), pp. 3010–3029. Available at: https://doi.org/10.1021/acscatal.6b03625.

[14] Bhosale, S.H., Rao, M.B. and Deshpande, V.V. (1996) "Molecular and industrial aspects of glucose isomerase," Microbiological Reviews, 60(2), pp. 280–300. Available at: https://doi.org/10.1128/mr.60.2.280-300.1996.

[15] Zhao, H. et al. (2016) "Enhancement of glucose isomerase activity by immobilizing on silica/chitosan hybrid microspheres," Journal of Molecular Catalysis B: Enzymatic, 126, pp. 18–23. Available at: https://doi.org/10.1016/j.molcatb.2016.01.013.

[16] Delidovich, I. and Palkovits, R. (2016) "Catalytic isomerization of biomass-derived aldoses: A Review," ChemSusChem, 9(6), pp. 547–561. Available at: https://doi.org/10.1002/cssc.201501577.

[17] Venkatasubramanian, K. and Harrow, L.S. (1979) "Design and operation of a commercial immobilized glucose isomerase reactor system," Annals of the New York Academy of Sciences, 326(1 Biochemical E), pp. 141–153. Available at: https://doi.org/10.1111/j.1749-6632.1979.tb14158.x.

[18] Ventura, M., Mazarío, J. and Domine, M.E. (2021) "Isomerization of glucose-to-fructose in water over a continuous flow reactor using ca−AL mixed oxide as heterogeneous catalyst," ChemCatChem, 14(3). Available at: https://doi.org/10.1002/cctc.202101229.

[19] Carraher, J.M., Fleitman, C.N. and Tessonnier, J.-P. (2015) "Kinetic and mechanistic study of glucose isomerization using homogeneous organic brønsted base catalysts in water," ACS Catalysis, 5(6), pp. 3162–3173. Available at: https://doi.org/10.1021/acscatal.5b00316.

[20] Liu, C. et al. (2014) "Selective base-catalyzed isomerization of glucose to fructose," ACS Catalysis, 4(12), pp. 4295–4298. Available at: https://doi.org/10.1021/cs501197w.

[21] Yabushita, M. et al. (2019) "Selective glucose-to-fructose isomerization in ethanol catalyzed by hydrotalcites," ACS Catalysis, 9(3), pp. 2101–2109. Available at: https://doi.org/10.1021/acscatal.8b05145.

[22] Xu, H. et al. (2018) "Hydrothermal synthesis of sn-beta zeolites in F−-free medium," Inorganic Chemistry Frontiers, 5(11), pp. 2763–2771. Available at: https://doi.org/10.1039/c8qi00672e.

[23] Botti, L. et al. (2020) "Solvent-activated hafnium-containing zeolites enable selective and continuous glucose–fructose isomerisation," Angewandte Chemie International Edition, 59(45), pp. 20017–20023. Available at: https://doi.org/10.1002/anie.202006718.

[24] Caudle, B. et al. (no date) "Modeling Phase Equilibrium of Common Sugars Glucose, Fructose, and Sucrose in Mixed Solvents," Battelle Savannah River Alliance (BSRA), LLC [Preprint].

[25] van Putten, R.-J. et al. (2014) "Experimental and modeling studies on the solubility of d-arabinose, d-fructose, d-glucose, d-mannose, sucrose and d-xylose in methanol and methanol–water mixtures," Industrial & Engineering Chemistry Research, 53(19), pp. 8285–8290. Available at: https://doi.org/10.1021/ie500576q.
Leckenby, R. J. (2005) Dynamic characterisation and fluid flow modelling of fractured reservoirs. PhD thesis. Imperial College London.

[26] Botti, Luca. (2020) Metal-incorporated beta zeolites: A versatile class of catalysts for continuous glucose upgrading. PhD Thesis, Cardiff University.

[27] Compare 2022 business water rates (no date) Business Electricity Prices. Available at: https://www.businesselectricityprices.org.uk/water-prices/ (Accessed: December 15, 2022).

[28] Alves, B. (2022) U.S. Industrial Retail Electricity Price 2021, Statista. Available at: https://www.statista.com/statistics/190680/us-industrial-consumer-price-estimates-for-retail-electricity-since-1970/ (Accessed: December 15, 2022).

[29] Moliner, M., Román-Leshkov, Y. and Davis, M.E. (2010) "Tin-containing zeolites are highly active catalysts for the isomerization of glucose in water," Proceedings of the National Academy of Sciences, 107(14), pp. 6164–6168. Available at: https://doi.org/10.1073/pnas.1002358107.

[30] Padovan, D., Botti, L. and Hammond, C. (2018) "Active site hydration governs the stability of sn-beta during continuous glucose conversion," ACS Catalysis, 8(8), pp. 7131–7140. Available at: https://doi.org/10.1021/acscatal.8b01759

[31] Motagamwala, A.H. et al. (2019) "Solvent system for effective near-term production of Hydroxymethylfurfural (HMF) with potential for long-term process improvement," Energy & Environmental Science, 12(7), pp. 2212–2222. Available at: https://doi.org/10.1039/c9ee00447e.

[32] Li, B., Relue, P. and Varanasi*, S. (2012) "Simultaneous isomerization and reactive extraction of

biomass sugars for high yield production of ketose sugars," Green Chemistry, 14(9), p. 2436. Available at: https://doi.org/10.1039/c2gc35533g.

[33] Sugar and sweeteners yearbook tables (no date) USDA ERS - Sugar and Sweeteners Yearbook Tables. Available at: https://www.ers.usda.gov/data-products/sugar-and-sweeteners-yearbook-tables/ (Accessed: December 15, 2022).

[34] Wunsch, N.-G. (2022) Production volume of high fructose corn syrup in the U.S. 2020, Statista. Available at: https://www.statista.com/statistics/496475/high-fructose-corn-syrup-production-in-the-us/ (Accessed: December 15, 2022).

Electrothermal Energy Storage: A simulation for thermal storage system modelling and part-load operation

Meriem Chennoufi and Christina Theochari
Department of Chemical Engineering, Imperial College London, U.K.

Abstract

Electrothermal energy storage (ETES) is a novel bulk energy storage system which utilizes the closed Brayton cycle with supercritical carbon dioxide as a working fluid. This study focuses on proposing a realistic and accurate model that analyses part-load operations to address fluctuations in renewable energy supplied to power grids. An exploration of exisiting ETES research was first performed to better understand the impact that each key parameter had on roundtrip efficiency. Key relationships were noted and brought forward to the simulation stage of the analysis. This paper also focusses on the modelling and the implementation of the turbomachinery performance maps in the simulation. Ultimately, this study proposes an optimal ETES design that can perform at a roundtrip efficiency of 51%. The part-load evaluation reveals a round-trip efficiency drop of 9% when power supplied undergoes fluctuations of 10%.

Keywords: Electrothermal Energy Storage, supercritical CO_2 Brayton cycle, part-load, turbomachinery performance maps

Introduction

Due to their fluctuating nature, renewable energies such as wind and solar power cannot be scheduled at the request of power grid. Currently, in cases of electricity production exceeding the power grid demand, curtailment of renewable energy is the only available option due to lack of efficient energy storage systems. Research shows that amidst the UK energy crisis between 2021 and 2022, 1300 GWh of energy have gone to waste from wind turbines alone. [18] When electricity demand is greater than renewable source energy production, carbon-emitting dispatchable generators are used. This has a vast negative environmental and economical impact. Curtailment of renewable energy and the use of gas for energy compensation during lapses cost an additional £390 million during the energy crisis [18]. This contributedg to a surge in energy prices affecting millions of households. As long as countries remain dependent on the import of oil and gas- arguably more costly than renewables - energy insecurity will remain prevalent and households will continue to struggle with ever-fluctuating energy prices. This stresses the importance of building reliable renewable storage systems that can enable countries to make full use of their natural resources and gain energy independence to stabilize their energy prices.

The energy industry has recognised the importance of such research and are advancing their understanding through extensive research on energy storage systems. The most promising established technologies that currently exist are lithium-ion batteries, pumped hydro storage (PHS) and compressed air energy storage (CAES). Lithium-ion batteries have high efficiencies, but they suffer from costly materials and low capacities. Pumped hydro storage and CAES benefit from high storage capacity. Though, they have geographical restrictions, as pumped hydro requires high altitude water reservoirs and CAES requires underground caverns for energy storage. Additionally, CAES relies on natural gas combustion as a heat source.

Therefore, there is a need for location and fossil fuels-independent storage solutions with high efficiencies, long charging hours and a long lifetime. The innovative technology of electrothermal energy storage (ETES) is believed to meet all the above criteria. It operates as a "Carnot battery"; when energy demand is low, it stores electricity in the form of a thermal energy in a working fluid. Meanwhile, when energy demand is high, this thermal energy can be converted back into electricity. The system also benefits from a cold storage unit which cools down during charging, when heat is being stored. Most studies suggest that this energy storage system can reach efficiencies of 60%. It does not require specific location characteristics nor costly materials as in the case of lithium-ion batteries. Its cost solely

depend on the turbomachinery, the working fluid and the widely accessible storage fluids used. It is, also, believed to not possess any lifetime limitations, as degraded materials can be readily replaced .

Simplified simulations of the ETES systems have been realized in the past, though, they typically lack a realistic representation, as they include turbomachinery simplifications and assumptions such as the constant supply of the design electricity input into the system. The aim of this analysis was to construct a robust and reliable simulation of the process. The simulation included the design of thermal storages and considered their impact on reversibility. It, also, implemented turbomachinery characteristic performance maps which reflected the most current prediction models for these novel components. Most importantly, this study assessed the operation with a variable, lower electricity input than the design on, referred to as part-load operation. This was essential to consider the innate characteristic of renewable sources, which is intermittency. Finally, future recommendations will be provided to potentially overcome the key inefficiencies observed.

1- Background

1.1 Working Principle of the Thermodynamic Cycle

ETES is composed of two thermal storage tanks, two compressors, a turbine, an expander, and a working fluid that acts as a carrier for the exchange of energy from one form to another. This system includes a charging and discharging cycle which can be approximated to follow a Brayton reversible cycle.

The working principle behind an ETES system is that during low electricity demand, excess energy supplied will be used to drive the charging cycle. This charging cycle transfers thermal energy to the hot storage through the process presented in Figure 1. The cycles starts when excess energy is directed towards an electrical motor. This motor is linked to the charging cycle compressor, which elevates the working fluid's temperature and pressure. The compressed fluid is then passed through a counter current heat exchanger where its heat is transferred into the hot storage fluid. After this thermal exchange, the fluid's pressure is lowered via an expander. Since the expander is mechanically coupled to the compressor, work from the expander will be recycled to the compressor and reduce the net work input into the system. It will then enter the heat exchanger coupled with the cold storage arrangement where it gains thermal energy and cools down the cold storage material before being recycled back to the compressor. (Figure 1) When additional reserves of power are necessary, the stored heat is converted back to electricity through a discharging cycle – which is a reverse process to the one described above.

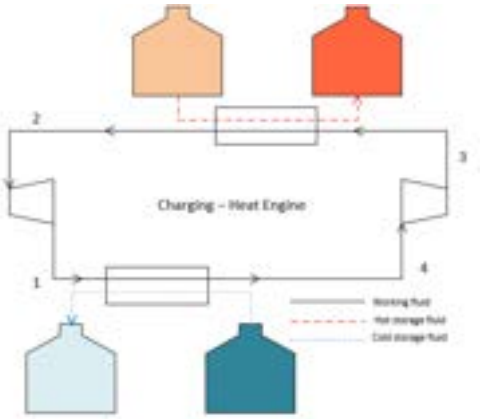


Figure 1: ETES Charging cycle

1.2 Key Performance Indices

By convention, the following performance indices are used to quantify performance of refrigeration cycles, including those used for the purpose of energy storage. These same ratios will be used as indicators to measure performance of the simulated ETES system.

Work ratio is defined as the fraction of the expander work output consumed by the compressor during the charging cycle. Indeed, the compressor and the expander are mechanically coupled to ensure the transfer of the power generated by the expander to the compressor (Equation 1). $W_{compressor}$ can therefore be written as the sum of the net charging work input W_{net} and the power generated by the expander $W_{expander}$. A high work ratio is preferred as it indicates a smaller required net work input. [1]

$$W_{ratio} = \frac{W_{compressor}}{W_{expander}} = \frac{W_{net} + W_{expander}}{W_{expander}}$$

Roundtrip efficiency is defined as the ratio between the work recovered during discharge and the work inputted during charge.

$$\eta_{round-trip} = \frac{W_{discharge}}{W_{charge}}$$

Work inputted during the charging cycle can be expressed as the consumption of the compressor reduced by the expander power, while the work recovered during discharge corresponds to the turbine power subtracted by the power supplied to the compressor. A high roundtrip efficiency is preferred as it is an indicator of small energy losses between the charging and discharging processes.

1.3 Transcritical vs Supercritical Operation

Supercritical operation occurs when the entire operation of a system follows a supercritical Brayton refrigeration cycle. This means that the working fluid remains in the supercritical region throughout the entire charging and discharging cycles, exchanging sensible heat with the storage material. In contrast, transcritical operation describes a transcritical Brayton refrigeration cycle, exchanging latent heat and causing phase change of the storage fluid.

Transcritical cycles benefit from high efficiencies, due to low losses in their heat transfer. However, they face some practical challenges [2]. Since a phase transition happens during the heat transfer, it would be complex to ensure a consistent flow of the freezing, solid, material used for the cold storage. The use of sensible heat storage, usually in the liquid state, is therefore preferred. In this case, smaller temperature differences within the heat exchangers should be chosen to reduce losses. Transcritical cycles also require turbomachinery in the liquid region of working fluid, which are lacking in efficiency. Supercritical cycles, in contrast, are less sensitive to turbomachinery and can lead to very high round-trip efficiency, assuming that the temperature difference of the working fluid within the heat exchangers is kept at a minimum [3]. Hence, a supercritical Brayton cycle was chosen for the ETES system simulation.

1.4 Integration of Regenerative Heat Exchange

A higher temperature difference between the hot and the cold thermal storage implies a higher round-trip efficiency and energy density [1]. Though, there are several challenges that emerge when attempting to achieve this.

First, the operating temperature of the hot thermal storage must be greater than the freezing temperature of molten salt and smaller than its degradation temperature. Hence, the possible heat absorbed from the hot thermal storages must be limited. Additionally, it is important to note that it is essential for the temperature difference within the heat exchangers to be kept small to minimise losses. Finally, the expander work increases as its inlet temperature increases. The expander power output, therefore, becomes a higher fraction of the compression work, which decreases work ratio and round-trip efficiency.

The addition of a regenerative heat exchanger (Figure 2) can resolve all the issues described above. The molten salts are maintained in the liquid state, temperature input to the expander is decreased, thus, leading to a high work ratio and ensuring the largest possible temperature difference between the hot and cold thermal storages. On the cold side of the cycle, the sCO_2 is sufficiently heated to reach the required temperature inlet to the compressor, while maintaining the temperature difference within the cold storage heat exchanger at a minimum.

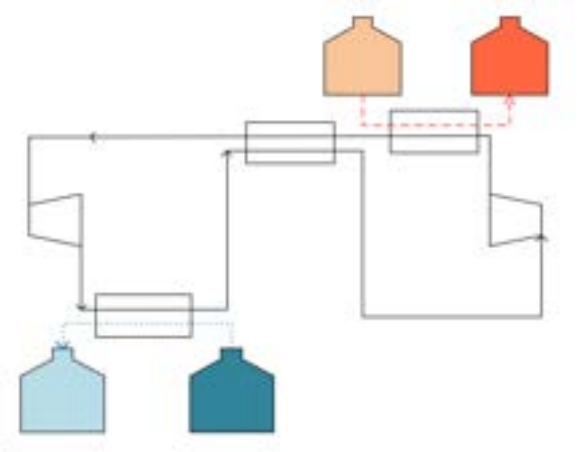


Figure 2 ETES cycle with regenerative heat exchanger

1.5 Working fluid selection

The choice of working fluid for the heat storage cycle is crucial. For the configuration and simulations conducted, CO2 was chosen as a working fluid. The use of carbon dioxide (R744 in refrigerant nomenclature) has multiple benefits. Its pressure in the critical region is amongst the lowest compared to other common working fluids [2]. It is beneficial for its low cost, exceptionally low critical point and possesses excellent thermal

properties. Additionally, it is non-toxic and has a low global warming potential compared to other refrigerants. [2].

1.6 Thermal storage material selection

Molten salts were chosen as the ideal hot thermal storage material due to their high heat capacity and their high operating temperatures. A comparison between the most widely used molten salts was conducted based on the following properties: degradation temperatures, heat capacity (averaged over their operational temperatures) and price (Table 1). Solar salt is observed to be the optimal molten salt for the ETES system designed, for its high degradation temperature, low cost, and sufficient heat capacity.

Table 1: Comparison of commercially available molten salts as thermal storage material

Molten Salt	Degradation T [°C]	Cost [$/kg]	Heat Capacity [kJ/kg K]
Solar Salt	600	0.5	1.50
HITEC	550	0.9	1.49
HITEC XL	500	1.1	1.45
LiNaK	550	1.1	1.55

When considering fluids for the cold thermal storage, water was selected for to its high heat capacity and its thermal properties. It is also cost competitive, has a low impact on the environment and is safe to use [2].

2- Methodology

Aspen Hysys was selected to model this simulation for its ability to compute the performance of the system as well as measure the behaviour of all components during both charging and discharging cycles . Since the systems in question are both closed loop in nature and highly dependent on one another, a very robust methodology was constructed for simulations. The methodology is listed below.

2.1 Thermal storage material selection

A property package analysis was essential prior to the modelling of the system to ensure that the simulation data collected were sensible.

Supercritical fluid simulation can vary significantly due to the complexity of property behaviour within the critical region. Therefore, an extensive investigation took place to determine the optimal property packages. The most widely used Aspen properties for pure supercritical CO_2 are RefProp, GERG2008, Peng-Robinson and Lee-Kesler-Plocker. Upon comparison with the values provided by NIST, Lee-Kesler-Plocker had entropy differences that were the closest to experimental data. Additionally, as literature suggests, Lee-Kesler-Plocker provides the best predictions near the critical point and has superior performance when operated at high temperature and pressures [10]. The National Energy Technology Laboratory, also, recommends the Lee-Kesler-Plocker property package due to its higher consistency in the critical region [8]. Therefore, it was the property package chosen.

For the simulation of water, NRTL was used, as it accurately predicts polar components. For the case of solar salt, it was modelled using a mixture of 60wt% $NaNO_3$ and 40wt% KNO_3. Multiple property packages were investigated. Despite being suggested for ionic compounds by Aspen Hysys, Electrolyte NRTL predicts molten salt as a vapor in the operational temperature range chosen. This is further indicated by the small value obtained for mass density in these conditions (Table 2). For this reason, Peng-Robinson was chosen, as it provided the closest prediction of the essential properties and accurately predicted the state of the molten salt, as shown below.

Table 22: Property comparison of molten salt using different property packages at the conditions of the cold side of the hot thermal storage system.

Property Package	Heat Capacity (kJ/kgC)	Mass Density (kg/m^3)
Electrolyte NRTL	1.061	1.730
ENRTL-HG	1.162	409.6
Peng-Robinson	1.236	409.6
Experimental values	1.468	1823

The difference in heat capacity and density values between the Aspen and those obtained through literature remains significant, especially for density. Hence, the density value used for tank sizing was the one obtained from literature. To avoid overestimation of the tank size using the mass flowrates extracted from Aspen, which were higher than the actual ones due to the difference in

heat capacity, a correction factor of 19% was be applied.

2.2 Charging cycle simulation

As this paper corresponds to a preliminary study of the part-load operation of the ETES technology, a pilot sized system with a compressor input of 10MW was first simulated. As the process dealt with is a closed loop, the cycle was disconnected, initially, at the inlet of the compressor. The compressor input conditions were assumed to be in accordance with the design described in the ETES project with regenerative heat exchanger [4] and the polytropic efficiency of the turbomachinery was fixed at 90% [1]. The remaining conditions were determined with the aim of maximizing the temperature difference between the hot and the cold storage systems, the importance of which was highlighted in Section 1.4, while increasing the work ratio (see section 1.2). The steps listed below were followed and, ultimately, the loop was closed.

2.3 Turbomachinery Modelling

In an effort to understanding the effect that fluctuating power supply had on the performance of the system, it is essential to predict the impact these fluctuations had on the turbomachinery . To estimate the compressor and expander efficiencies at optimal design and off-design conditions, the performance maps of the turbomachinery are modelled. The sCO2 compressor performance map utilized in this study is first based on the characteristic performance curves of a typical compressor. These were expressed in normalized mass flowrates and presents an optimal operating point at 4Kg/s and 3700RPM [5]. However, the Aspen simulation of the optimal case described in section 3.2.1 indicates that the required sCO2 mass flowrate in the charging cycle is 60Kg/s. Therefore, in accordance with the fan law which states that the volumetric flowrate is proportional to the speed of the compressor, the mass flowrates and speeds were multiplied by a factor of 60/4 in order to stay in line with the design point of the system. Finally, as present studies were able to confirm that the efficiency of these machines was in the high eighties, the efficiency curves were adjusted so that the optimal design point was set at 85%.

After implementation of the compressor performance map in Aspen, a sensitivity analysis was run to determine the power input resulting in the highest compressor efficiency. The optimum design operating conditions of the charging cycle were considered reached for this power supply. In parallel, the compressor outlet conditions were monitored with the varying power input in order to facilitate the realistic temperature definition of the molten salt storage system.

A similar approach based on was adopted to model the characteristic performance curves of the expander [5]. Its outlet conditions were utilized to determine the temperatures of the cold thermal storage tanks.

2.4 Heat Exchanger Modelling

The next component type that required simulation was the heat exchangers of the system. The first important step was the selection of temperatures for the cold and hot storage tanks. The upper bound of the molten storage tanks is influenced by the degradation temperature of the solar salt used, which is 600ºC. The achieved compressor outlet temperature was 570 ºC. Hence, the hot side of the molten salt storage system was defined at 560 ºC, maintaining the chemical integrity of the material. On the cold side, the temperature of 420ºC was chosen, as it is widely cited by previous research in this field and it ensures a small temperature difference within the heat exchanger for minimum exergy losses.

With regards to the regenerative heat exchanger, the outlet of the cold side was defined with the same conditions as the compressor inlet to accommodate the closing of the loop. Then, the outlet of the hot side was manipulated, and the compressor outlet, at maximum capacity was monitored. The aim was to reach a sufficiently small temperature, while remaining within the supercritical region, to maximize the temperature difference between the heat source and heat sink.

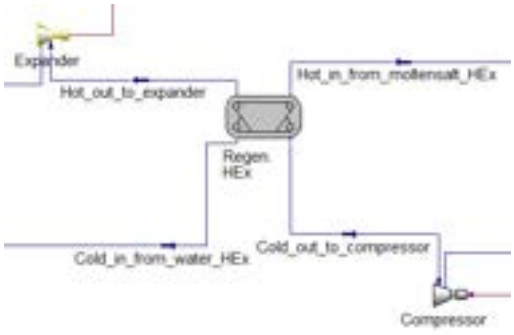


Figure 3: A diagram of the regenerative heat exchanger of the Aspen Hysys charging configuration

As the water thermal storage system heats up the working fluid, the cold tank was defined at 40°C, which is lower than the temperature of the outlet of the expander. On the hot side, the tank was kept at 80 °C, keeping the temperature difference small and the water below its boiling point.

Currently, in industry, the most efficient heat exchangers for systems involving supercritical fluids are plate counter current heat exchangers [2]. Hence, the two storage heat exchangers and the regenerative heat exchanger were modelled as such. The corresponding pressure loss in each one, on the CO_2 side, was kept at 1% as indicated by prior research conducted [1]. On the liquid side of molten salt and water the pressure loss was assumed negligible.

In both charging and discharging simulations conducted, the temperatures of the storage tanks were kept constant. This would, effectively, correspond to a storage fluid flow control system in real life operation. Such decision was important in order for the operation of the discharge cycle to not vary. The optimal conditions and flowrate of the discharge cycle will be retained, ensuring a constant design efficiency and hence, the highest round-trip efficiency. The sole variability will be the operating time in each run, depending on the amount of molten salt and cold water collected in the respective charging cycle.

2.5 Discharging cycle modelling

The discharging cycle operates in reverse to the charging cycle. There is the sole addition of a cooling unit to the ambient to ensure inefficiencies expressed as heat accumulation in the system can be removed and reversibility retained.

The cycle was modelled to closely approach the charging cycle in temperature and pressure. The optimal storage temperature values from charging were input into the discharging cycle. The storage fluid temperatures would effectively push the discharge thermodynamic cycle within the charging one. Moreover, it was essential to specify the operating times for these cycles. The charging time was chosen to be 8 hours and the discharging time 8 hours. These time periods overlap with the low demand dip and the high demand peak of the "duck cycle" [19]. The molten salt flowrate in the discharging cycle was therefore matched with the charging cycle molten salt flowrate in a one to one ratio. These specifications would be possible through the implementation of a temperature control system on the hot storage. Ultimately, it would maintain the discharging cycle conditions, including flowrates and energy exchange rates, unchanged regardless of part load operation. As a result, the efficiency of turbomachinery in the discharging cycle was considered non-variable. To guarantee the simulation is as realistic as possible, the efficiency of the turbine and the compressor were defined as the maximum efficiency obtained by the expander and the compressor of the charging cycle respectively.

2.6 Selection of design specifications and part load analysis

After identifying the optimal power input, different amounts of power were supplied to the compressor to simulate the part load operation of the system. The effect of part-load operation on the overall performance of the process was studied via an analysis of the roundtrip efficiency.

During each part-load operation run, the net power input in the charging cycle which corresponds to the power required from the compressor reduced by the power recovered from the expander is collected. The respective power in the discharging cycle is independent of part-load due to the temperature controls (see section 2.5). Moreover, the storage tank flowrates during part-load were used to calculate the amount of fluid heated or cooled respectively during the eight-hour charging time which will, hence, be available during discharge. the molten salt mass flowrate is kept constant, the collected fluid mass was used to calculate the updated discharge time. The following roundtrip equation was used, converting the power to work using the running times of the two cycles.

$$\eta_{round-trip} = \frac{W_{discharge}}{W_{charge}}$$

$$\eta_{round-trip} = \frac{W_{turbine} - W_{compressor}}{W_{compressor} - W_{expander}}$$

2.7 Storage tank sizing and operating times

Based on the temperatures chosen, during design operation, the mass flowrate between the hot storage tanks was noted. With a targeted 8 hour running time for the charging cycle, the molten salt tanks were sized accordingly to correspond to the fluid capacity required.

The modelling of the discharge cycle enables the sizing of the cold storage tanks. The water storage flowrate required for the optimal operation of the discharge cycle was noted and the water storage tanks were sized to allow for efficient capacity of cold water for discharge of a duration of 6 hours.

3.1 Performance maps

The modelling of the turbomachinery maps made it possible to evaluate its performance when operating at part-load conditions. A required shaft speed between 50 and 55 Krpm was predicted to reach the high efficiency region. These values agree with the range of sCO2 compressor speed cited by TiTech and SNL [6]. The simulation was therefore carried out with a constant corrected speed of 50 Krpm. The curves were also validated by comparison with the experimental and numerical results of the project sCO2-HeRo[7].

A representation of the compressor's characteristic curves in figure 5 shows the constant efficiency and constant impeller speed lines. The region to the right of the surge line is highly undesirable as it is characterized by a backflow of gas through the device. The operating point should therefore always remain on its right but must not be located at very high flowrates and low head values to avoid high power losses.

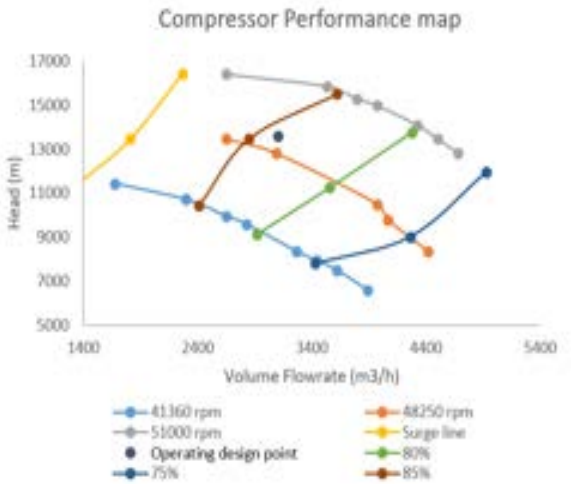


Figure 4 Compressor Performance map

Sensitivity analysis were performed to monitor the position of the operating point in the compressor and expander maps for different power inputs into the compressor. As the operating point is at the surge line at 6.5 MW and exceeds the maximum compressor flowrate at 9MW, only this power range can be further considered as power input. Figure 3 displays the efficiency of the compressor in this range. However the expander performance map indicates that the sCO2 flowrate exceeds its operational limits for a power supply to the compressor of 8.15MW. Therefore, the part load analysis will be carried in the range of 6.5 to 9 MW.

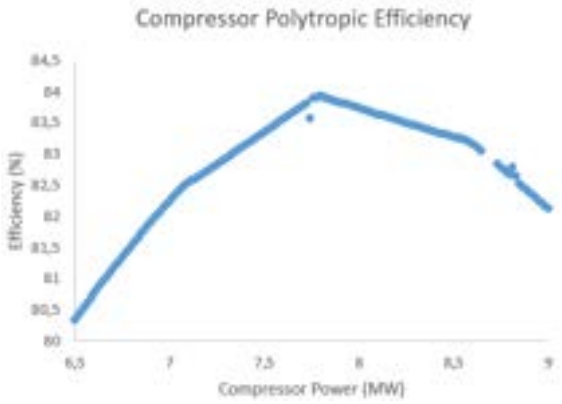


Figure 5 Compressor Polytropic Efficiency

3.2 Thermodynamic Cycle

Following the simulation of the charging and discharging cycle, the thermodynamic cycles displayed in figure 6 were obtained for the design operation of the process.

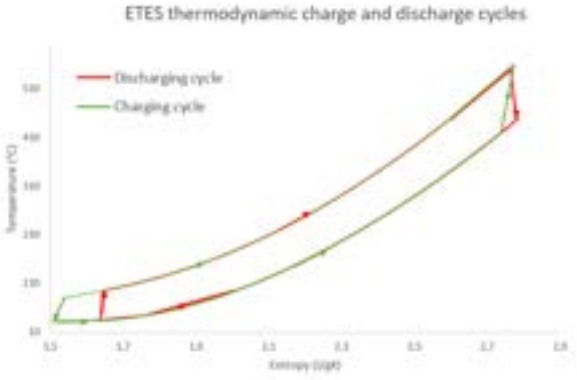


Figure 6 Thermodynamic charge and discharge cycles

At its biggest part a very good fit was achieved. This was the closest approximation to reversibility that the charging and discharging cycle conditions obtained from the Aspen Hysys simulations could provide. As expected, the discharging cycle is constrained by the charging one both in temperature and in pressure. The horizontal curved lines closely approach isobars, as plate heat exchangers, which were the ones used, benefit from a small pressure drop.

The gap between the temperatures of the working fluid on the cold side during charging and discharging is due to conditions approaching the

critical point. Isobars converging near the critical point. Therefore, despite the temperature difference between cycles being quite small, the corresponding difference in entropies are much larger.

3.3 Total and part-load operation

The design power supply to the compressor was chosen to be 7.79MW as it corresponds to the best turbomachinery efficiency (see figure 7). As the molten salt flowrate corresponds to 45.3Kg/s, a hot storage flowrate of 60.4Kg/s was implemented in the discharging cycle simulation. Following the methodology detailed in section 2.6, the roundtrip efficiency was calculated at different part loads as shown in figure 7. It is important to note that compressor power input at values higher than the initial design power chosen, the discharging time will not be curtailed at 6 hours. The charging and discharging temperatures maintained their 1:1 relationship and the operating time of the discharge was adjusted to ensure that all hot molten salt was utilized. This was done as a verification that the true optimal design power supply to the compressor was chosen.

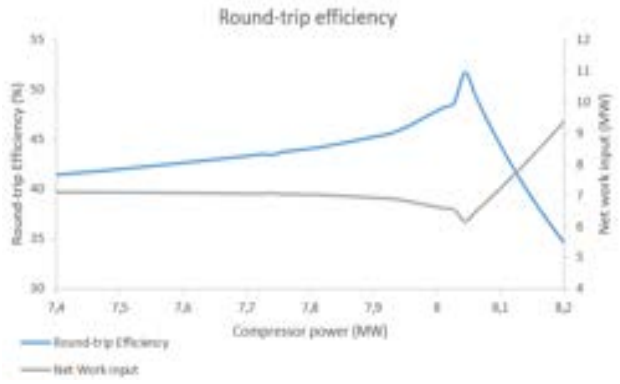


Figure 7 Round-trip Efficiency at Part-load operation

Simulation data shows that an increase in power input to the compressor enables an increase in the required hot storage flowrate. As a direct consequence, during part-load, less molten salt is accumulated in the hot molten salt tank and the discharging cycle running time is reduced. As the discharge power is constant, the discharge work will decrease. In the case of charging work, as evident in the graph the net input power is increasing, and the operating time is constant at 8 hours. Hence, the charging work increases. Therefore, their quotient, which is equivalent to overall round-trip efficiency. This loss of energy is in accordance with the drop of compressor efficiency at power values below 7.79MW, as showcased in Figure 7.

Though, it was observed that there is a peak in efficiency at a value that corresponds to a higher duty that the one of maximum turbomachinery efficiency. The mathematical explanation for this occurrence is the reverse of the previous case, as discharging times are increasing and net input power decreasing.

Table 33: Power associated with the turbomachinery during design operation of the systems configured

	Charging	Discharging
Compressor	8.043MW	4.442MW
Expander/Turbine	1.894 MW	1.251MW

3.4 Sizing of the thermal reservoirs

For a charging operating time of 8 hours and a discharging operating time of 6 hours the following tank sizes were calculated.

Table 4: Sizes required for hot and cold thermal storage tanks and the associated capital expenditure for their contents

	Tank size (m^3)
Molten Salt	880
Water	460

As observed, the size of the tanks for the pilot plant design power input is small enough for an ETES storage system to be implemented within cities. Placing energy storage at a short distance from the point of demand is very beneficial for overall power grid efficiency. Transmission over long distances creates power losses in the range of 8% to 15%. [20]. Local ETES systems can help support sustainable transportation by decreasing the transmission line.

Though, during operation, a complexity arises with the use of the water tanks. The required flow of water during charging is significantly smaller than the respective required one during discharge. Hence, after the charging cycle is completed, water will not have fully filled the cold side of the storage at 40^{o}C, which is the amount required for the efficient 6-hour discharge operation. A potential solution is the draining of the remaining water at 80^{o}C to the 40^{o}C. The cold tank will be open to the atmosphere to dissipate the excess heat and a heating body will be placed to ensure the temperature stays at the required temperature. Assuming the plant operates at a location with a

warm climate, the heat required is negligible. Alternatively, the surplus of hot water could be repurposed, for example for household use. Then, a fresh supply of water would be required to fill the cold water tank.

4. Conclusion

This study collected some key findings that will accommodate future implementation of the electrothermal energy storage. The maximum round trip efficiency for a pilot plant using current research based supercritical CO_2 compressor characteristics is 51%. It was observed that variability in CO_2 flowrate, caused by the fluctuating power input during part-load, restricts the operating range of the system. It also decreases the roundtrip efficiency to as low as 37% at minimum compressor power input. Externally increasing the CO2 flowrate during part load via a control system could counteract these challenges.

The tank sizes required for a pilot scale plant are compact, at 880 m^3 for the molten salt and at 460 m^3 for the water tanks. They could fit, even, within cities, assuming ones with availability of natural resources. A suggested suitable location for the system would be an island in the Mediterranean Sea. Though, it is important to note that additional heating and, potentially, water supply is required for the cold water tank. This could reduce round trip efficiency.

In order for ETES to become a reliable and widely applicable energy solution, further studies are essential. Primarily, as models for turbomachinery for supercritical fluids become increasingly realistic, the ETES simulations need to be updated accordingly.

This will ensure an up-to-date prediction for plant efficiency, as well as, capital and operational expenditure. Additionally, more robust heat exchanger modelling is essential. In this study, an assumption of 1% pressure drop within heat exchangers was made. It is very important for accurate pressure drop models to be constructed for the materials used. This will lead to an additional layer of complexity and realism in the simulations. Lastly, it is essential for the system to be designed to respond to the fluctuating nature of renewable energy, via a variable power input within the same charging cycle. Hence, an in-depth analysis of the dynamic operation of the process should be performed. This will enable the design of highly important control systems to the overall efficiency, such as CO_2 flowrate and thermal storage temperature control.

5. Bibliography

1. McTigue, J.D., et al., *Supercritical CO2 heat pumps and power cycles for concentrating solar power.* AIP Conference Proceedings, 2022. **2445**(1): p. 090006.
2. Mercangöz, M., et al., *Electrothermal energy storage with transcritical CO2 cycles.* Energy, 2012. **45**(1): p. 407-415.
3. McTigue, J.D., et al., *Supercritical CO2 heat pumps and power cycles for concentrating solar power.* AIP Conference Proceedings, 2022. **2445**(1).
4. Mercangoez, F.B.H., *Thermoelectric energy storage system with regenerative heat exchange and method for storing thermoelectric energy.* 2011.
5. Niu, L., et al., *Off-design performance analysis of cryogenic turbo-expander based on mathematic prediction and experiment research.* Applied Thermal Engineering, 2018. **138**: p. 873-887.
6. Saravi, S.S. and S.A. Tassou, *An investigation into sCO2 compressor performance prediction in the supercritical region for power systems.* Energy Procedia, 2019. **161**: p. 403-411.
7. Hofer, M., et al., *Simulation, analysis and control of a self-propelling heat removal system using supercritical CO2 under varying boundary conditions.* Energy, 2022. **247**: p. 123500.
8. Zoelle, A., *Quality Guidelines for Energy System Studies: Process Modeling Design Parameters*. 2019: United States. p. Medium: ED.
9. Mecheri, M. and Y. Le Moullec, *Supercritical CO2 Brayton cycles for coal-fired power plants.* Energy, 2016. **103**: p. 758-771.
10. Bertini, M., et al., *Evaluation of the property methods for pure and mixture of CO2 for power cycles analysis.* Energy Conversion and Management, 2021. **245**: p. 114568.
11. Kenisarin, M.M., *High-temperature phase change materials for thermal energy*

storage. Renewable and Sustainable Energy Reviews, 2010. **14**(3): p. 955-970.
12. Bonk, A., et al., *Advanced heat transfer fluids for direct molten salt line-focusing CSP plants.* Progress in Energy and Combustion Science, 2018. **67**: p. 69-87.
13. Vignarooban, K., et al., *Heat transfer fluids for concentrating solar power systems – A review.* Applied Energy, 2015. **146**: p. 383-396.
14. Gimenez, P. and S. Fereres, *Effect of Heating Rates and Composition on the Thermal Decomposition of Nitrate Based Molten Salts.* Energy Procedia, 2015. **69**: p. 654-662.
15. Bernagozzi, M., A.S. Panesar, and R. Morgan, *Molten salt selection methodology for medium temperature liquid air energy storage application.* Applied Energy, 2019. **248**: p. 500-511.
16. T. Conboy, J.P., D.Fleming, *Control of supercritical CO2 Recompression Brayton Cycle demonstration Loop*.
17. McTigue, J.D., et al., *Supercritical CO2 heat pumps and power cycles for concentrating solar power.* AIP Conference Proceedings, 2022. **2445**(1).
18. "New Analysis Reveals the Scale of Wasted Renewable Energy in the UK." *Highview Power*, Highview Power, 7 Sept. 2022, https://highviewpower.com/news_announcement/uk-energy-security-undermined-by-lack-of-energy-storage-capabilities/.
19. IEA, The California Duck Curve, IEA, Paris https://www.iea.org/data-and-statistics/charts/the-california-duck-curve, IEA. Licence: CC BY 4.0
20. Commission, I.E., *Efficient Electrical Energy Transmission and Distribution*. 2007: IEC.

Characterisation of Peptide Adsorption Mechanisms in Reversed-Phase High-Performance Liquid Chromatography

Jinhong Peng and Yuxin Sun
Department of Chemical Engineering, Imperial College London, U.K.

Abstract

Reversed-phase high-performance liquid chromatography (RP-HPLC) is considered the most widely employed method of purification for peptide-based drugs. A variety of research has been done on developing methods to improve HPLC performance and achieve higher yields. However, the lack of association with the peptide adsorption mechanisms derived from a deeper understanding leads to a great consumption of time and resources during method development. In this study, the fundamental mechanisms in peptide adsorption using RP-HPLC are investigated in detail. This was achieved by performing adsorption experiments in the HPLC using caffeine and different types of peptides as analytes, and cross-comparing the performances of columns Luna C18(2), Luna C8 and Luna phenyl-hexyl. The adsorption isotherms were characterised by employing different techniques such as frontal analysis (FA) and peak maxima (PM). Different mathematical models were fitted to the isotherms, leading to a more thorough investigation. The formation of multiple plateaus across all experiments with caffeine using FA indicates the existence of different adsorption sites within the columns. FA is deemed suitable for estimating the maximum column capacity due to column saturation being achieved at equilibrium. However, due to the low analyte quantity requirement and the minimal risk of gelation, using PM is considered more practical to examine the characteristics of peptide adsorption.

1. Introduction

Peptide-based drugs have gone through huge development over the past decades. In 2019, 11 of the top 20 drugs by global sales were peptide-based, including peptides and proteins such as antibodies and fragment antigens (Blankenship, 2019). Peptides are becoming more popular than proteins due to their simpler structure and comparable specificity. Additionally, the possibility of synthesising peptides through chemical synthesis, without the need for recombinant technologies required for protein, makes the production of these biomolecules more rudimental (Allee, 2017). However, since the by-products from chemical synthesis are analogous to the main product of interest, particular attention must be focused on the purification step.

High-performance liquid chromatography (HPLC) is the most common technique in purification for such types of peptides in the biopharmaceutical industry. The principal chromatography modes used for this application include normal-phase (NP-HPLC), size-exclusion (SEC), ion-exchange (IEX) and reversed-phase (RP-HPLC). Constant efforts have been made to improve the performance of all these modes to optimise HPLC purification steps (Žuvela *et al.*, 2019).

SEC has received high attention for producing resins capable of separating tiny molecules. However, the current improved technologies remain a limited minimum molecular weight cut-off of 5-6 kDa, which coincide with the upper limit of the average size for most peptides. Additionally, in the case of purification of peptide mixtures derived from chemical synthesis, since most contaminants are derivative species of the target sequence with comparable sizes, SEC mode is not suitable to be utilised (Mant *et al.*, 2007).

Although IEX provides good resolution for most applications, it requires extensive method development to ensure binding to the surface and programmed elution (Aguilar, 2004). Another disadvantage is that the selectivity between peptides with similar or close isoelectric points is poor, which leads to the requirement for a second purification step.

Due to the relatively high hydrophobic properties of peptides, RP-HPLC is considered the best purification method and most frequently used mode. In this mode, separation of the mixture of organic compounds and impurities is achieved based on the difference between their chemical properties and affinities to the stationary phase, where the more hydrophobic analytes are adsorbed stronger onto the stationary phase. Elution is achieved by changing the solvent strength of the mobile phase. Moreover, the resolution and selectivity of this technique are sufficient to separate complex mixtures of peptides with a relatively simple system (Žuvela *et al.*, 2019).

By adjusting the composition of the aqueous mobile phase, solute retention and selectivity can be manipulated. Hence, even though the experimental window of solvent concentration required for protein and peptide elution is very narrow, fine-tuning can lead to satisfactory purifications. The three most employed organic solvents to increase the solvent strength of the aqueous mobile phase in RP-HPLC are acetonitrile, methanol, and 2-propanol, which all exhibit high optical transparency under the detection wavelengths used for peptide and protein analysis (Aguilar, 2004). Due to the interaction with the stationary phase, these different organic solvents form an organic-rich adsorbed layer with a unique thickness, which has a direct effect on the adsorption ability of the stationary phase (Gritti & Guiochon, 2005).

During method development, columns are tested at different mobile phase compositions, normally through a gradient mode where the organic solvent is mixed gradually with water until the analyte is eluted. This iteration process is time-consuming, and the final method only applies to the column and analyte tested to the point where even columns with similar characteristics show small variations. The complexity of the RP-HPLC separation reproducibility result from inconsistencies during manufacturing of the resins.

Silica-based resin is currently the optimal choice for RP-HPLC packing materials, as its pore structure and morphology allow silica particles to be mass-produced

with a certain degree of reproducibility while maintaining a rapid mass transfer and decent loading ability (Žuvela *et al.*, 2019). During synthesis, condensation or polymerisation between silanol (Si-OH) groups is required to create siloxane bridges (Si-O-Si) (Rahman & Padavettan, 2012). This type of reaction leaves free silanol groups at the surface of the resin, which is then functionalised with any desired ligand. However, the efficiency of this reaction is affected by the type of silanol present on the surface (isolated, vicinal or geminal), pore accessibility and steric hindrance (Bracho *et al.*, 2012).

Residual silanols can produce numerous intramolecular interactions such as hydrogen bonding, van der Waals forces and London dispersion forces, with not only the analyte but also mobile phase molecules (Bocian *et al.*, 2010). Thus, the presence of these undesired silanols can interfere with the hydrophobic interaction between analyte and stationary phase during a chromatographic run depending on the coverage density of the resin ligand. A customary strategy to counter the effect of residual silanols is by adding an ion-pairing reagent, frequently trifluoracetic acid (TFA), as an additive in the mobile phase in RP-HPLC for the purification of peptides and proteins. These molecules contain both ionic and hydrophobic functional groups; this characteristic allows them to control the pH by providing an acidic environment for adsorption, form a complex with oppositely charged ionic groups to enhance retention, and suppress the ionic interactions between peptides and silanol groups on the silica (Supelco, 2002).

These complex interactions can be investigated through the measurement of adsorption isotherms which describe the relationship between the adsorbate in the liquid phase and the adsorbate on the surface of the adsorbent at equilibrium. A variety of methods for the determination of adsorption data are available, such as frontal analysis (FA), frontal analysis by characteristic point (FACP), peak maxima (PM), elution by characteristic point (ECP), and the inverse method (IM) (Marchetti *et al.*, 2009). A method is selected depending on the availability of the analyte and the precision of measurement required. Nonlinear behaviour of adsorption isotherms is usually observed in the purification of peptides or proteins because of the properties of the solute, mobile phase and stationary phase, such as the nature of the distribution equilibria of the solutes between phases, chemical reactions or equilibria in the column, solubility limitations, and viscosity effects in liquids.

The experimental adsorption data can be fitted to mathematical models based on different fundamental assumptions to characterise and understand the fundamentals of the adsorption mechanisms. These models include Freundlich, Langmuir, Brunauer-Emmett-Teller (BET) and so on. However, the degree to which these models represent the characteristics of the stationary phase and its interactions with the analyte is determined by the accuracy of the adsorption data.

The main objectives of this research is to thoroughly investigate the fundamentals of the complicated adsorption mechanisms for peptides. This was achieved by first examining the performance of frontal analysis and peak maxima using a relatively simple compound, caffeine. This leads to gaining an understanding of the adsorption mechanisms for small molecules with relatively simple interactions to the column stationary phase. The adsorption isotherms were determined for peptides of different sequences and chemical properties. In this way, the adsorption mechanism of peptides were better understood with model fitting and comparison with that of caffeine.

2. Background

Over the years, along with the extensive use of RP-HPLC for purification of proteins and peptides, the majority of research has shown the tendency to investigate methods to enhance the HPLC performance for higher yields to be achieved as per industrial requirements. For instance, it is demonstrated that a mixed-mode reversed-phase/ weak anion-exchange technique with the utilisation of two columns with different stationary phases can provide better selectivity based on the differences in hydrophobicity and charge between peptides and impurities (Nogueira *et al*., 2005). However, there has rarely been any research investigating the fundamentals of the adsorption mechanism in RP-HPLC for the separation of proteins and peptides. Without a deeper understanding of such mechanisms, a large amount of resources and time may be consumed using merely trial-and-error to improve the performance of RP-HPLC.

In a study published by Gritti & Guiochon (2005), models were successfully developed to analyse the adsorption isotherms and the adsorption energy distributions of caffeine and phenyl using a variety of organic solvents. This provided insight into the adsorption mechanisms in RP-HPLC. In the case where methanol was used as an organic modifier in the mobile phase, strictly convex upward isotherms were established, indicating that an adsorbed analyte monolayer was formed in the organic-rich film on the stationary phase surface. On the other hand, s-shaped isotherms were observed when acetonitrile was used in the mobile phase, suggesting the formatting of upper layers. It can be concluded from these results that acetonitrile is a stronger eluent for low-molecular-mass polar compounds compared to methanol. It is also demonstrated that the behaviours of adsorption isotherms in RP-HPLC are highly dependent on the nature of the organic mobile phase (Gritti & Guiochon, 2005).

To explore the fundamentals of peptide separation in RP-HPLC, the correlation between the hydrophobicity of peptides and their retention time was well investigated by Krokhin &Spicer (2009). In their work, a hydrophobicity index (HI) parameter was measured in RP-HPLC under isocratic conditions and fitted by the retention prediction model. This value is capable of indicating the hydrophobicity and concentration of the organic modifier, as well as the fixed retention factor for each peptide. Regardless of the lack of further investigation into the adsorption isotherm behaviour, this study well characterised peptide hydrophobicity and

provided a foresight in methodology development, which could be further utilised.

The adsorption isotherm of a tripeptide (LLL) was investigated experimentally by frontal analysis under both controlled and uncontrolled pH conditions. By fitting the calculated profile from the inverse method (IM) of isotherm determination to the experimental breakthrough profile, the comparison was made and showed the existence of two different adsorption isotherms between the two pH conditions. With further analytical procedures, this study concluded that the pH of the elution environment and the quantity of ion-pairing agent added to the mobile phase have a significant influence on the retention mechanisms of peptides, as well as the non-linear isotherm behaviour (Andrzejewska *et al.*, 2009).

3. Methodology

3.1 Materials

Caffeine (1,3,7–Trimethylxanthine) powder was purchased from Sigma-Aldrich (ReagentPlus). Bovine Serum Albumin (BSA) lyophilized powder (>=96%) and Lysozyme (from chicken egg white) crystalline powder were purchased from Sigma-Aldrich. Peptides P1 (LGGGGGGDGSR), P2 (LGGGGGGDFR), P3 (LLGGGGDFR), P4 (LLLGGDFR), P5 (LLLLDFR) and P6 (LLLLLDFR), developed by Krokhin and Spicer, were purchased from GenScript with a minimum purity of 96% and no termini modification. All organic solvents used were HPLC grade. Methanol (>=99.9%) and acetonitrile (>=99.8%) were purchased from Fisher Scientific. The ion-pairing agent, Trifluoroacetic Acid (TFA, >=99.0%, 100mL) was purchased from Sigma-Aldrich. Deionised (DI) water with quality of 18MΩ Ohm was obtained from Elga Purelab Chorus unit.

3.2 Equipment

A modular HPLC prominence LC-40 (Shimadzu, Japan) was used to carry out all chromatographic sorption experiments. Mobile phase solutions were continuously filtered with a 10 µm in-line filter and degassed using a DGU-405 degassing unit and an LC-40D pump. The chromatography data was recorded during injection and elution by placing a dual channel SPD-40 UV/Vis detector and an SPD-M40 photodiode array detector (PDA) before and after the column, respectively. The temperature was controlled using a CTO-40C column oven. All samples were injected using a SIL-40C autosampler.

HPLC columns Luna C18(2) (150 x 4.6 mm ID,), Luna C8 and Luna phenyl-hexyl were purchased from Phenomenex (USA) with similar dimensions of 150 x 4.6 mm ID and average particle size of 5 µm**.**

3.3 Determination of mobile phase volume

The volume of the mobile phase, also known as the thermodynamic void volume (V_M), was obtained from the excess isotherms measured by the minor disturbance method. The HPLC pump was used to mix water (line A) and acetonitrile (line B) in a step series from 0% to 100%B. Each step was run for 20 mL, allowing an equilibration volume of 10 mL between steps. After equilibration, a 1 µL injection of pure acetonitrile was injected into the column to maintain infinite dilution of the organic modifier in the mobile phase. The retention volume of the first resulting disturbance peak, either positive or negative, detected by the post-column detector was adjusted by the retention volume of the injection peak detected by the pre-column detector.

The integration of these retention volumes over the concentration range was computed to determine V_M, as explained by the following equation:

$$v_M = \frac{\int_{C_B^l(0\%)}^{C_B^l(100\%)} v_r dC_B^l}{C_B^l(100\%)} \qquad (1)$$

where C_B^l is the concentration of ACN in the bulk liquid, and the percentage in parenthesis refers to the concentration of ACN in equilibrium.

3.4 Experimental procedure for caffeine adsorption

Caffeine was used as a reference adsorbate to compare the agreement between FA and PM methods. Both methods were carried out under 40 °C and with a mobile phase solution of methanol/water (30/70, v/v) in accordance with the Tanaka method (McHale *et al*., 2021) and to ensure desorption of the analyte.

FA experiments were performed with a saturated mother batch of 35 g/L caffeine in methanol/water (30/70, v/v) and another mobile phase solution of methanol/water (30/70, v/v) to adjust the composition.

For PM experiments, caffeine in methanol/water (30/70, v/v) solution of the following concentrations were prepared in 2 ml vials respectively and placed into the Auto Sampler in HPLC: 0.5, 1, 2.5, 5, 7, 8, 9, 10, 15, 20, 25, 30 mg/mL. The following injection volumes were taken for each vial: 1µL, 10 µL, 25 µL, 50 µL, 100 µL. Each volume of injection was repeated by triplicates for result accuracy. The same method was carried out for all three target columns and with an additional run without any column for comparison. The system was initially equilibrated using the mobile phase solution in between testing for each column.

3.5 Experimental procedure for peptide adsorption

3.5.1 Isocratic test

For all six types of peptides, 200 µL samples of 2 mg/mL peptide in water solution were prepared, transferred into vials, and placed in the Auto-Sampler, separately. Isocratic tests for P1, P2, P3, P4, P5, P6 were carried out in the ranges of 0%-20%, 5%-25%, 10%-30%, 15%-35%, 20%-40%, 25%-45% volume of ACN in water, respectively. Each injection was completed with increments of 2.5% of ACN.

3.5.2 Adsorption of peptides by PM

Solutions of ACN with 0.1%(vol) TFA and water with 0.1%(vol) TFA were prepared and connected to line B and line A of the HPLC inlet, respectively. The concentrations of the ACN in water mobile phase solution used for P1, P2, P3, P4, P5 and P6 were 5%, 12.5%, 17.5%, 22.5%, 27.5% and 30%, accordingly.

For P1-P4, 1 mL samples of 75 mg/mL peptide in water solution were prepared, transferred into vials, and placed in the Auto-Sampler. The 1 mL samples for P5 and P6 were prepared with 4 mg/mL concentration of

peptide in ACN/water (25/75, v/v) and ACN/water (50/50, v/v), respectively.

Injection volumes of 0.5, 1, 2.5, 5, 7.5, 10, 20, 30, and 50 µL were taken from the P1-P4 vials. Injections of 9.4, 18.7, 46.9 and 93.8 µL were taken for P5 and P6. All injection volumes were performed by triplicates. Same procedure was carried out on all three columns.

3.6 Measurements of Adsorption Isotherms

3.6.1 Frontal Analysis (FA)

The breakthrough curves of analyte before and after the column were measured by two detectors. Signal intensities were transformed to the concentration of analytes and the volume of the mobile was subtracted from the post-column detector signal. Area between two curves on the front part was calculated and represented to the mass of analytes adsorbed. Equilibrium concentration was calculated by following equation:

$$q_e = \frac{m_a}{V_s} = \frac{m_a}{V_C - V_m} \quad (2)$$

where m_a is the mass of analyte adsorbed, V_s is the volume of stationary phase, V_C is the volume of the column, V_m is the void volume of column.

3.6.2 Peak Maxima (PM)

The retention time was obtained from the main peak in chromatograph. Equilibrium concentration was calculated by following equations:

$$V_r = t_r * F_m \quad (3)$$

$$q_e = \int \frac{V_r - V_m}{V_s}\, dc \quad (4)$$

where V_r is the retention volume, F_m is the flowrate of mobile phase, t_r is the retention time, V_s is the volume of stationary phase, V_m is the void volume of column

3.7 Modelling of Adsorption Isotherms

Adsorption data were fitted via multiple-objective optimisation using the following models:

Freundlich model:

$$q_e = K_F {C_e}^{1/n} \quad (5)$$

Redlich-peterson model:

$$q_e = \frac{K_{RP} C_e}{1 + \alpha_{RP} C_e^g} \quad (6)$$

Langmuir model:

$$q_e = \frac{q_m K_L C_e}{1 + K_L C_e} \quad (7)$$

BET model:

$$q_e = \frac{q_m K_{BET1} C_e}{(1 - K_{BET2} C_e)(1 - K_{BET2} C_e + K_{BET1} C_e)} \quad (8)$$

Anti-Langmuir model:

$$q_e = \frac{\hat{q}_m \hat{K}_L C_e}{1 - \hat{K}_L C_e} \quad (9)$$

The best-fitting model was selected by minimising the mean squared error.

4. Results and Discussion

4.1 Determination of V_M

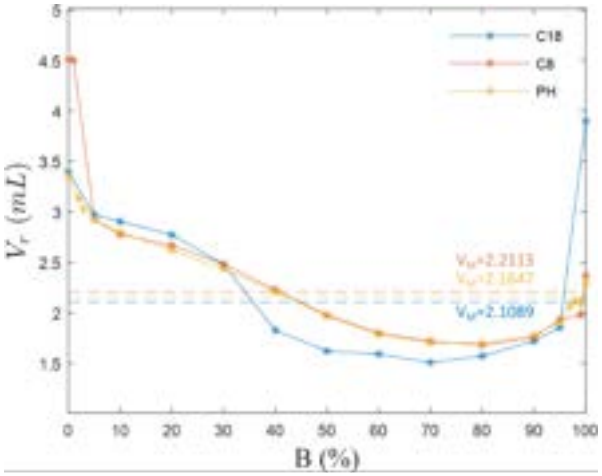


Figure 1: Estimation of V_M values in acetonitrile/water mixtures. The organic modifier in the mobile phase, acetonitrile, for this case, is represented as B. Percentages are shown as volume/volume ratios. V_M values were calculated using *Eq.* 1.

The mobile phase volume of a column (V_M) is a crucial parameter in chromatography used to adjust retention volumes to obtain simple retention factors (k') for system suitability assessment, theoretical descriptors, prediction of retention times, and determination of thermodynamic properties responsible for chromatographic retention, partitioning and sorption processes. It represents the approximate volume of liquid inside a column during chromatographic conditions. In other words, it describes the minimum amount of liquid required to fill the empty spaces in the column, including the pore and the interstitial volumes. Approximations of V_M are experimentally performed by measuring the minimum amount of mobile phase required to elute an unretained molecule, customarily referred to as void volume (V_0). Under ideal conditions, V_0 and V_M are identical; thus, they are frequently used interchangeably in the literature. However, reports have shown that no completely unretained marker exists for RPLC columns due to the complex surfaces of these silica particles (Luo & Cheng, 2005). Variations of V_0 are attributed to the mobile phase composition and marker properties.

Alternatively, V_M can be defined thermodynamically by the Gibbs excess isotherms of organic/water mixtures in contact with an adsorbent defined by *Eq. 1*.

Figure 1 depicts the retention volumes of a peak formed from a minor disturbance of the mobile phase at equilibrium for the three columns tested. The V_M has values of 2.2123, 2.1647, and 2.1089 mL for C8, PH and C18, respectively. The order is not surprising since the apparent size of the ligand is proportional to the length of the ligand when fully extended; thus, the V_M values are inversely proportional. In other words, since C18 occupies more space than PH and C8, the volume available to the mobile phase reduces. However, this inference is only valid when the ligand surface coverage and the packing density are the same amongst columns. Hereinafter these V_M values were considered for all calculations unless otherwise stated.

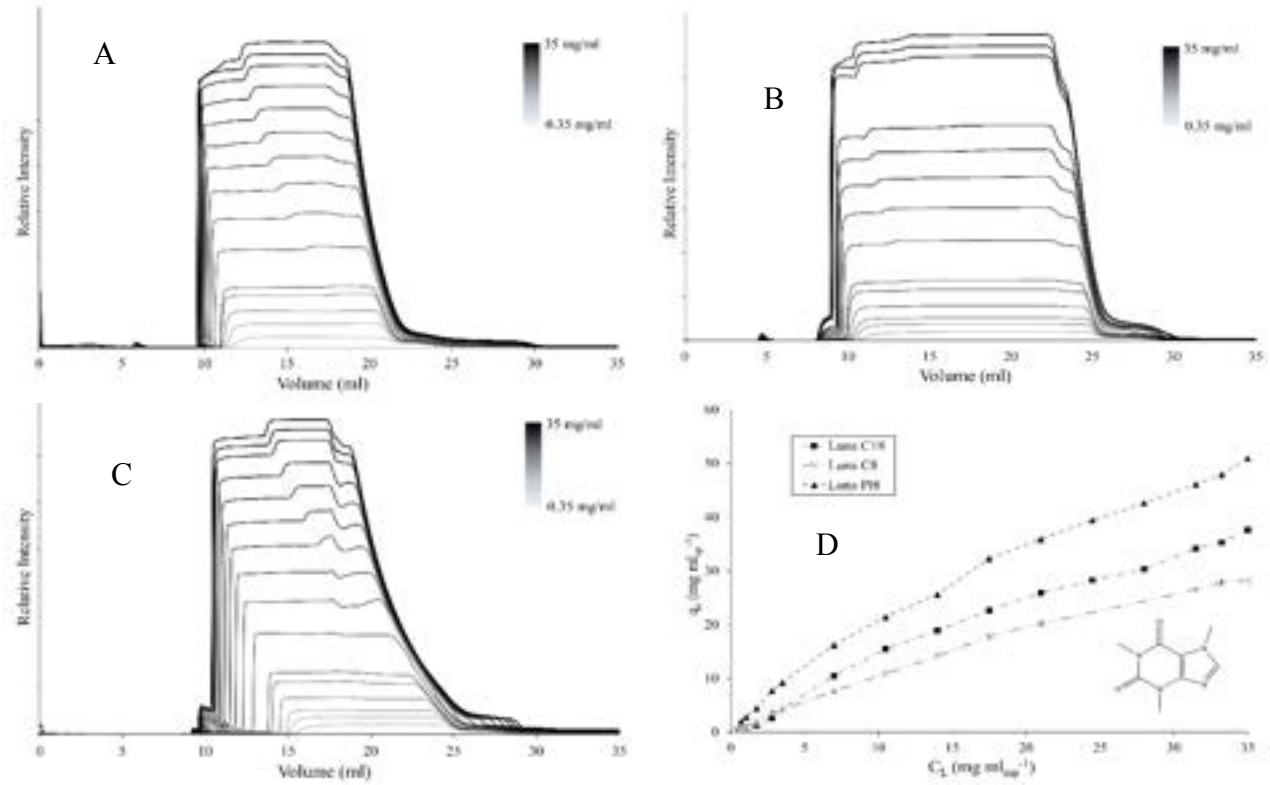


Figure 2: Caffeine breakthrough curves recorded by PDA detector at λ = 308nm in column (A) C18, (B) column C8, (C) column PH, and (D) the adsorption isotherm data of caffeine determined by FA for all three columns and its molecular structure

4.2 Analysis of adsorption data of caffeine

4.2.1 Breakthrough curve and adsorption isotherm by FA

As per the principle of FA, saturation of column is achieved to measure the overall loading capacity at equilibrium, thus a plateau is formed once the column is fully saturated in that specific concentration. This can be observed for columns C8, C18 and PH in *Figure 2A, 2B* and *2C*, respectively. If the analyte follows an ideal adsorption mechanism with only a single adsorption site, a single plateau would be expected (Marchetti *et al*., 2009). Contrastingly, all columns tested for this project, showed the appearance of a second plateau, suggesting the presence of another adsorption mechanism.

As the concentration of analytes increases at equilibrium, the earlier and higher this plateau forms. So the first assumption made to this phenomenon the is formation of second or multiple layers are formed on the stationary phase. A plateau is also seen during desorption at the rear part of the breakthrough curves, which suggests another assumption that the existence of complete different adsorption site dependent on the analyte-surface interactions such as hydrophobic interactions at the resin ligand, hydrogen bonding at the silanols, electrostatic interactions at protonated silanols, π-π interactions between aromatic rings. However, the breakthrough curves are not capable to justify which assumption is correct, further justification is carried out in *Section 4.2.3*.

The adsorption isotherms have also been calculated and plot in *Figure 2D*. Since columns C18 and C8 have alkyl chains with different length covalently attached to the silica surface of the resin, similar hydrophobic and electrostatic interactions between the analytes and stationary phase would be expected, implying that the distribution of analytes and loading capacity in high-energy sites would also be similar. The experimental adsorption data shows a good agreement to this theoretical statement, the equilibrium concentration of caffeine of column C18 has the same trend and comparable adsorption at low concentrations of analyte. In higher range of concentration of analytes, column C18 has a longer alkyl chain which provides more high-energy sites for analytes to be adsorbed and relatively higher overall loading capacity. Size exclusion effect is not considering, since caffeine should be small enough to access all the pores (Miyabe & Guiochon, 2004).

The isotherm for column PH shows the highest equilibrium concentration among all three columns in the entire range of concentration. This result can be explained by the π-π interaction between analytes, caffeine, and the aromatic ring on the ligand of the silica surface, so stronger attraction from the stationary phase contributes greater loading capacity on the high-energy site, as well as the overall capacity of adsorption.

4.2.2 Breakthrough curve and adsorption isotherm by PM

The peak maxima method has also been used to analyse the adsorption mechanism of columns. As per the principle of peak maxima, only a small amount of analyte is injected into the system, and the analytes are adsorbed onto stationary, and then eluted away. Therefore, with the same condition, the equilibrium concentration should also be the same as the volume of injection increase. However, instead of remaining the same shape, the adsorption isotherms curve downward as the injection volume rise as shown in *Figure 3A*, which means that the retention time of analytes is shorter.

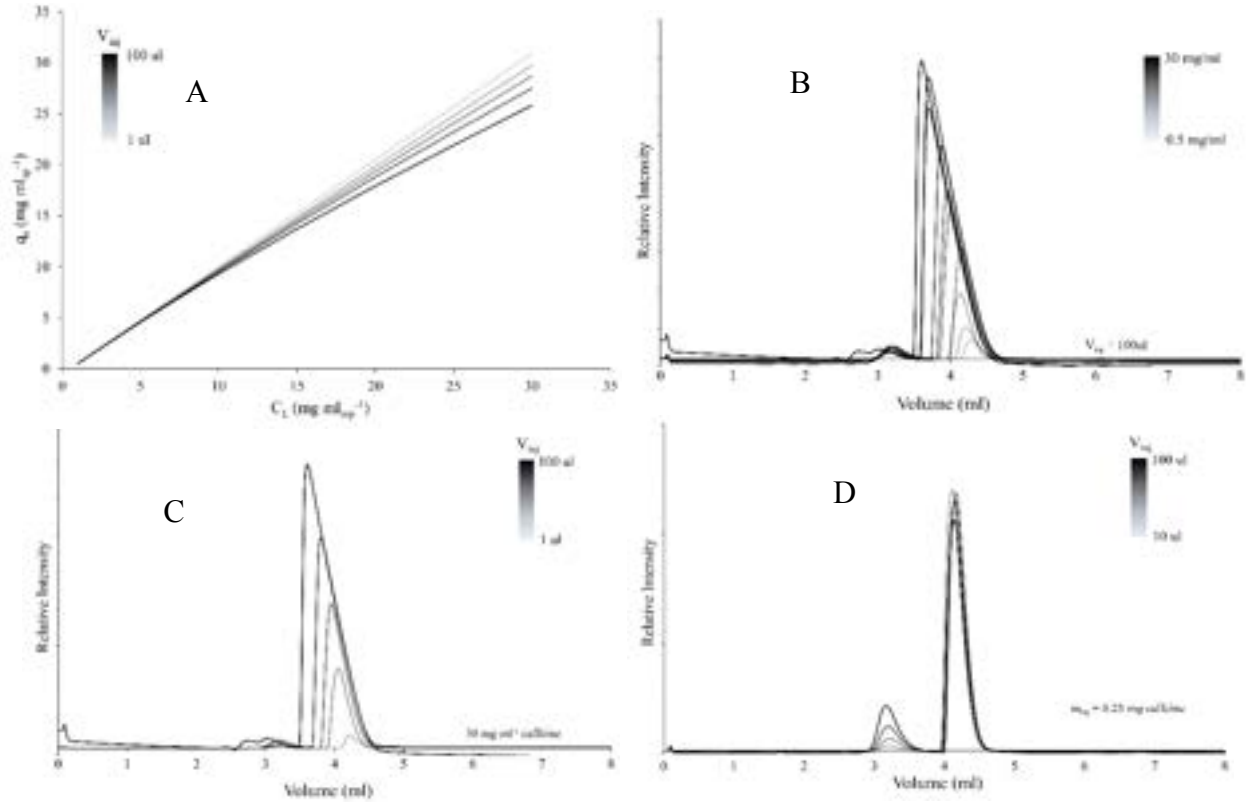


Figure 3: Adsorption data of caffeine determined by PM with different injection volume (A); The chromatographs of peak recorded by PDA detector at λ = 308nm for varying concentration (B), of varying injection volume (C), of same injection mass (D).

To understand this isotherm behaviour, a comparison between peak properties with varying injection parameters was carried out. *Figure 3A* shows that the peak is higher, broader, and earlier with an increase in injection volume for the same sample concentration. The same observation is also observed with an increase in concentration and the same injection volume as shown in Figure 3B. This result indicates that the more mass of analytes is injected into the column, the elution would start earlier and have a longer duration. Thus, the peak behaviour is only dependent on the mass of injection. This statement is confirmed in *Figure 3C* as the same peak shape is observed (V_r = 4.157mL ±0.028; Height = 152.9E+03mV ±6.66E+03; W0.5h =0.58mL± 0.014) for the same mass of injection regardless of the volume of injection and sample concentration.

From these results, it can be speculated that multiple adsorptions and desorption stages occur along the column. When a small amount of analyte is injected, the compounds are momentarily adsorbed onto the stationary phase, desorbed into the mobile phase, and re-adsorbed onto the stationary phase repeatedly as the peak band moves along the length of the column until full elution can be observed. With higher mass of injection, a higher surface coverage is achieved, preventing re-adsorption. Thus, these adsorption-desorption stages would complete faster, resulting in an earlier and broader peak.

4.2.3 Comparison between isotherms of FA and PM

The adsorption isotherms determined by PM have shown that column PH has the highest equilibrium concentration and C8 has the lowest, which is consistent with the isotherms by FA and verifies the theoretical assumption mentioned in *Section 4.2.1*. When both methods are compared to each other, PM gives more linear adsorption behaviours and intersects their FA counterparts. However, an overestimation to the equilibrium concentration produced by the PM method is expected because the mass of analytes injected ($m_{inj,max}$=3 mg) is significantly lower than that for FA ($m_{inj,max}$= 25 mg). The analytes adsorbed onto the stationary phase in the latter would not saturate the column, hence equilibrium concentration is never achieved.

The Langmuir model (*Eq. 7*) was chosen to fit the adsorption data determined by PM and FA, and isotherm parameters were obtained in *Table 1*. The great difference between isotherm parameters q_m which represent the maximum adsorption capacity of the column for PM and FA has proved the over-estimation of equilibrium concentration by PM.

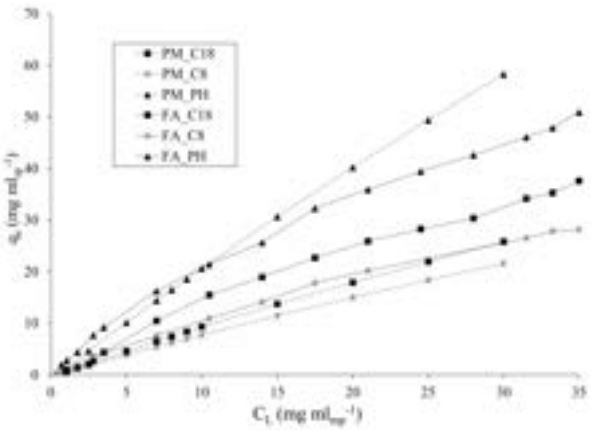


Figure 4: the adsorption isotherm data determined by FA and PM for three columns

Table 1: Isotherm parameters acquired from the adsorption of caffeine by FA and PM in three columns

	FA			PM		
	C18	C8	PH	C18	C8	PH
q_m	108.309	82.251	110.239	258.894	228.527	748.744
K_L	0.015	0.015	0.023	0.004	0.003	0.003
SSE	10.432	1.120	8.429	0.326	0.203	1.459

The BET model was selected to justify the formation of a second adsorption layer mentioned in *Section 4.2.1*, due to its derivation from a multi-layer physical adsorption mechanism. When the model was fitted (*Eq. 8),* the parameter K_{BET2} nullifies meaning that the model becomes Langmuir, hence there is no more loading capacity apart from the first layer, which denies the 'second layer' assumption. Gritti and Guiochon concluded that the organic-rich layer that forms from the excess isotherm of MeOH/H2O mixtures has a thickness of 4 Å, which is not sufficiently thick to allow the analytes to form an additional layer. On the other hand, they also proved that the organic-rich layer formed from ACN/H2O mixtures measures 13 Å which provides enough space to create a second layer showing a BET adsorption behaviour for caffeine (Gritti & Guiochon, 2005).

Similarly, when the Redlich-Peterson model was conducted (*Eq. 6*), the exponential parameter g turned to 1, and the model turned into Langmuir model which is an indication of macroscopic homogeneous adsorption. However, the chemistry of silica particles contradicts this idea as residual silanols from the condensation reaction prevail after end-capping. These silanols cannot be capped with current technologies (Bracho *et al*., 2012), which means that there should always be at least two energy sites, alkyl ligands and silanols, on the surface of the resin.

Even though the silica material should be heterogenous in theory, as the greater difference in adsorption energy between energy sites and high coverage of end-capping, the resin could be considered a homogenous surface mathematically. Analytes which are only affected by hydrophobic interaction or silica resin with low end-capping coverage can be used in the same experiment to verify the assumption above.

4.3 Analysis of adsorption data of peptides

4.3.1 Isocratic Test

The amino acids present in the peptide sequence determine the chemical properties of the biomolecule. For instance, when comparing peptides P1 and P2, a glycine (Gly, G), a neutral amino acid, and a serine (Ser, S), a polar uncharged amino acid, was substituted from the former for one hydrophobic phenylalanine (Phe, F) in the latter. This subtle substitution increases the relative hydrophobicity of the sequence. This behaviour also occurs when two glycine side chains are substituted by a leucine (Leu, L) another hydrophobic amino acid. On that account, the hydrophobicity increases from P1 to P6.

The same trend is observed in the elution order of RPLC separation processes, as shown in Figure 5. Since a higher content of ACN, which means stronger solvent strength, is needed to split the hydrophobic interaction between peptides with higher hydrophobicity and stationary phase and cause elution.

The same explanation is applicable to the retention behaviour of individual peptides. The retention volume of P1 remains constant when the percentage of ACN is greater than 10% and starts retaining only within the range of 5-10%. As the content of ACN decreases in the mobile phase, the solvent strength reduces, favouring the adsorption of P1 onto the stationary phase and displacement from the mobile phase. Additionally, no elution occurs when the percentage of ACN is lower than 5%. Therefore, the adsorption range of each peptide can be inferred from the data presented in *Figure 5*.

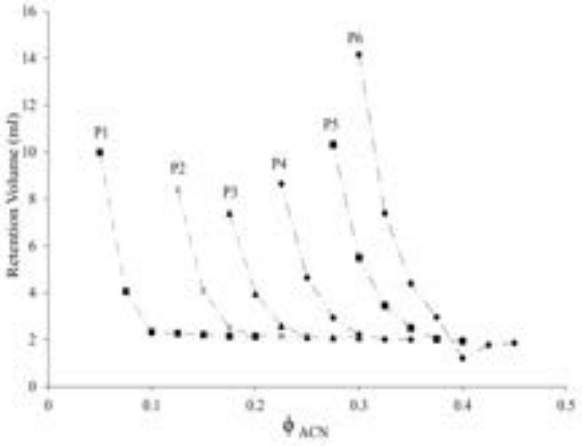


Figure 5: Retention volume of P1 to P6 in correspond percentage of ACN

4.3.2 Isotherm behaviour under adsorption conditions

In the adsorption of peptides, high-energy sites adjacent to the ligand on the stationary phase becomes inaccessible because the size of the peptide is greater than the separation between alkyl ligands. Thus, peptides only interact with the tail of the ligands and cause adsorption on the low-energy sites.

With this theory, similar adsorption isotherm behaviours of P1 amongst different columns are expected, as shown in figure 6A. As there is no aromatic ring in molecules, so the effect of π-π interaction is not observed in the case of P1, and only hydrophobic interactions affect the adsorption. Contrastingly, for P2 to P6 (*Figure 6B-F*), where an aromatic ring exists at the serine side chain, a lower equilibrium concentration in column PH, compared to column C8 and C18 is observe. As the discussion in *Section 4.2.1*, π-π interaction should have synergy with the hydrophobic interaction between analytes and stationary phase. This result indicates that a repulsive π-π interaction is dominant instead of attraction, it has a significant interference with the attractive hydrophobic interaction.

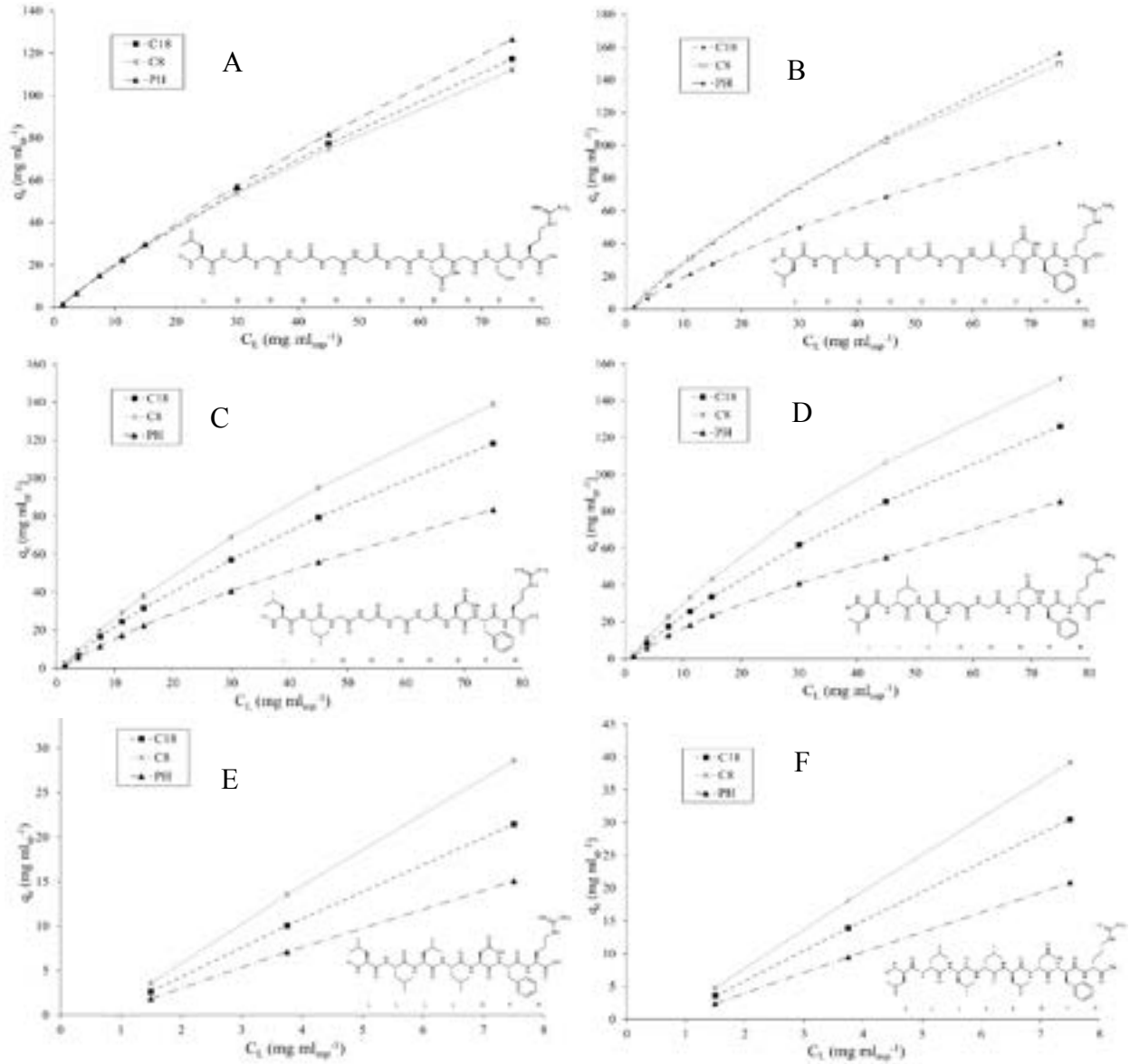


Figure 6: The adsorption isotherm data of P1(A), P2(B), P3(C), P4(D), P5(E), P6(F) determined by PM under adsorption conditions for three columns and corresponding peptide sequences

Another major observation is that column C8 appears to have higher adsorption than column C18 for P3, and this difference becomes greater from P4 to P6. This result could be explained by considering the properties of both the peptides and the column resins.

A relative ligand coverage can be calculated by dividing the carbon load of the column by the number of linear carbons in the ligand. As the carbon load for C8 is 13.5% and 17.5% for C18 and PH, according to the supplier, the relative coverage for each column would be 1.69, 0.97 and 1.94 for C8, C18 and PH, respectively. If the coverage of ligands on the surface is low, the residual silanols and other end-capping species become more accessible to peptides, resulting in contributions from hydrophilic and ionic interactions during adsorption. So, a higher ligand coverage generates a relatively more homogenous surface which enhances the hydrophobic interaction with peptides. Therefore, the difference in coverage between columns C8 and C18 could explain higher adsorption for more hydrophobic peptides, as observed from P2 to P6 (*Figure 6B-F*).

Another consideration with regard to the efficiency of adsorption is the accessibility of pores. By taking into account the properties of P1 and P2, where both peptides have comparable sizes (approximately 10 nm) but distinct hydrophobicity, a similar concentration of adsorbed peptide onto both columns shows that the driving force governing adsorption might be the size and not hydrophobicity. As the three resins have an average pore size of 10 nm, according to the manufacturer, one molecule of P1 and P2 can effectively occupy the whole opening of the pore, reducing the surface area for adsorption, causing a size exclusion effect. As the size of a peptide decreases, more pores can be accessed and enhance the adsorption of peptides. This can be extended to P3 to P6, where this statement occurs. There have been reports that ligands close to the opening of the pores can reduce the pore diameter or completely block it (Carr *et al*., 2011). Assuming a kinetic diameter of 0.7 nm per carbon (Aguilar-Armenta & Dıaz-Jimenez, 2000), two molecules of C18 at opposite sites in the pore aperture will reduce the diameter to 8 nm. A shorter length of alkyl chains can reduce the possibility of

Table 2: Isotherm parameters acquired from the adsorption of peptides P1-P6 under adsorption conditions in three columns

	P1			**P2**			**P3**		
	C18	C8	PH	C18	C8	PH	C18	C8	PH
q_m	502.796	411.360	674.262	562.140	442.257	312.016	397.566	427.420	271.199
K_L	0.081	0.100	0.062	0.102	0.136	0.006	0.113	0.128	0.006
SSE	1.897	2.515	2.841	4.385	4.789	3.677	4.468	3.733	2.319
	P4			P5			P6		
	C18	C8	PH	C18	C8	PH	C18	C8	PH
q_m	415.425	409.862		-	-	-	-	-	-
K_L	0.116	0.157		-	-	-	-	-	-
SSE	2.901	4.576		-	-	-	-	-	-

having a blocked pore on the surface, thus, adsorption is enhanced.

All adsorption data of P1 to P4 are fitted with the Langmuir model (*Eq. 7*), and corresponding isotherm parameters are shown in *Table 2*. Due to the solubility of P5 and P6 is low, data is not sufficient to be fitted by any model.

4.3.3 Isotherm behaviour under elution conditions

Ideally, no adsorption should occur under elution conditions, as the solvent strength is greater than the interaction between peptides and the stationary phase. However, adsorption is still observed and presents an anti-Langmuir behaviour, as shown in *Figure 7*.

As the peptides are injected into the column with an organic-rich mobile phase, compounds can still be adsorbed onto the stationary phase and cause mere retention if they get close enough to the surface. With a higher mass of injection, more compounds are possible to be adsorbed, so greater retention is achieved.

Higher adsorption is still observed in column C8 under elution, which may prove that the surface of the stationary phase in column C8 is more homogenous and provides stronger hydrophobic interaction with peptides.

Adsorption data is fitted with the anti-Langmuir model, and corresponding isotherm parameters are shown in *Table 3*.

5. Conclusions

In conclusion, since the unique molecular structure and physicochemical properties of peptides, the interaction between peptides and phases in PR-HPLC system is more complicated caffeine. Therefore, more factors are considered to characterise the adsorption mechanism of peptides, such as size exclusion effect, hydrophobicity of peptides, pore structure and coverage of stationary phase, solvent strength of mobile phase, interactions with residual silanols and so on.

FA is a more suitable method to analyse the complete adsorption isotherm and estimate the maximum column capacity since full column saturation is achieved at equilibrium. However, this method requires large amounts of analyte compared to PM. This is important for the investigation of peptides whose availability is low. It is also curial to note that the experimental conditions of FA are limited to the solubility of the peptide, because of the risk of gelation inside the column.

So, PM is particularly useful in this case because the amount of analyte required is significantly lower. Even though the PM method cannot provide accurate adsorption parameters, it can still characterise the adsorption mechanism sufficiently to compare different peptides onto different columns.

In the future, the adsorption of phenol can be experimented with to compare with the data of caffeine and analyse the effects of hydrogen bonding on adsorption. The same experiment can be carried out for peptides with NP-HPLC to investigate the interaction between residual silanols and peptides. Larger or more hydrophobic peptides can also be used in the same experiment to enlarge the size exclusion effect or enhance hydrophobic interaction.

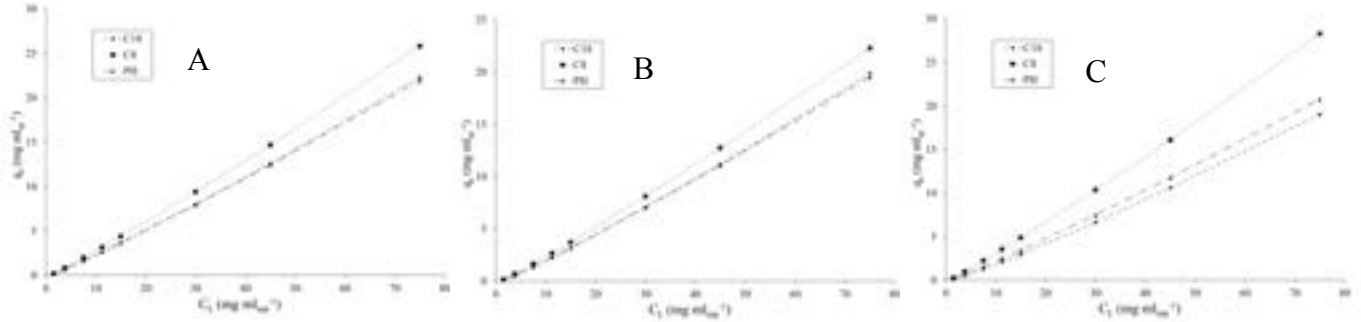


Figure 7: The adsorption isotherm data of P2(A), P3(B), P4(C) determined by PM under elution conditions for three columns

Table 3: Isotherm parameters acquired from the adsorption of peptides P2-P4 under elution conditions in three columns

elution condition	P2			P3			P4		
	C18	C8	PH	C18	C8	PH	C18	C8	PH
$\widehat{q_m}$	93.789	120.621	119.515	72.143	167.701	101.458	261.117	140.909	115.327
$\widehat{K_L}$	0.046	0.042	0.037	0.056	0.039	0.045	0.012	0.041	0.043
SSE	0.449	0.329	0.331	0.350	0.223	0.198	0.110	0.272	0.200

6. Acknowledgement

We would like to express our utmost gratitude to Oscar Mercado Valenzo for his continuous support and guidance throughout the period of this research project.

7. References

Aguilar-Armenta, G. & Díaz-Jiménez, L. (2001). Characterization of the porous structure of two naturally occurring materials through N2-adsorption (77 K) and gas chromatographic methods. *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, 176 (2-3), 245–252.

Aguilar, M.-I. (2004). Reversed-Phase High-Performance Liquid Chromatography. *HPLC of Peptides and Proteins*, 251, 9–22.

Andrzejewska, A., Gritti, F. & Guiochon, G. (2009). Investigation of the adsorption mechanism of a peptide in reversed phase liquid chromatography, from pH controlled and uncontrolled solutions. *Journal of Chromatography A*, 1216 (18), 3992–4004.

Blankenship, K. (2020). *The top 20 drugs by global sales in 2019*. [online] Available at: https://www.fiercepharma.com/special-report/top-20-drugs-by-global-sales-2019. [Accessed 16th December 2022].

Bocian, S., Vajda, P., Felinger, A. & Buszewski, B. (2010). Effect of End-Capping and Surface Coverage on the Mechanism of Solvent Adsorption. *Chromatographia*, 71 (S1), 5–11.

Bracho, D., Dougnac, V.N., Palza, H. & Quijada, R. (2012). Functionalization of Silica Nanoparticles for Polypropylene Nanocomposite Applications. *Journal of Nanomaterials*, 2012, 1–8.

Carr, P.W., Dolan, J.W., Neue, U.D. & Snyder, L.R. (2011). Contributions to reversed-phase column selectivity. I. Steric interaction. *Journal of Chromatography A*, 1218 (13), 1724–1742.

Gritti, F. & Guiochon, G. (2005). Adsorption Mechanism in RPLC. Effect of the Nature of the Organic Modifier. *Analytical Chemistry*, 77 (13), 4257–4272.

Hannappel, M. (2017). Biopharmaceuticals: From peptide to drug. *AIP Conference Proceedings*, 1871 (1), doi:10.1063/1.4996533.

Krokhin, O. V. & Spicer, V. (2009). Peptide Retention Standards and Hydrophobicity Indexes in Reversed-Phase High-Performance Liquid Chromatography of Peptides. *Analytical Chemistry*, 81 (22), 9522–9530.

Luo, H. & Cheng, Y.-K. (2006). A comparative study of void volume markers in immobilized-artificial-membrane and reversed-phase liquid chromatography. *Journal of Chromatography A*, 1103 (2), 356–361.

Marchetti, N., Cavazzini, A., Pasti, L. & Dondi, F. (2009). Determination of adsorption isotherms by means of HPLC: Adsorption mechanism elucidation and separation optimization. *Journal of Separation Science*, 32 (5-6), 727-741.

McHale, C., Soliven, A. & Schuster, S. (2021). A simple approach for reversed phase column comparisons via the Tanaka test. *Microchemical Journal*, 162.

Miyabe, K. & Guiochon, G. (2004). Characterization of monolithic columns for HPLC. *Journal of Separation Science*, 27 (10-11), 853–873.

Nogueira, R., Lämmerhofer, M. & Lindner, W. (2005). Alternative high-performance liquid chromatographic peptide separation and purification concept using a new mixed-mode reversed-phase/weak anion-exchange type stationary phase. *Journal of Chromatography A*, 1089 (1-2), 158–169.

Rahman, I.A. & Padavettan, V. (2012). Synthesis of Silica Nanoparticles by Sol-Gel: Size-Dependent Properties, Surface Modification, and Applications in Silica-Polymer Nanocomposites—A Review. *Journal of Nanomaterials*, 2012, 1–15.

Shibue, M., Mant, C.T. & Hodges, R.S. (2005). Effect of anionic ion-pairing reagent hydrophobicity on selectivity of peptide separations by reversed-phase liquid chromatography. *Journal of Chromatography A*, 1080 (1), 68–75.

Supelco. (2002) *Application Note 168 Eliminate TFA and Improve Sensitivity of Peptide Analyses by LC/MS*. [online] Available at: https://www.sigmaaldrich.com/deepweb/assets/sigmaaldrich/marketing/global/documents/129/989/11547.pdf [Accessed 15th December 2022].

Wang, J. & Guo, X. (2020). Adsorption isotherm models: Classification, physical meaning, application and solving method. *Chemosphere*, 258, 127279.

Žuvela, P., Skoczylas, M., Liu, J., Bączek, T., Kaliszan, R., Wong, M.W. & Buszewski, B. (2019). Column Characterization and Selection Systems in Reversed-Phase High-Performance Liquid Chromatography. *Chemical Reviews*, 119 (6), 3674–3729.

Developing a Chiral Alanine and Water Ternary Phase Diagram and investigation of NRTL Model Applications

Sarah Gunnery and Bastiaan Geurtz

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

Abstract

Ternary diagrams are pertinent in the development of preferential crystallisation techniques, yet current experimental techniques to develop these are slow, and modelling is limited to a few applications of the NRTL model. Therefore, the focus of this research will be to reduce the time required. Techniques to reduce experimental time will be explored and, while doing so, novel ternary data for a chiral alanine and water system will be collected. Investigation into the NRTL model will also be undertaken, exploring reduction of the data input required. Significant reductions in time were successfully made at the cost of high precision by applying the assumption that behaviour between different enantiomers and a solvent can be considered the same. The modelling investigation has shown that without extensive data to highlight patterns ternary data will remain very reliant on experimental procedures. Novel ternary data has also shown alanine is a racemic forming compound.

1. Introduction

Chiral compounds are a type of compound which have two non-superimposable mirror image enantiomers, they are abundant in biological systems and are used throughout industry. As of 2021 the global chiral chemical market was valued at USD 58.82 billion, an increase of 47% on the valuation made in 2015, USD 39.79 billion, and is expected to continue growing with an estimated value of USD 149.95 billion by 2030 ((NMSC), 2022) (GVR, 2022).

The pharmaceutical, agrochemical, food, and cosmetic industries all contribute to this market and regulations for these bodies are developing to create safer and more environmentally conscious products regarding chirality. For instance, in Sweden there is incentive to reduce environmental loading, and hence the use of racemic mixtures in the agrochemical industry, through the implementation of a tax on the weight of the active components (Williams, 1996). To use the single active isomer product which would be more desirable in this situation, production techniques require development.

Significant caution is required, specifically in the pharmaceutical industry where biological structures such as proteins, sugars, amino acids, and nucleic acids are also often chiral (Shaffer, 2022). As a result, enantiomeric drugs produce different effects, ranging from the desired therapeutic effect to severe side effects – one of the most prominent cases of this was thalidomide; of which one enantiomer has a sedative effect compared with the other enantiomer resulting in a teratogenic effect (Zhang et al., 2019). As of 2006 56% of the drugs in use were chiral, of which 88% were racemic mixtures (Nguyen et al., 2006). Hence, pharmaceutical industries require intensive research surrounding chiral compounds. They are tasked with investigating the differing pharmacokinetics, pharmacodynamics, and toxicology of enantiomers to determine whether a racemic mixture is acceptable for drug delivery or if an enantiopure compound is required for safe or improved delivery. FDA policy states 'manufacturers should develop quantitative assays for individual enantiomers in *in vivo* samples early in drug development' (FDA, 1992). This requires the capacity to produce pure enantiomers both for investigation and the production of those which are found to be safe or improved in pure enantiomeric form.

Two available routes for producing enantiomerically pure compounds are asymmetric synthesis and separation. Asymmetric synthesis requires, either a chiral pool (reactant), chiral auxiliary or chiral catalyst, and is reaction specific. Separation routes are also especially difficult because of the similarities between enantiomeric properties.

The aim of this paper will be to consider how preliminary research required for development of ternary diagrams for the application in preferential crystallisation of chiral compounds can be improved; both by reducing the time and therefore economic input. The process used for gathering experimental data to create a ternary diagram will be evaluated, this will include considering assumptions which can contribute to the reduction of experimental time required. The experimental procedure will be completed upon a chiral alanine and water system as there is currently no ternary data available from a literature review for this system. Additionally, an implementation of an NRTL model in Julia will be evaluated which will consider how efficiently binary data can be used to create ternary diagrams in chiral compound and solvent systems.

2. Background

A ternary phase diagram is the graphical depiction of the solubility of two compounds at varying compositions and temperatures in a solvent. There are two distinct types of chiral compounds which can undergo preferential crystallisation, these are conglomerates and racemic compounds resulting in two types of ternary diagrams as shown in Figure 1. Racemic forming compounds form crystals with both enantiomers within the crystal lattice, conglomerates however form separate crystals for each enantiomer therefore creating a physical mixture instead.

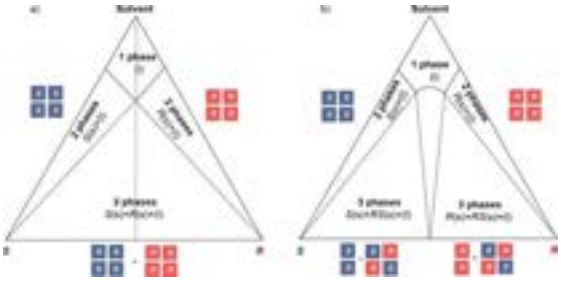


Figure 1: "Ternary solubility phase diagrams for (a) a conglomerate and (b) a racemic compound-forming system, with S and R, the two enantiomers of a chiral compound, RS, the racemic compound, and s and l, indication of the solid and liquid state, respectively" (Cascella et al. 2020).

Racemic systems are a particularly important focus as they account for 90% of chiral systems (Wang et al., 2005). The ternary diagrams regions within racemic forming systems and in particular the eutectic points (the two points at which the three phase, two phase and one phase region meet) determine viable preferential crystallisation limits (yields and purity) and are pertinent to development of preferential crystallisation separation techniques. Hence, significant importance of this report is derived from the application to preferential crystallisation, however preferential crystallisation is not covered in further detail in this report, two significant papers in this field to consider for further information are referenced at the end of this paper are (Cascella et al. 2020) and (Gänsch et al. 2021).

Furthermore, classical crystallisation techniques are present in at least one stage in over 80% of pharmaceutical separation processes (VAISALA n.d). Crystallisation techniques are a well-known technique in industry and would allow for new developments with preferential crystallisation techniques to be much more simply adopted than entirely novel processes.

For modelling purposes, ternary solubility equilibria have been determined for multiple chiral systems in solvents such as Mandelic Acid in Water and Threonine in a Water/Ethanol mixture (Lorenz et al 2003). Earlier work by Worlitschek et al (2004) describes the determination of Trögers base in Ethanol where Trögers base is used as model system in chiral chromatography. The NRTL activity coefficients are then used in solubility equations to build a ternary diagram. The enantiomer-enantiomer binary system is also investigated. More recently, N-methylphedrine has been determined in two different chiral solvents (Kaemmerer et al 2010) Since then, Mandelic Acid has received considerable attention because of both its role as pharmaceutical precursor and its favourable crystallisation characteristics for experimental procedures. Ternary solubility phase diagrams for Mandelic Acid in chiral solvents have also been constructed (Tulashie et al. 2010).

3. Methodology

This research has two main focuses, the experimental research determining a preliminary ternary diagram of alanine and the computational research to evaluate and improve the modelling of ternary diagrams available.

3.1 Experimental

3.1.1 Rig Design

The rig design included a jacketed vessel to contain the solution attached to a cooling system. The cooling system worked via feedback control cooling or heating water in response to a temperature probe inserted into the solution. This was pumped through a sintered metal filter to remove crystals and subsequently through a second PTFE 0.22µm screw filter to remove air bubbles. This was followed by the pump (Jasco PU-1585), which was set to a 1ml/min flowrate when in use, a densitometer (Anton Paar mPDS 1000) and a polarimeter (Advanced Laser Polarimeter PDR-Chiral Inc) as seen pictured in Figure 2. Additionally, there was a second temperature probe within the vessel attached to a data logger used to log the temperature, voltage from the densitometer and the voltage from the polarimeter. A magnetic stirrer was used in the vessel at 700rpm to ensure mixing of the solution.

In addition, a turbidity and imaging probe was used to evaluate the turbidity when evaluating the metastable zone width.

Figure 2: Image of the rig design used for the experimental collection of ternary chiral alanine and water data.

The materials in use are:

L-Alanine (Biosynth/Carbosynth >98.5%)
D-Alanine (Biosynth/Carbosynth >98.0%)
DL-Alanine (Alfa Aesar 99.0%)

3.1.2 Metastable Zone Width

Two solutions were made for the metastable zone width evaluations, the first containing 100ml of water and 20.16g of L-alanine and the second 100ml of water and 20.92g of DL-alanine. Each was heated to 55°C whilst the magnetic stirrer was set to 700rpm to dissolve the initial crystals. The imaging probe and temperature data logger probe were inserted into the solution. The imaging software was run before starting the experiment to check if surface bubbles or the stirrer was in view of the probe and adjusted if necessary. Once the probe was reading a steady turbidity and the images processed showed no crystals, bubbles, or the stirrer, the temperature control was dropped to 10°C. This temperature was allowed to reduce until crystallisation occurred (determined by a recorded increase in turbidity) at which point the temperature was increased to 55°C. When the turbidity returned to its initial steady value, the crystals were fully dissolved. The two temperature values of importance are those at which the crystallisation first began and the temperature at which the crystals fully dissolved; these signify the start and end of the metastable zone width for this solution. The exact temperatures for these two points were determined using the data logger information at the times recorded from the probe for the point of the beginning of crystallisation and end of dissolving. It is important to note that the crystallisation and dissolving should happen while the temperature is changing i.e., not yet at a steady state temperature, when considering the metastable zone width.

3.1.3 Densitometer Calibration

To calibrate the densitometer five 50ml solutions of L-alanine and water were created with varying concentrations (10, 20, 30, 40 and 60 g/L) which were below the literature saturation solubility at 25℃ of 164g/L (Yalkowsky & Dannenfelser, 1991). Each of these were continuously stirred by the magnetic stirrer to ensure mixing. The solution was pumped through the system at 1ml/min to a waste jar. The voltage output of the densitometer was recorded, and the value taken once it had reached a steady value; this was repeated three times for each solution. A linear calibration was created linking the density input and the voltage output as seen in Figure 3.

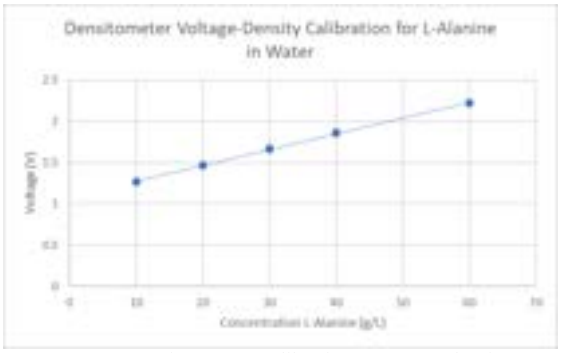


Figure 3: Densitometer calibration curve, densitometer voltage(V) vs known density of Alanine(g/L) with a linear regression.

The calibration line followed Equation 1

$$Voltage(V) = 0.0191 * Concentration(g/L) + 1.0857 \quad (1)$$

The linear regression (R^2) value of this calibration curve is 0.9991 with all three repeats considered.

3.1.4 Ternary Diagram

The Ternary Diagram data collected consisted of eighteen distinct points made up of, six different enantiomeric excesses (100%, 80%, 60%, 40%, 20% and 0%) each of which were evaluated at three different temperatures (10°C, 15°C, 20°C) to create three isotherms. The solution in the jacketed vessel was mixed using the magnetic stirrer throughout the experiments and the data logger in use. To begin, pure L-alanine and water was mixed for a 100% enantiomeric excess and was heated to 55°C, the crystals fully dissolved and then the temperature reduced to 20°C. The solution was allowed 30 minutes to begin crystallising and the pump was then turned on at 1ml/min. To enable the crystallisation to finish the solution was allowed to recycle through the system for 2 more hours. The pump was then turned off and the temperature returned to 55°C for the crystals to once again dissolve. This full process was then repeated substituting 15°C and 10°C as the temperature the solution was reduced to.

Visually the data logger represents this as seen in Figure 4, the increase in temperature to dissolve all crystals happens after the initial experiment for 20°C has ended and the voltage stabilised, 30 minutes are allowed to pass before the pump is restarted which here results in a spike as the crystallisation has not fully finished and the next experiment is commenced.

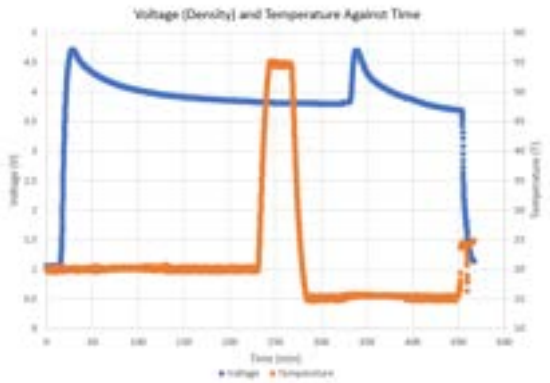


Figure 4: Voltage (V) and Temperature (ºC) against Time (mins) of the alanine system when running until crystallisation has finished.

The full process is repeated for each new enantiomeric excess which are created by adding D-alanine to the original solution. As alanine is a racemic forming compound, the solubility of the solution does decrease beyond the eutectic point, a new solution had to be made from scratch with a lower total alanine content to dissolve at 55°C once the eutectic point was passed and the solubility increased too much. This solution then underwent the same addition of D-alanine for the remaining points. The final solution created was the racemic solution itself which was created from the racemic compound rather than a mixture of L-Alanine and D-alanine, this was used for increased accuracy of the racemate point.

3.2 Modelling

Solid- liquid equilibria for racemic enantiomers in a solvent can be represented with two model equations, namely the Schröder and van Laar and the Prigogine and Defay equations. The SolMod.jl *(github.com/RGambarini/SolMod.jl)* package in Julia was used to solve the solubility predictions.

The Schröder and van Laar equation, shown below in Equation 2, determines the equilibrium between a single enantiomer and the solvent at a given composition x_i and activity coefficient γ_i.

$$ln(x_i\gamma_i) = \frac{\Delta_{fus}H_i}{R}\left(\frac{1}{T_{m,i}} - \frac{1}{T}\right) \qquad (2)$$

For L-alanine an enthalpy of fusion ($\Delta_{fus}H_i$) of 75.33 kJ/mol and temperature of melting($T_{m,i}$) of 581.95 K were found using differential scanning calorimetry (DSC) (O'Brien, no date) Alternative values of 22±5 kJ/mol and 608±9 K respectively are also reported in another paper which are found using fast scanning calorimetry. Hence, a range of values were tested in our modelling evaluation. It is worth noting that the literature does also highlight issues arising due to the decomposition of L-alanine during measurements of these values and is likely the contributing factor to this uncertainty (Zen Chua et al., 2018).

The equation by Prigogine and Defay (1973) shown in Equation 3 considers the interactions between the two enantiomers using both enantiomeric compositions and activity coefficients. This gives rise to the racemic forming compound behaviour within the ternary diagram.

$$ln(4x_i\gamma_i x_j\gamma_j) = \frac{\Delta_{fus}H_{rac}}{R}\left(\frac{1}{T_{m,rac}} - \frac{1}{T}\right) \qquad (3)$$

Here the enthalpy of fusion and temperature of melting also change to reflect the focus on the racemic compound rather than the pure enantiomer. The racemic form, DL-alanine, has a quoted enthalpy of fusion ($\Delta_{fus}H_{rac}$) of 113 kJ/mol and a temperature of melting ($T_{m,rac}$) of 562.15 K.

An example outcome of these two equations is shown in Figure 5 below, with the linear sections attributed to the Schröder and van Laar equation and the non-linear line a result of the Prigogine and Defay equation. The two points of intersection shown are the eutectic points. To create the ternary diagram, the Schröder and van Laar results are used at enantiomeric excesses higher than that of the eutectic points (the outer edges of the ternary diagram) and the Prigogine and Defay results are used at enantiomeric excess values between the two eutectic points.

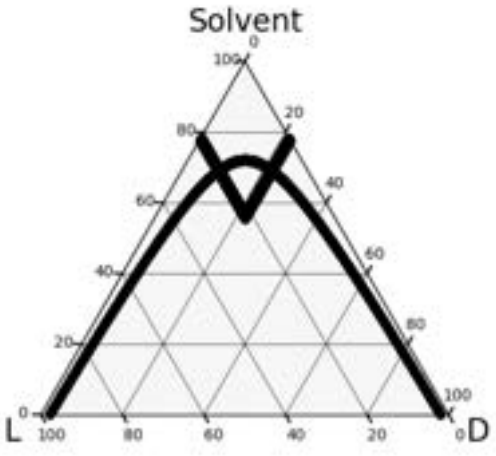


Figure 5: Example Ternary Diagram for a racemic forming system, for the linear component see Equation 2 and non-linear component Equation 3.

It has been shown that values on the RHS of both equations are determined by experimental

procedures or found via literature which quotes experimental procedures. The composition and temperature of the isotherm are inputs to the model. All that is left to consider is the activity coefficients, here is where the NRTL modelling method is required. The Non-Random Two-Liquid (NRTL) Gibbs excess model is applied to calculate the non-ideal solvent-solute interactions. In a system of c components the activity coefficient is given by Equation 4:

$$(\gamma_i) = \frac{\sum_{j=1}^{c} \tau_{ji} G_{ji} x_j}{\sum_{j=1}^{c} G_{ji} x_j} + \sum_{j=1}^{c} \frac{x_j G_{ij}}{\sum_{j=1}^{c} x_k G_{kj}} \left(\tau_{ij} - \frac{\sum_{j=1}^{c} x_k \tau_{kj} G_{kj}}{\sum_{j=1}^{c} x_k G_{kj}} \right) \quad (4)$$

In which the values of τ and G are determined by the Equation 5 and Equation 6 where α_{ji} is the non-randomness parameter introduced into the model. These variables express the temperature dependency of the activity coefficients in the NRTL model:

$$\tau_{ji} = \frac{g_{ji} - g_{ii}}{RT}, \ \tau_{ij} = \frac{g_{ij} - g_{ii}}{RT} \quad (5)$$

$$G_{ji} = exp\left(-\alpha_{ji}\tau_{ji}\right), \ G_{ij} = exp(-\alpha_{ij}\tau_{ij}) \quad (6)$$

In the case of the ternary system of alanine in a solvent (c=3) the interaction parameters between each enantiomer (i=1 or 2 for L/D) and the solvent (i=3) are considered symmetrical as shown in Equation 7 and Equation 8. The nonidealities are then identical with:

$$g_{13} = g_{23}, \ g_{31} = g_{32} \quad (7)$$

$$\alpha_{13} = \alpha_{23} = \alpha_{31} = \alpha_{32} \quad (8)$$

An additional assumption in the first instance is that the heterochiral interactions between the enantiomers are considered negligible so that:

$$g_{12} = g_{21} = 0 \quad (9)$$

The three final initial parameters which then still need determining beyond these assumptions are g_{13}, g_{31} and α_{13}.

Correlations for the binary parameters can be found in Aspen Plus databases of regression information. NISTV110 NIST-IG provides values for L-alanine and water of g_{13}= -3'526 kJ/mol, g_{31}= 11'019 kJ/mol and α_{13}= 0.1. The required binary parameters for L-alanine at 100% EE can be determined from the experimental procedure carried out for alanine and the heterochiral interaction parameters g_{12}, $g_{21}$and α_{12} were calculated using the objective function described by Haida et al. (2010). The objective function uses a least-squares method using MATLAB to further refine the parameters. The same function implemented in Julia 1.8.2 was used to iterate the parameters found from Aspen.

4. Results and Discussion

4.1 Experimental

4.1.1 Alanine Ternary Diagram

As outlined prior, the primary focus of the experimental procedure was to build a ternary diagram for chiral alanine and water. The output of the method outlined above is the voltage measurements from the densitometer corresponding to each solubility point. The values found from the experiment are shown in Table 1, the voltage output and hence the density calculated using Equation 1.

Table 1: Solubility values determined at varying L-Alanine enantiomeric excess values and varying isotherms.

Enantiomeric Excess (%) (L-Alanine)	Temperature (K)	Densitometer Reading (V)	Density (w_1+w_2) (g/L) (Calibration curve found in section 3.1.3)
100	293.15	3.77	140.54
100	288.15	3.65	134.26
100	283.15	3.52	127.45
80	293.15	4.17	161.48
80	288.15	4.03	1 54.15
80	283.15	3.89	146.82
60	293.15	4.39	173.00
60	288.15	4.26	166.19
60	283.15	4.12	158.86
40	293.15	4.18	162.01
40	288.15	3.98	151.53
40	283.15	3.82	143.16
20	293.15	3.89	146.82
20	288.15	3.7	136.87
20	283.15	3.56	129.54
0	293.15	3.83	143.68
0	288.15	3.69	136.35
0	283.15	3.53	127.97

Using the density of the alanine above, water density taken at 1000g/L and the enantiomeric excess values the compositions can be calculated. Considering w_1 and w_2 as the weight of L-Alanine and D-alanine respectively the solubility given as a mass fraction can be found via a mass balance as shown in Equation 10. Where the enantiomeric excess reported above is found through Equation 11.

$$W_{1\%} + W_{2\%} = \frac{w_1 + w_2}{w_1 + w_2 + 1000} \quad (10)$$

$$EE = \frac{w_1 - w_2}{w_1 + w_2} * 100 \quad (11)$$

Plotting this results in the ternary diagram shown in Figure 6.

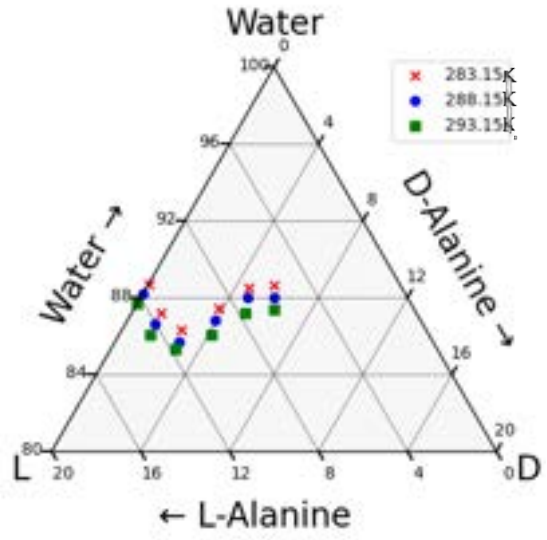


Figure 6: Ternary phase diagram of the alanine enantiomers in water at enantiomeric excess varying between 0% - 100% for three isotherms.

From this, Alanine shows behaviour of a racemic crystal forming compound (this was the likely outcome considering as stated in the background 90% of compounds create racemic forming systems) the solubility on average increases by 5.9g/L of alanine per 5 degree increase of temperature.

Assumptions to extend the available data can allow for development beyond the experimental results. In total there are 9 possible combinations of binary interactions between the two chiral alanine components and the solvent. One possible assumption is that the L and D-alanine – water interactions are identical because of the similarity of the compounds involved. Beyond that there are the heterochiral interactions – these are the interactions between L-alanine and D-alanine. If for a preliminary evaluation these are considered equal, asymmetry can be ignored, therefore this diagram can be extended to show the ternary diagram through the full range of L-alanine 100% EE to D-alanine 100% EE. This is shown below in Figure 7.

Using Figure 7, the eutectic point is equal for both D and L alanine and can have an initial evaluation made through a line of best fit through the given data points. The expected form of the line of best fit is determined by the typical solubility diagram of a racemic forming system. The eutectic point is shown to be at an EE of 53.3% for 10°C, 53.8% for 15°C and 54.2% at 20°C. With a range of 0.9%EE across 10°C it would be a reasonable to expect that a eutectic of approximately 54% is valid in a range of temperatures surrounding the investigated ones.

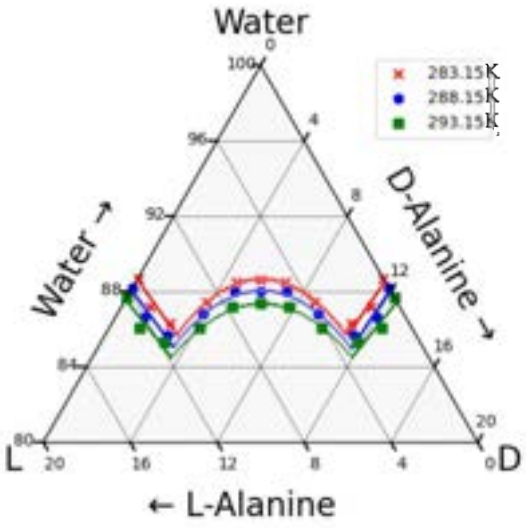


Figure 7: Ternary phase diagram of the alanine enantiomers in water at enantiomeric excess varying between 0% - 100% for three isotherms mirrored at the racemic and with lines of best fit.

The total data collection time is approximately halved using the assumptions above. If more accuracy is required, the use of the assumptions can act as a starting point to reduce the range in which investigation to find the eutectic needs to take place. It could also reduce the amount of the solute required by using values closer to the solubility limit while maintaining a saturated solution.

4.1.2 Enantiomeric Excess

As stated in the methodology the final determination of the enantiomeric excess, as graphed above, was based on the mass balances entering the system. However, this does not account for small amounts of mass being lost through cleaning of the rig which was required daily. The mass lost was minimised by using the densitometer to register the residence time and change between the recycling to waste and vice versa at a more accurate time. Whilst use of a check of enantiomeric excess was explored, the polarimeter in the rig setup was not used in the methodology because of the small optical rotation of alanine – 14 degrees (sigma-aldrich, n.d.). Hence, the polarimeter was not sensitive enough and would have needed a larger chamber to accurately determine the EE via optical rotation.

An alternative consideration was to use HPLC; this involved selection of solvents, IPA, methanol, ethanol, acetonitrile, and cyclohexane were tested. Unfortunately, solvents required in HPLC use is specific to the compounds and though data available directed these as possible options in binary combination with alanine when combined with both chiral alanine and water all solvents tested except cyclohexane acted as antisolvents. Hence, further investigation of HPLC during this experiment was hindered due to access to available solvents.

This does lead to the conclusion that the enantiomeric excess will have some error in the quoted values. However, throughout the project varying solubilities meant the current solution was above saturation at 55°C and to resolve this issue three solutions had to be made from scratch: 100% EE, 40% EE and 0% EE. With both the addition of new solutions and preferential crystallisation being neglected due to the fast crystallisation these points can be taken with higher degree of certainty; hence the lines of best fit and the eutectic point can be considered with more certainty.

4.1.3 Metastable Zone Width

The metastable zone width evaluation was carried out at two distinct points rather than evaluating it across the range at which the experiment was performed. The focus of this was to determine why the experiments nearer to the racemic composition were slower to crystallise than those at 100% EE. The two values found are shown in Table 2

From this it is clear to see that the metastable zone width is substantially larger for the DL-alanine than the pure L-alanine and hence was a determining factor in the decrease speed of the crystallisation.

Table 2: Values for the metastable zone width

Concentration (g/L)	Enantiomeric Excess (L-alanine) (%)	Temperature to begin crystallisation (°C)	Temperature to fully dissolve (°C)	MSZW (°C)
201.6	100	18.01	45.85	27.84
219.2	0	11.44	54.63	43.19

4.2 Modelling

In this section, the results of simulating the three-equation model introduced in the previous sequence will be discussed. Three specific scenarios were modelled and will be discussed here.

Table 3: Binary interaction parameters of L-Alanine and Water with the Parameters, Scenario 1 and Scenario 2 parameters respectively.

	Scenario 1	Scenario 2
g_{13}	-1.256•10^6 kJ/mol	-3.526 kJ/mol
g_{31}	8.669•10^5 kJ/mol	1.102•10^4 kJ/mol
α_{13}	0.004	0.1
g_{12}	-	2.786•10^4 kJ/mol
g_{21}	-	-1.004•10^4 kJ/mol
α_{12}	-	0.15

In the first scenario, the binary interaction parameters shown in Table 3 were established by fitting the experimental alanine data to the Schröder and van Laar equation, assuming hundred percent enantiomeric excess.

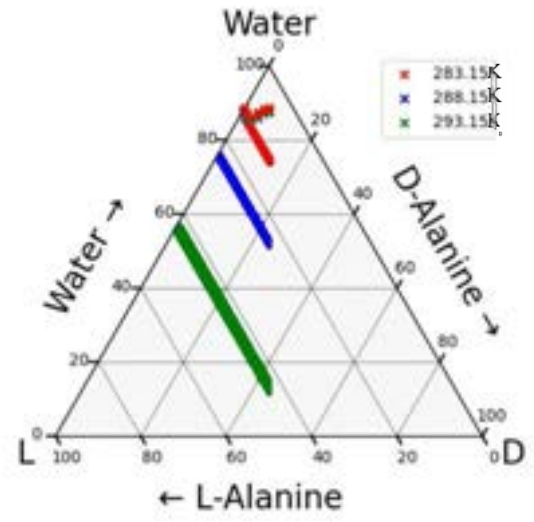


Figure 8: Predicted ternary phase diagram of the alanine enantiomers in water for scenario 1, according to the NRTL model and measurement data for three solubility isotherms.

Figure 8 shows the resulting ternary solubility diagram with the NRTL model - predictions. Clear deviation from the experimental values is shown. The 283.15K modelled isotherm shows good correlation on the left side of the triangle as the parameters were calculated using the L-alanine and water solubility data. Approaching the eutectic point however the model continues downwards. The parameters are at first trained without racemic contribution, hence a conglomerate forming system is expected as the initially presented result. The higher two isotherms at 288.15K and 293.15K do not resemble the experimental data, the shift in solubility is due to the initial high heat of fusion parameter for L-alanine used in the model. Comparisons of Equation 2 between L-alanine in water and literature data for L-Mandelic Acid in

Diethyl Tartrate (Tulashie *et al.*, 2010) show magnitude 10 difference in the activity coefficient needed to satisfy the system. The decomposition of amino acids at higher temperatures requires optimisation of the enthalpy of fusion. The isotherm starting solubility i.e., the binary solubility, is thereby better fitted to the experimental curve.

In the second scenario, the energy parameters g_{13}, g_{31} and α_{13} shown in Table 3 were ***Table 3*** were obtained from literature and were used to model the solubility. At an enthalpy of fusion of 75.33 kJ/mol, the NRTL model failed to create a model. Lowering the enthalpy to 10.96 kJ/mol showed a preliminary model. The updated enthalpy parameter did not provide a racemic fit, instead a conglomerate system formed.

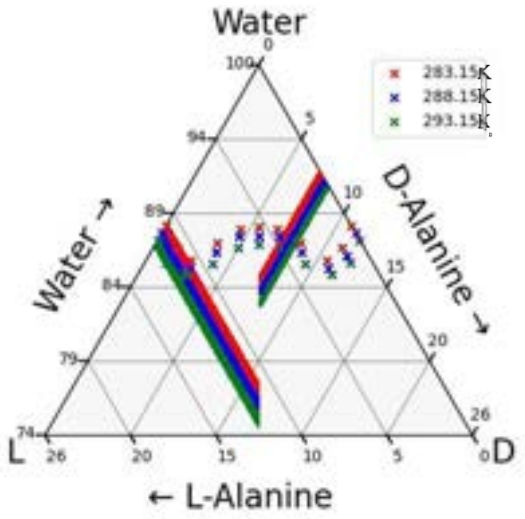


Figure 9: Predicted ternary phase diagram of the alanine enantiomers in water for scenario 2, according to the NRTL model and measurement data for three solubility isotherms.

The heterochiral interactions for alanine were calculated through the racemic solubility equation resulting in energy parameters g_{12}, g_{21} and α_{12} shown in Table 3. The objective function parametrization was only partially feasible. Figure 9 shows the heterochiral interaction by comparison of the left side with the isotherms starting at solubility values close to experimental and the right side with the isotherms shifted towards lower solubility. This is in direct contradiction with the symmetrical enantiomer assumption. Both sides should mirror at each isotherm. With these parameters the conglomerate system has different eutectic points at the same isotherm, indicating a discontinuity, making the model infeasible again.

For both Figure 8 and Figure 9 no correct fit was found with only the initial shape being satisfied around the L-enantiomer and water. Better model fit would be achieved by incorporating ternary data in parametrization at different enantiomeric excess using Equation 3 to model racemic compounds

Finally, in the third scenario, the experimental data and binary energy parameters for Mandelic Acid in Diethyl Tartrate and Ethyl Lactate as described by Tulashie et al. (2010) were replicated.

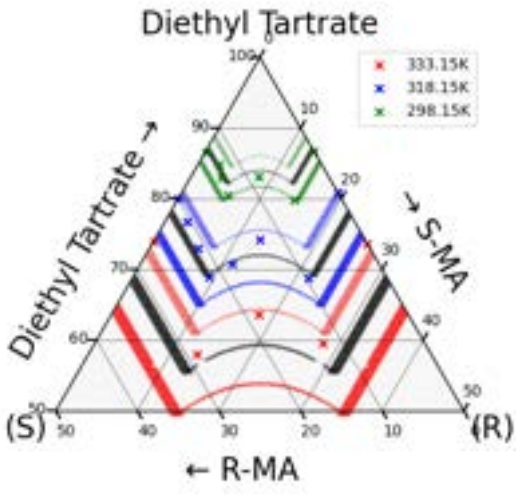


Figure 10: Predicted ternary phase diagram of the mandelic acid enantiomers in diethyl tartrate according to measurement data for three solubility isotherms and the NRTL model in black, colour isotherms above the black represent parameter decrease of 10% and colour isotherms under the black represent parameter increase of 10%.

The sensitivity of the binary parameters with Diethyl Tartrate were studied in Figure 10. The original output of the parameters in black decrease in accuracy with increasing temperature. Increasing the solvent – solute energy parameter interactions by 10% caused a marked increase in solubility reported, with the isotherms shifting downwards. The opposite happened with a decrease of 10% where the predictions made for lower solubility than experiments showed. Varying α_{ij} had no impact on model output, in line with it being the non-randomness parameter.

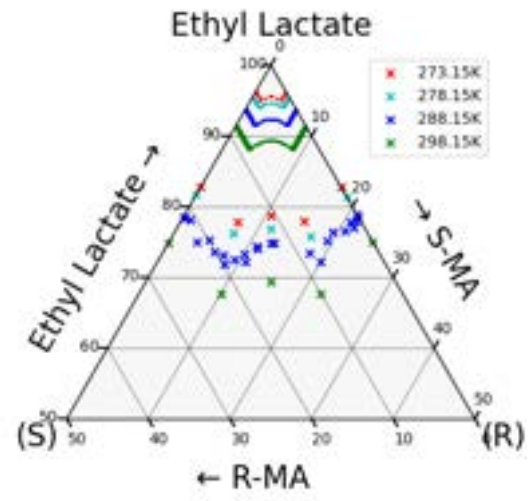


Figure 11: Predicted ternary phase diagram of the mandelic acid enantiomers in ethyl lactate according to the NRTL model and measurement data for four solubility isotherms.

When available, extracting parameters from literature would speed up the process of constructing ternary diagrams. Figure 11 however shows the discrepancy for Mandelic Acid in Ethyl Lactate between reported and modelled results. The calculated solubility is off by more than 25% at any point. This contradicts the correct implementation of the Diethyl Tartrate system. The equation implementation in Julia is therefore valid. The lack of good fit for alanine being due to unsuccessful objective function optimisation. No errors were discovered in the open-source code that could indicate the model being optimised to run Diethyl Tartrate correctly. The belief is therefore that the results may have been misreported in the final journal article.

5. Conclusion

Throughout the experimental and modelling procedure the need for improved techniques to build ternary diagrams has been clear both through the time taken experimentally and the failings of the model.

The experimental procedure and results have shown how the understanding of the metastable zone width could be used for more effective manipulation of temperatures for crystallisation speed and, if compound is limited, a reduction of excess compound beyond the upper solubility limit. Additionally, by choosing appropriate chiral compound behaviour assumptions, the time taken to build an experimental ternary diagram can be significantly reduced (approximately halved) by using a line of symmetry at 0% EE.

The ternary diagram built has determined alanine is a racemic forming compound, with a eutectic at approximately 54% enantiomeric excess and hence could be used to determine the possible preferential crystallisation routes for alanine in future work.

Modelling the behaviour of alanine using the NRTL model has proven challenging. The racemic compound forming shown from solubility data was not replicated in the model. Therefore, more investigation is needed in the behaviour of the model around the eutectic point with the second scenario already providing a better fit. The importance of the calorimetric properties was confirmed by the difference in energy parameters between scenario 1 and 2. The sensitivity of the literature parameters showed robustness for Mandelic Acid in Diethyl Tartrate in contrary to Mandelic Acid in Ethyl Lactate which failed to approach the experimental data.

6. Outlook

The outlook requires consideration of points discussed throughout the report and new options exploring. Further work would have a focus directed towards outlining a fast and accurate method to determine solubility with a modelling method less reliant on ternary experimental data. It would be insightful to see how the model performs in regions which have not been used in training the energy parameters and hence see how the model extends beyond its fed data. This would require more experimental data to be available. These results could be compared between NRTL models and other models such as UNIQUAC or COSMO-RS not explored in this paper.

Another consideration which is, as of current, limited by the lack of experimental ternary data available is to use machine learning on large samples of ternary data and evaluate the output. This could be through either reinforcement learning techniques or alternatively through a level of supervised/unsupervised machine learning.

Experimentally, the significance of the time taken for crystallisation and the impact of the metastable zone width have highlighted routes for improvement. Consideration towards how antisolvents/additives could be used in systems to manipulate the solubility limits particularly for application in preferential crystallisation would be useful. This was particularly highlighted by the antisolvent response to HPLC solvents tested.

7. References

Cascella, F., Seidel-Morgenstern, A. & Lorenz, H. (2020) Exploiting Ternary Solubility Phase Diagrams for Resolution of Enantiomers: An Instructive Example. *Chemical Engineering & Technology*. 43 (2), 329–336 Available from: doi:10.1002/ceat.201900421.

Chua, Y.Z., Do, H.T., Schick, C., Zaitsau, D., et al. (2018) New experimental melting properties as access for predicting amino-acid solubility. *RSC Advances*. [Online] 8 (12), 6365–6372. Available from: doi:10.1039/c8ra00334c.

Dannenfelser, RM & Yalkowsky, SH 1991, 'Data base of aqueous solubility for organic non-electrolytes', Science of the Total Environment, The, vol. 109-110, no. C, pp. 625-628. Available at: https://doi.org/10.1016/0048-9697(91)90214-Y

FDA (1992) *Development of new stereoisomeric drugs*, *U.S. Food and Drug Administration*. FDA. Available at: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/development-new-stereoisomeric-drugs (Accessed: December 5, 2022.

Gänsch, J., Huskova, N., Kerst, K., Temmel, E., Lorenz, H., Mangold, M., Janiga, G. & Seidel-Morgenstern, A. (2021) Continuous enantioselective crystallization of chiral compounds in coupled fluidized beds. *Chemical Engineering Journal*. 422, 129627. Available from: doi: 10.1016/j.cej.2021.129627.

GVR, G.V.R. (2022) *Chiral chemicals market size, share: Industry Report, 2024*, *Chiral Chemicals Market Size, Share | Industry Report, 2024*. Available at: https://www.grandviewresearch.com/industry-analysis/chiral-chemicals-market (Accessed: December 2, 2022).

Haida, H., Kaemmerer, H., Lorenz, H. & Seidel-Morgenstern, A. (2010) Estimation of reliable parameters for solid-liquid equilibrium description of chiral systems. *Chemical Engineering & Technology*. [Online] 33 (5), 767–774. Available at: doi:10.1002/ceat.200900612.

Kaemmerer, H., Seidel-Morgenstern, A. & Lorenz, H. (2013) Chiral separation of systems of high eutectic composition by a combined process: Case study of serine enantiomers. *Chemical Engineering and Processing: Process Intensification*. [Online] 6771–79. Available from: doi: 10.1016/j.cep.2012.11.002.

Lorenz, H., Sapoundjiev, D. & Seidel-Morgenstern, A. (2003) Solubility equilibria in chiral systems and their importance for enantioseparation. *Engineering in Life Sciences*. [Online] 3 (3), 132–136. Available from: doi:10.1002/elsc.200390016.

Nguyen, L.A., He, H. and Pham-Huy, C. (2006) 'Chiral Drugs: An Overview', International Journal of Biomedical Science : IJBS, 2(2), pp. 85–100.

N.M.S.C. (2022) *Chiral chemicals market by separation technique (high-performance liquid chromatography (HPLC), ultra-high performance liquid chromatography (UHPLC), Supercritical Fluid Chromatography (SFC), and simulated moving bed (SMB)), by technology (traditional separation method, asymmetric preparation method, biological separation method, and others), and by application (pharmaceuticals, Agrochemicals, Food & Beverage, cosmetics, and others) – global opportunity analysis and industry forecast 2022-2030*, *Chiral Chemicals Market Size, Share, Forecast, Industry Analysis Report | 2022 - 2030*. Available at: https://www.nextmsc.com/report/chiral-chemicals-market#:~:text=The%20Chiral%20Chemicals%20Market%20size,atom%20in%20the%20molecule's%20center (Accessed: December 2, 2022).

O'Brien, C.D. (no date) *'Enantioselective Crystallisation of Racemic Forming Compounds in Continuous and Batch Processes'* Unpublished

Prigogine, I. & Defay, R. (1973) *Chemical thermodynamics*. London, Longman.

Shaffer, C. (2022) *Chirality in biochemistry*, *News-Medical.net*. Available at: https://www.azolifesciences.com/article/Chirality-in-Biochemistry.aspx (Accessed: October 25, 2022).

Sigma-Aldrich (n.d.) *A7627pis - sigma-aldrich*, *sigma-aldrich*. sigma-aldrich. Available at: https://www.sigmaaldrich.com/deepweb/assets/sigmaaldrich/product/documents/398/268/a7627pis.pdf (Accessed: October 15, 2022).

Tulashie, S.K., Kaemmerer, H., Lorenz, H. & Seidel-Morgenstern, A. (2009) Solid−liquid equilibria of mandelic acid enantiomers in two chiral solvents: Experimental determination and model correlation. *Journal of Chemical & Engineering Data*. [Online] 55 (1), 333–340. Available from: doi:10.1021/je900353b.

VAISALA (no date) *Pharmaceutical crystallization*, *Vaisala*. Available at: https://www.vaisala.com/en/industries-applications/life-science/pharmaceutical-drug-manufacturing-and-biotechnology-processing/pharmaceutical-crystallization (Accessed: December 10, 2022.

Wang, Y. et al. (2005) 'Eutectic Composition of a Chiral Mixture Containing a Racemic Compound', Organic Process Research & Development, 9(5), pp. 670–676. Available at: https://doi.org/10.1021/op0501038.

Williams, A. (1996) 'Opportunities for chiral agrochemicals', Pesticide Science, 46(1), pp. 3–9. Available at: https://doi.org/10.1002/(SICI)1096-9063(199601)46:1<3::AID-PS337>3.0.CO;2-J

Zhang, J.D., Mohibul Kabir, K.M. and Donald, W.A. (2019) 'Chapter Three - Ion-Mobility Mass Spectrometry for Chiral Analysis of Small Molecules', in W.A. Donald and J.S. Prell (eds) Comprehensive Analytical Chemistry. Elsevier (Advances in Ion Mobility-Mass Spectrometry: Fundamentals, Instrumentation and Applications), pp. 51–81. Available at: https://doi.org/10.1016/bs.coac.2018.08.009.

IMPERIAL COLLEGE LONDON

DEPARTMENT OF CHEMICAL ENGINEERING

ADVANCED CHEMICAL ENGINEERING PRACTICE

Recyclability of base catalysts for the production of biodiesel from rapeseed oil

Authors: Natalia Arapoglou, Naomi Lim

Abstract

The ability of a catalyst to be recycled without noticeably losing its catalytic activity or efficiency is a highly desirable component of every catalytic process. This paper discusses the recyclability of two different base catalysts, sodium methoxide (CH_3ONa) and sodium hydroxide (NaOH), in the transesterification of rapeseed oil (RO) with methanol to produce biodiesel. The key findings of this study are that the catalysts are not destroyed in reaction, and that both catalysts are recyclable for 2 times using the approach detailed in this paper. The glycerol phase containing all the recovered catalyst is utilised in the subsequent recycle while keeping the oil:methanol:catalyst ratio constant for every recycle. The mass of catalyst recovered is used as the reference value to calculate the scaled-down masses of RO and methanol. Results indicate that although recyclability of the base catalyst via the method detailed in this paper is possible, it is not favourable. This is due to the glycerol buildup unexpectedly competing with methanol to react with some of the RO, forming monoglycerides and diglycerides. As this study proves that the base catalysts are not destroyed in reaction and are hence recyclable, further work can be done to examine other methods for recycling the catalyst that do not involve the buildup of glycerol in order to avoid glycerol interference with the transesterification process.

Keywords: Biodiesel, Base Catalysts, Recyclability, Glycerol

1 Introduction

Biodiesel is a widely used, renewable, clean biofuel which can be produced from vegetable oils[1], animal fats, waste cooking oil and microalgal oil[2] in a transesterification reaction with alcohols such as methanol and ethanol[3]. The worldwide biofuel market was estimated to be worth $ 131.85 billion in 2021, and is projected to grow to $ 331.89 billion by 2030, with a Compound Annual Growth Rate(CAGR) of 11.9% from 2022 to 2030[4]. With a 28.34% share in 2021, vegetable oil dominated the worldwide biofuel market.[5] Biodiesel is set to become well-established in the coming decades as a measure to reduce greenhouse emissions after the Paris Agreement.

The most popular method for producing biodiesel, or Fatty Acid Methyl Esters (FAME), is via transesterification of fats and oils with excess methanol in the presence of a catalyst:

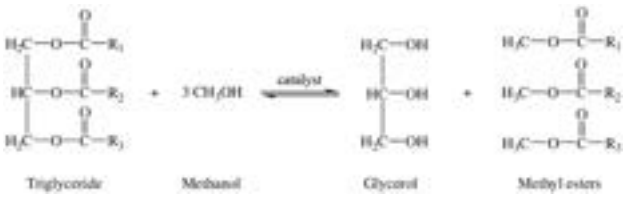


Fig. 1 Transesterification of triglyceride with methanol in the presence of a catalyst to produce methyl esters and glycerol

Triglycerides are transesterified batch-wise or continuously in multistep reactors at atmospheric pressure and a

temperature of roughly 60-70°C. At the conclusion of the reaction, the mixture is given time to settle. The upper methyl ester layer is washed to remove impurities while the lower glycerol layer is pulled off.[6]

An equilibrium between the reactants is what defines the transesterification reaction. 1 mole of triglyceride reacts with 3 moles of alcohol based on stoichiometry in a sequence of three reversible reactions which have intermediate products of monoglycerides and diglycerides. To achieve larger yields of the products and improve phase separation between the produced esters and glycerol, an excess of alcohol is typically utilized in the process for a high level of triglyceride to ester conversion. The unreacted transesterification agent is then taken out of the mixture.[7]

Both basic and acidic catalysts have been widely studied for their efficiency in transesterification of vegetable oils in order to create FAMEs. In early reports, homogeneous liquid acids were widely used as acidic catalysts[8]. Later, numerous heterogeneous acidic catalysts, such as sulfated silica, cationic exchange resin, zeolites, sulfonic acid-functionalized carbon materials, heteropoly acids, and others, were also examined for transesterification. Acid catalysts are advantageous in that they do not result in saponification, however the rates of reaction associated with the use of acid catalysts are too slow. Base catalysts, which demonstrate remarkable effectiveness as well as a higher rate of reaction for the transesterification process, have been widely studied for their efficiency in transesterification reactions. However, little research has been done to understand the recyclability of the base catalysts. This aspect of the process is especially important and worth looking into because catalyst recyclability reduces carbon emissions and resources associated with extracting, refining, transporting, and processing catalysts, which ultimately contribute to economical and environmental impacts.

This study examines the recyclability of homogenous base catalysts. Given the scarcity of research done on the recycling process using base catalysts, this paper presents a preliminary glimpse into this novel development that could be further looked into. In this study, the glycerol phase containing all the recovered catalyst is brought forward to the subsequent recycle. To keep the oil:methanol:catalyst ratio constant for every recycle, the mass of catalyst recovered is used as the reference value to calculate the scaled-down masses of RO and methanol.

2 Materials and methods

2.1 Experimental setup

Two base catalysts, sodium methoxide (CH_3ONa) and sodium hydroxide (NaOH), were examined for their recyclability. Experiments for both catalysts were performed using rapeseed oil (RO) and methanol for the initial reactions. Biodiesel and glycerol were synthesized in house via transesterification, and the products were left to separate overnight. The produced glycerol phase, which contained all the recovered catalyst and some methanol, was brought forward to the subsequent recycles. This was repeated until no more pure biodiesel was obtained.

A 500 mL two-neck round-bottom flask was immersed in a water bath, heated, and stirred using an IKA Magnetic Stirrer RH digital. One end had an ETS-D5 electronic thermometer, connected to a needle septum stopper, to ensure temperature control in the mixture. On the other end was a serpentine condenser, capped with a glass stopper that was used to prevent the loss of methanol through evaporation, as seen in figure 2.

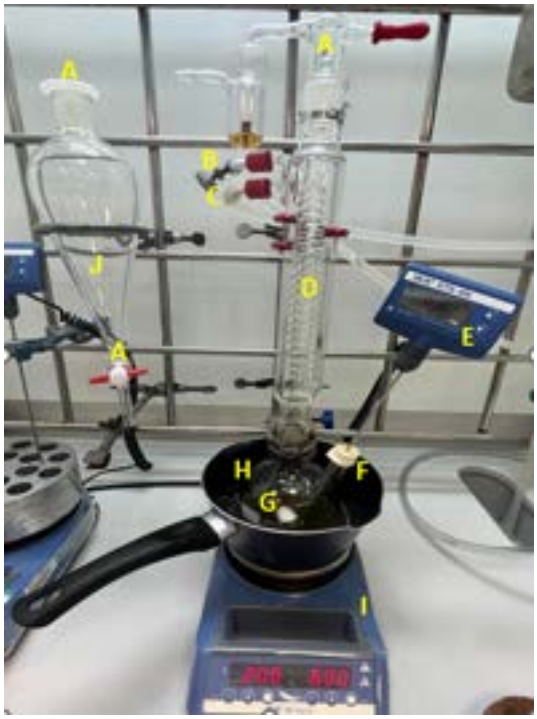


Fig. 2 Experimental setup showing A) stopper, B) water outlet, C) water inlet, D) serpentine condenser, E) thermometer, F) needle septum, G) magnetic bar, H) two-neck round-bottom flask, I) magnetic stirrer, J) separatory funnel

2.2 Reaction conditions

For this experiment, edible RO was selected as feedstock for biodiesel synthesis as it is a common vegetable oil that is liquid at room temperature and hence easy to measure and handle. To prevent extraneous factors, such as the levels of water and free fatty acids (FFA), from influencing the results, oil from the same batch was utilised. The RO was heated until the desired temperature of 63°C was reached, after which the methanol and catalyst was added. The reaction temperature was kept constant at 63°C to maximise the oil conversion while keeping the reaction below the boiling point of methanol (64.7°C)[9]. The stirring rate was kept at 600 RPM to ensure sufficiently mixing of the reaction mixture. Methanol:oil ratio was kept at 6:1, which is twice the 3:1 stoichiometric ratio, so that methanol was in excess to faciliate complete conversion. Catalyst loading of 1.12 wt% and 1.00 wt% was used for CH_3ONa and NaOH respectively.[10]

2.3 Standardisation of solutions for titration

To prepare an accurate concentration of NaOH, standardisation was done by autotitrating against a known amount of potassium hydrogen phthalate (KHP) with three drops of phenolphthalein. Three samples of KHP were titrated and the results were averaged to find the correct concentration using:

$$C_{NaOH} = \frac{m_{KHP}}{M_{KHP} V_{NaOH} 0.001}, \tag{1}$$

where m_{KHP} is the mass of KHP, M_{KHP} is the molar mass of KHP (204.22 g/mol), and V_{NaOH} is the volume of NaOH used in mL until the solution turned pink. The concentration of NaOH was calculated to be 0.0186M.

This concentration was then used to standardise the HCl solution based on:

$$C_{HCl} = \frac{V_{NaOH} C_{NaOH}}{V_{HCl}}, \tag{2}$$

where V_{NaOH} is the volume of NaOH in mL used until the solution turned pink with V_{HCl}, which was a known volume of 5 mL of HCl. The concentration of HCl was calculated to be 0.0272M.

2.4 Calculation for mass of base catalyst recovered

After separation of the transesterification products, titration was carried out in order to determine the amount of base recovered in each phase.

For titration of each phase, a known volume (5mL for biodiesel titration and 10mL for glycerol titration) of 0.0272 mol/L HCl was pipetted into a 100mL container of a known mass of sample (within the range of 0.1g - 0.5g). The HCl reacted with all of the base in the sample (if any), and the remaining unreacted HCl was titrated against 0.0186 mol/L of NaOH. Using the volume of NaOH titrated, the concentration of base in each phase could be calculated as follows:

$$n_{base} = \frac{n_{HCl} - n_{NaOH}}{m_{sample}} m_{phase} \tag{3}$$

where n_{base} is the moles of base recovered, m_{sample} is the mass of biodiesel sample or glycerol sample, m_{phase} is the mass of the biodiesel phase or glycerol phase, and n_{HCl} and n_{NaOH} are the moles of HCl and NaOH respectively, calculated as:

$$n_x = C_x V_x, \tag{4}$$

where x is either HCl or NaOH, C_x is the concentration of x in mol/L, and V_x is the volume of x in L. The mass of base can then be calculated as follows:

$$m_{base} = n_{base} M_{base}, \tag{5}$$

where n_{base} is the moles of base in each phase and M_{base} is the molar mass of the base used for each experiment. To reduce the likelihood of errors or anomalous results, at least 3 samples were taken from each phase and the results were averaged. This was done for every recycle.

The mass of base recovered was used as the reference value to calculate the scaled-down masses of RO and methanol for the subsequent recycle.

2.5 ^{1}H NMR spectroscopy

One drop of glycerol in 0.5 mL of deuterium oxide (D_2O) and one drop of biodiesel in 0.5 ml of chloroform ($CDCl_3$) were tested with NMR spectroscopy. NMR peaks were analysed to obtain i) the mass of methanol in the glycerol phase as detailed in section 2.6, and i) the biodiesel content in the upper phase as detailed in section 2.8.

2.6 Calculation for mass of methanol

The total methanol for all recycles was made up of i) fresh methanol, and ii) the existing methanol in the glycerol phase from the previous cycle. By analyzing the NMR spectrum of the glycerol phase, the mass of methanol in the glycerol phase was first calculated. This mass was then deducted from the calculated mass of methanol needed for the recycle reaction, to give the amount of fresh methanol needed.

As seen in Fig. 3 the signals corresponding to the

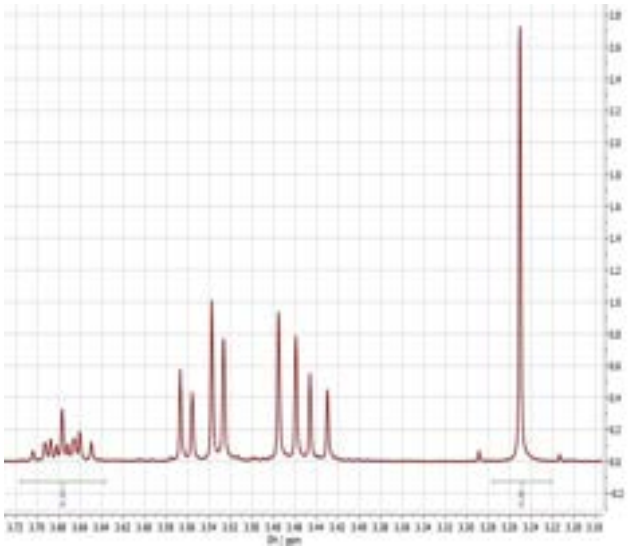


Fig. 3 H-NMR spectrum of glycerol phase, 1st recycle, CH_3ONa catalyst. Integrations of the signals around 3.25ppm and 3.68ppm show values of 0.25 and 0.14 respectively.

methyl protons of methanol (around 3.25ppm) and the proton of the centre carbon of glycerol (around 3.68ppm) were first integrated. The percentage of methanol in the glycerol phase was then calculated by comparing the areas associated with the protons of methanol and glycerol. In a case where methanol and glycerol are equimolar, the ratio of the areas of the 3H/1H signals would be 50%. Hence, the following can be derived:

$$\%_{MeOH} = \frac{\frac{area_{3H}}{area_{1H}}}{3} \times 50\% \quad (6)$$

where $\%_{MeOH}$ is the percentage of methanol in the glycerol phase, $area_{3H}$ is the area of the 3H signal, and $area_{1H}$ is the area of the 1H signal.

By using the mass of the glycerol phase and the molar masses of methanol and glycerol, the mass of methanol in the glycerol phase can be obtained as follows:

$$m_{MeOH} = \frac{m_{glycerol\,phase} \times (\%_{MeOH} M_{MeOH})}{\%_{MeOH} M_{MeOH} + (100\% - \%_{MeOH}) M_{glycerol}} \quad (7)$$

where m_{MeOH} is the mass of methanol in the glycerol phase, $m_{glycerol\,phase}$ is the mass of the glycerol phase, $\%_{MeOH}$ is the percentage of methanol in the glycerol phase, M_{MeOH} is the molar mass of methanol (32.04 g/mol), and $M_{glycerol}$ is the molar mass of glycerol (92.09 g/mol). This calculation method was applied for all recycles of the reactions for both catalysts.

2.7 Washing and drying of biodiesel

The biodiesel phase was washed repeatedly with a 10% brine solution until a pH of 7 was obtained in order to remove any methanol and impurities left in the mixture. Anhydrous sodium sulphate ($NaSO_4$) was then used to dry the mixture overnight to ensure all the water was removed. $NaSO_4$ was chosen over magnesium sulphate (Mg_2SO_4) as a drying agent due to better results. A 0.8mm hypodermic needle was then used to extract the purified biodiesel.

2.8 Calculation for mass of biodiesel obtained

In the chemical structure of a FAME molecule, The OCH_3 group is highly specific to the ester, whereas the CH_2 group is common in both the triglycerides and the ester. By analyzing the NMR spectrum of the biodiesel phase and comparing the areas of the signals associated with the protons of these groups, the mass of biodiesel obtained can be calculated as detailed below.

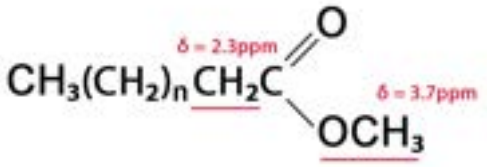


Fig. 4 Chemical structure of a FAME molecule. The methoxy (OCH_3) protons at the end of the molecule are associated with a H-NMR signal of around 3.7ppm and the methylene (CH_2) protons are associated with a H-NMR signal of around 2.3ppm

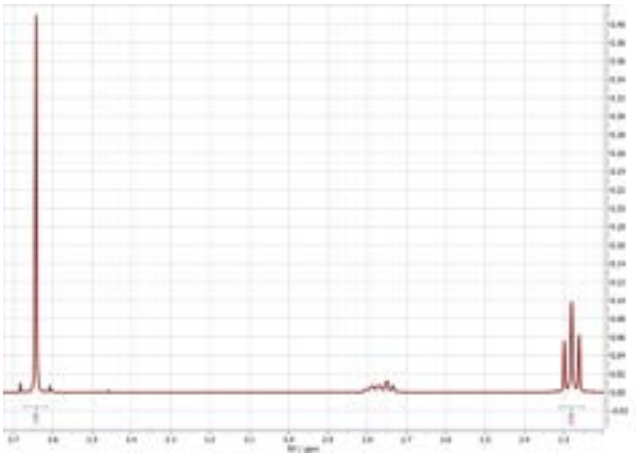


Fig. 5 H-NMR spectrum of biodiesel phase, original reaction, CH_3ONa catalyst. Integrations of the signals around 2.3ppm and 3.7ppm show values of 0.68 and 1.00 respectively.

As seen in Fig. 5, the signal around 3.7ppm corresponding to the methoxy (OCH_3) protons and the signal

around 2.3ppm corresponding to the methylene (CH_2) protons were integrated. The percentage of biodiesel content can then be calculated by normalising the values of integration according to the number of protons contributing to each peak as follows:

$$\%_{biodiesel\ content} = \frac{2area_{methoxy}}{3area_{methylene}} \times 100\% \quad (8)$$

where $\%_{biodiesel\ content}$ is the percentage of biodiesel content in the upper phase, $area_{methoxy}$ is the area of the OCH_3 signal, and $area_{methylene}$ is the area of the CH_2 signal.

The mass of the upper phase after washing and drying was also weighed and recorded. The mass of pure biodiesel in the upper phase was thus calculated as follows:

$$m_{biodiesel} = \%_{biodiesel\ content} \times m_{upper\ phase} \quad (9)$$

where $m_{biodiesel}$ is the mass of pure biodiesel in the upper phase, $\%_{biodiesel\ content}$ is the percentage of biodiesel content in the upper phase, and $m_{upper\ phase}$ is the mass of the upper phase. This calculation method was applied for all recycles of the reactions for both catalysts.

3 Results and Discussion

3.1 Mass of catalyst recovered

3.1.1 Biodiesel phase

From titration of the biodiesel phase, it was found that the concentration of base in the biodiesel phase was 0 for all reactions for both catalysts. This indicated that the base was not in biodiesel phase, and that all the base was in the glycerol phase.

3.1.2 Glycerol phase

From titration of the glycerol phase, it was found that base catalyst was present in the glycerol phase for all reactions for both catalysts.

As seen in Fig. 6, there was a general decrease in the mass of catalyst present in glycerol at the end of every recycle. For CH_3ONa, 66.8% of the base was recovered from the initial reaction, followed by a 69.5% recovery in the 1st recycle, a 70.0% recovery in the 2nd recycle, and a 81.5% recovery in the 3rd recycle. For NaOH, 71.0% of the base was recovered from the initial reaction, followed by a 58.9% recovery in the 1st recycle, a 78.4% recovery in the 2nd recycle, and a 66.0% recovery in the 3rd recycle. The portion of base lost is attributed to saponification.

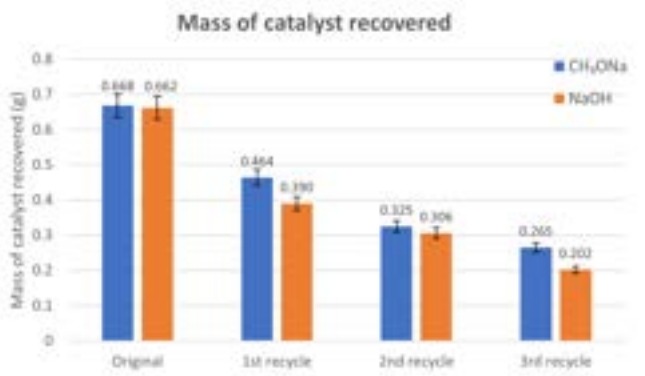


Fig. 6 Mass of base catalyst in the glycerol phase at the end of every cycle, for CH_3ONa and NaOH

These results showed that the base catalyst is not destroyed in the reaction, and can hence be reused as catalyst for subsequent reactions.

3.2 Mass of glycerol obtained

The masses of the glycerol phase after every cycle was weighed. This mass was made up of glycerol and methanol. By deducting the mass of methanol (calculated using the methodology detailed in section 2.6), the mass of glycerol was obtained as follows:

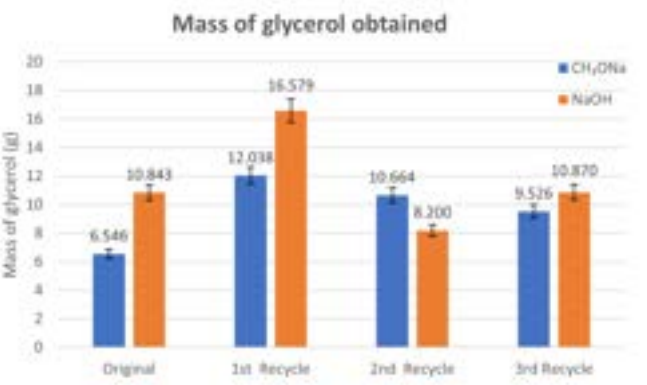


Fig. 7 Mass of glycerol obtained at the end of every cycle, for reactions catalysed by CH_3ONa or NaOH

In principle, it would be expected that the mass of glycerol increases by recycle, due to new amounts of glycerol produced via transesterification. However, the results showed that there was an increase in mass of glycerol only up until the 1st recycle. The mass of glycerol decreased after the 2nd recycle and showed no clear trend, as seen in Fig. 7.

This was because glycerol, instead of methanol, had unexpectedly reacted with some of the triglycerides (to form monoglycerides and diglycerides). Due to the structure of this research that necessitated all of the recovered

base to be used in the subsequent recycle, the amount of glycerol added to each recycle was not limited, and thus eventually built up to be in excess. The glycerol reacting with the RO had been unanticipated as glycerol is insoluble in vegetable oil, and the methanol had been used in an excess 6:1 ratio. However, it was found that when the reaction was in equilibrium, some of the glycerol had competed with methanol to react with some of the RO in a side reaction.

Fig. 8 Glycerolysis of triglycerides to form monolycerides (MAG) and diglycerides (DAG)

To validate that glycerol had reacted with the triglycerides to form mono- and diglycerides, the H-NMR spectrum of biodiesel was analysed for signs of mono- and diglycerides. This was done by looking for the signal for glycerol, as the signals for mono- and diglycerides would be complex signals in the region similar to that of glycerol. This was found to be around 4.1ppm to 4.3ppm.

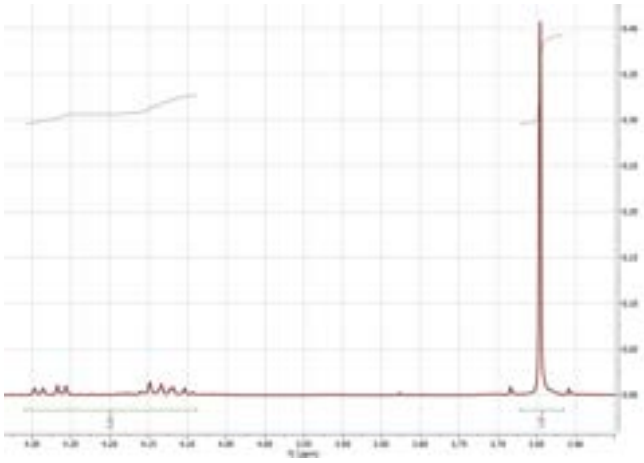


Fig. 9 NMR spectrum of biodiesel, 2nd recycle, CH_3ONa catalyst. Integrations of the signals around 3.7ppm and 4.2ppm show values of 1.00 and 0.32 respectively.

As seen in Fig. 9, the signal around 3.7ppm corresponding to the OCH_3 protons of the biodiesel, as well as the signal around 4.1 to 4.3ppm corresponding to mono- and diglycerides, were integrated. The percentage of mono- and diglycerides can then be calculated by comparing the areas of these selected signals.

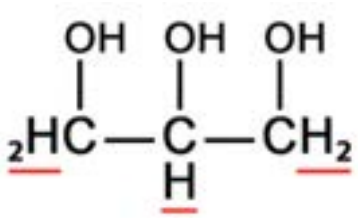


Fig. 10 Chemical structure of a glycerol molecule showing the 5 protons that contribute to the signal

As mentioned above, the signals for mono- and diglycerides is taken to be similar to that of glycerol. In the chemical structure of glycerol, there are 5 protons contributing to the signal, whereas in the -OCH_3 group specific to the biodiesel, there are 3 contributing protons. By normalising the values of integration according to number of protons contributing to each signal, the percentage of mono- and diglycerides in biodiesel can be obtained:

$$\%_{\text{MAG and DAG}} = \frac{\frac{A_1}{5}}{\frac{A_1}{5} + \frac{A_2}{3}} \times 100\% \qquad (10)$$

where $\%_{\text{MAG and DAG}}$ is the percentage of mono- and diglycerides in the biodiesel, A_1 is the area of the mono- and diglycerides signal, and A_2 is the area of the OCH_3 peak.

Results demonstrated that the amount of mono- and diglycerides increased in each consequent recycle, as seen in Fig. 11. . This further validated that glycerol had reacted with the triglycerides in a side reaction to form monoglycerides and diglycerides.

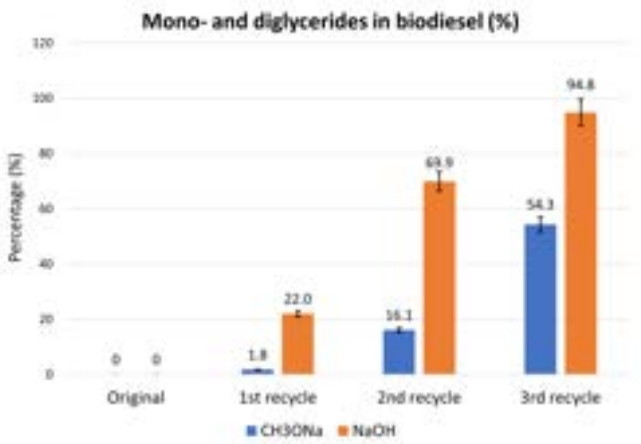


Fig. 11 Percentage of mono- and diglycerides in the biodiesel, for reactions catalysed by CH_3ONa or NaOH

3.3 Mass of biodiesel obtained

From the NMR results and using equation 8 as detailed in section 2.8, the content of biodiesel in the upper phase was found to be as follows:

Table 1 Percentage of biodiesel content in the upper phase, for reactions using CH_3ONa or NaOH as catalyst

	Percentage of biodiesel content (%)	
	CH_3ONa	NaOH
original	98.0	98.0
1st recycle	93.9	42.6
2nd recycle	69.4	10.2

The mass of the upper phase after washing and drying was also weighed and recorded as follows:

Table 2 Mass of the upper phase after washing and drying, for reactions using CH_3ONa or NaOH as catalyst

	Mass of upper phase (g)	
	CH_3ONa	NaOH
original	45.2447	56.6859
1st recycle	19.6184	26.4604
2nd recycle	3.441	14.6892

Using equation 9 from section 2.8, the mass of pure biodiesel in the upper phase was thus found to be as follows:

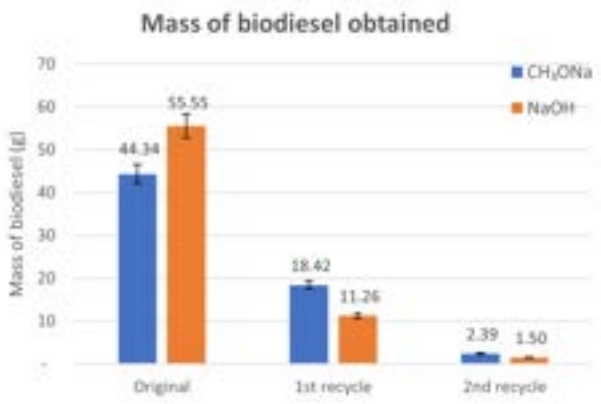


Fig. 12 Mass of pure biodiesel obtained, for reactions using CH_3ONa or NaOH as catalyst

As seen in Fig. 12, there is a general decrease in the mass of pure biodiesel obtained after every subsequent recycle. In the initial reaction, the mass of biodiesel obtained was higher when catalysed by NaOH compared to that of CH_3ONa; for the recycles, the mass of biodiesel obtained was higher when catalysed by CH_3ONa. In the 3rd recycle, no pure biodiesel was obtained.

3.3.1 Comparison with theoretical values

To keep the oil:methanol:catalyst ratio constant for every recycle, the oil input for every subsequent recycle was scaled down according to the recovered catalyst mass. The mass of triglycerides available for transesterification was thus decreased by every recycle. Hence, one might wonder whether the general decrease in the mass of biodiesel obtained was caused by the decreasing oil input. To accurately examine this, the experimental mass of biodiesel obtained was compared with the theoretical mass of biodiesel expected from stoichiometry:

$$\%yield = \frac{\text{mass of biodiesel obtained}}{\text{mass of biodiesel expected}} \times 100\% \qquad (11)$$

The results obtained were then plotted. As seen in Fig. 13, there is a general decrease in the percentage yield as the recycle goes on, so the amount of pure biodiesel obtained does decreases with every recycle.

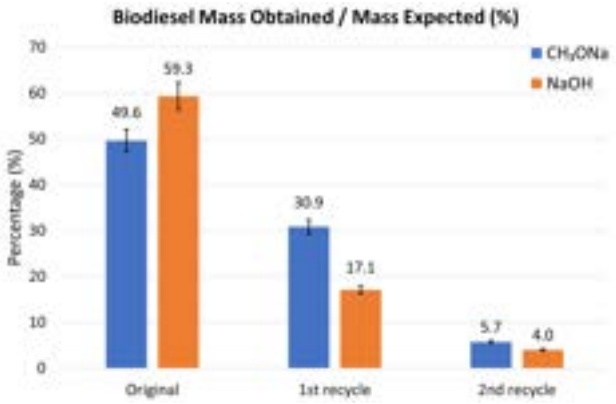


Fig. 13 Mass of biodiesel obtained/mass of biodiesel expected

The highest yield obtained in this study was 59.3% when catalysed by NaOH and 49.6% when catalysed by CH_3ONa. This was lower than expected, as a significant amount of biodiesel was lost during washing due to limitations in the methods of purification available for this study. The reason for the low yield is also partly attributed to the phenomenon of glycerol competing with methanol as detailed in section 3.2.

3.4 Additional approaches

Other approaches of carrying out the transesterification reaction were also investigated.

3.4.1 Amberlyst-A21 as catalyst

A separate reaction was conducted to test the feasibility of the weak base Amberlyst-A21 resin as a heterogenous catalyst. This was examined because in principle, the

Amberlyst-A21 would be very easy to recycle just by filtration as it is a solid catalyst.

To keep the base concentration similar to that of the reactions using CH_3ONa and NaOH as catalyst, the capacity of the Amberlyst-A21 (1.3 meq/mL by wetted bed volume) was used to calculate the amount of resin required. Other than the change in catalyst, all other factors of the reaction were kept constant. Compared to CH_3ONa and NaOH, the Amberlyst-A21 was much more difficult to handle, especially during transfer, as the resin stuck to the walls of the containers.

Unfortunately, the Amberlyst-A21 did not present any activity for this reaction. There was no production of glycerol, and no transeterification results were obtained.

3.4.2 No stirring or heating

To investigate the feasibility of the reaction by diffusion alone without energy input, the reaction using CH_3ONa was replicated with all factors kept constant; the only difference was that there was no stirring or heating involved.

Results indicate that initially, the reaction was slow and showed an indistinguishable mixture. After 10 days, a distinct, separate glycerol layer was observed. The yield of pure biodiesel obtained was found to be the same as that obtained when stirring and heating was involved. This showed that the reaction is feasible without energy input, albeit with a longer lag period.

An advantage in this approach was that the equilibrium was shifted by the phase separation with methanol. Although the phases are immiscible, at the interface of phases where transesterification occurs, monoglycerides and diglycerides are formed which can carry the base and the methanol to the bulk of oil, starting the reaction. Hence no energy input was needed, although the time taken was much longer. Further research can be done to examine this aspect of the compromise between energy expended and time taken.

3.4.3 Days left to sit

To examine whether the biodiesel yield was influenced by the time period for which the mixture was left to sit, the reacted mixture was left to sit for 3 days instead of 1 day. This was carried out to investigate if more time would facilitate better separation between the biodiesel and the glycerol phase. The products were then separated using a separatory funnel. This was performed for both CH_3ONa and NaOH, with all other factors kept constant.

There was no significant difference in the biodiesel yield or the amount of base recovered, compared to the reactions with 1 day of sitting. This demonstrated that the time for which the reacted mixture is left to sit does not affect the experimental results.

4 Conclusion

This research demonstrated that the base catalysts CH_3ONa and NaOH are not destroyed in reaction, and that they are recyclable for 2 times. Although recyclability of the base catalyst via the method detailed in this paper is possible, it is not favourable due to the excess glycerol competing with methanol to react with some of the triglycerides. This lead to the formation of mono- and diglycerides, which are not the products this study focuses on. The yield of biodiesel was lower than expected as a significant amount of the mass was lost in purification steps. Under more ideal purification methods, it is expected that the mass of pure biodiesel obtained would be higher than the values discussed in this paper. The low yield is also partly attributed to the phenomenon of glycerol interfering with the transesterification process.

To test for more environmentally friendly and economically viable options, this research proved that the transesterification reaction can be achieved solely by diffusion, without energy input. The compromise between energy expended and time taken is an development worth looking into. A separate reaction using the resin Amberlyst-A21 unfortunately showed no activity for this reaction. It was also found that the time left for the reaction mixture to sit did not affect biodiesel yield or catalyst recyclability.

5 Limitations and Outlook

Over the course of the project, several issues linked to the method used were identified. Given the time constraints imposed, these were not acted upon but only discussed as below.

Due to time and equipment restrictions, no pretreatment of the oil was done, resulting in high content of free fatty acids in the solution, which in turn increased the possibility of secondary reactions during transesterification.[11] The yield of pure biodiesel was also limited by the method used for the washing of biodiesel, resulting in considerable loss of the desired product.

For future work, it would be advisable to select a way to pre-treat the oil before the reaction. More optimal methods of purification could also be looked into. Alternatively, the focus could also be put on the formation of MAG and DAG as they could also be desired products

due to their applications. MAG and DAG are widely used in food, pharmaceutical and cosmetics as they enhance emulsion stability when mixed types are employed, stabilize ingredients, prevent separation, improve food texture, and lengthen product shelf life.[12]The global market was worth $8.26 billion in 2021 and is projected to be worth $15.23 billion by 2029.[13]

As this study proves that the base catalysts are not destroyed in reaction and are hence recyclable, further work can be done to examine other methods for recycling the catalyst that do not involve the buildup of glycerol, such as the possible recyclability of heterogeneous base catalysts that could be extracted from the products.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The authors would like to thank Dr. Roberto Rinaldi for his advice and guidance throughout the research and Raul Rinken for his assistance with ^{1}H NMR facilities.

Notes and references

1 N. K. Patel and S. N. Shah, *Food, Energy, and Water*, Elsevier, Boston, 2015, pp. 277–307.

2 *U.S. Energy Information Administration - EIA - independent statistics and analysis*, https://www.eia.gov/energyexplained/biofuels/biodiesel-rd-other-basics.php.

3 K. Malins, J. Brinks, V. Kampars and I. Malina, *Applied Catalysis A: General*, 2016, **519**, 99–106.

4 *Biofuel Market (Type: Biodiesel, Bioethanol, Bio-Heavy Oil, And Others; And Feedstock: Corn, Sugarcane, Vegetable Oil, Palm Oil, And Others) - Global Industry Analysis, Share, Growth, Regional Outlook And Forecasts, 2022-2030*, 2022.

5 *Latest Study on "Biofuel Market Size, Share, Trends, Growth, Production, Consumption, Revenue, Company Analysis and Forecast 2022-2030"*.

6 A. Zarli, *Catalysis, Green Chemistry and Sustainable Energy*, Elsevier, 2020, vol. 179, pp. 77–95.

7 M. Ehsan and M. T. H. Chowdhury, *Procedia Engineering*, 2015, **105**, 638–645.

8 E. Malewska, K. Polaczek and M. Kurańska, *Materials*, 2022, **15**, 7807.

9 *Methanol*, https://commonchemistry.cas.org/detail?cas_rn=67-56-1(retrieved2022-12-23)(CASRN:67-56-1).

10 J. M. Encinar, J. F. González, A. Pardal and G. Martinez, *G. Martinez*.

11 I. M. Rizwanul Fattah, H. C. Ong, T. M. Mahlia, M. Mofijur, A. S. Silitonga, S. M. Rahman and A. Ahmad, *Frontiers in Energy Research*, 2020, **8**, year.

12 E. Subroto, R. Indiarto, A. Pangawikan, E. Lembong and R. Hadiyanti, *Advances in Science Technology and Engineering Systems Journal*, 2021, **6**, 612–618.

13 *Global Mono and Diglycerides and Derivatives Market– Industry Trends and Forecast to 2029*, https://www.databridgemarketresearch.com/reports/global-mono-and-diglycerides-and-derivatives-ma

Modelling Hydrogen Emissions from the Australia-Japan Liquid Hydrogen Supply Chain to Assess Climate Impact

Tianpeng Lao, Yusuf Ogazi-Khan

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

Abstract

Hydrogen could be a key source of energy as the world transitions to net-zero emissions – hydrogen supply chains inherently have emissions of their own. This paper studies the emissions from an established hydrogen supply chain and how hydrogen acts as a harmful indirect greenhouse gas. In this paper a detailed analysis of the hydrogen emissions for the pilot phase of the Hydrogen Energy Supply Chain project was conducted in order to quantify the long-term impact on the climate. The emissions at different stages of the supply chain have been calculated, providing insights into emissions reduction in the highest emitting sectors. The climate impact quantified using Global Warming Potentials over a time scale of 100 years of 12.8±5.2 $kg_{CO2,\,eq.}/kg_{H2}$. A total emissions intensity of 12.57% was calculated for the Hydrogen Energy Supply Chain supply chain, corresponding to 120.83 tonnes of $CO_{2,\,eq.}$ being emitted. These emissions were scaled to assess the impact as the supply chain was scaled to its commercial scale of 225,000 tonnes of LH_2.

1. Introduction

Traditional fossil fuels like oil, coal and natural gas have supported modernization and economic growth across the world. Global efforts such as the Paris Agreement have set clear goals to limit global warming to well below 2 degrees Celsius compared with pre-industrial levels [31]. Japan and Australia are two of the ratifying countries that have submitted their Nationally Determined Contributions to demonstrate their commitments. By 2030 Japan aims to reduce greenhouse gas emissions to 26% below 2013 levels and Australia aims to reduce its emissions to 26% below 2005 levels [1]. Hydrogen offers many advantages as an alternative energy source, however, its effects on the environment in the long term are not completely understood. This paper aims to model hydrogen emissions of a liquid hydrogen (LH_2) supply chain to assess its wider impact on the global climate.

Japan's limitations in domestic natural resources created a dependency on a foreign import of energy to fuel its livelihood and industry. As a result, Japan's energy transition strategies must look to achieve a reduction in greenhouse gas emissions whilst simultaneously ensuring its future energy security. Japan looks to realize its strategy through energy security, economic efficiency, environment, and safety, known as the "3E+S" energy policy [2]. For a long-term solution through to 2050, and into the latter half of the century, it must reform its current energy supply structure and transition to a new clean energy system. Hydrogen is one such alternative, it is plentiful and contains no carbon. Its similarity in characteristics to light natural gas (LNG) allows hydrogen to complement approaches to decarbonization. It is estimated that hydrogen has the potential to account for up to one fifth of global energy consumption reducing global emissions by 6 Gigatonnes of $CO_{2,\,eq.}$ [3].

In 2017, Japan became the first country to adopt a national hydrogen framework in the Basic Hydrogen Strategy. As a highly industrialised country, it has a severe lack of hydrocarbon resources, developing a strong hydrogen economy can provide energy security and industrial competitiveness as well as reducing carbon emissions. Since hydrogen is an unrestricted energy source anyone can act as a consumer or supplier regardless of geographic advantages some countries may have. Japan have been trialling hydrogen supply chain projects, most notably from Australia, Brunei and Saudi Arabia [2].

Historically, Australia has been the biggest supplier of energy and key minerals to Japan. Australia provides around two-thirds of Japan's coal, and a third of its natural gas [4].

Australia's competitiveness as a producer of raw materials has cemented it as a safe and reliable trade partner. In the transition from fossil fuels Japan looks to achieve net-zero in its current supply chains; its lack of carbon capture infrastructure means that building stronger synergies with Australia will be essential.

The future hydrogen industry is expected to grow to $4 trillion globally by 2050 [5], Japan and Australia both look to be huge players in this market. A consortium of Australian and Japanese companies and government has raised $500 million to fund the Hydrogen Energy Supply Chain (HESC) [3] project where hydrogen gas will be produced in Australia and shipped as liquid hydrogen to Japan. The world's first overseas shipment of liquid hydrogen, transporting 75 tonnes of liquid hydrogen from Australia to Japan, was completed in January 2022 as part of the pilot phase of this project. Following this, it will be moving onto the commercial phase where the 225 kilotonnes of liquid hydrogen will be produced for shipment to Japan by 2030 [6].

Hydrogen has great potential as an effective alternative energy carrier to support the energy transitions in the future. It is a zero-emission fuel that has many production routes from available resources such as fossil fuels, biomass and water. Gaseous hydrogen's energy content by weight, 33.6 kWh/kg, is greater than diesel, 12-14 kWh/kg, however it has a lower energy content by volume. Therefore, throughout supply chains hydrogen can be transported as liquid organic hydrogen carriers, ammonia or liquid hydrogen [7]. Liquid hydrogen is 800 times more dense than gaseous hydrogen and does not require as much further processing in the downstream of supply chains [3].

Hydrogen presents many advantages but it is considered as an indirect greenhouse gas [8]. Hydrogen has an atmospheric lifetime of 1.4-2.1 years based on modern estimates, after which it completely oxidizes with the hydroxyl radicals. This hydrogen oxidation leads to a reduction of hydroxyl radicals and formation of ozone in the troposphere, as well as formation of water vapour in the stratosphere. These will have adverse consequences on the climate. Yet, limited research has been conducted to assess the long-term climate impact of hydrogen emissions [8].

Most recently, the consequences of hydrogen leakages were modelled in literature by Ocko and Hamburg., 2022 across different timescales [9]. Additional studies are needed to assess the real effect of hydrogen releases on atmospheric warming. The authors are aware of only one paper that researches hydrogen emissions along supply chains, by Cooper et al., 2022 [10]. This study aims to estimate hydrogen emissions across different stages from the HESC project with a wider range of modelling methods to more accurately assess the impact of these emissions on global warming as the supply chain enters its commercial stage.

2. Background

2.1 H_2 production, brown coal gasification with carbon capture

This supply chain is based on the HESC project, the first project in the world to produce, process and transport liquid hydrogen by sea to an international market. Hydrogen is produced through gasification of coal and biomass and refined in Loy Yang, Latrobe Valley. Whilst carbon credits were used to offset CO_2 emissions in the project's pilot phase, a carbon capture system will be implemented in the commercial phase during refining. Hydrogen gas is then compressed and transported to the Port of Hastings by road for liquefaction [3].

2.2 H_2 production, green electrolysis

The supply chain is based on a research trial in Queensland shipping hydrogen as a liquid organic hydrogen carrier (LOHC) to Tokyo, Japan [11]. Hydrogen is extracted through solar powered electrolysis of treated non-drinking water in Redland, Australia is modelled for its production stage only, to understand the hydrogen emissions associated with a green hydrogen production route. The extracted hydrogen is converted to methylcyclohexane (MCH) and exported to Tokyo, Japan. As this work only models shipments for liquified hydrogen, only the production and processing stage of this supply chain is considered [11]. A limited amount of data is available to make an informed assessment of hydrogen emissions along the shipping and distribution stages of this supply chain.

2.3 Liquefaction

High purity hydrogen is brought to the Port of Hastings and loaded into a pilot liquefaction plant. The specific hydrogen liquefaction process used in the HESC pilot project is unknown. Yin., 2019 reviews the design and optimization from a wide range of liquefaction processes [12]. This source is the key literature piece used to make a suitable assumption of the liquefaction process used in the Port of Hastings. It considers the performance of a wide range of pre-cooled systems and cascade systems.

2.4 Shipping

The ship studied in this research is the inaugural Suiso Frontier - the world's first liquid hydrogen carrier. The vessel has a gross tonnage of 8,000 t and is 116 m long [3]. Given the lack of data on the ship specifications the IMO Type-C limits were used when applicable [13]. The Software used to calculate the Boil-off rate in the Suiso Frontier's 1,250 m^3 Stainless Steel tank, is called BoilFast [14]. This software was developed by the University of Western Australia and has been tested extensively at Nasa's Glenn Research Centre. The software required specifications for the tank was unavailable. To address this lack of granular data IMO type-C values were used to determine the thresholds for certain tank specifications and estimated appropriate values.

2.5 Unloading and regasification

Liquid hydrogen was successfully shipped and unloaded at the Port of Kobe. Here, it is unloaded into Japan's largest liquid hydrogen storage tank where it can be transported for further processing such as regasification or for distribution and consumption. There is a lack of specific process data for this stage.

3. Methodology

H_2 emissions are estimated across different stages of the HESC supply chain through various modelling methods. The supply chain is segmented into four stages, production and processing, liquefaction, shipping, unloading and regasification. Hydrogen emissions are also modelled from an alternative green hydrogen production route, a large scale solar powered electrolysis plant in Gladstone for comparison with brown coal gasification.

The production and processing stages are estimated using emissions factors found in literature which is scaled to the relevant throughput of the system in consideration. This method was repeated for the unloading and regasification stage due to a lack of process specifications. A combination of methods found in literature and Aspen Plus is used to model the liquefaction stage. A software developed by the University of Western Australia called BoilFast is used to model fugitive emissions in the shipping stage.

Throughout this paper hydrogen emissions are defined as hydrogen directly lost to the atmosphere such as through fugitive emissions, venting and incomplete combustion. Whereas hydrogen consumption is defined as hydrogen's energy content used to power the system.

3.1 Emissions factors

For the production and regasification stages of the supply chain specific process data was unavailable so a method found in literature was used to calculate the relevant hydrogen emissions [10]. This method calculates the quantity of hydrogen emitted using emission factors which takes the average hydrogen emissions as a percentage of total hydrogen throughput in the systems considered.

Where E_i is the emissions per supply chain *i*, A_i is the total throughput in supply chain *i*, EF_i is the emissions factor of supply chain stage *i*.

$$E_i = A_i \times EF_i \quad (1)$$

$$E = \sum_{i=1}^{N} E_i \quad (2)$$

The total emissions across a supply chain can subsequently be calculated by summing E_i across *N* number of stages.

3.2 Aspen Plus for Hydrogen Liquefaction

Hydrogen gas is liquefied when it is cooled to below -253°C at 101.325 kPa [3]. Hydrogen liquefaction is typically a low efficiency process which consumes high amounts of energy. There is a lack of data in literature on the specific liquefaction process used in Latrobe Valley. A review of current hydrogen liquefaction facilities from literature was conducted to decide a base process, considering specific energy consumption (SEC) and exergy efficiency (EXE). A precooled dual-pressure Linde-Hampson process was selected for modelling due to its high exergy efficiencies and relatively low specific energy consumption.

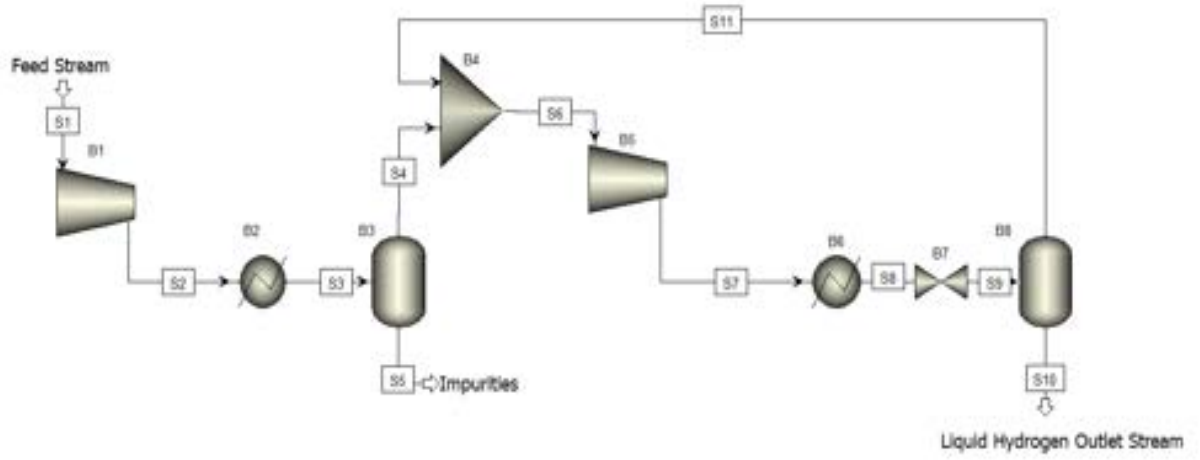


Figure 1 - Pre-cooled Dual Pressure Linde-Hampson Process modelled in Aspen Plus

Aspen Plus software is used to model this pre-cooled dual pressure Linde-Hampson process as seen in Figure 1. This process involves four stages:

i. Pre-cooling from 300K to 80K
ii. Compression from 20bar to 80bar
iii. Cooling from 80K to 20K
iv. Expansion From 80 bar to 1 atm (1.01325 bar)

The feed stream has a temperature of 300K, this must be precooled to hydrogen's inversion temperature 80K at 20bar. This is necessary for the temperature to decrease when the cooled hydrogen is finally liquefied through a Joule-Thompson expansion from the valve into the separator [15]. The LH_2 output is set to be 10.4kg/hr for the HESC project [3].

A Peng-Robinson equation of state is implemented in the simulation. The feed stream was simulated with 99.99% purity of hydrogen with the remaining composition assumed to be carbon monoxide. The mass flowrate of the feed is adjusted to ensure a liquid hydrogen output of 0.25 tonnes/day of LH_2 as specified in the HESC project [3]. It is assumed that both compressors operate under isentropic conditions with an isentropic efficiency of 0.72 and mechanical efficiency of 1.0 as recommended by Aspen Plus.

Values retrieved from simulations of this model is used to evaluate the systems performance through specific energy consumption (SEC) and exergy efficiency (EXE) are calculated by the following equations:

$$SEC = \frac{\dot{W}}{\dot{m}_{LH2}} \quad (3)$$

$$EXE = \frac{\dot{W}_{rev}}{\dot{W}_{act}} \quad (4)$$

Where $\dot{W}$ is the net power of the entire system and $\dot{m}_{LH2}$ is the mass flow rate of liquid hydrogen. For the exergy efficiency is calculated as a ratio of $\dot{W}_{rev}$, the ideal reversable liquefaction work and $\dot{W}_{act}$, the actual liquefaction work. The amount of hydrogen energy consumed in the system is also calculated to demonstrate the inefficiency of this stage.

3.3 Shipping

The ship's containment system has a capacity of 1,250 m3 (ClassNK Classification Registration, 2022). As one cubic metre of liquified hydrogen weighs 70kg, it can carry 87.5 tonnes of hydrogen fully laden, however the operating capacity is 75 tonnes of Hydrogen. The ship can be compared to a small LNG tanker and the tank is based on IMO-type C Specifications. In the HESC Pilot project, the Suiso Frontier was used to ship blue hydrogen from the Port of Hastings last year in December and arrived at the Kobe Port in Japan in January, a journey of 4,907 nm (Sea-distances, 2022), taking 23 days at a speed on 13 knots.

The major source of emissions contributors during the shipping phase of the supply chain is boil off. Boil-off in the cryogenic storage tank in the Suiso Frontier was modelled as being vented to the atmosphere and the only

emissions involved with this stage of the supply chain.

We inputted the journey distance, and the software then yielded the boil-off rate curve below and the boil-off rates across the journey with intervals of 1 hour. Uncertainty bounds of 10% were assumed for total journey time of 377 hours.

3.4 Global Warming Potential

The climate impact of the hydrogen emissions over different time scales can be quantified through global warming potentials (GWP). Different values for the GWP_{100} of hydrogen can be found in literature can be defined as the amount of energy the emissions of 1 kg of gas will absorb over 100 years compared with the emissions of 1 kg of $CO_{2,eq}$. The most recent study from Nov 2022 [16] calculates GWP_{100} for hydrogen to be 12.8±5.2 $kgCO_{2,eq}/kg_{H2}$. A government study from April 2022 found the GWP_{100} of hydrogen to be 11±5 $kg_{CO2,eq}/kg_{H2}$ [16].

Most of the GWP_{100} uncertainty in both papers is due to uncertainty in the magnitude of the hydrogen soil sink and radiative forcing scaling factors. These modern studies consider stratospheric effects more accurately, this value is significantly larger than previous estimates made in 2006, where GWP_{100} for hydrogen was 5.8 $kgCO_{2,eq}/kg_{H2}$ [8]. For this study the most recent literature GWP_{100}value of 12.8±5.2 $kgCO_{2,eq}/kg_{H2}$ is used in this paper to calculate $CO_{2,eq}$ emitted along the HESC supply chain to understand the effect hydrogen emissions in the long term.

4. Results & Discussion

To compare the results from each stage of the chosen supply chain, hydrogen emissions were expressed as a ratio of tonnes of hydrogen emitted per tonne hydrogen. A GWP_{100} value of 12.8±5.2 $kgCO_{2,eq}/kg_{H2}$ is then used to calculate tonnes of $CO_{2,eq}$ emitted to compare with the current emissions profile of Japan and Australia. The contribution of emissions reduction achieved by the HESC project can also be estimated. Through these models it was found that the heaviest emitting stage of the HESC supply chain is shipping, twice as much as liquefaction, second highest emitting stage. The relevant hydrogen emissions of each stage is shown in Table 1. The total emissions intensity of the supply chain is calculated to be 12.57% of total hydrogen in the supply chain, using values in Table 1. Therefore, to ship 75 tonnes of liquid hydrogen from Australia to Japan 9.44 tonnes of hydrogen is emitted equivalent to 120.83 tonnes of $CO_{2,eq}$.

4.1 Production and Processing

The chosen production route of the HESC project was brown coal gasification, mostly lignite, with carbon capture to abate greenhouse gas emissions [16]. The exact process for gasification is unknown so a hydrogen emissions factor of 0.55% from literature is used to calculate the hydrogen emissions for this stage. The Latrobe Valley gasification plant has a capacity to produce 5000 tonnes of hydrogen per year, the pilot phase saw 75 tonnes of this shipped to Japan. It is assumed the throughput to achieve this is in this stage is 78 tonnes of hydrogen, relevant emissions data for this stage can be seen in Table 1.

Brown coal gasification is typically an intensive process with low efficiency. According to

	H_2 throughput (tonnes)	H_2 emitted (%)	H_2 emitted ($tonnes_{H2}$)	$CO_{2,eq}$ emitted ($tonnes_{CO2,eq}$)
Brown H_2 production with CCS	78.0	0.55	0.44	5.6
Green H_2 production via. electrolysis*	78.0	2.05	1.60	20.5
Liquefaction	77.6	3.30	2.56	32.8
Shipping	75.0	**6.69**	5.02	64.3
Unloading and regasification	70.0	2.03	1.42	18.2

Table 1 – Emissions per stage of the supply chain, * denotes a supply chain stage not part of the HESC project

literature values [18], gasification processes typically produce up to 0.17kg of hydrogen from 1kg of coal. For a typical fixed bed reactor processing lignite, up to 0.418 m^3 of carbon dioxide could be released per kg of the synthesis gas produced. Assuming the same performance for brown coal gasification in the HESC project, 458.8 tonnes of coal would be needed to produce 78.00 tonnes of hydrogen. Therefore, should this technology be scaled to withstand the commercial scale of the HESC project of hydrogen to Japan and minimise impact on climate, carbon capture systems must be implemented to limit direct greenhouse gas emissions from this stage. Or alternative production routes which does not rely on carbon-based fuels should be considered. Therefore, a large-scale electrolysis plant in Gladstone preparing for export to Tokyo Japan is modelled to compare with the HESC production route [19]. The relevant hydrogen emissions factor is found to be 2.1%. The plant is part of a research trial in Queensland which will produce up to 365kt of renewable hydrogen per year for [20]. If this scale is successfully achieved, 7.5 ktonnes of hydrogen will be emitted per year, equivalent to 95.8 ktonnes of $CO_{2,eq}$ using a GWP_{100} for hydrogen of 12.8±5.2.

4.2 Liquefaction

From model simulations it is found that to attain 75 tonnes of liquid hydrogen the system requires 92.3 kW of power which corresponds to a SEC of 9.1 kW/kg_{LH2} and EXE of 26%. The higher heating value of LH_2 is found to be 141.9 $MJ/_{kgLH2}$ from literature [21]. It is calculated that 31.9% of hydrogen's energy content will be consumed to output 75 tonnes of hydrogen. Liquefaction is by far the most energy intensive stage of the supply chain so a small improvement to the plant's efficiency will have a significant impact on overall emissions for liquid hydrogen supply chains. A sensitivity analysis was conducted to see how specific energy consumption and percentage of hydrogen consumed changes as hydrogen feed rate is varied, this can be seen in Figure 2 and Figure 3. The specific energy consumption significantly improves as the feed rate is increased, because when the feed rate is low there is the compressors and heat exchangers are disproportionately intensive. However, when the HESC project scales to its commercial stage the necessary hydrogen liquefaction rate will increase significantly to a point where the process has an improved specific energy consumption.

The main source of fugitive hydrogen emissions in the liquefaction stage is due to boil off in the unloading of LH_2 from the process.

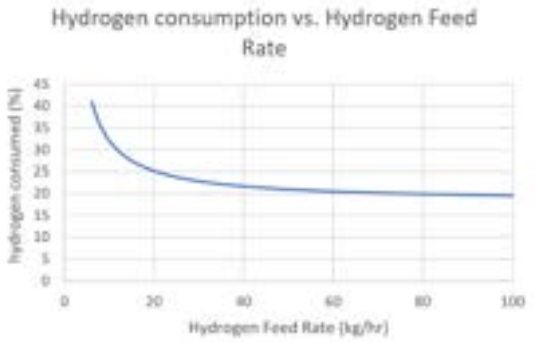


Figure 2 – Hydrogen consumption vs. hydrogen feed rate

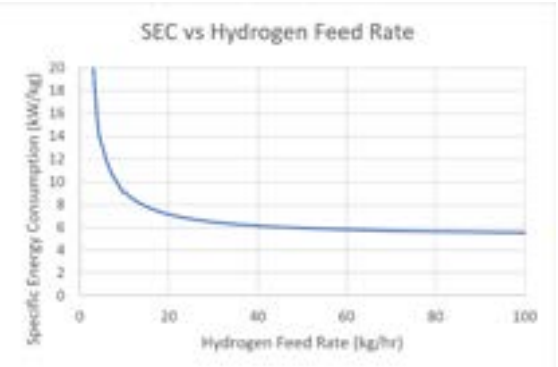


Figure 3 – Specific energy consumption vs. hydrogen feed rate

The boil off emissions in this stage is 3.3% of unloaded hydrogen is gathered from literature [22]. For the pilot stage of the HESC, 2.6 tonnes of hydrogen will be emitted in liquefaction. Should this be vented it is equivalent to 21.86 tonnes of $CO_{2,eq}$ being emitted for the pilot phase of the project. These emissions could be either recycled back into the liquefaction system or flared off as it combusts cleanly.

Despite the relatively low boil off rate, inefficiencies in the system are due to the high energy intensity of liquefaction. This is due to the complexity of the process and the low operating temperatures and high pressures. In typical liquefaction plants the lower limit of energy consumption is 30% of hydrogens higher heating value [23]. This is a significant factor causing H_2 losses in the supply chain and demonstrates the need for optimization of this process.

Hydrogen exists in the form of two spin isomers, ortho and para hydrogen, the composition of the two changes with temperature. As the

temperature of gaseous hydrogen is decreased in liquefaction, ortho-hydrogen will be converted to para-hydrogen until the composition of the liquid hydrogen output stream is almost 100% para-hydrogen. This conversion releases heat of 670 kJ/kg, which is greater than hydrogen's latent heat of evaporation of 452 kJ/kg [12]. Due to this phenomenon, the deal heat of transformation will cause 50% of liquid hydrogen stored to be evaporated within 10 days. Therefore, to increase overall efficiency of the liquefaction system, it is recommended that a catalytic ortho-para hydrogen conversion should be implemented in the precooling stage to prevent excess boil off in the unloading of the system.

4.3 Shipping

The Suiso Frontier uses a propulsion system consisting of three Daihatsu DE-23 1,320kW diesel engines and two 1,360kW electric motors enabling the vessel to sail at speeds of up to 13 knots.

During transportation LH2 is stored in cryogenic tanks to minimise heat loss. However, heat ingress is unavoidable and heat ingress into the liquid and vapour stages causes local convection at tank walls and heat conducted through the liquid-vapour interface results in thermal stratification which causes a temperature gradient on the top layer of the liquid. These two factors result in the formation of Boil-Off gas, which forms in the tank and causes an increase in pressure known as 'self-pressurisation'. During self-pressurisation, both the liquid–vapor interface temperature and vapor pressure continue to rise until the tank pressure reaches the pressure relief set point and the BOG is vented to the atmosphere [24].

Shipping was the heaviest emitting section of the supply chain, but it should be noted that there are already methods to reduce boil-off gas emissions, that could be adopted should the liquefied hydrogen market become mainstream. The BOG generated in the tank does not have to be directly vented to the atmosphere, the other options include: utilising the boil-off gas to fuel the propulsion system, flaring the gas as it is vented. The excess gas can also be burned in a gasification unit, but this means there will be waste of materials and energy

[30]. The boil off gas can also be reliquefied and sent back to the tank – however this is a complex process and requires many components. [25] The calculated emissions data for this stage can be found in Table 1.

4.4 Unloading and Regasification

This stage of the HESC project had a severe lack of data available. A hydrogen emissions factor was found to be 2.05% of the total throughput [3]. The relevant emissions for this stage can be found in Table 1. To minimise these emissions, it is recommended liquid hydrogen should not be stored for longer than necessary to ensure minimal boil-off emissions.

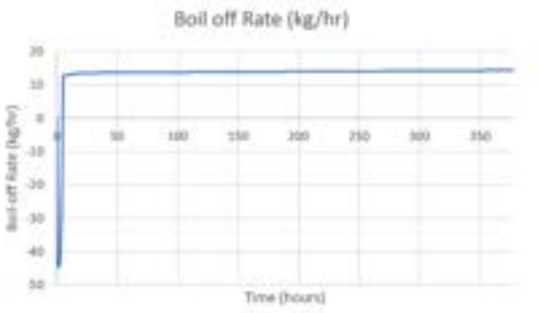


Figure 4 - Boil off rate over time as liquid hydrogen is shipped to Japan

4.5 Alternatives

Other transportation methods could also be considered. Compressed hydrogen pipelines do currently exist but only on a small scale to transport chemicals between facilities. The pipeline pumps for this consumes a lot of hydrogen's energy content to power the compressors, at least 1.4% of hydrogen flow is consumed per 150km of pipeline. This makes pipeline transport very unfavourable if Japan continues to rely on a foreign import of hydrogen [23].

Hydrogen can alternatively be transformed into liquid organic hydrogen carriers such and methylcyclohexane (MCH) for high density hydrogen storage and transport. Projects producing this through green electrolysis and grey natural gas in Australia and Brunei respectively both ship MCH to Japan. Although MCH has a lower energy density than liquid hydrogen, 47 kg H_2/m^3 and 70 kg H_2/m^3 respectively. MCH can be stored at ambient conditions, compared with liquid hydrogen, which is stored at 20K, there is no need for the energy intensive liquefaction process. MCH does however require heavier downstream processing as it has a high enthalpy of dehydrogenation, 73.6MJ/mol_{H2}. The largest barrier for MCH to be successfully adopted at large scale is its infrastructure. Dedicated hydrogenation and dehydrogenation process

can be optimized to improve the efficiency of MCH supply chains [26].

Ammonia is another great hydrogen carrier that should be considered. The high volumetric density of hydrogen, 107 kg_{h2}/m^3 and mature synthesis and distribution infrastructure are just a few of its advantages. Therefore, it must be considered in the discussion of likely hydrogen carriers for the future. There are a few drawbacks in the high temperature needed for ammonia decomposition and incomplete separation allows residual ammonia to poison polymer electrolyte membranes in fuel cells. The typical Ammonia production route is the well-established Haber-Bosch process which emits roughly 2.9 tonnes of CO_2 per tonne of ammonia, this process consumes 1-2% of yearly global energy demand. Electrochemical routes to production offer a green alternative to the Haber-Bosch process and can complement electrification well. Ammonia does have disadvantages with lower flammability and increased difficulty in handling due to being toxic to humans and marine life. A wider assessment of the environmental impact of ammonia must be considered.

4.6 The climate impact of hydrogen emissions

By 2030 for Japan to meet its emissions reduction goal of 46% from benchmark 2013 levels, 1,395$MtCO_{2,eq}$ an annual reduction of 642$MtCO_{2,eq}$ must be achieved. As of 2021, Japan's has already achieved a 17.6% reduction in emissions. For Australia, to reduce its emissions by 43% from 2005 levels, an annual reduction of 240$MtCO_{2,eq}$ needs to be achieved and as of 2021 Australia has achieved a 12.7% reduction in emissions.

In this study models of hydrogen emissions along the supply chain that follows a 75-tonne liquid hydrogen pilot shipment. The quantity of emissions is modelled to be 120.83 tonnes of $CO_{2,eq}$. Considering the liquid hydrogen to have an energy content of 120 MJ/kg_{H2}, this pilot project provides 9 million MJ. It is subsequently calculated that at the pilot scale the HESC project will emit 13.4 g $CO_{2,eq}/MJ_{H2,HHV}$ of energy delivered [7]. Comparisons can be drawn with a literature value that found methane emissions from Australia to Japan LNG supply chains is 3.2 g $CO_{2,eq}/MJ_{LNG,HHV}$.

The HESC project looks to be scaled to its commercial phase of producing 225 ktonnes of liquid hydrogen per year for shipping to Japan by 2030 [6]. Should this production rate be achieved, the associated emissions for each stage of this modelling study can be scaled to assess the climate impact of the projects commercial phase. A total of 28.3 ktonnes of hydrogen will be emitted, equivalent to 361.98 ktonnes of CO2,eq.

This ambitious project's pilot stage has created 400 jobs, and thousands more when the supply chain is successfully commercialised. This will also act as a great opportunity to upskill and retrain workers for renewable energy systems. This project estimated to reduce CO_2 emissions by 1.8 million tonnes per year, equivalent to emissions of 350,000 cars or 0.129% reduction of Japan's annual emissions compared with 2013 levels. By 2030 Japan looks to boost its demand for hydrogen to 3 million tonnes per year, the commercial phase of the HESC supply chain can meet 7.5% of this forecasted demand [27] [28].

4.7 Limitations

This study has uncertainties due to lack of relevant process data. Production and processing, unloading and regasification is all modelled using emissions factors from literature. Errors within this method arise from discrepancies, and because a proxy of LNG is used, where leakage rates of hydrogen is calculated by drawing parallels with existing LNG supply chains. In liquefaction the uncertainty is from a cubic equation of state being used to model quantum fluids, as well as specific compressor duty and efficiency data. The uncertainty in the shipping stage is mostly due to lack of specific data on storage tank specifications.

In these estimates of climate impact using GWP_{100} values, error and uncertainty is calculated using error bounds of ±5 kg $CO_{2,eq}/kg_{H2}$. Uncertainty comes from unknown atmospheric distribution of hydrogen and other gases, unknown size of a hydrogen soil sink. These factors vary greatly depending on geographic location and environment.

5. Conclusion & Future Work

If hydrogen becomes a main-stream source of energy production – it will be essential to carry out further research on its emissions. Some conclusions and suggestions for future work are outlined in this section:

1.) Production and processing route must shift away from brown coal. Work should be done to understand different

supply chains. Other dense hydrogen carriers such as LOHCs and ammonia should also be researched. A variety of distribution methods should be considered as well, such as transporting the hydrogen via pipeline or trucks.

2.) Liquefaction processes must be optimized to improve efficiency of the supply chain.
3.) Emissions from the shipping stage can be significantly reduced through the utilization of boil-off gas in the ship's propulsion system, or through flaring the gas or reliquefying it.
4.) lack of specific process data for large-scale electrolysis prohibited this research therefore more in-depth modelling of the green hydrogen production route should be conducted.
5.) For future modelling of hydrogen emissions using Aspen software, it should be considered that hydrogen is a quantum fluid and that cubic equations of state, such as the Peng-Robinson model, has yielded poor predictions of thermodynamic properties [29].
6.) Should the HESC project continue to rely on brown coal gasification, a dependency on sequestration technologies would be the only way to offset greenhouse gas emissions. At a commercial scale the effects on climate this will have could be prevented by using green routes.
7.) To inform policy, future studies should be conducted on the downstream distribution pathways and future hydrogen demand should be forecasted. A cost-benefit analysis would be useful to determine the economic incentives for adoption of and investment into this supply chain.

References

[1] Intended Nationally Determined Contributions (INDC): Greenhouse Gas Emission Reduction Target in FY2030 (no date). Available at: https://www.mofa.go.jp/ic/ch/page1we_000104.html.

[2] Basic Hydrogen Strategy | ESCAP Policy Documents Managment (no date). Available at: https://policy.asiapacificenergy.org/node/3698.

[3] Hydrogen Energy Supply Chain (HESC) Project (2022b) Supply Chain. Available at: https://www.hydrogenenergysupplychain.com/supply-chain/.

[4] Smith, M. (2022) Australia assures Japan it can count on energy supply. Available at: https://www.afr.com/world/asia/australia-assures-japan-it-can-count-on-energy-supply-20221009-p5bodd.

[5] A Quarter of Global Hydrogen Set for Trading by 2050 (2022). Available at: https://www.irena.org/news/pressreleases/2022/Jul/A-Quarter-of-Global-Hydrogen-Set--for-Trading-by-2050.

[6] Gillespie, M. (2022) Hydrogen Energy Supply Chain - Pilot Project. Available at: https://research.csiro.au/hyresource/hydrogen-energy-supply-chain-pilot-project/.

[7] Slanger, D. (2022) Run on Less with Hydrogen Fuel Cells. Available at: https://rmi.org/run-on-less-with-hydrogen-fuel-cells/.

[8] Derwent, R. (2018) HYDROGEN FOR HEATING: ATMOSPHERIC IMPACTS. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/760538/Hydrogen_atmospheric_impact_report.pdf.

[9] Ocko, I. B. and Hamburg, S. P.: Climate consequences of hydrogen emissions, Atmos. Chem. Phys., 22, 9349–9368, https://doi.org/10.5194/acp-22-9349-2022, 2022.

[10] Cooper, J., Dubey, L., Bakkaloglu, S., & Hawkes, A. (2022). Hydrogen emissions from the hydrogen value chain-emissions profile and impact to global warming. Science of The Total Environment, 830, 154624. https://doi.org/10.1016/J.SCITOTENV.2022.154624

[11] QUT (2018) "QUT leads new hydrogen pilot plant," QUT, 5 September. Available at: https://www.qut.edu.au/research/article?id=135488.

[12] Yin, L. (2019) "Review on the design and optimization of hydrogen liquefaction processes," SpringerLink, 25 December. Available at: https://link.springer.com/article/10.1007/s11708-019-0657-4?error=cookies_not_supported&code=79ec3fdb-3210-44a6-b741-f290fe5dad22.

[13] ClassNK Register of Ships - M/S SUISO FRONTIER(CNo.210624) (no date). Available at: https://www.classnk.or.jp/register/regships/one_dsp.aspx?imo=9860154.

[14] BoilFAST (no date). Available at: https://www.fsr.ecm.uwa.edu.au/software/boilfast/.

[15] Connelly, E. and Penev, M. (2019) "DOE Hydrogen and Fuel Cells Program Record," https://www.hydrogen.energy.gov/pdfs/1

9001 hydrogen liquefaction costs.pdf, 9 September.
[16] Warwick, N. et al. (2022) "Atmospheric implications of increased hydrogen use," https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment data/file/1067144/atmospheric-implications-of-increased-hydrogen-use.pdf, April.
[17] Hydrogen production begins successfully at the Hydrogen Energy Supply Chain project (2021). Available at: https://ifrf.net/ifrf-blog/hydrogen-production-begins-successfully-at-the-hydrogen-energy-supply-chain-project/.
[18] Midilli, A., Kucuk, H., Topal, M. E., Akbulut, U., & Dincer, I. (2021). A comprehensive review on hydrogen production from coal gasification: Challenges and Opportunities. International Journal of Hydrogen Energy, 46(50), 25385–25412. https://doi.org/10.1016/J.IJHYDENE.2021.05.088
[19] FEED Contract for Hydrogen Production Plant in Australia (2022). Available at: https://www.sumitomocorp.com/en/africa/news/release/2021/group/14270.
[20] Central Queensland Hydrogen Project (2022). Available at: https://research.csiro.au/hyresource/central-queensland-hydrogen-project/.
[21] Committee on Alternatives and Strategies for Future Hydrogen Production and Use; Board on Energy and Environmental Systems; Division on Engineering and Physical Sciences; National Research Council; National Academy of Engineering (no date) Read "The Hydrogen Economy: Opportunities, Costs, Barriers, and R&D Needs" at NAP.edu. Available at: https://nap.nationalacademies.org/read/10922/chapter/21.
[22] Petitpas, G. (2018) Boil-off losses along LH2 pathway. Available at: https://www.osti.gov/biblio/1466121/.
[23] Bossel, U. and Eliasson, B. (no date) "Energy and the Hydrogen Economy," https://afdc.energy.gov/files/pdfs/hyd e conomy bossel eliasson.pdf.
[24] al Ghafri, S. Z. S., Perez, F., Heum Park, K., Gallagher, L., Warr, L., Stroda, A., Siahvashi, A., Ryu, Y., Kim, S., Kim, S. G., Seo, Y., Johns, M. L., & May, E. F. (2021). Advanced boil-off gas studies for liquefied natural gas. Applied Thermal Engineering, 189, 116735. https://doi.org/10.1016/J.APPLTHERMALENG.2021.116735
[25] LNG: what is boil-off gas and what does it do? (2022). Available at: https://www.fluenta.com/lng-boil-off-gas/.
[26] Makepeace, J. W., He, T., Weidenthaler, C., Jensen, T. R., Chang, F., Vegge, T., Ngene, P., Kojima, Y., de Jongh, P. E., Chen, P., & David, W. I. F. (2019). Reversible ammonia-based and liquid organic hydrogen carriers for high-density hydrogen storage: Recent progress. International Journal of Hydrogen Energy, 44(15), 7746–7767. https://doi.org/10.1016/J.IJHYDENE.2019.01.144
[27] Reuters (2022) More than 20 countries agree to boost low-emission hydrogen output by 2030. Available at: https://www.reuters.com/business/sustainable-business/more-than-20-countries-agree-boost-low-emission-hydrogen-output-by-2030-2022-09-26/.
[28] Whittaker, J. (2021) First hydrogen produced from Latrobe Valley coal generates export hopes, emissions fears. Available at: https://www.abc.net.au/news/2021-03-12/hydrogen-from-coal-production-begins-la-trobe-valley/13241482.
[29] Aasen, A., Hammer, M., Lasala, S., Jaubert, J. N., & Wilhelmsen, Ø. (2020). Accurate quantum-corrected cubic equations of state for helium, neon, hydrogen, deuterium and their mixtures. Fluid Phase Equilibria, 524, 112790. https://doi.org/10.1016/J.FLUID.2020.112790
[30] Kalikatzarakis, M., Theotokatos, G., Coraddu, A., Sayan, P., & Wong, S. Y. (2022). Model based analysis of the boil-off gas management and control for LNG fuelled vessels. *Energy*, *251*, 123872. https://doi.org/10.1016/J.ENERGY.2022.123872
[31] IPCC, 2018: Global warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty [V. Masson-Delmotte, P. Zhai, H. O. Pörtner, D. Roberts, J. Skea, P.R. Shukla, A. Pirani, W. Moufouma-Okia, C. Péan, R. Pidcock, S. Connors, J. B. R. Matthews, Y. Chen, X. Zhou, M. I. Gomis, E. Lonnoy, T. Maycock, M. Tignor, T. Waterfield (eds.)]. In Press.

Deep Reinforcement Learning Algorithms to Optimize Supply Chain Processes with Uncertain Demand

Nabeel Ali and Tayyib Tahir
Department of Chemical Engineering, Imperial College London, U.K.

Abstract This study aims to utilise reinforcement learning (RL), a subfield of artificial intelligence, to solve the supply chain management and optimization problem. Widespread problems through periods of global shortages and uncertainties from geopolitical tensions have rendered traditional supply chain policies ineffective. Recent advancements in RL are a key area for the future operation and development of sustainable industrial production systems. This study leverages these techniques to better optimise these systems by maximising the reward function subject to key parametric industry constraints. These include inventory storage limits as well as simulating stochastic and seasonal demands with both continuous and discrete products. A generalised multi-period 2-echelon supply chain environment was implemented through a custom OpenAI gym environment. The approach investigates the applicability and performance of model-free policy optimization RL algorithms. Specifically, the Advantage Actor-Critic (A2C) and Proximal Policy Optimization (PPO) algorithms. It was observed both proposed methods outperformed the baseline simple agent policy and show promising results for the supply chain optimization problem. The loss function of each method showed convergence over large timesteps with PPO, using a 0.2 epsilon clipping, converging to the optimal policy significantly faster. Seasonal demand provided greater volatility and paired with the continuous dataset, a larger converging reward per unit of time.

Introduction

1. Supply Chain Optimization

Supply chain management is the centralized management of the flow of goods and services of a supplier, monitoring each touchpoint. Over the last few decades, the conventionally linear supply chain with entities in direct series with one another has evolved into an increasingly complex and uncertain integrated supply network. With so many opportunities to enhance value along the supply chain, proper management can drive an increase in profit margins, improve customer service and reduce the environmental impact of suppliers (Fernando, 2022). Supply chain optimization exploits technology and resources such as Artificial Intelligence, IoT and blockchain (IBM, 2022) to improve the efficiency and performance of supply networks. They help to address issues with data that is siloed, supply disruptions, and sustainability and can even help to build a competitive advantage.

Global supply chain issues were prominent during the outbreak and spread of the COVID-19 pandemic, due to national lockdowns and shifts in demand, which brought challenges to governments, enterprises, medical institutions and citizens. There have been sudden shortages in various sectors including consumer goods, metals, food and chemicals (Chase, 2022). More recently, the Russia-Ukraine conflict and wider geopolitical issues continue to escalate supply issues.

Traditional methods to address supply chain optimization include branch and bound (Karimi & Davoudpour, 2014), Tabu search (Melo, et al., 2012), genetic algorithms (Govindan, et al., 2010) and linear programming (Piedro, et al., 2010). These approaches have been widely used in supply chain management and have achieved remarkable results. However, for large-scale and complex supply chain systems, traditional methods still face many difficulties in practical application.

Firstly, the solution space is often very large. Many modern supply chain scenarios involve many nodes, and complex network relationships, and therefore require a long period to be solved online. A second challenge is that of being able to cope with large operational uncertainty. There are various uncertain factors in the operation of the supply chain, not only internal operation, such as demand, price fluctuations, and production uncertainty but also external uncertainties, including the risk of disruption caused by unexpected events.

2. Reinforcement Learning

RL is a subfield of Artificial Intelligence specialising in sequential decision-making which trains an agent how to take optimal actions to maximise reward over time. RL was designed to address the optimization of stochastic, sequential decision-making processes, and it turns out to supply chains are exactly this type of system.

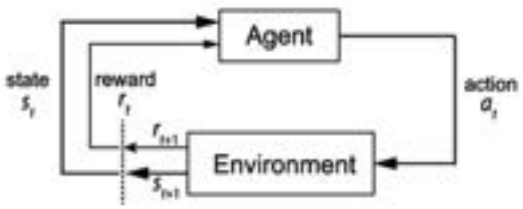


Figure 1: The agent-environment interaction in RL

The typical RL setting is illustrated in Figure 1 and can be described by an agent interacting with and exploring the environment in which it resides. At each time step t, the agent receives a state s_t and selects an action a_t from some set of possible actions $\mathcal{A}$ according to its policy π, where π is a mapping from states s_t to actions a_t. In return, the agent receives the next state $s_{t+1}$and receives a scalar reward r_t. This process continues to loop around until the agent arrives at a terminal state, at which point the episode resets and restarts.

Formally, RL is a framework enabling us to solve problems that can be described as a Markov Decision Process (MDP). The MDP serves as the flexible framework for goal-directed learning that can be described as a tuple:

$$< \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma > \quad (1)$$

Where $\mathcal{S}$ is a finite set of states, $\mathcal{A}$is a finite set of actions, $\mathcal{P}$ is a state transition probability matrix such that $\mathcal{P}_{ss'}^{a} = \mathbb{P}[\mathcal{S}_{t+1} = s' | \mathcal{S}_t = s, A_t = a]$, $\mathcal{R}$ is a reward function such that $\mathcal{R}_s^a = [\mathcal{R}_{t+1} = s' | \mathcal{S}_t = s, A_t = a]$, and γ is a discount factor, $\gamma \in [0,1]$.

The goal of the RL agent is to find the optimal policy $\pi: \mathcal{S} \times \mathcal{A}$ that maps states into actions so that the cumulative expected return over the time horizon is maximised.

In many scenarios, it can be ambiguous whether the action performed by the agent contributed to the gained reward, so an n-step discounted return is applied, where the cumulative rewards for the action a_t for n steps are exponentially weighted by a discount factor γ.

As supply chains evolve and further increase in complexity, action and solution spaces continue to grow and RL algorithms alone become ineffective. Deep RL is a more recent advancement that pairs an RL algorithm with an artificial neuron network to approximate Q values instead of storing all the state and value pairs in a table. Agents can make decisions for huge, unstructured data sets. This increases the manageability of the solution space and allows the RL agent to generalise the values of states, that have not even been encountered during the training phase, based on past experiences.

3. Literature Review

There have been some preliminary studies portraying great promise for RL to solve many of the current supply chain challenges. RL was first applied to supply chain optimization and inventory management over twenty years ago (Giannoccaro & Pontrandolfo, 2002). This was the first published paper to use RL to optimize supply chain processes, as they recognized traditional methods are struggling to optimize supply chains as system and action spaces grow.

More recently, Deep RL-based methods have also been proposed to solve supply chain optimization problems. A Deep Multi-Agent RL technique was proposed to solve supply chain optimization problems (Fuji, et al., 2018).. Multi-agent RL employs numerous artificial intelligence agents that collectively learn, collaborate and interact with each other whilst cohabitating in an environment. This technique allows for faster training and greater exploration of the environment, and the agents update the network's weights to reinforce the probability of actions with positive rewards and weaken the inclination to take actions with negative rewards. The study used Deep-Neural-Network-Weight-Evolution to optimize processes in a beer distribution game and managed to achieve 80.0% lower costs in the game than expert players.

An innovative, cooperative multi-agent RL approach for a resource balancing problem on a simulated ocean transportation service was also proposed and compared with various Multi-agent RL methods (Li, et al., 2019). The study found that all multi-agent techniques outperformed the baseline methods and the Diplomatic Awareness Multi-agent RL technique performed best.

From the literature review, past studies show RL outperforms traditional methods and is a promising solution for the optimization of supply chain optimization problems. However, these studies usually deal with smaller, simple supply chain networks. The literature also focuses on only discrete action spaces and rarely considers capacity constraints.

In this study, we focus on multi-period supply chain optimization problems with:

- **Discrete and continuous action spaces**
- **Capacity constraints**
- **Demand uncertainty (stochastic and seasonal)**

Two DRL-based methods (PPO and A2C) are proposed to solve the supply chain optimization problem on a two-echelon supply chain model that has been easily generalized and can easily be modified to increase the complexity.

The rest of this paper is organized as follows: section two describes the inventory management problem statement. Section three presents the methodology of building a custom OpenAI gym environment and the decision-making behind selecting the Deep RL-based methods selected to solve the optimization

problem. Section four presents the results of the proposed methods, and section five follows with the discussion. Finally, section six outlines the conclusions and prospects of the study.

Problem Statement

In this paper, we focus on the multi-period supply chain optimization problem. The supply chain is structured as Figure 2 and consists of a plant, two warehouses, a single retailer and consumers with both stochastic and seasonal demands.

Figure 2: Supply chain structure with two echelons

The structure of the optimization problem is divided into a set of periods with the same length of 365. At the beginning of each time period, the agent reviews the current state of on-hand inventory and backorders of the plant warehouses and the retailer. It then uses its policy to return a number of products to produce and deliver between each echelon. Due to the capacity constraints set, there are upper limits on the number of products the plant can produce, and each warehouse and retailer can store. The products produced by plants are assumed to be stored onsite at the plant warehouses in one action and therefore not subject to any delays here. A waiting time is applied between each echelon, from Echelon 1 to Echelon 2 and Echelon 2 to the consumers. This waiting time is represented through a gaussian distribution for each independent echelon transfer. The supply of raw material is assumed to be adequate and stable and so negligible to the optimization. The demand from consumers is stochastic and potentially seasonal and the decision maker satisfies the demand to the fullest extent by the on-hand inventory. In addition, the overstocked products or unsatisfied demands are carried over to the next period, which means that the decision made in a period will affect the inventory levels in future periods. In the case of excess inventory, a storage cost is incurred per unit of overstocked products. Otherwise, in the case of deficient inventory, a penalty cost is incurred per unit of unsatisfied demand. The process is fulfilled till the end of the full-time period. The initial inventory is set constant as well as an initial order to start the supply chain environment.

The aim is to maximize the total profit taking into consideration revenue from sold products, production costs, storage costs, penalty costs and transportation costs incurred in during all periods. The demands across the periods are independent, though not necessarily identically distributed. This problem encapsulates the dilemma of matching supply with volatile demand in the presence of capacity constraints and distributed waiting times. The supply chain optimization problem can be formulated as follows:

$$\boldsymbol{max}\sum_{t=1}^{T}\begin{cases} v_1\sum_{j=1}^{K} Dem_{j,t} - v_2 p_t - v_5\sum_{j=1}^{K}[\frac{d_{j,t}}{\gamma}] \\ \quad -v_s\sum_{j=1}^{K} max\{inv_{j,t},0\} \\ \quad +v_4\sum_{j=1}^{K} min\{inv_{j,t},0\} \end{cases} \tag{2}$$

Subject to

$$0 \le p_t \le P_{max},\ \forall t \in \{1,\dots,T\} \tag{3}$$

$$\sum_{j=1}^{K} d_{j,t} \le inv_{j=0,t}, \forall t \in \{1,\dots,T\} \tag{4}$$

$$inv_{j=0,t} + p_t \le C_{j=0,t}, \forall t \in \{1,\dots,T\} \tag{5}$$

$$inv_{j,t} + d_{j,t} \le C_{j,t}, \forall j \in \{1,\dots,K\}, \forall t \in \{1,\dots,T\} \tag{6}$$

$$inv_{j=0,t+1} = inv_{j=0,t} + p_t - \sum_{j=1}^{K} d_{j,t}, \forall t \in \{1,\dots,T\} \tag{7}$$

$$inv_{j,t+1} = inv_{j,t} + d_{jt} - Dem_{j,t}, \forall j \in \{1,\dots,K\},\ \forall t \in \{1,\dots,T\} \tag{8}$$

Where v_1 is the revenue from sold products, v_2 is the production costs, v_3 is the storage costs, v_4 is the penalty cost and v_5 is transportation cost. p_t is production target in time period t, $d_{j,t}$ signifies the products delivered to retailer j, $inv_{j,t}$ is the inventory level of the plant and warehouse $(j = 0)$ and retailers $(j = 1)$. When backorder occurs and there is inadequate inventory the value of $inv_{j,t}$ will become negative.

There is a production capacity limit for the plant as a constraint (2) and the total delivered amount should be no more than the current inventory level of the plant warehouse as a constraint (3). Constraints (4) and (5) represent the storage capacity constraints of the plant warehouse and retailers. In the plant warehouse, the sum of current inventory and newly produced products should not exceed the capacity limit. For retailers, the sum of the current inventory level and newly delivered products should not exceed the capacity limit. Constraints (6) and (7) represent the material balance for plant warehouses and retailers.

The objectives of the project are outlined below:

1. Build an Open AI GYM environment to model a 2-echelon supply chain process as the base case to test agent performances
   - Outline and then apply the parametric constraints of the supply chain process
2. Using A2C and PPO StableBaslines3 RL models, train agents on the environment
3. Tune the hyperparameters and cross-validate to maximise our model rewards and accuracy
   - Vary the demand type, vary between discrete and continuous
   - Obtain reward graphs, loss graphs

Methodology

1. Generalised Supply Chain Model

For this study a simple 2-echelon supply chain was modelled and implemented into a custom OpenAI gym environment, that can be easily generalized. The environment can be described by the main supply chain class with a gym wrapper to simulate the initialisation, reset of the environment and step function for each given period. The gym wrapper is key to model StableBaselines3 algorithms; it provides easy implementations to policies, setting up and resetting environments as well as being easily able to make changes to the supply chain. This includes varying the key echelon parameters sent to the supply chain such as product sale/cost, lead times, storage costs, storage capacities and material costs. However, more importantly, it also allows for changes to the entire supply chain class including the number of echelons, nodes per echelon and the given number of products a customer can demand.

The structure of how a simulated supply chain is run is similar to the structure of any RL algorithm. The environment is observed within the observation space. Below is the supply chain state which is comprised of all existing inventory of all echelons with the given demand of products concatenated.

$$\begin{pmatrix} I_1 & d_{1,1} & d_{2,1} & \dots & d_{i,1} \\ I_2 & d_{1,2} & d_{2,2} & \dots & d_{i,2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ I_n & d_{1,n} & d_{2,n} & \dots & d_{i,n} \end{pmatrix} \quad (9)$$

Where $d_{i,n}$ represents the demand for product i of echelon n and I_n represents the array of inventories at echelon n. The agent, using a trained policy or value-based function (or both), observes the state and reward to output an action that becomes an order request sent to the previous echelon. This action is within the limits of the action space and holds the demand for a given product.

$$[O_{1,2}, O_{2,3}, O_{3,4} \cdots O_{n-1,n}] \quad (10)$$

Where for a given product order $O_{n-1,n}$ represents the order from echelon $n-1$ to n. It is important to note both action space and observation spaces are different which tailors the RL algorithms that can be applied, with more details in the following section.

The action is sent via the step function which updates the environment, returning the new state (observation) and reward from the reward function. Appendix A can be referred to and is the gym environment used for this study. (Note the appendix environment setting is set for discrete constraints, taking stochastic demand for our base policy.) The environment is as described in the problem statement.

In this process, the demand was described as both stochastic and seasonal. Randomness can be applied in programs via the use of pseudorandom number generators. The program uses packages *random* and *gauss* to help achieve the continuous demands. This is supported by random seeding to ensure values are stochastic in nature. Seasonal demand is achieved using a sinusoidal curve to map the bounded demands to this curve. This helps achieve rising and lowering demands multiple times throughout the entire period. Whilst the stochastic demand exhibits stochastic and stationary behaviours seasonal also exhibits along with periodicity, stochastic behaviour.

Continuous and discrete are also calibrated throughout the environment. How the environment changes from discrete to continuous is the following changes. First, changing the actions and observation space constraint to *float* data type. Also changing our discrete stochastic and seasonal demand to take floating values. On top of this, the state function should hold continuous values and the lead times can be selected via a continuous dataset. The action data type provided by the agent must be changed via the hyperparameters to output an order within its limits to any degree of accuracy.

Rule-based methods are a simple solution for the supply chain optimization problem, where for example a fixed quantity will be ordered when the inventory position drops to the reorder point. These policies are easy to understand and implement, so they have been widely used in practice. Therefore, the simple agent policy, shown below, will be used as a baseline to compare the performance of the RL algorithms.

```
orders=(n_echelon)*demand
if storage>2*demand:
    orders[0]=int(demand*0.8)
return orders
```

Figure 3: Simple agent policy used as baseline case

The simple agent observes the current storage and if the demand is greater than the storage it places an order for the number of echelons multiplied by the demand. Otherwise, if the storage is greater than two times the demand the agent places an order for 80% of the demand. This rule-based method is functional; however, it is not efficient and it's clear the storage costs and losses incurred can be optimized.

As the aim of this study is supply chain optimization, model-free policy optimization RL algorithms were adapted and evaluated. Specifically, this paper investigates the Advantage Actor-Critic (A2C) and Proximal Policy Optimization (PPO) algorithms. The decision-making process behind selecting these algorithms is outlined below.

2. Reinforcement Learning Taxonomy

The landscape of algorithms in RL is vast and an exhaustive list of all the current methods would be futile. Therefore, a taxonomy of a few distinguished modern RL algorithms is illustrated in figure 4.

Figure 4 - A non-exhaustive, but useful taxonomy of algorithms in modern RL (OpenAI, 2018)

Further research into the algorithms was conducted and is outlined below to help decide which of the algorithms would be best to adapt and implement into the supply chain model.

2.1 Model-Free v Model-Based

There are two main groups of RL algorithms: Model-based algorithms and model-Free algorithms.

Model-based algorithms use the transition and reward functions to estimate the optimal policy. They are utilised in situations where the agent has access to a model of the environment and has full knowledge of how it responds to different actions, with probabilities and subsequent rewards attached. Model-based algorithms allow the RL agent to plan ahead, and therefore for static environments where everything is fixed model-based RL algorithms are best suited.

Model-free algorithms on the other hand estimate the optimal policy directly from the agent's experience and interaction with the environment. The agent learns by altering its behaviour and observing the different kinds of rewards it receives. They can function without access to any transition or reward functions, and with limited knowledge regarding the dynamics of the environment. Model-free RL is more applicable in circumstances involving incomplete information about the environment. In the real world, we rarely have fixed environments. Supply chains have a dynamic environment with many internal and external uncertainties, with stochastic and unpredictable demands. Often model-free, Deep RL based algorithms do not require the transition probability distribution to operate. Hence, they are able to make decisions without a thorough model of the environment. In such scenarios, model-free algorithms outperform other techniques.

2.2 Policy Optimization v Q-Learning

Delving into model-free RL there are a further two key groups of algorithms: Q-learning and Policy Optimization.

Q-learning is an off-policy method, implying the policy the algorithm optimizes differs from the policy the agent uses to select an action. The algorithms learn the action-value function and determines the policy using it. They are also deterministic, meaning that the methods will always give the same output, given the same input. Due to this, some sort of $\in$-greedy policy needs to be implemented to allow exploration of the environment.

Off-policy algorithms often tend to lead to severe instabilities when optimizing policies and using data off-policy only tends to be useful if the environment in which the agent resides is slow (OpenAI, 2018), so you want to squeeze as much information from each experience the agent has.

Policy gradient algorithms are on-policy methods that directly optimize the same policy that the agent uses to select the actions it takes. The agents update the network's weights to reinforce the probability of actions with positive rewards and weaken the inclination to take actions with negative rewards. They are also stochastic so unlike off-policy methods, they can give different outputs given the same input. The action is sampled from the distribution outputted from the network (Salwiczek, 2021). Therefore, exploration of the environment is not an issue and no epsilon greedy strategy is required.

Policy gradient algorithms tend to be less hyperparameter sensitive and have more stable convergence properties (OpenAI, 2018). If the agent is learning from a fast environment so that observations are easy to obtain or if you can run many instances of your environment, like in the

supply chain optimization model, then an on-policy method is the preferred method.

2.3 Policy Optimization Algorithms

There are four major modern policy optimization Deep RL algorithms: (Reinforce) policy gradient, A2C, PPO and Trust Region Policy Optimization (TRPO).

Policy gradient functions employ stochastic gradient descent to converge to the optimal policy. After looking into the specifics of the algorithms we found it was sample efficient. One rate-limiting aspect of the reinforce policy gradient algorithm is that the gradient estimate at each step is only valid for the current policy and therefore only one step of gradient descent can be carried out per trajectory batch, otherwise the policy can fluctuate wildly and destroy training.

The TRPO algorithm was designed, motivated by this and is able to do multiple gradient descent steps by looking at the KL divergence of its distributions between policies and constraining the steps of optimization close to the original policy. TRPO does this by determining a maximum step size to be explored around the original policy and imposing a hard trust region constraint which has to be solved with second-order methods (Hui, 2018), whereas the PPO algorithm does this by doing first-order optimization and clips the objective function between a range eliminating reasons for the new policy to fluctuate drastically away from the original policy. (OpenAI, 2018).

Therefore, the reinforce policy gradient is eliminated in favour of the other algorithms for this problem. Looking into the algorithms even deeper, it was found that TRPO, although suitable when dealing with continuous control tasks, is not effective with algorithms that share parameters between a policy and value function (Gujar, 2018) and lacks a faster convergence rate compared to PPO. Therefore, TRPO was also eliminated and A2C and PPO were selected as the RL algorithms that would be adapted into the supply chain model.

2.4 A2C and PPO

The two core types of RL algorithms are policy-based and value-based. Actor-Critic algorithms merge these methods by explicitly representing the policy independent of the value function. The actor outputs the best action by taking the state as an input, whereas the 'critic' evaluates the action and portrays how good it is to be in this state.

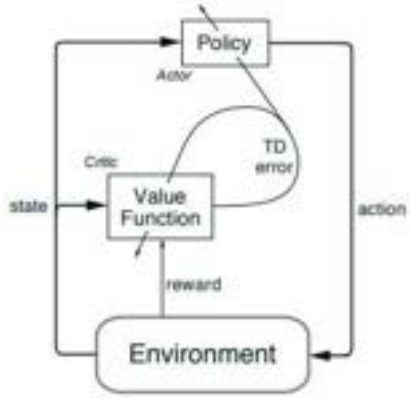


Figure 5 - Actor-Critic Model with TD error

The critic computes the action value function (Q-value), and outputs the resultant temporal difference (TD) error, which in turn informs the actor how to adjust. If the TD error is positive the inclination to take that action is reinforced for the future, and if the TD error is negative the action is discredited.

Both actor and critic are composed of separate function approximators, and the training is performed using gradient ascent to find the global maximum and update their individual weights at each step.

Two popular improvements of Actor-Critic models are the A2C and PPO models where the function approximators are non-linear artificial neural networks. Action value functions can be decomposed into the state value function $V(s)$ and the advantage value $A(s,a)$. Advantage functions depict how better an action is compared to others at a given state and help stabilize the model by having the critic learn the advantage values instead of the action value, which reduces the high variance of policy networks. A2C uses these advantage estimates to calculate the value proposition for each action state pair (Mnih, et al., 2016), resulting in faster training.

The major distinction between A2C and PPO is the loss function (Lisi, 2021). Policy optimization algorithms employ stochastic gradient descent to try to optimize a policy objective function. PPO incorporates a clipped objective function:

$$L^{CLIP}(\theta) = \hat{\mathrm{E}}_t[\min(r_t(\theta)\hat{\mathrm{A}}_t, clip(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\hat{\mathrm{A}}_t)] \quad (11)$$

Where θ is the policy parameter, $\hat{\mathrm{E}}_t$ is the empirical expectation over timesteps, r_t is the ratio of the probability under the new and old policies (from importance sampling), $\hat{\mathrm{A}}_t$ is the estimated advantage over time and ε is a hyperparameter (usually 0.1-0.2).

By introducing the clip function PPO directly improves the training stability of the policy and aims to make the largest possible improvement of the original policy in a single step, without overdoing it and taking a huge leap that could potentially destroy the training. The algorithm uses a ratio that indicates the difference between our current and old policies:

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta old}(a_t|s_t)} \quad (12)$$

and then clips this ratio between a specific range $(1-\varepsilon)(1+\varepsilon)$. Empirically, it's clear smaller steps in policy updates during training are more likely to converge towards an optimal policy. A large step in a policy update can result in requiring a long time to stabilise and recover, if it ever does.

Both models are implemented via StableBaselines3 and trained up to a maximum of 2 million timesteps, with models saved periodically and uploaded to a tensor board. The key here is being able to cross-validate and hyper-tune the necessary parameters for the best model for our environment. Documentation around this library supports an open-source hyperparameter optimization framework to automate hyperparameter search (Optuna, 2022).

Results

To conduct an assessment on the performance of the two proposed methods, the environment of the supply chain model is run on different settings. These settings include changing three main factors. This includes exploring how the seasonal demand varies alongside stochastic demands, observing the difference when all constraints, inputs and outputs are compared from continuous and discrete and lastly, how both proposed algorithms perform against the baseline simple agent.

These are run on our multi-echelon environment for up to two million time steps and terminate when the loss function converges and is minimal. Time complexity is a limitation, as models are run under CPU. Maximum timesteps per episode were set at 200 with the number of time steps per batch set to 2048.

Firstly, hyper tuning the parameters is crucial. They were tuned for each model, using a random sampler and median pruner, 2 parallel jobs with a budget of 1000 trials and a maximum of 50000 steps. The general parameters for our network were as follows.

Table 1 – Hyperparameters of our MLP network

Hyperparameter	Tuned Value
Learning Rate	0.00003
Activation Function (Output Layer)	ReLU
Hidden Layer Activation Function	ReLU
Policy	MLP
Number of episodes	10

For our Actor-Critic models, using the tensorbaord cross-validation method, the further following Actor-Critic parameters were tuned.

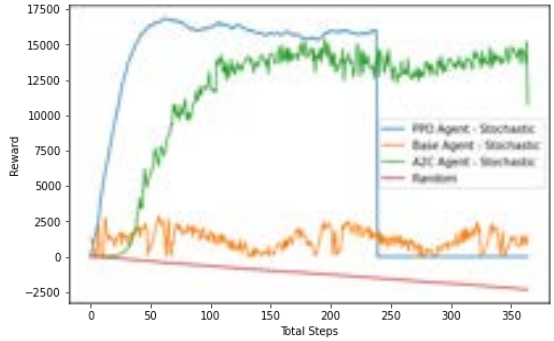
Table 2 – Hyperparameters of our MLP network

Hyperparameter	Tuned Value
Loss Function	MSE
Clipping Fraction Epsilon (PPO)	0.2
Batch Size	2048 (continuous) 512 (Discrete)
Epochs	10
Number of episodes	10

These hyperparameters were varied and tested to value the model that helped produce the highest cumulative reward, quickest convergence, and lower loss values per period. Using *tensorboard* extensive features, the mean reward per period illustrated that many time steps could achieve a high reward model. Most models with time steps above 200,000 had similar rewards.

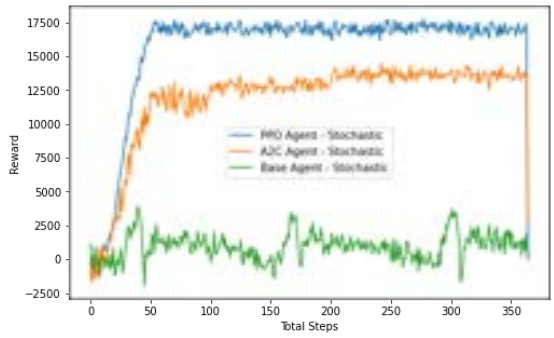
The following case is for discrete data sets in an environment with stochastic demand. (Note: the random agent is a control variable for reference of the performance on our environment if a random action from the action space was to be taken.)

Graph 1 – Reward per step, using discrete data and stochastic demand



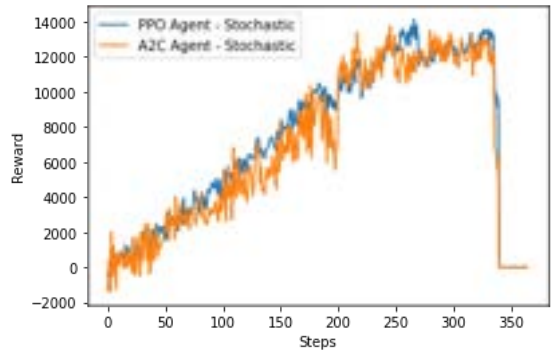
The following case is for the continuous data on all agents on the environment with stochastic demand

Graph 2 – Reward per step, using d data and stochastic demand



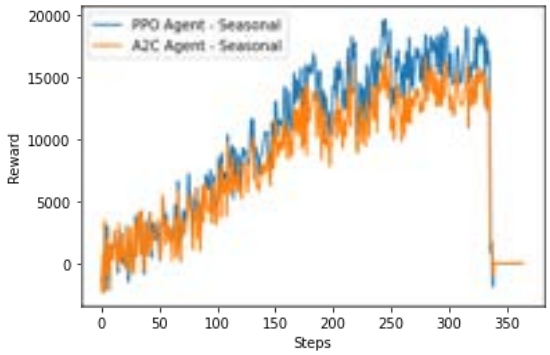
The following case is for the discrete data on both RL agents on the environment with seasonal demand.

Graph 3 – Reward per step, using discrete data and seasonal demand



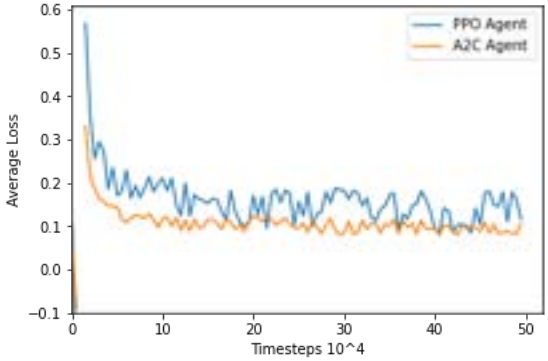
The following case is for the continuous data on both RL agents on the environment with seasonal demand.

Graph 4 - Reward be step, using continuous data and seasonal demand



Using the loss function built into PPO and creating a similar graph for A2C the following is observed. (Note: similar trends are established in all settings.)

Graph 5 – Loss function of both A2C and PPO models trained over 500,000 timesteps



Discussion

The results clearly show in all models and tests that the model-free RL algorithms outperform the simple agent policy. This is mainly due to the interaction with the environment and the hyperparameters used.

Both RL models show rewards higher than the base model, with a plateau as the losses start to converge and the best policy is found. This is the attempted solution to the optimality equations.

In many cases, A2C has higher volatility, but performance is somewhat similar to PPO due to the similarity in nature of both algorithms. Since the difference is the fact that PPO clips the objective function so that step changes from the previous policy are minimal, it is able to maintain a lower volatility in its performance across the time steps. Over larger timesteps, both models obtain similar rewards.

The base model seems to show the highest volatility in its rewards due to its limited functionality. When more extreme demands are made the inventory will be full of existing backlog orders and the rewards are

hindered in the penalty aspect of the function. This causes periods of extreme drops in rewards. Similarly, when demands are not as extreme it matches the orders exactly to the per demands with sufficient inventory. This is exhibited in both seasonal and stochastic demands. But the rewards for this throughout the period are capped. Therefore, the highest reward is limited by the rewards function where the weights of both the sale and cost of inventory heavily negate themselves.

The seasonal demands compared to the stochastic show better performance but are less consistent with higher volatility. This is because seasonal demands have a greater shift in the extreme values of the action space, ultimately having a bigger impact on reward considering the policy each model learns in the previous step. The PPO algorithm, due to the clipping function, shows an even more stable performance when compared to A2C.

Conclusion

In this paper, an investigation is carried out on the use of reinforcement algorithms, specifically PPO and A2C, on a supply chain optimisation problem. A simple 2-echelon supply chain process is built under an OpenAI gym wrapper. This allowed us to vary three key settings to test the performance of each RL model against the base model. These settings include changing three main factors and exploring how the seasonal demand varies alongside stochastic demands from customers, and observing the difference when all constraints, inputs and outputs are compared from continuous and discrete testing. Key conclusions are highlighted as follows.

1. Both RL models in all cases outperform the base simple agent. PPO performs best with the greatest reward per unit step and less volatility due to its clipping function.
2. Over long periods PPO and A2C, due to their similar nature, perform evenly.
3. Hyperparameters play a crucial role in optimising StableBaselines3 algorithms, as they allow for better learning and heavily influence rewards.
4. Seasonal demand seems to have a greater impact on the volatility of the reward functions on any model.

The results from the overall study are promising and demonstrates that both the two RL methods always converge to a better policy than the simple agent policy, in all the different settings mentioned above.

References

Chase, J. M., 2022. *What's behind the Global Supply Chain Crisis?*. [Online]
Available at: https://www.jpmorgan.com/insights/research/global-supply-chain-issues

Fernando, J., 2022. *Supply Chain Management (SCM): How it works and why it is important.* [Online]
Available at: https://www.investopedia.com/terms/s/scm.asp

Fuji, T., Ito, K., Matsumoto, K. & Yano, K., 2018. *Deep Multi-Agent Reinforcement Learning using DNN-Weight Evolution to Optimize Supply Chain Performance,* s.l.: s.n.

Giannoccaro, I. & Pontrandolfo, P., 2002. *Inventory management in Supply Chains: A reinforcement learn,* s.l.: s.n.

Govindan, K., Sasikumar, P. & Devika, K., 2010. *A genetic algorithm approach for solving a closed loop supply chain ...,* s.l.: s.n.

Gujar, S., 2018. *Trust Region Policy Optimization (TRPO) and Proximal Policy Optimization (PPO).* [Online]
Available at: https://medium.com/@sanketgujar95/trust-region-policy-optimization-trpo-and-proximal-policy-optimization-ppo-e6e7075f39ed
[Accessed 29 11 2022].

Hui, J., 2018. *RL - trust region policy optimization (TRPO) explained.* [Online]
Available at: https://jonathan-hui.medium.com/rl-trust-region-policy-optimization-trpo-explained-a6ee04eeeee9
[Accessed 25 11 2022].

IBM, 2022. *What is supply chain optimization?*. [Online]
Available at: https://www.ibm.com/uk-en/topics/supply-chain-optimization

Karimi, N. & Davoudpour, H., 2014. *A branch and bound method for solving multi-factory supply chain scheduling with batch delivery,* s.l.: Pergamon.

Lisi, A., 2021. *Beating Pong using Reinforcement Learning — Part 2 A2C and PPO.* [Online]
Available at: https://medium.com/analytics-vidhya/beating-pong-using-reinforcement-learning-part-2-a2c-and-ppo-b83391dd3657#:~:text=The%20only%20difference%20between%20A2C%20and%20PPO%20is%20in%20the%20loss%20function
[Accessed 2 12 2022].

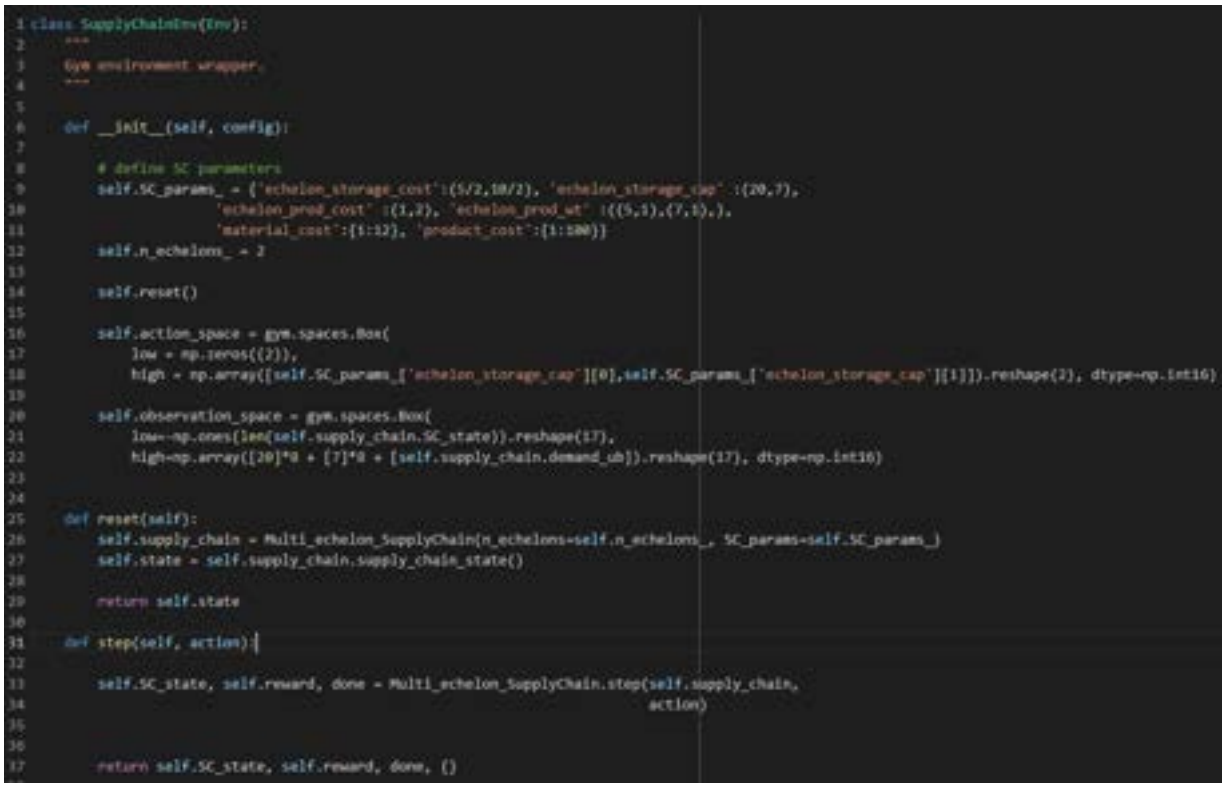
Li, X. et al., 2019. *A Cooperative Multi-Agent Reinforcement Learning Framework for Resource Balancing in Complex Logistics Network,* s.l.: s.n.

Melo, T., Nickel, S. & Saldanha-da-Gama, F., 2012. *A Tabu search heuristic for redesigning a multi-echelon supply chain ...,* s.l.: s.n.

OpenAI, 2018. *Kinds of RL Algotihms.* [Online]
Available at:
https://spinningup.openai.com/en/latest/spinningup/rl_intro2.html
[Accessed 29 November 2022].
Optuna, 2022. *Optuna.* [Online]
Available at: https://optuna.org/
[Accessed 2022].

Piedro, D., Mula, J., Jimenez, M. & Botella, M. d. M., 2010. *A fuzzy linear programming based approach for tactical supply chain ...,* s.l.: s.n.
Salwiczek, B., 2021. *Off-Policy vs On-Policy in Reinforcement Learning ~ Simplified.* [Online]
Available at:
https://medium.com/@bsalwiczek/simplified-off-policy-vs-on-policy-in-reinforcement-learning-3ed113e68a32
[Accessed 10 12 2022].

Appendix A



```
class SupplyChainEnv(Env):
    """
    Gym environment wrapper.
    """

    def __init__(self, config):

        # define SC parameters
        self.SC_params_ = {'echelon_storage_cost':(5/2,10/2), 'echelon_storage_cap' :(20,7),
                           'echelon_prod_cost' :(1,2), 'echelon_prod_wt' :((5,1),(7,1),),
                           'material_cost':{1:12}, 'product_cost':{1:500}}
        self.n_echelons_ = 2

        self.reset()

        self.action_space = gym.spaces.Box(
            low = np.zeros((2)),
            high = np.array([self.SC_params_['echelon_storage_cap'][0],self.SC_params_['echelon_storage_cap'][1]]).reshape(2), dtype=np.int16)

        self.observation_space = gym.spaces.Box(
            low=-np.ones(len(self.supply_chain.SC_state)).reshape(17),
            high=np.array([20]*8 + [7]*8 + [self.supply_chain.demand_ub]).reshape(17), dtype=np.int16)


    def reset(self):
        self.supply_chain = Multi_echelon_SupplyChain(n_echelons=self.n_echelons_, SC_params=self.SC_params_)
        self.state = self.supply_chain.supply_chain_state()

        return self.state

    def step(self, action):

        self.SC_state, self.reward, done = Multi_echelon_SupplyChain.step(self.supply_chain,
                                                                          action)


        return self.SC_state, self.reward, done, {}
```

Figure 6 – Gym wrapper of discrete environment with the base model

Imperial College London

Effect of Additives and Surfactants on the Properties of Shake-Gels

Andrew Hepper and Luke Terry
Imperial College London, Department of Chemical Engineering

Abstract

Aqueous silica particles in colloidal suspension in mixtures of polyethylene oxide (PEO) and water, widely known as shake-gels have unique rheological properties. This mixture of materials can undergo a discontinuous step increase in viscosity of up to ten times its original value when subject to shear stress (such as shaking). This transforms the initial liquid equilibrium state into a gelled material, with viscoelasticity sufficient to support its weight. After the gelled state is reached, and no more shear is applied, the gelled state reverses back into a liquid, with this process known as relaxation. As interesting as these materials might sound, they have found limited application in both industrial and domestic settings. To promote increased applicability of shake gels, a series of experiments were conducted to evaluate the rheological and qualitative properties of shake gels when a quaternary component (or additive) was introduced. Additives assessed were the surfactant sodium dodecyl sulfate (SDS); vanillin, a phenolic aldehyde responsible for the taste and aroma of vanilla; and an ionic salt, sodium chloride.

At certain concentration ranges of additives, shake-gels formed with additional qualitative properties to that of 'normal' shake-gels, including change in critical shear rate, relaxation time, foaming, scent, and even gelatine-like properties. The rheological behaviour of these new shake gels was most dramatically impacted by the addition of SDS, where viscosity was subject to a significant increase at all measured concentrations with relaxation time dramatically increased and the critical shear rate decreased.

Controlling the various aspects of the shake-gel such as the critical shear rate (CSR), relaxation time, smell, and even colour allows for a broader range of future applications, perhaps opening research to places previously not thought of. Injecting additives such as SDS in shake gels have a varying effect on the physical properties of a shake gel. Rheological measurements such as critical shear rate and relaxation time can be used as benchmarks to test these varying conditions. Unfortunately, there is no widespread use for shake-gels at the moment, but there certainly could be a place in domestic products.

Introduction

Shake-gels, first discovered in the early 2000s, are mixtures containing a high molecular weight polymer with a colloidal suspension of a silica-based compound in an aqueous environment. These shake gels show non-Newtonian shear-thickening behaviour in the liquid state but differ from the classic behaviour of a shear-thickening fluid. This is presented in shake-gels by a jump discontinuity in viscosity when subjected to a shear rate sufficient to activate the gelation process, whereby the low viscosity, almost colourless liquid transforms into a highly viscous gel with high elasticity, sufficient to support its weight. This gelatinous state also shows opposing non-Newtonian behaviour to that of the liquid state in that shear-thinning behaviour is observed. When shear stresses are removed, the gelation event is followed by a period of relaxation whereby the gel reverses back to the low-viscosity liquid in varying periods ranging from seconds to weeks; unlike the gelation event which shows a jump in viscosity, relaxation is a gradual process.

The literature on relaxation time is well established with the impact of mixture composition, temperature and pH being the common focus of studies; however, one area of shake-gels lacking research is the impact of additives on the properties of shake-gels. Having limited application in a few areas such as the domestic market; with the addition of additives to shake-gels, the potential for application skyrockets as these gels may be tailored and fine-tuned to meet specific requirements in various industries and household environments.

Upon activation of this 'gelation' event by shaking, one may notice a stimulating effect, accompanied by satisfaction, which is achieved by converting a transparent, low-viscosity liquid to an elastic type gel in a matter of seconds; so much so, that during this research project, visitors that entered the lab started to 'play' with the shake gels, many if not all wanted to continue shaking/playing with these gels.

A potential application of a shake gel is as a children's soap, hence we explored how the addition of sodium dodecyl sulfate (SDS), an anionic surfactant, commonly found in many soaps, toothpaste, shampoos, etc., due to its ability to act as a foaming agent and its thickening ability [28]. For the varying array of potential domestic use, testing the effect of SDS as an additive in shake gels is of great interest. Also tested as additives were vanillin the molecule responsible for scent in vanilla extract and NaCl for its use as a preservative. in designing new household products in the children's sector, not only will these products be fun to play with for a child, but they may pose as a solution to bathing in children with sensory issues, with the soap viewed as a friendly toy rather than an adversary, a common observation of children with these types of conditions.

Background

To convert a silica-PEO mixture into a gel, an applied shear force is needed. This phenomenon shown in shake-gel is a unique type of shear thickening. Shear thickening fluids are those that show an increase in viscosity when subject to an increasing shear rate.

When a shear force is applied to a shake-gel mixture, the PEO chains that exist in solution are highly coiled due to the interaction with water [30], a shear force enables the expansion of the coiled chains, into elongated polymer chains that make available more active adsorption sites to adsorb onto the silanol groups present on the surface of silica particles. Once the PEO is adsorbed onto the silica particles, further shear allows access for more PEO chains to be adsorbed to the silica, leading to

cross-linking and polymer bridging, which results in large clusters of silica-PEO that make up a silica-PEO network that expands the entire solution resulting in the gelled form of the shake gel.

Once the shear force is removed from the shake-gel, the PEO desorbs from the silica reducing the extent of bridging and the polymer chains return to their original coiled form gradually until the solution reaches its original liquid equilibrium state. The time that is taken to transition back from the gelled to the liquid state is known as the relaxation time which is dependent on several factors such as temperature, the concentration of PEO and of silica, pH, and molecular weight, which has been studied extensively [2,7,8,9].

There are many papers exploring the effect of these factors on relaxation time, however, Cabane, et al. (1997) [1] is thought to have been the first to discuss the shake-gel and postulate that there should be further structural studies into shake-gels since their properties can be manipulated quite easily.

Due to the formation of interconnected 3D aggregates and the bridging effect, shake gels can sometimes support their weight once a shear force is applied. As the coverage of PEO increases, the laponite particles become saturated, and it no longer forms a shake-gel. The behaviours observed with laponite-PEO mixtures are also applicable to silica-PEO mixtures, except laponite holds its structure better [2].

Experiments can be conducted using these laponite-PEO mixtures to determine the effect of PEO molecular weight on critical surface coverage. It can be concluded that shear-induced gelation occurs at laponite concentrations of above 1.25 wt.%, and the amount of PEO required for formation is between 0.2 wt.% and 0.3 wt.%. Below the formation range of PEO, the mixture says a liquid. A higher molecular weight PEO mixture decreases the surface coverage and creates a 'weaker' overall shake-gel. Also, increasing the molecular weight of PEO above $6x10^5$ g/mol does not severely affect the physical properties of the shake gel [3].

Two sources of silica are described in the literature, those being Laponite possessing disk shape silica particles and Ludox which has spherical-like silica particles. Particle shape impacts the formation of shake gels, with shake gels containing laponite existing at narrower concentration ranges compared to that of Ludox.

At different PEO concentrations, a weak gel could be formed, or perhaps an irreversible phase separation (gel + water). It is suggested that extensional shear is more important in shake-gel formation than simple shear and can be investigated by sending the shake-gel through a fine needle [4].

There is a distinct transitional state when a shear force is applied to a shake-gel between it being a liquid and a gel and increasing temperature can decrease this gelation time[2].

Increasing the applied shear rate also decreases the gelation time and decreases viscosity [4,5,10]. During this transitional state, the viscosity of the mixture increases by several orders of magnitude while forming a shake-gel.

When given a shake-gel mixture at a low pH, it becomes flocculated and adhesive, and the shake-gels do not form due to the weak interparticle repulsion and the presence of too many hydrogen atoms [6]. The mixture does not exhibit the same transition at higher pH since there are fewer hydrogen atoms. Relaxation time also happens to be longer at lower pH, and shear thickening occurs at a pH range of 8-9.9 but not outside this range.

Increasing the concentration of PEO in the solution has a directly proportional effect on relaxation time unless the concentration is between 0.1 wt.% and 0.2 wt.% [7]. Increasing silica concentration also exponentially increases relaxation time, and it seems as if increasing temperature decreases relaxation time linearly [9]. Making use of small angle X-Ray scattering analysis (SAXS) and dynamic light scattering analysis to track the structural development of the silica-PEO mixtures showed that the correlation length of PEO chains returned to their original states after 10 to 20 minutes [8].

There exists a specific set of concentrations that produce excellent shake-gels, specifically 15 wt.% to 35 wt.% of silica, and 0.1 wt.% to 0.5 wt.% of PEO. It is also understood that the effect of PEO concentration on the half-life of the shake-gel follows a similar curve, which can be approximated by the formula $813.09x^2-420.78x+59.011$. The relaxatio time of the shake-gel also seems to increase with higher silica concentration. As for temperature, higher temperatures lower relaxation time since bonds are more easily breakable [9]. It is also known that high PEO molecular weights and high silica concentrations promote the formation shake gelsels. Tnhis evidence further supports the theory that polymer bridging of the PEO is the primary mechanism behind the formation of shake gels [10].

This research project aims to augment the properties of the shake-gels to find the ideal blend of Ludox and PEO to create a fascinating children's soap using different additives such as salts and surfactants. These shake-gels will be created using Ludox TM-50 colloidal silica mixture, and average molecular weights of PEO varying from $6x10^5$ g/mol to $2x10^6$ g/mol.

Methodology

All shake-gel samples were formulated with the three most common components found in the literature: colloidal silica as Ludox TM-50® [11] purchased from Sigma-Aldrich, containing 49.9 wt.% of silica, a molecular weight of 60.08 g/mol, and density of 1.4 g/mL at 25°C [11]; pH of Ludox was measured as 9.0 in a lab with ambient temperature ~ 25°C. PEO is also from Sigma-Aldrich with an average molecular weight of 900,000 g/mol, the density of 1.21 g/mL at 25°C and containing an inhibitor BHT in the range of 200-500 ppm which prevents the autoxidation of the polymer [12,13]. While the water was present in the Ludox suspension at 50.1 wt.%, additional water was needed to reach the required concentrations of each component; deionised water sourced from Millipore-Q System located in the department of Chemical Engineering, Imperial College London, was used for this purpose.

In addition to the base shake-gels, the additives explored were: the surfactant sodium dodecyl sulphate (SDS) purchased from Sigma-Aldrich, with a purity $\geq$ 98.5%, SDS has a molecular

weight of 288.38 g/mol and a critical micelle concentration in the range of 7 – 10 mM at 20 – 25°C [14]; sodium chloride purchased from BDH chemicals at a purity of 99.9%, density of 2.5 g/ml at 25°C and a molecular weight of 58.44 g/mol; vanillin (the component that pertains to the aroma of vanilla) purchased from sigma Aldrich with a purity of 99%.

All materials purchased were used as-is from the manufacturer with no further processing or refinement.
The choice of additives relates to the potential application of shake-gels for use as children's soap. A surfactant sodium dodecyl sulfate (SDS) was investigated for its foaming ability, an essential property required for soaps to maximise the surface contact of the soap with the skin. Vanillin was tested as an additive for its scenting ability since soap requires a pleasant smell. NaCl was the final additive selected, it was noticed when formulating shake gels w/o additives, that over an extended period shake gels in sealed containers lost their properties as a shake gel, with NaCl being a common preservative in food products for its ability to reduce the amount of water unbounded to allow for the growth of biological microbes

Samples were weighed on an OHAUS Pioneer Analytical balance precise to 3 decimal places (i.e., an absolute error of ± $5x10^{-4}$ g). First, high-concentration PEO solution was prepared at 1.6 wt.% which was later diluted further before formulating the shake gels. This was achieved by the slow (to prevent agglomeration) addition of PEO to DI water stirred by a magnetic stirrer and a high rotational velocity.

Additives were also first dissolved into a base solution before the formulation of the shake gel in the same manner as the PEO solutions. The additive solutions were formed at SDS solutions of 20 wt.%, a Vanillin solution of 0.9422 wt.% and a NaCl solution of 0.2 wt.%.

All solutions used to prepare the samples were kept for no longer than 48 hours to prevent concentration of the solutes due to evaporation of water, to combat this issue within the 48-hour window, parafilm was placed over containers and sealed from the atmosphere to again prevent evaporation. Solutions were kept at room temperature to prevent degradation of the polymer.

Before formulating the shake gel, the PEO base solution was diluted down to 0.4 wt.% with DI water; next, for shake gels containing additives the additive solution was added, then was mixed, and left to settle. The Ludox suspension was then added, which formulated the shake gel. It was found that following the initial mixing of Ludox into the solutions, the shake gel did not form but after being left in a closed container for ~24 hours, the shake gel solutions were established.

Designing a product for children's use in domestic environments, the aim was to formulate a base shake gel with a fast relaxation time for maximum appeasement, hence the aim was to create a shake gel with the lowest fraction of Ludox and the highest fraction of water, i.e., the most 'dilute' shake gel. Not only does this have the benefit of a fast relaxation time, but there are also positive implications on economic feasibility; a significant cost fraction of the shake gel comes from the Ludox/silica source, minimising this component concerning the total volume of shake-gel formulated, which helps to minimise production costs.

A base shake-gel (BSG) was used as a reference sample to compare with the additive containing shake gels To find the 'most dilute' shake gel composition, an intermediate weight fraction (between the established bounds of 0.25 to 0.5 wt.% of PEO was nominated at 0.4 wt.% and kept constant. Silica fraction was then decreased by substitution with DI water to make several shake-gels until a shake gel no longer formed, to establish a minimum concentration of Ludox needed to form a shake-gel at the selected PEO weight fraction.

The final composition of the BSG was 0.400 wt.%, 20.615 wt.% and 78.985 wt.% of PEO, silica, and water, respectively. To account for the change in composition upon the presence of a fourth additive, the mass of the additive was substituted by the reduction of the water content, for example, a sample containing 0.5 wt.% of SDS meant the water mass fraction would be reduced to 78.485 wt.%. Reduction of the water content over the reduction of silica or PEO content was in an attempt to minimise the distortion in shake-gel properties due to a deviation from BSG composition. Water was not only the most abundant component in the BSG meaning a change in composition led to the lowest relative change in composition, but also water has the least impact on shake-gel properties compared to silica and PEO.

All rheological measurements were performed on a Thermo Scientific™ (previously HAAKE™) MARS™ 60 modular rheometer platform with the *recessed bottom coaxial cylinder (CC20 Ti)* measuring geometry. This geometry consists of a serrated cylindrical rotor, that during measurements, is placed concentrically inside an also serrated cylindrical cup (CCB26-C32/SE). The rotor possesses a recessed bottom, trapping air as the rotor is vertically moved into the cup containing the sample, this trapping of air prevents contact of the sample with the bottom portion of the rotor negating torque measurement influence, which would arise from contact to the bottom surface. A top recession is also present on the rotor and provides an overflow region for cases where sample volume exceeds the specified 17.2 ml, the excess volume (up to the point where the overflow is filled) is trapped in the top recession preventing any influence on torque measurements on that region.

A double gap CC27 geometry was also a candidate for use in viscosity measurements, but its use was neglected as it is believed to have required a longer period to activate the gelation event. The CC20 Rotor and the CCB26-C32/SE Cup were chosen for the rheometer due to the serrations in both the rotor and cup, allowing for more shear stress so that the shake-gel can form in a reasonable amount of time.

Calculation of the viscosity in rotational rheometry is performed using the following procedure: The shear stress τ is given by the equation:

$$\tau = A \cdot M_d \qquad (1)$$

Where M_d is the torque applied to the rotor and A is a geometry factor given by the equation:

$$A = \frac{1}{2 \cdot \pi \cdot R_i^2 \cdot L} \qquad (2)$$

Where R_i is the outer radius of the rotor and L is the height of the cylindrical part of the rotor. This geometry factor specific to the recessed coaxial cylinder geometry used in the experiments, with the variation of this factor pertaining to the modularity of the equipment.

The shear rate (or strain-rate) $\dot{\gamma}$ can then be determined by the following:

$$\dot{\gamma} = M \cdot \Omega \qquad (3)$$

Where Ω is the angular velocity of the rotor and M is a second geometry factor again specific to the recessed coaxial cylinder measuring geometry and is expressed in the following:

$$M = \frac{2 \cdot \delta^2}{\delta^2 - 1} \qquad (4)$$

With δ denoting the ratio between the inner radius of the cup and the outer radius of the rotor.

Finally, viscosity η can be related to the shear stress and shear rate in the following:

$$\eta = \frac{\tau}{\dot{\gamma}} \qquad (5)$$

[24]

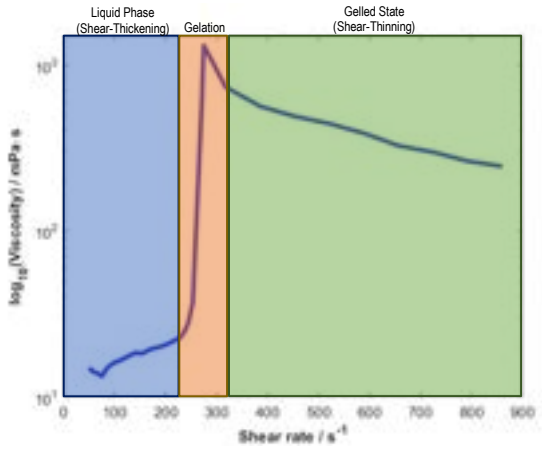


***Fig. 1**. Example of a viscosity shear rate plot for a shake gel sample when subjected to a linear, continuous increase in shear rate. Showing the initial liquid, gelation and gelled states split by distinct regions.*

Typically, studies that measure gelation point nominate a constant shear rate and measure the time taken for the liquid-gel transition to occur. However, the problem faced in this scenario was that a dramatic change in rheological properties arose due to the presence of certain additives (such as SDS) which lead to an issue in nominating a shear rate to measure gelation time.

When subjected to a shear rate above 500 s^{-1}, shake gels containing SDS instantly gelled and no distinct step change in viscosity was observed. Below a shear rate of ~900 s^{-1}, the base shake-gel containing no additives and the other additive-containing shake-gels would not gel independent of the amount of time the sample was subjected to shear.

To combat this, a ramp (or linear) increase in shear rate was set up on the rheometer to measure the point at which a discontinuous jump from a liquid to a gel occurred this is known as the critical shear rate (CSR). Keeping the rate of increase of shear rate constant w.r.t time at 1.4 s^{-2}, allowed for a comparison of relative minimum shear.

Due to the sudden change in viscosity at the gelation point, the data acquisition needed to be sufficiently fast such that the viscosity change shows as an almost vertical line on a plot of viscosity-shear rate; the data acquisition used for these measurements was every 0.5 s; important to note is that while this rate of data acquisition was sufficient for the rate of shear rate increase used in these experiments if testing at a greater increase rate it is likely that a faster data acquisition would be required.

Keeping the rate of increase in shear rate constant allows for the comparison of the shear rate value at which the gelation event occurs. An essential point to note is that time is also a crucial factor in the occurrence of gelation, thus running the experiments at a lower rate of increase in shear would likely result in the gelation occurring at a lower shear rate.

To process the data and attain a shear rate at which the gelation occurred, the raw viscometer data was processed in a MATLAB script that used the gradient function. This script uses a central difference approximation between subsequent data points (except the first and last points where a simple difference is calculated, allowing for the output of vectors with dimensions equal to that of the input). From the output data, the maximum gradient (or differential) value for change in viscosity w.r.t. shear rate was extracted along with the corresponding shear rate at which this maximum gradient took place, this shear rate was deemed as the CSR. Figure 1 shows an example of the viscosity-shear rate profile obtained during a CSR determination; in this example, the CSR would have been determined to lie somewhere on the steep upwards gradient shown in the orange (gelation) section

Due to the high data acquisition required to observe a 'sharp' step change, the data was subject to noise in the form of oscillations. To combat this, at points (other than where the step occurred), data were averaged to create a smooth plot.

Once the critical shear rate for each sample was determined, a new template job was created that assessed the respective relaxation time of all shake-gel samples. To achieve this, the following procedure was created: the shear rate was initially set to 10 s^{-1} for three minutes to attain a baseline measurement of viscosity in which the sample is known to be fully relaxed; the shear was stepped up to 1100 s^{-1} (a shear rate that exceeded all critical shear rates previously determined) and kept constant for five minutes; finally, the shear rate was stepped back down to 10 s^{-1} and left to run, once the viscosity equalizes and returned to the initial baseline measurement, the job was terminated.

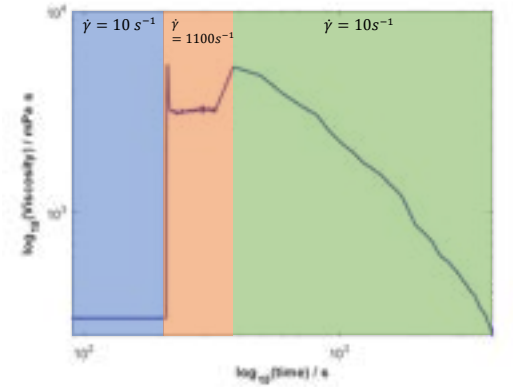


Fig. 2. *Example of a viscosity time plot for a shake gel sample when subjected to a multi-step change in shear rate. Showing the initial liquid baseline viscosity measurement, the gelation induction and relaxation time measurement.*

The period from which the shear was stepped back down to 10 s^{-1} to the point where the viscosity reached the initial baseline level was deemed as the relaxation time shown graphically in figure 2. This was determined mathematically using a MATLAB script that performed the following operations: first, the period in which overcoming of inertial effects was shown was extracted to minimise the error of the baseline measurement; noise that was present in the baseline measurement was accounted for by averaging of all points, hence the horizontal line show in the blue region in figure 2. Then, the baseline measurement average was extended in the time dimension to find the intersection with the relaxation curve, the time between the step down back to 10 s^{-1} and the time at which the intersection point was found was quantified as the relaxation time.

Results and Discussion

The presence of all tested additives allowed for the formation of shake-gels, i.e., mixtures that possess the same unique rheological properties found in 'normal' shake-gels. As expected, there exists a finite concentration of each additive for which the typical shake-gel properties are maintained.

The BSG showed qualitative properties consistent with that found in the literature such that, the colour of the liquid form saw a highly translucent white solution at low concentrations which upon application of shear a transformation to an opaque white gelled agglomerate with the relaxation time qualitatively occurring in a matter of seconds as desired.

Contrary to the rheometer data which observed the relaxation time of the BSG as 1005.79 s, after hand shaking the base shake-gel, the gel seemed to relax within seconds of formation when shear was no longer applied. This disparity is likely due to the rheometer shear rate of 10 s^{-1} exceeding a shear rate that enables the maximum rate of relaxation. Other factors that explain the disparity in visual observation of relaxation vs. quantitative measures involve the geometry of the rheometer apparatus, such that the U-shape slowed down the relaxation due to a higher surface area to volume ratio meaning the desorption of PEO from the silica surfaces was slowed.

At the BSG mixture composition tested in this work, SDS had a maximum concentration of 0.120 wt.%; above this point, the liquid form became too viscous to distinguish from the gelled phase with the gelation event not sufficiently evident to be classed as a shake-gel.

As expected, the addition of SDS to the shake gels increased the viscosity of the liquid form of the shake gel at all concentrations with the viscosity seeming to be independent of the concentration of SDS in the first four samples containing weight fractions of SDS in the range (0.2, 0.4), but the samples containing an increase in SDS at 0.8 and 1.2 wt.% showed a dramatic increase in liquid phase viscosity.

This research aimed to establish the potential for shake-gel in domestic applications, with one of those potential applications being as a soap for children. To be an effective soap, foaming is required to increase the surface contact between the soap and the area of the body that is being washed, with SDS being a common additive in domestic products to allow this foaming ability it was tested in shake

Another interesting factor of the SDS containing shake-gels was observations in colour. In their liquid form, all the samples containing SDS were colourless, i.e., no hint of a white hue was shown to be present in the base shake-gel. After shaking and forming the gel, the SDS samples also showed a significant increase in transparency (decrease in white colour) as compared to the base shake-gel, with samples containing 0.20, 0.24 and 0.28 wt.% showing a slight white tinge that decreased with increasing SDS concentration with the samples at 0.40, 0.80 and 1.2 wt.% having no observable colour in their gelled form.

An increase in SDS concentration above 0.4 wt.% showed a noticeable decrease in gel quality, in the sense that upon hand shaking a sample, while a gelation event did still occur, the cohesivity of the gel or formation of one large agglomerate was not as evident as in the base shake-gels and the SDS containing shake-gels at lower concentrations.

As a point of comparison, the quantitative results of the BSG were a CSR of 991.81 s^{-1} and a relaxation time of 1005.79 s.

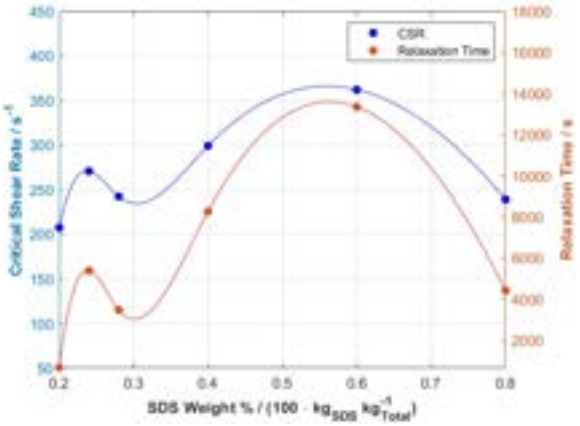


Fig. 3. A co-plot of both critical shear rate and relaxation time against SDS weight fraction (shown as a percentage).

The trend shown in Fig. 3 can be split into three distinct regions:

The first region is bound by the critical point shown as a maxima at the SDS concentration of 0.24 wt.% which can be defined as the critical concentration for the formation of polymer-bound micelles ($c_p$mc), below this point the concentration of SDS

is not sufficient to form polymer bound micelles, leaving the SDS anions freely existing in solution.

The second region is bound by the $c_p$mc and a second critical point is shown as a second (larger) maxima at an SDS concentration of 0.4 wt.%. This point can be defined as the polymer-bound micelle saturation point ($s_p$mc). Above the $c_p$mc, the formation of polymer-bound micelles occurs; these micelles present a high-density anionic surface charge from the organosulfate head forming the outer region of the micelle as shown in Figure x. The polymer-bound micelles electrostatically repel one another resulting in the coiled form of PEO expanding/extending and elongating. This expansion occurs up to the $s_p$mc at which point the polymer chain is saturated with micelles.

Region 3 defined concentrations at concentrations above the $s_p$mc, further addition of SDS results in the formation of free (unbounded) micelles. These unbounded micelles also present an anionic surface charge and electrostatically repel the polymer-bound micelles resulting in a contraction of the PEO chain back to its original shape. [23]

Research conducted by Witte and Engberts [25] found the values of the $c_p$mc and the $s_p$mc at 0.4 wt.% and 0.85 wt.% SDS respectively. However, these points were established from plots of both viscosity-concentration and viscoelasticity-concentration.

The value of these points for the SDS shake-gels tested in this study were found to be $c_p$mc = 0.240 wt.% and $s_p$mc = 0.600 wt.%. The lower value found for the $c_p$mc compared to that found in the literature for PEO-SDS mixtures is likely due to the presence of silica, or more specifically the presence of the silanol groups on the surface of the silica particles. These silanol groups dissociate a proton when in an aqueous environment, in the following manner:

$$-SiOH \leftrightarrow -SiO^- + H^+$$

leaving an anionic surface charge on the silica particles. Adsorption between silica and PEO occurs due to electrostatic attractions that form between these anionic silanol surface groups and the PEO chains. While partially shielded by the presence of Na^+ ions in the Ludox suspensions and dissociated from SDS when in an aqueous environment, the anionic silanol groups explain the increase in $c_p$mc whereby substitution of the silica adsorbed to the PEO chains requires an increased concentration of SDS to overcome the attraction between silica and PEO.

The reduced value of $s_p$mc is also likely explained by the anionic silanol groups whereby substitution of silica particles by the SDS micelles results in free (or less adsorbed) silica, electrostatically repelling the polymer-bound micelles contracting the polymer chains, meaning the point at which the contraction begins to occur may be lower than the point at which the polymer is saturated with SDS micelles.

This mode of action posed as an explanation for the increased viscosity and viscoelasticity in SDS-PEO solutions with increasing SDS concentration, which may also explain the change in CSR. For the gelation event to occur, shear is required to extend the PEO molecules from the coiled equilibrium state, allowing more area for adsorption to silica surfaces. At SDS concentrations between the $c_p$mc and the $s_p$mc, the expansion of the PEO coils due to the presence of polymer-bound micelles reduces the required amount of shear to fully expand the PEO chains allowing for adsorption with silica surface at significantly lower shear rates.

The polymer-micelle complex is broken down when subject to shear, this occurs at a shear rate when the hydrodynamic drag exerted on the polymer-bound micelle exceeds the force that binds the micelle to the polymer, essentially ripping the micelle off the polymer chain, analogous to how a sufficient wind speed is required for leaves to detach from the branch of a tree [15].

The amount of force that binds a micelle to the polymer depends on the total number of micelles bound to the polymer; increasing the SDS concentration from the $c_p$mc to the $s_p$mc leads to a decrease in the attraction between individual micelle-polymer forces.

Figure 3 shows a double plot of the CSR and the relaxation time of SDS containing shake gels. There is a strong relationship between critical shear rate and relaxation time. To quantify the correlation, the following statistical equation was used to calculate the correlation coefficient:

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2\Sigma(y-\bar{y})^2}} \qquad (6)$$

A correlation coefficient of r = 0.993 was calculated, suggesting a very strong relationship between CSR and the relaxation time. This implies that the interaction between SDS and PEO that accompany the gelation process occurs in reverse sequential order in the relaxation process.

Upon the initial formulation of the vanillin containing shake gels, 'good' shake gels formed such that upon shaking they had excellent agglomeration into a gelled ball that relaxed quickly, and this was the case at all vanillin concentrations tested for shake gels. The colour of the shake gel on the initial mixing shake gels was similar to that of the BSG i.e., a translucent white in the liquid form transforming to an opaque white in the gelled state.

Another observation following 24 hours of creating these shake-gels was a change in colour of all samples in the liquid state, where they had been found to have transitioned to a light red/brown colour; with the intensity of this browning increasing for the samples containing increasing concentrations of vanillin, hence it was evident that the colour change was due to a reaction occurring with the vanillin.

This colour change occurred due to the oxidation of vanillin which is known to happen in the presence of an alkaline environment [27]. This oxidation reaction may also explain why the shake gels containing large quantities of vanillin were permanently gelled the day following formulation.

Not only was a colour change observed, but a change in rheological properties also accompanied this reaction. After 48 hrs in the lab, a sample containing a greater concentration of vanillin that tested (at 0.317 wt.%) formed an inviscid permanent gel in its resting/equilibrium state those of which were excluded from any rheological measurements.

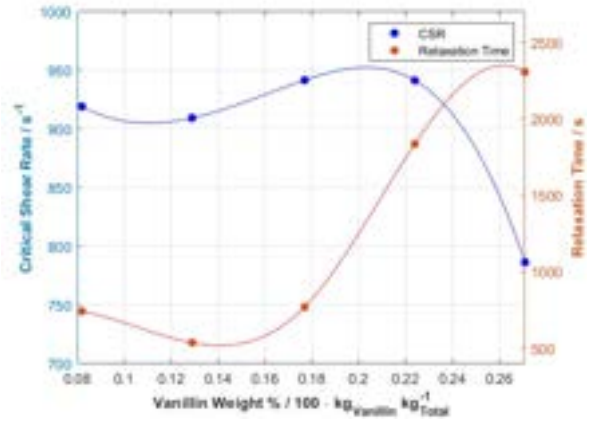


Fig. 4. A co-plot of both critical shear rate and relaxation time against Vanillin weight fraction (shown as a percentage).

Vanillin present showed a slight impact on the CSR values as compared to the BSG, at concentrations of 0.082 and 0.129 wt.% an approximate 10% decrease was seen; at intermediate concentrations of 0.177 and 0.224 wt.% a decrease of ~ 5% was shown; at the highest vanillin concentration measured at 0.271 wt.% a more significant decrease of ~ 20% was shown.

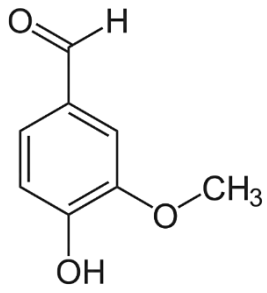


Fig. 5. Skeletal structure of Vanillin

The interaction of vanillin with PEO and vanillin with silica has not been assessed in the currently available literature, but insights can be postulated from the assessment of its chemical structure shown in Figure 5. The presence of a hydrophobic aromatic ring is likely to interact with the hydrophobic regions present in PEO. This interaction may reduce the extent of coiling PEO at equilibrium in solution, leading to a reduction in shear required to fully extend the chain easing the formation of the silica-PEO network that pertains to the gelled state of shake gels.

The low solubility of vanillin in water is reported as 11 mg/mL [26] and talks to the aforementioned hydrophobic nature of the vanillin molecule. F. Shakeel et al. measured the solubility of vanillin in different solvents including PEO-400 a polymer solvent containing molecules of PEO at a molecular weight of 400 g/mol; they report an increase in solubility of vanillin in PEO-400 at 4.29 x 10^{-1} as compared to that in the water of 11 mg/mL at 25°C [26]. This increased solubility of vanillin in a PEO solvent (as compared to water) affirms the speculation of hydrophobic PEO-vanillin interactions impact on the CSR.

Relaxation time, at low vanillin concentration, was shown to have reduced compared to that of the BSG by ~ 20 to 40%, but at concentrations above 0.224 wt.%, a large increase in relaxation time is observed. A reason for this behaviour is that at high concentrations of vanillin the hydrophobic interactions between vanillin and PEO may present themselves in the gelled state and interrupt the breakage of the polymer-silica network that occurs during relaxation.

Shake-gels would only form with the NaCl additive at low concentrations and this is likely due to the shielding of charges on the silanol group by Na^+ ions, preventing the formation of the silica-PEO complex seen in the gelled state.

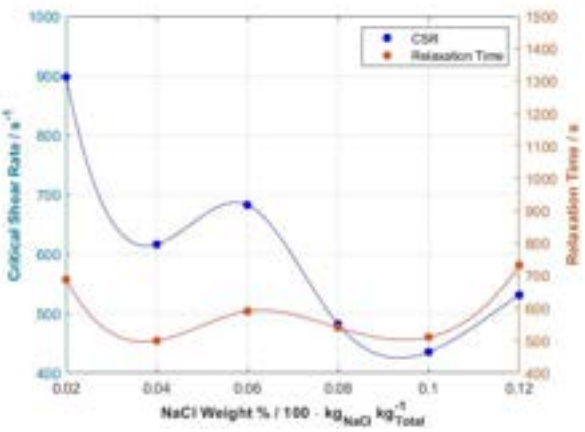


Fig. 6. A co-plot of both critical shear rate and relaxation time against NaCl weight fraction (shown as a percentage).

The salt concentrations assessed were in the range (0.02, 0.12 wt.%), and the upper limit was determined by an inability to form shake gels at greater salt concentrations. The intensity of the white colour associated with the liquid state, increased with increasing salt concentrations up to 0.12 wt.%.

Above a salt concentration of 0.12 wt.% and below 0.60 wt.%, a milky white emulsion formed that when subjected to shear saw no transition into the gelled state.

At salt concentrations exceeding 0.60 wt.%, a phase separation was observed, whereby a single white agglomerate coexisted with a colourless liquid; this phase separation is known as the cloud point and is dependent on the salt concentration [17]. In non-ionic surfactants and glycols such as PEO, glycol solutions generally show a reduction in cloud point in more saline fluids, thus a shake gel at 0.6 wt.% reduces the cloud point of the shake-gel mixture to approximately 25 °C (the lab temperature in which the samples were prepared), which is consistent to what was found at high concentrations of NaCl in shake gels.

One observation made in this study with shake-gels that is seldom mentioned in the literature is the loss of shake-gel properties over time in those formed without additives.

24 hours after the initial formulation (of shake gels without additives), the expected shake gel properties are seen with the change from a low viscosity to an almost colourless liquid to a highly viscous agglomerated white gel. When left to sit for an extended period of time (approximately 2-3 weeks), found a loss in this ability to undergo gelation, with shake gels remaining in their liquid state when subjected to shear and this was true for all samples of shake-gels tested without additives. This is of concern when considering a shake-gel as a marketable product that may be held in stores for an extended period of time, as a loss in functionality leads to the loss of the selling point of the product.

However, in the case of the salt and the SDS containing shake-gel samples, this loss of shake-gel properties was not observed. These samples preserved their shake-gel ability over a kept period of several weeks. This is an essential finding when considering applications of shake-gels and furthers the potential of its use for both domestic purposes and industrial applications.

The error associated with weighing the samples arises due to the ± 0.0005 g inaccuracy of the balance. To quantify the relative error associated with the mass balance, the following is an example of a relative error calculation associated with the measurement of 5.007 g of PEO solution (the amount used in each shake-gel sample).:

$$relative\ error = \frac{0.0005\ g \times 2}{5.007\ g} \times 100 = 0.020\ \%$$

An error this small is unlikely to have impacted either the qualitative or quantitative properties of any shake-gel sample. This error can be further reduced by the measurement of larger quantities of individual components of the shake-gel.

The absolute error associated with Thermo Scientific rheometer measurements is ± 0.000005 Pa·s. To put this error into perspective, for a viscosity measurement of 10.20489 Pa·s (lowest recorded viscosity in all of the samples), the relative error is in the order of 10^{-5}.

According to Sigma-Aldrich's website, the Ludox TM-50 suspension has a 'quality level' of 200. This denotes increased control over a substance with a standard 100 quality level (with the scale ranging between 100-600 in steps of 100). Sigma-Aldrich gives the purity of the Ludox TM-50 suspension a range between 49 to 51%. To put this error into perspective, the quantity of Ludox used in each sample was 8.246g meaning the silica content ranges by ± 0.165 g. resulting in a relative error of 4%. This is somewhat significant in the case of comparing shake-gel properties when formulated with different batches of Ludox but for comparison of the same batch, the error is irrelevant.

The 900,000 molecular weight PEO powder has the same 200 'quality level' as the Ludox.

The Sigma-Aldrich SDS powder has a quality level of 200, and a purity listed as ≥ 98.5%. The greatest error associated with SDS purity concerns the shake-gel with the greatest quantity of SDS; this shake-gel contained 0.200 g of SDS resulting in an error of ± 0.003 g and a relative error of 1.5%, this is also a significant error when dealing with small quantities. This relative error could have been reduced by formulating larger samples.

The sodium chloride supplied by VWR Chemicals has a ≥ 99.9% purity according to the label. This led to a negligible error.

The Vanillin powder, the Sigma-Aldrich website lists a quality level of 200, but this time the purity is also shown at ≥ 98.5%.

Sigma-Aldrich lists the molecular weight of the supplied PEO as a nominal value, there is no way to know (with our current resources) if the listed MW of 900,000 is accurate. Since the coaxial CCB26 cup has vertical serrations, leftover residue was impossible to clean fully out of the serrations, this may have led to slight errors in viscosity measurements.

It is also known that the Ludox solution contains differing silica particle sizes for different batches of Ludox, possibly distorting results [5].

Conclusion

We have demonstrated that shake-gels can be formulated when in the presence three different quaternary additives: SDS, vanillin and NaCl, with additive containing shake-gels showing the same unique properties to that found in standard shake gels.

SDS shake-gels exhibited the desired foaming when subject to shear and acted as a preservative, maintaining the shake-gel properties over an extended period of time. The vanillin shake-gels possessed a significant aroma at very low concentrations, but the vanillin was subject to oxidation, disrupting the quality of the shake-gels. The NaCl shake-gels, similar to the SDS shake-gels showed a preservation in shake-gel properties over an extended period of time.

The rheological properties of shake-gels containing SDS are consistent with those found in the literature studying the interaction between SDS and PEO.

Outlook

Formulation of a shake-gel that possesses the combined properties of scent, foaming ability and preserved quality over time would be required to decide the suitability as a children soap and to potentially start the process of bringing such a product to market.

Testing of shake-gels that possesses two or even all three additives hopefully combine all the beneficial properties of each additive, which should be the next step, following this study.

References

1. Cabane et al. (1997). Shear induced gelation of colloidal dispersions. *Journal of Rheology* [online], 41(3), pp.531-547. Available from: Journal of Rheology [accessed 14 December 2022].
2. Zebrowski et al. (2003). Shake-gels: shear-induced gelation of laponite–PEO mixtures. *Colloids and Surfaces A: Physicochemical and Engineering Aspects* [online], 213(2-3), pp.189-197. Available from: ScienceDirect [accessed 14 December 2022].
3. Can, V. and Okay, O. (2005). Shake gels based on Laponite–PEO mixtures: effect of polymer molecular weight. *Designed Monomers and Polymers* [online], 8(5), pp.453-462. Available from: Taylor & Francis Online [accessed 14 December 2022].
4. Ramos-Tejada, M. and Luckham, P. (2015). Shaken but not stirred: The formation of reversible particle – polymer gels under shear. *Colloids and Surfaces A: Physicochemical and Engineering* Aspects [online], 471(1), pp.164-169. Available from: ScienceDirect [accessed 14 December 2022].
5. Collini et al. (2018). The effects of polymer concentration, shear rate and temperature on the gelation time of aqueous Silica-Poly(ethylene-oxide) "Shake-gels". *Journal of Colloid and Interface Science* [online], 517(1), pp.1-8. Available from: ScienceDirect [accessed 14 December 2022].
6. Kawasaki, S. and Kobayashi, M. (2018). Affirmation of the effect of pH on shake-gel and shear thickening of a mixed suspension of polyethylene oxide and silica nanoparticles. *Colloids and Surfaces A: Physicochemical and Engineering Aspects* [online], 537(1), pp.236-242. Available from: ScienceDirect [accessed 14 December 2022].
7. Huang, Y. and Kobayashi, M. (2020). Direct Observation of Relaxation of Aqueous Shake-Gel Consisting of Silica Nanoparticles and Polyethylene Oxide. *Polymers and Nanomaterials: Interactions and Applications* [online], 12(5), pp.1-13. Available from: MDPI [accessed 14 December 2022].
8. Tian et al. (2021). Quantitative analysis of the structural relaxation of silica-PEO shake gel by X-ray and light scattering. *Polymer Testing* [online], 104(1), pp.1-6. Available from: ScienceDirect [accessed 14 December 2022].
9. Hong, Y. and Zhao, T. (2021). Factors that affect Relaxation time of Shake-Gels. *Department of Chemical Engineering, Imperial College London*. pp.1-9.
10. (Huang et al., 2022). Conditions for Shake-Gel Formation: The Relationship between the Size of Poly(Ethylene Oxide) and the Distance between Silica Particles. *Sol-Gel Functional* Materials [online], 27(22), pp. 1-10. Available from: MDPI [accessed 14 December 2022].
11. Sigma-Aldrich. LUDOX® TM-50 colloidal silica [online]. Available from: https://www.sigmaaldrich.com/GB/en/product/aldrich/420778 [accessed 14 December 2022].
12. Sigma-Aldrich. Poly(ethylene oxide) [online]. Available from: https://www.sigmaaldrich.com/GB/en/product/aldrich/189456 [accessed 14 December 2022].
13. Yehye et al. (2015). Understanding the chemistry behind the antioxidant activities of butylated hydroxytoluene (BHT): A review. *European Journal of Medicinal Chemistry* [online], 101(1), pp.295-312. Available from: ScienceDirect [accessed 14 December 2022].
14. Hammouda, B. (2013). Temperature Effect on the Nanostructure of SDS Micelles in Water. *Journal of Research of the National Institute of Standards and Technology* [online], 118(1), pp.151-167. Available from: National Library of Medicine [accessed 14 December 2022].
15. Bird et al. (1987). Dynamics of polymeric liquids. Vol. 1&2. 2nd ed. New York: Wiley.
16. Cabane, B. and Duplessix, R. (1985). Neutron scattering study of water-soluble polymers adsorbed on surfactant micelles. *Colloids and Surfaces* [online], 13(1), pp. 19-33. Available from: ScienceDirect [accessed 14 December 2022].
17. Kumar et al. (2012). Electrical conduction mechanism in NaCl complexed PEO/PVP polymer blend electrolytes. *Journal of Non-Crystalline Solids* [online], 358(23), pp.3205-3211. Available from: ScienceDirect [accessed 14 December 2022].
18. Rubio, J. and Kitchener, J. (1976). The mechanism of adsorption of poly(ethylene oxide) flocculant on silica. Journal of Colloid and Interface Science [online], 57(1), pp.132-142. Available from: ScienceDirect [accessed 14 December 2022].
19. Barclay-Nichols, S. (2016). *Understanding The Vanillin Villain.* Available from: https://www.wholesalesuppliesplus.com/handmade101/learn-to-make-articles/understanding-the-vanillin-villain.aspx
20. Anklam et al. (1997). Oxidation behaviour of vanillin in dairy products. *Food Chemistry* [online], 60(1), pp.43-51. Available from: ScienceDirect [accessed 14 December 2022].
21. Rhee et al. (2006). Effect of flavors on the viscosity and gelling point of aqueous poloxamer solution
vanillin binding to the end chains of PEO. *Archives of Pharmacal Research* [online], 29(1), pp.1171-1178. Available from: SpringerLink [accessed 14 December 2022].
22. Shakeel et al. (2015). Solubility and thermodynamic function of vanillin in ten different environmentally benign solvents. *Food Chemistry* [online], 180(1), pp.244-248. Available from: ScienceDirect [accessed 14 December 2022].
23. Brackman, J. (1991). Sodium Dodecyl Sulfate Induced Enhancement of the Viscosity and Viscoelasticity of Aqueous Solutions of Poly(ethylene oxide). A Rheological Study on Polymer-Micelle Interaction. Vol. 7. 3rd ed. Washington, D.C.: American Chemical Society Publications.
24. Thermo Scientific, (2014). *HAAKE MARS Rheometer Instruction Manual*. Available from: ResearchGate [accessed 14 December 2022].
25. Witte, M. and Engberts, J. (1989). Micelle—polymer complexes: Aggregation numbers, micellar rate effects and factors determining the complexation process. Colloids and Surfaces [online], 36(3), pp.417-426. Available from: ScienceDirect [accessed 14 December 2022].
26. Yalkowsky et al. (2010). Handbook of Aqueous Solubility Data. 2nd ed. Boca Raton: CRC Press, Boca, pp.480
27. Frenkel, C and Havkin-Frenkel, D. (2006). The physics and chemistry of vanillin. *Perfumer & Flavorist* [online], 31(1), pp.28-35. Available from: ResearchGate [accessed 14 December 2022].

28. Lionti-Addad, S. and Di Meglio, J.M., (1992). Stabilization of aqueous foam by hydrosoluble polymers. Sodium dodecyl sulfate-poly(ethylene oxide) system. Vol. 8. 1st ed. Washington, D.C.: American Chemical Society Publications.
29. Fennema et al. (1996). Preservation and Physical Property Roles of Sodium in Foods. *Strategies to Reduce Sodium Intake in the United States* [online]. Available from: National Library of Medicine [accessed 14 December 2022].
30. Hammouda, B. and Ho, D.L. (2007) "Insight into chain dimensions in PEO/Water Solutions," Journal of Polymer Science Part B: Polymer Physics, 45(16), pp. 2196–2200. Available at: https://doi.org/10.1002/polb.21221.

Separation of Biopharmaceuticals using Nano-templates

Asjad Ahmed Khan and Raksha Lalwani
Department of Chemical Engineering, Imperial College London, U.K.

Abstract Crystallization is a widely used technique across various industries today. Controlling the nucleation stage, the onset of crystallization remains a challenge especially for peptide crystallization across the biopharmaceutical industry. Heterogeneous template nucleation method is being researched for this purpose as they not only assist with polymorph selection but also have an impact on induction times. This report investigates the effect of surface porosity on the induction time via the use of silica nano templates of 50, 30, 10, 6nm pore diameters. A supersaturation of 1.2 was used across all experiments and the silica mass loading was kept at a constant 10% for each set of heterogeneous nucleation experiments. Induction times were determined from the de-supersaturation curves obtained by monitoring the concentration via use of the IR sensor in-situ. A linear trend between induction times and pore diameter was obtained where the best results were 5.5 and 6.1 times faster induction times with 10nm and 6nm pores respectively compared to the time taken for homogeneous nucleation. The results obtained can be used to increase the efficiency of separation of biopharmaceuticals by reducing process times and increasing their cost effectiveness by reducing operational costs of batch crystallization separation techniques currently employed in the pharmaceutical sector. Underlying mechanisms behind hetero-template nucleation like surface porosity, surface chemistry and epitaxy were also discussed and suggestions for further analysis were provided.

1.0 Introduction

Crystallization is a technique used in a wide range of industries like pharmaceutical, nutraceutical, paints and even semiconductor industries.[1] This process consists of two main steps: nucleation and growth phase. The nucleation phase, the rate determining step, consists of the formation of a new phase due to the supersaturation of the medium. This is followed by the growth phase which is characterized by the evolution and agglomeration. Crystallization is one of the most crucial final steps in active pharmaceutical ingredient manufacturing (API) as the crystals' shape, size and structure determines the downstream operations needed to achieve >99% purity. This high-quality requirement is key in the pharmaceutical industry for drug manufacturing as the presence of any impurities can weaken the therapeutic effect.

The pharma market size was estimated to about 1.42 trillion USD in 2021 which is an almost 4-fold increase from the 390 billion USD market size in 2001[2]. The COVID-19 pandemic was one of the many triggers for this significant increase in the past two decades. The growing need for more advanced drugs is a huge motivation to make pharma processes more cost efficient and sustainable. Separation and purifications steps often characterize a large proportion of the total manufacturing costs and hence can be a key area of improvement. Traditionally techniques like liquid chromatography have been used for downstream purification because crystallization does not allow the control of crystal type and size.

The surface chemistry, pore size and amount of Hetero-seeding affect the level and time taken for nucleation. They can be used to better control the crystallization process.

Diglycine form three polymorphs: α, β, and γ. α is the easiest and thermodynamically stable polymorph to crystallize while β and γ require many recrystallizations.

The potential of heterogenous crystallization in enabling better control of crystals' characteristics using nano templates as seeds will guide this study.

2.0 Background

2.1 Crystallization

Crystals are solids comprising of orderly arranged atoms, ions or molecules repeated in three dimensional arrays. The angles between faces of crystals of the same compound are identical and characteristic of that material. Crystallization has cemented itself as an important separation technique capable of producing highly pure products even from solutions with significant impurities. This importance is further reinforced by the low energy input in its operation compared to other separation units like distillation which are much more energy intensive.[3] The conventional separation for peptides involves chromatography which is not only more expensive but also much slower compared to crystallization. Crystallization's driving force is supersaturation which can be described as a thermodynamically unstable state at which the solution contains more solute than that present at saturation.

The first step of crystallization is Nucleation, which can be described as the genesis of crystalline nuclei forming a site upon which additional particles deposit resulting in crystal growth and can further evolve by aggregation and agglomeration. The degree of supersaturation influences both nucleation and crystal growth.

Nucleation can be classified into two main routes, primary and secondary. Primary nucleation is when crystals formation is driven solely by solution properties in the absence of crystals (of the material itself). Primary nucleation can be further classified

into homogeneous and heterogeneous. Homogeneous nucleation is when crystals form due to the chemical potential resulting from supersaturation alone. Heterogeneous nucleation results from the presence of insoluble materials in the system which could involve the walls of a vessel, the impeller or as in this report's case, heterogeneous seeds. Secondary nucleation happens in the presence of crystals or homo-seeds and involves attrition which is the breakage of existing crystals into smaller fragments due to impact from the impeller for instance. These fragments then act as nuclei which grow to form more crystals.

Nucleation is a very important stage in the process as it governs the physical characteristics of the crystalline structure like habit, morphology, size, number, and flaws. Therefore, control of nucleation is regarded as paramount because if nucleation could be controlled then all the aforementioned characteristics too could be controlled. Traditional models for nucleation include the famed Classical Nucleation Theory (CNT) and the newly adapted Two-step Nucleation model (TSN). According to CNT, the particles of the substance begin aggregating together until they reach a critical radius, beyond which the aggregate starts functioning as nuclei and crystal growth occurs. The CNT makes two assumptions: the nuclei are spherical, and that the interfacial surface tension is isotropic. However, recently, the TSN (Vekilov, 2010), explained the process of nucleation with two-steps instead of the one in CNT. TSN's rate determining step in homogeneous nucleation is characterized by the formation of pre-nucleation clusters of size in the range of several hundred nanometers within dense liquid droplets suspended in the solution. In the second step of TSN, these pre-nucleation metastable structures act as hetero surfaces for the onset of nucleation similar to the heterogenous case. So, the homogeneous nucleation is not quite homogenous as was previously expected. This is seen as promising as it also resolves some longstanding mysteries of protein crystallization and also opens up an avenue for nucleation control. For example, on of the long standing mysteries was why the theoretically predicted nucleation rates were much larger than the ones actually measured through experiments. This can be explained through the presence of two steps in the nucleation mechanism compared to just one in classical theory. The first step is theorized to be the formation of pre nucleation clusters which is much slower than the step after which is the formation of stable nuclei. This first step is the rate determining step and if the second step alone is considered, then there is a match between actual and predicted nucleation rates (Parambil, 2019).

This places an even stronger emphasis on the mechanistic understanding on the role of surfaces and how they influence nucleation. As this appears to be the only way to accurately model crystallization heterogeneously. These mechanisms include Surface Epitaxy, Chemistry and Porosity.

2.2 Surface chemistry

This mechanism refers to solute-template and solvent-template intermolecular interactions. (Parambil, 2019) The surface chemistry heavily influences both these interactions which can impact template assisted nucleation (Frank & Matzger, 2017). Some of these specific interactions between the functional groups of template surface and the crystallizing molecule can reduce the interfacial free energy required for nucleation to occur. This can result in increased nucleation rates. Moreover, research has been done on diglycine-silica surface chemistry by (Vivek et al., 2021). The research studied the hydrogen bond donor (HBD) and hydrogen bond acceptor sites (HBA) present in diglycine. There were 2 HBDs and 3 HBAs found within diglycine where it was also discovered that there existed a hydrogen bonding complementarity between the diglycine and silica surface. Furthermore, the hydrogen bond lifetime (10 – 70 ns) was compared to the time needed for one molecule to be added to the crystal (21 ps). As the latter was much shorter, this implied that there was a higher probability to form stable crystal nuclei. These two effects combined resulted in improved nucleation rates for diglycine with silica template.

2.3 Epitaxy

The interplay of similar molecular arrangements between the molecule and surface is considered for this mechanism. Once a good match between the two is struck, interfacial free energy is reduced which favors nucleation. A study specific to diglycine has not yet been conducted, making its effect uncertain. This mechanism also has a profound effect on polymorph selection (Parambil et al., 2019)

2.4 Surface Topography

Topography covers a broad range of factors including surface geometry, confinement and pores. Some of these factors like surface geometry have overlaps with mechanisms like epitaxy. This report, however, will only focus on the surface porosity aspect of topography. Surface Porosity in the context of nano templates refers to the presence of nanosized pores on the hetero-surface and encompasses both number of pores as well as their size distribution. Pores have been theorized to help trigger nucleation through generation of local 'zones' of elevated supersaturation. Additionally, pores of different sizes have also exhibited selectivity favouring the nucleation of some molecules over others mainly based on their size. This mechanism unlike surface chemistry remains largely unexplored specially for

diglycine. Surface porosity will be the main area of investigation in this report. (Shah, 2012)

3.0 Methodology

3.1 Experimental Set-up

The experimental setup comprised of Mettler Toledo EasyMax 102 which had two reactors of volume, 20ml and 50ml. A temperature probe and a Mettler Toledo ReactIR 15 IR probe was inserted into the head of the vessel. This allowed automated in situ measurements of temperature and concentration in fixed intervals of temperature, so no external disturbance to the system, i.e., no sampling and handling was required.

Prior to every set of experiments, the IR probe was calibrated to obtain a spectrum for the background atmosphere. This was done to ensure that the probe head was clean, was only measuring the sample and not any contaminants. A bottom stirrer bar was inserted in the reactor instead of an overhead stirrer to avoid any collision with the IR probe. Additionally, the Mettler Toldeo ReactIR 15 apparatus was topped up with liquid nitrogen every 12 hours to cool the internal optical fibres to prevent overheating causing potential malfunctions.

A saturated 40ml solution of diglycine in water (solvent) was prepared at 40℃ using the solubility data solubility data (Guo et al., 2022) with a concentration of 284.28mg ml^{-1}. To ensure the complete dissolution of the sample, it was placed in the 50ml reactor for an hour at 60℃ followed by filtration with the 200nm Nylon membrane syringe filter and added to a fresh reacting vessel. Filtration was performed as a preventative measure against secondary nucleation due to undissolved diglycine particles.

3.2 Mass Loading of Silica

For the heterogeneous nucleation experiments with silica nano-templates of pore sizes 6nm, 10nm, 30nm, and 50nm, the mass loading (m_L) was kept constant at 10% of the maximum theoretical amount of silica expected to crystallize out of the solution (m_{crys}) at 1.2 supersaturation. To calculate, m_L: solubility of diglycine at 40℃ (S_{40}) and at 32.7℃ ($S_{32.7}$) and the total volume of the solution (V) is required.

$$m_{crys}\,[mg] = (S_{40} - S_{32.7})\,[mg\ ml^{-1}] \cdot V[ml]$$

$$m_{crys} = (284.28 - 236.90) \cdot 40 = 1900\,mg$$

$$m_L\,[mg] = 0.1 \times m_{crys} = 190mg$$

3.3 Script for Rig control

The iControl software that controls the rig required a script with the sequence of steps to be provided. The saturated solution was set and maintained at 45℃ for 120 minutes to allow the system to equilibrate and ensure complete dissolution of diglycine. The sample was then cooled to 32.7℃ over 25 minutes to achieve a supersaturation of 1.2 (Guo et al., 2022) allowing to operate within the metastable zone, hence enabling nucleation. This supersaturation was desirable as smaller supersaturations yield longer induction times consequently leading to larger measurable difference in induction times with use of nano templates. To obtain the induction time marking the onset of nucleation and the complete de-supersaturation curve as diglycine crystalises out of the solution, the temperature was maintained at 32.7℃ for 300-400 minutes. The steps were designed to conduct three repeats of every experiment to minimise the error.

3.4 Interpretation of the Infrared (IR) Spectrum for Transmittance

IR spectrum provides a fingerprint region for chemicals by measuring the transmittance of functional groups which is characteristic/unique for every compound. The experiment was set to record the IR spectrum for the sample every 10 seconds. The spectra for both water and diglycine solution were obtained and superimposed for comparison.

Initially, the two peaks of the diglycine spectra, at 1384cm^{-1} and 1262cm^{-1} that corresponds to the flat lined region in the water spectrum from 1000 – 1500 cm^{-1} were observed to determine the peak corresponding to diglycine and that corresponding to silica. The area under the peak at 1384cm^{-1} changed with time, while that under the peak at 1262cm^{-1} remained constant. Hence, the 1384 cm^{-1} peak was identified as the diglycine peak and the 1262cm^{-1} was identified as the silica peak as the silica concentration in solution remains constant.

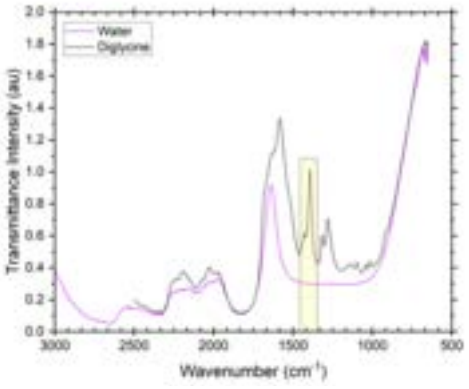


Figure 1: IR Spectrum superimposed for Water (pink) and Diglycine Solution (black) with the peak at 1384cm^{-1} highlighted as the diglycine peak

The area under the diglycine peak (highlighted) observed in **Figure 1** was recorded from the diglycine solution's spectrum produced every 10 seconds. Area under the peak is directly proportional to the concentration of diglycine. This area was then plotted against time to obtain the 'S' shaped supersaturation curves as the diglycine continued to crystalise, hence reducing the concentration.

3.5 Induction Time Determination

Figure 2 shows the tangent method used to determine the induction time 't_{ind}' from the supersaturation curve. This involved drawing a tangent of the greatest slope to the curved part of the graph and intersecting it with the straight line in the beginning. The time at which the temperature stabilizes to the set-point of 32.7℃ , was taken as the t_i. The point of intersection of the tangents from the region of the greatest slope and the stable part before nucleation was taken as t_n, time at onset of nucleation.

Here, $t_{ind} = t_n - t_i$

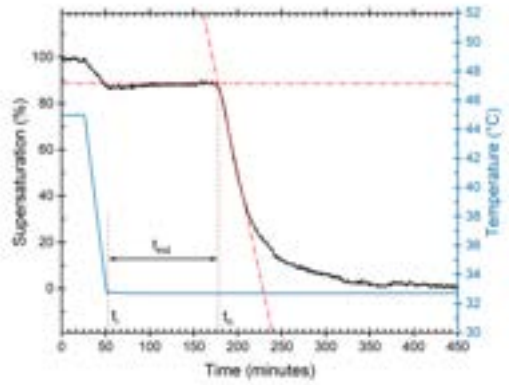


Figure 2: Tangent Method for Induction time measurement shown for the example of Homogeneous nucleation experiment (induction time found to be 185 minutes)

3.7 Hydrodynamic Diameter determination

Litesizer 500 by Malvern works based on the principle that particles in solution that are in Brownian motion diffract the incident beam of light causing a shift in its frequency which is recorded by the diffractometer. The Disposable PS cuvette was used for the experiment. This data alongside correlation curves was then used to generate a particle size distribution graph as well as the average particle size. This technique was utilised to determine the hydrodynamic diameter of diglycine.

3.8 Powder X-Ray Diffraction (PXRD)

The Powder X-ray diffractograms were recorded on XPRT-PRO diffractometer system with PANalytical measurement program using a copper anode as radiation source (λ= 1.541nm) at 40 mA and 40 kV. The scans were performed at a frequency 0.013° 2θ min^{-1} in the range between 5° and 35°. This technique results in the generation of a patterns of intensity against angle (2θ) for the silica, homogeneously and heterogeneously formed diglycine. This ultimately aided in the identification of the polymorph of diglycine formed by a comparison with previously obtained polymorphic patterns.

3.9 Microscope Imaging

The diglycine crystals obtained after using a vacuum filter to separate the crystals from the solution, were visually analysed using an Olympus CX-41 microscope (Essex, UK) under magnifications of 5x and 10x. The GT Vision GXCAM HiChrome MET display (Suffolk, UK) connected to the microscope, was used to capture the images of the crystals, determining the crystal habit and size. These results with the aid of PXRD results were then used to confirm the obtained polymorph.

4.0 Results and Discussion

4.1 Solid State Characterization

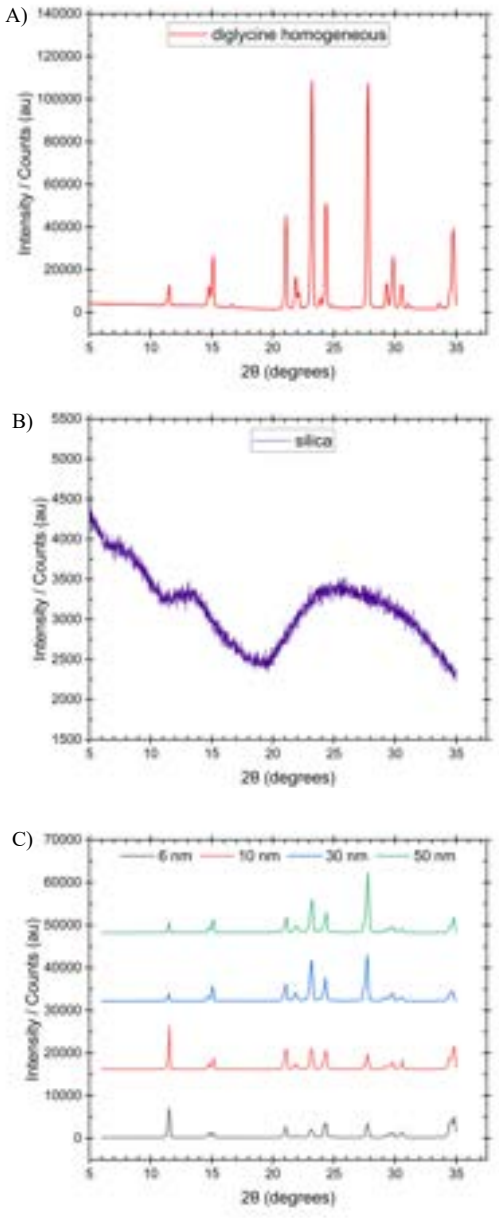


Figure 3: Diffractograms obtained from PXRD for:
A) Crystals obtained in Homogeneous experiments
B) Silica beads
C) Crystals obtained for heterogeneous experiments

A total of 6 diffractograms were obtained for this characterisation pertaining to the homogeneously formed diglycine as shown in **Figure 3.**

The silica diffractogram was hollow and showed no characteristic peaks which is consistent with its amorphous nature, this ensures that it does not interfere with the heterogeneously formed diglycine patterns. The diffractograms, belonging to all diglycine crystals investigated, had peaks at the same 2θ values. From this it can be inferred that they had the same preferred orientations. Some variance was observed in relative peak intensities between the 50nm, 30 nm and the 10nm, 6 nm samples. This could be attributed to the non-uniform consistency of the powder sampled and could be mitigated by grounding the powder more finely in future experiments. Nevertheless, upon comparison of the diffractograms to literature, it was confirmed that the polymorph obtained was stable α-form of Diglycine.

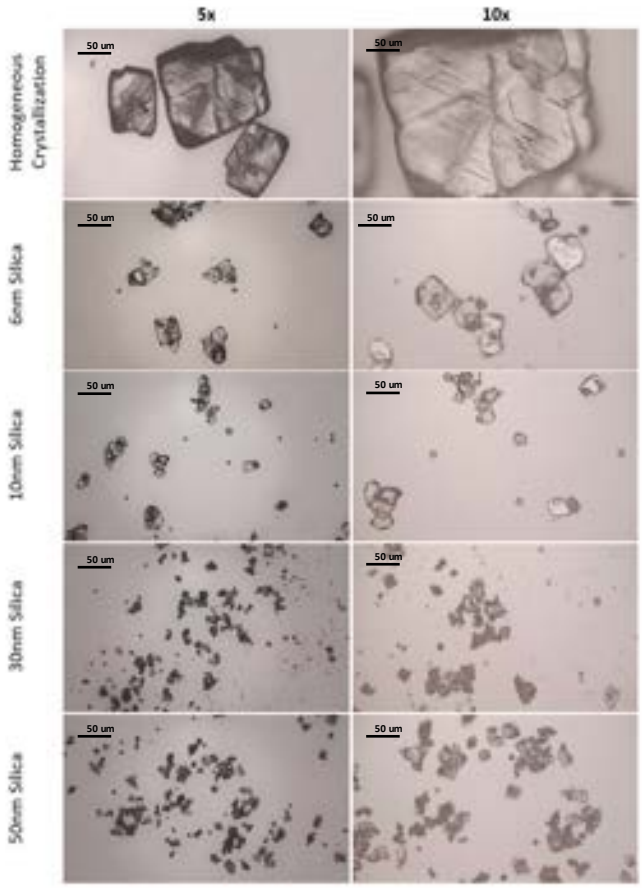


Figure 4: Microscope images obtained for all the experiments' crystals obtained

The obtained crystals were further observed under an inverted light microscope, **Figure 4**. A plate like crystal shape was observed. This observation could be made more clearly with the homogeneous sample due to its larger size. This is consistent with the α-Diglycine polymorphic structure. Another observation was that the size of crystals was smaller for larger pores which could have implications in applications where crystal size distribution (CSD) is important and should therefore be further investigated.

4.2 S-shaped Curves

Figure 5 displays all the de-supersaturation 'S' shaped curves stacked for comparison. Homogenous nucleation shows the longest induction and overall de-supersaturation times. This clearly demonstrates that the addition of seeds of any diameter positively impacts the induction time. Decreasing the pore diameter of the silica seeds results in the graphs shifting to the left indicating further reduction in induction times.

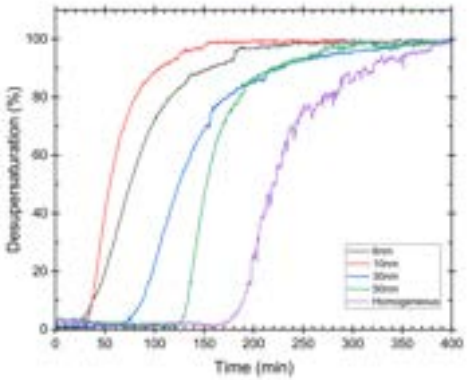


Figure 5: De-supersaturation curves plotted against time for the homogeneous case (purple), heterogeneous cases using silica with pore size 6nm (black), 10nm (red), 30nm (blue), and 50nm (green)

The results extracted from **Figure 5** are shown in the **Table 1**. Additionally, the surface area (SA) corresponding to each bead was also calculated using specific surface area (a) and mass loading (m_L) which is represented in the table.

$$SA\ [m^2] = a\ [m^2 g^{-1}] \cdot m_L\ [g]$$

For convenience, the induction times (t_{ind}) have also been interpreted as improvement factors (i) for all pore sizes

$$i = \frac{Homogeneous\ t_{ind}}{Heterogeneous\ t_{ind}}$$

Table 1: Surface area, Induction time, Error and Improvement Factors for all cases

	Surface area [m^2]	Induction time [min]	Error (St. Dev) [min]	Improvement Factor
Homo	-	165	22.9	-
50 nm	57	125	14.6	1.3
30 nm	152	87	18.6	1.9
10 nm	760	30	6.0	5.5
6 nm	855	27	8.0	6.1

All the experiments were conducted a minimum of 3 time while other were repeated 4 times. These repeats omit the anomalous results which were very short induction times. This could be due to internal solvent evaporation in vessel that can lead to crystal formation on the head of the reactor, potentially causing a crystal falling back into solution and acting as a homo-seed, resulting in a shorter induction time than expected. These results were taken as anomalous and only the results from other repeats were taken into consideration while calculating mean induction times and standard deviations.

A trend can be observed between surface area and improvement factors showing a positive correlation.

4.3 Induction Time vs. Pore diameter

To get a more graphical depiction of how induction time varies with pore diameter, **Figure 6** was constructed.

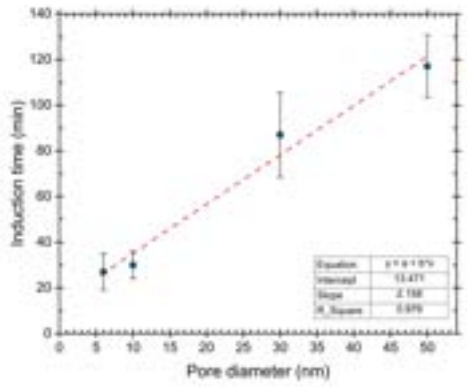


Figure 6: Induction Time vs. Pore diameter using values summarized in Table 1

The plot shows a positive linear trend between induction time and pore size. This trend is also rather strong reinforced by a 0.98 mean square error value. From the results obtained thus far it can be deduced that not only is the addition of silica better but also upon decreasing its pore size this effect is more profound. This can be attributed to not only a decrease in pore size but also presence of a greater number of pores due to an increase in surface area as seen in **Table 1**.

Due to this simultaneous increase in both variables, it could be said that surface porosity has a significant impact on induction times, however, one variable cannot be solely attributed as the reason behind this improvement. One way to further explore whether surface area or pore size had a more significant contribution, two sets of experiments could be performed.

The first would be to vary pore size and the mass loading in a way such that the surface area remains constant, and the other experiment should involve keeping the pore size constant but variation of surface area through a change in loading. This could provide a better indication of what factor; pore size or surface area has a more profound effect.

4.4 Hydrodynamic Diameter Measurement

To investigate the relation of pore size with a reduction in induction time, hydrodynamic diameter of diglycine was measured using a zeta sizer. The **Figure 7** shows the distribution obtained.

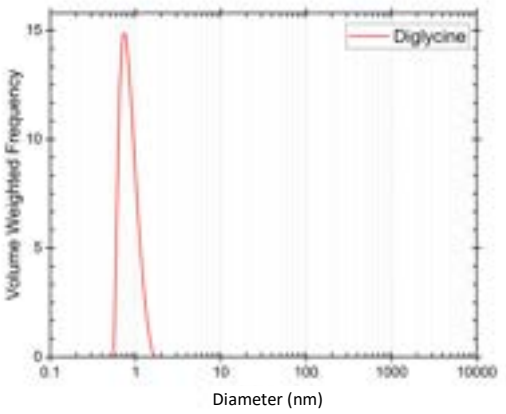


Figure 7: Hydrodynamic diameter determination from the Litesizer 500

A size of 0.94nm was obtained for the hydrodynamic diameter. The unit cell for diglycine comprises of 4 repeated units of diglycine molecule (Vivek et al., 2021) giving a rough size of 4nm. This size is very comparable to the pore sizes tested specially 6 and 10 nm which yielded the best results.

4.5 Mechanisms behind Nucleation

The mechanism by which pores facilitate nucleation through increased zones of local supersaturation can be visualised with the schematic below **Figure 8**.

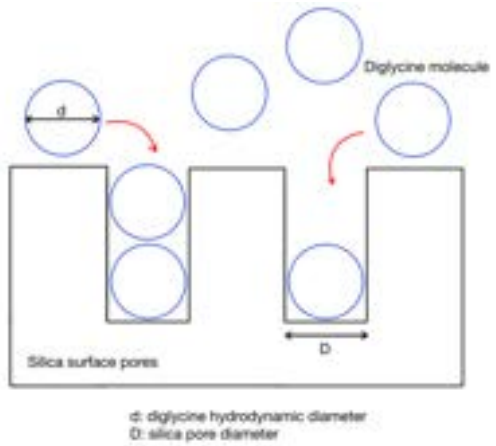


Figure 8: Schematic Representation of Silica Pores

It should be noted that the schematic in **Figure 8** uses simplified geometry to illustrate the mechanism at play and the pores are non-uniform and not cylindrical. Diglycine molecule is also represented as a circle for simplicity. The idea that pores generate these elevated supersaturation zones stems from the molecules getting deposited and stuck in the pores which results in a concentration differential near the surface and the bulk. This increases supersaturation close to the surface increasing nucleation rates. After the nucleation is initiated in the pores and a stable nuclei core is generated, crystal growth continues as more diglycine molecules get layered which can eventually engulf the seed itself. This can be visually observed in the microscopic images in **Figure 4** where it appeared that seeds were inside the heterogeneously formed crystals.

This effect from pores could be coupled with surface chemistry mechanisms as well (Diao, et al., 2011) Diglycine can form hydrogen bonds with the silica surface as it has both HBD and HBA sites. These hydrogen bond interactions lower the interfacial free energy needed for nucleation and help stabilise the formed nuclei core. This can be visualised in the **Figure 9** .

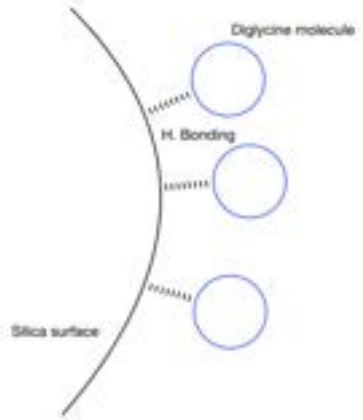


Figure 9: Schematic for HBD and HBA explaining the Surface chemistry mechanism

Research that has already been done on diglycine and silica has shown hydrogen bonding complementarity between them which allows them to interact even more effectively. Additionally, the lifetime effect also further enhances the nucleation as the lifetime of a hydrogen bond interaction (10-70ns) was much larger than the measured time for a molecule to attach to the growing crystal (21ps) (Vivek et al., 2021). This allows more molecules to attach overall resulting in a higher probability of forming a stable crystal nucleus. The hydrogen bond complementarity and the lifetime effect can therefore assist in the improvement factors of 1.3 and 1.9 for 50nm and 30nm samples respectively. These pore sizes were around 10 times larger than the unit cell size of 4nm and were unlikely to support the improvement through the pore mechanism described earlier. It could then be the case that surface chemistry mechanism dominated in those samples resulting in the improvement factors.

Lastly, epitaxy could also have played a role in the improvement. Epitaxy mechanism lowers the interfacial free energy needed to start nucleation upon a good match of molecular arrangement in molecule and surface. This could be thought of as a lock and key mechanism similar to then enzyme-substrate mechanics. This effect could improve nucleation rates but plays a more significant role in polymorph selection.

It is unlikely that only one of the three mechanisms discussed was responsible for the improvement in induction times. It would be a good assumption to assume they all played a role and contributed in combination with each other to yield the results. The contribution of each however is unknown specially at a quantitative level. It is essential to further investigate these mechanisms individually through a previously made suggestion. To run two set of experiments measuring induction times by varying surface area and pore diameter independently to see which had a more significant impact. Pore diameter would suggest that the surface topography mechanism influenced more heavily where surface area would suggest either or both the surface chemistry and epitaxy were at play. Quantification of these could support optimisation of operating conditions in biopharma industry for instance where then an appropriate mass of seeds would be loaded preventing waste and will result in shorter induction times allowing more batches to be processed. This would reduce the cost of running the process consequently making medicines more affordable.

5.0 Conclusion

This study showed that overall heterogenous nucleation has a shorter induction time than homogeneous nucleation for all pore sizes investigated. The effect of variation of surface porosity was observed by experimenting with four

pore sizes: 6nm, 10nm, 30nm, and 50nm which gave induction time improvement factors of 6.1, 5.5, 1.9, 1.3 respectively. Therefore, the smaller pore sizes of 6nm and 10nm were more efficient compared to the larger pore sizes of 30 and 50nm. This can be attributed to the comparability between the size of the small pores and the unit cell size of the α-polymorph observed, i.e. 4nm.

The polymorph formed was identified through PXRD and further confirmed by observing them under the microscope. Additionally, the size of crystals observed decrease with an increase in pore-size and homogeneous crystallization produced the biggest crystals. This can potentially be justified by the comparability between the pore size used and the size of α-polymorph nuclei. However, further experimentation is necessary to confirm this hypothesis.

Overall, the results from this study can be utilised to make biopharmaceutical separation of diglycine from mixtures more efficient through an effective reduction in operating time upon utilising nano-templates. The reduction in operational time would effectively reduce the operating cost of crystallization separation and purification techniques used in the pharmaceutical industry. This would effectively help reduce the cost of expensive drugs and make healthcare more affordable in the future.

6.0 Outlook

This study can be further expanded to gain a better understanding on ways to make the separation of diglycine in biopharmaceutical processes more efficient. Firstly, surface porosity of hetero-seeds should be investigated at higher supersaturations to check whether the improvement in induction time through hetero-seeding is still as significant when the supersaturation driving force increases. This information would be helpful in applications that have different operating conditions.

It would also be useful to find the crystal size distribution (CSD) for the homogeneous and heterogeneous sets of experiments. The information obtained through this would be useful to design the required down-processing steps like filtration, milling and grinding when diglycine in crystallised in industry.

Additionally, analytical experiments to realise the impact of surface chemistry and epitaxy on the nucleation of diglycine would provide a better insight into the mechanisms at play. Relative impact of these factors independently is not possible to explore since nucleation is a very complex process where these mechanisms of surface porosity, chemistry and epitaxy often work in combinations.

Lastly, since, pharmaceutical processes generally have more than one peptide present in mixtures that need to be separated, it is useful to conduct further experiments to find the selectivity of silica nano-templates to diglycine. This will allow the targeted crystallization of peptides thereby achieving better control over the industrial crystallization process.

7.0 References

[1] *Parambil, et al. (2019) Template-induced nucleation for controlling crystal polymorphism: from molecular mechanisms to applications in pharmaceutical processing, Royal Society of Chemistry. (Accessed: December 1, 2022).*

[2] The Statistics Portal (no date) Statista. Available at: https://www.statista.com/markets/412/topic/456/pharmaceutical-products-market/#statistic1 (Accessed: November 22, 2022).

[3] Crystallization from solutions and melts, Chemical Process Equipment (2009) Revised Second Edition). Gulf Professional Publishing. Available at: https://www.sciencedirect.com/science/article/pii/B9780123725066000186?via%3Dihub (Accessed: December 2, 2022).

[4] Vekilov, P.G. (2010) The two-step mechanism of nucleation of crystals in solution, Nanoscale. The Royal Society of Chemistry. Available at: https://pubs.rsc.org/en/content/articlelanding/2010/NR/c0nr00628a (Accessed: December 9, 2022).

[5] D. S. Frank and A. J. Matzger (2017) Crystal Growth Des. https://pubs.acs.org/doi/10.1021/acs.cgd.7b00593 (Accessed: December 7, 2022).

[6] Vivek Verma et al. (2021) Studying the impact of the pre-exponential factor on templated nucleation, Faraday Discussions. Royal Society of Chemistry. Available at: https://pubs.rsc.org/en/content/articlehtml/2022/fd/d1fd00101a (Accessed: December 9, 2022).

[7] Guo, M. et al. (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. Elsevier. (Accessed: October 25, 2022).

Time Series Prediction for Deep Learning Methods of Dynamical Systems in Chemical Engineering

Tigran Minasian and Ross Barrett
Department of Chemical Engineering, Imperial College London, U.K.

Abstract: The rapid advancement of machine learning and deep learning methods in recent years has provided an opportunity for a paradigm shift within the discipline of Chemical Engineering. Complex, non-linear dynamical systems with arbitrary relationships between inputs and outputs are not only difficult to predict and optimise but are also ever-present within Chemical Engineering, however, neural networks and their applications in deep, data-driven, learning provide researchers with potentially powerful tools to solve these problems. This paper investigates the applications of five different architectures of neural networks popular within data-driven learning, namely "Feed-Forward", "1 Dimensional Convoluted", and three forms of "Recurrent" neural networks and assesses their resilience to systems with a high degree of complexity generated through adding both stochastic and measurement noise to two known dynamical systems of differing linearity. The study found that the neural network architecture of the "LSTM", a "Recurrent" neural network variant, performed the best overall, with a high resilience to complex systems, but was extremely computationally expensive. "1 Dimensional Convoluted" and "Feed-Forward" neural networks performed well when predicting multivariate and univariate systems respectively and could be used as an alternative to the "LSTM" in cases where accuracy is less important. It is worth noting that this study recommends the further integration of mixed network systems, namely "Bayesian" and "Transformer" neural networks as these are popular in literature and could improve the resilience of predictive models of dynamical systems within Chemical Engineering by introducing probabilistic predictions and the ability to input sequences in parallel respectively. Additionally, further studies into the Kalman filter to use system estimate covariance when predicting complex, noisy data could further refine future predictive models.

Keywords: Data-Driven, Neural Networks, Dynamical Systems, Resilience, Complexity

1. Introduction and Objectives

1.1 Introduction and Motivations

Dynamical systems exist everywhere within the fields encapsulated by the acronym "STEM", in Chemical Engineering, these could be any one of a multitude of chemical reactions occurring within a variety of process systems. The ability to understand and model such systems is key to optimising systems such that, for example, more desired product is gained from the overall process system.

This understanding and modelling of systems manifests in the form of mathematical equations that analytically describe the system and its variables as a function of time, equations that were historically found by humans painstakingly combing through vast amounts of data obtained through observing interactions and environments within a system to find its common governing principles (Mussmann, 2021). This process can be arduous, especially when a system is complex and non-linear, resulting in a difficult to find analytical solution, a problem that is ubiquitous in Chemical Engineering (Hanyu Gao, 2022). An emerging solution to this common issue is the application of deep neural networks and data-driven learning, which is a range of computational methods that have the ability to learn the behaviours of a system more effectively than humans by understanding and interpreting patterns and trends within datasets.

Given the data-rich nature of Chemical Engineering, it seems strange therefore that computational technology has not been more of a leading factor within the field. However, this can be attributed to the fact that it is only recently that there has been a large amount of research into machine learning, deep learning, and other data-driven learning processes spurred on by the rapid advancement of computer technology and data storage infrastructure (Hanyu Gao, 2022).

1.2 Objectives

This paper will aim to further investigate the practical applications of data-driven learning using different neural network architectures to predict both simple and complex dynamical systems. A focus will be placed on assessing not only their overall ability to solve dynamical systems, but to do so while the data is purposefully altered to simulate both measurement noise and stochastic noise, reflecting potential real-life scenarios where data is wrong due to inaccurate measuring instruments and disturbances as a result of intrinsic system stochasticity. An emphasis will be placed upon finding neural networks that are resilient to poor data inputs and difficult data to assist in solving the issue of systems with complex and arbitrary links between inputs and outputs, a problem found throughout Chemical Engineering.

2. Background and Scope

2.1 Dynamical Systems

2.1.1 Dynamical Systems Overview

A dynamical system is described as a system which evolves over time according to a fixed rule and is typically described with an ordinary differential equation. The state conditions could be any number or variety of physical conditions, such as concentration or temperature for example. It is formally defined as a state space , a set of times , and a rule that specifies how the state evolves with time (Nykamp, 2008). Within this paper, deep learning is used to predict both simple linear and complex non-linear dynamical systems, with the following two systems chosen to for this purpose.

2.1.2 Mass-Spring System with Damping

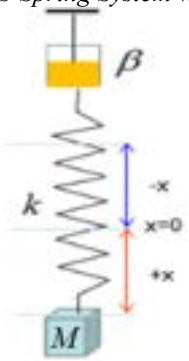


Figure 1: Illustration of the damped mass-spring system (ShareTechnote, 2015)

The simple, single-variate, linear system chosen is a mass attached to the end of a spring that oscillates vertically with a damping force that opposes the motion of the mass as seen in *figure 1*. This is represented as the following linear ODE:

$$M\frac{d^2x}{dt^2} + kx + \beta\frac{dx}{dt} = 0 \quad (1)$$

Where M is the mass of the suspended object, k is the spring constant, x is the displacement, is the damping coefficient, and t is time. The deep learning methods will be used to predict the displacement from the set point as a function of time.

This is a classically investigated dynamical system within the general field of deep learning applications to dynamical systems, thus selecting this made logical sense as it allowed for the results of this paper to be easily compared to other papers.

2.1.3 Biohydrogen Production System

The complex, multi-variate, non-linear system chosen as part of this investigation is the hydrogen production from naturally occurring cyanobacteria, Cyanothece 51142 (Figure 2), which turns solar energy into hydrogen under aerobic conditions.

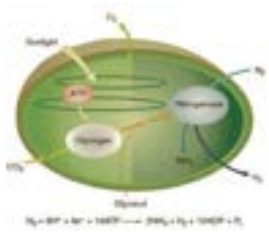


Figure 2: Cyanothece 51142 cell synthesizing Hydrogen (Urquhart, 2010)

The following equations have been adapted from "Optimal Operation Strategy for Biohydrogen Production" (Del Rio-Chanona, 2015) and will be used to predict the interlinked concentrations of Biomass (2), Extracellular Nitrogen (3), and Intracellular Nitrogen (3):

$$\frac{dx}{dt} = \mu_m + x + \frac{N}{N+K_N} - \mu_d \cdot x^2 \quad (2)$$

$$\frac{dN}{dt} = -Y_{nx} \cdot \mu_m \cdot x + \frac{N}{N+K_N} + F \quad (3)$$

$$\frac{dq}{dt} = -Y_{qx} \cdot \mu_m \cdot \frac{N}{N+K_N} - \mu_m (4)$$

Where N is nitrate concentration, q denotes normalised, x denotes biomass concentration, Y denotes yield, F denotes feed, and t denotes time.

This is an example of a potentially difficult system within Chemical Engineering in which the inputs and outputs have a difficult analytical relationship. This dynamical system will be investigated to assess how effective deep learning is when applied to more advanced and complex problems that are within the scope of Chemical Engineering.

2.2 Data-Driven Learning

Data-driven learning can be described as the process of learning how a specific task, system, or process works by using large amounts of data. In these processes, large amounts of example data, the "training data" is used to train the system that learns and refines predictive outputs (Shah, 2020). An increasingly popular example of this is "Chat GPT-3", an AI chatbot that uses machine and deep learning to generate responses to a wide range of prompts and questions. "Chat GPT-3" was trained using data from sources such as books, Wikipedia, and articles giving it an almost human-like response to many different questions (Mills, 2022). The data-driven methods explored in this paper revolve around the usage of artificial neural networks that resemble, and are intended to mimic, the human brain, a naturally occurring biological neural network.

2.3 Artificial Neural Networks

2.3.1 Artificial Neural Network Overview

The data-driven methods explored in this paper revolve around the usage of artificial neural networks, or ANN's, that resemble, and are intended to mimic, the human brain, a naturally occurring biological neural network (Wahlström & Dernsjö, 2021). Figure 3 illustrates this idea.

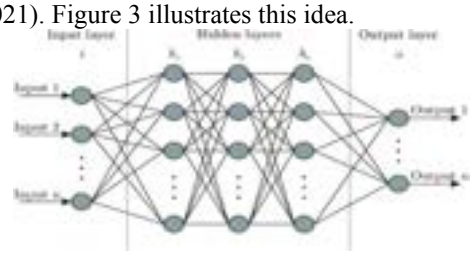


Figure 3: A basic feed-forward neural network (Facundo Bre, 2017)

Though there exist different artificial neural network architectures, there remains components that each have in common. The first is "input", which is data put into the model for learning. The second is "weight", this aids in organising each variable by impact of contribution on the output. The third is the "transfer function", which is where all the inputs are combined into a single output variable. The fourth is the "activation function", which decided if a specific neuron should be activated based on how important the neuron's input is to the prediction process. The fifth is the "bias", responsible for shifting the value given by the activation function. Finally, the last common feature is "layers", these refer to the general layout of individual neural network nodes arranged into different layers that receive raw data (input layers), output predictions (output layers), and all the layers in between (hidden layers) (H2O.ai, 2022).

The "learning" aspect of the process takes place through multiple iterations of forwards and backwards propagation of data. The initial inputs are run through the neural network and produce a predictive output, this is then compared to the desired output using a loss function such as the "mean squared error" (5).

$$MSE = \frac{1}{n} \cdot \sum(predictions - actual)^2 \quad (5)$$

the gradient of the loss with respect to the weights is then calculated using the chain rule. This is done by starting at the output layer and working backwards through the layers of the neural network, using the gradients of the loss with respect to the outputs at each layer to calculate the gradients with respect to the weights. The weights are then updated using an optimisation algorithm, such as stochastic gradient descent. This reduces the loss and improves the performance of the model (Brownlee, Machine Learning Mastery, 2016).

Neural networks take on a variety of architectures, and it is understood that within the scope of time series prediction, feed-forward neural networks (FNN's), convolutional neural networks (CNN's) and recurrent neural networks (RNN's) are effective for even the more complex dynamical systems with multiple inputs that have arbitrary mappings to outputs (Brownlee, Deep Learning for Time Series Forecasting, 2018). It is for that reason that these architectures have been chosen to predict the two chosen systems within this paper.

2.3.2 Feed-Forward Neural Networks

Multi-layer feed-forward neural networks, trained with a backpropagation learning algorithm, are reportedly the most widely used neural network, and its general structure can be seen in figure 3, with each neuron in one layer receiving an input from every neuron in the previous layer. In a feedforward neural network, the data flows only in one direction, from the input layer to the output layer, and there are no loops or connections between neurons in different layers. This makes feedforward neural networks easier to train and faster to run compared to other types of neural networks, such as recurrent neural networks, which have loops in their architectures. However, due to the lack of connections between neurons, FNN's may be less well suited to time series prediction as they are unable to retain data from previous time steps (Daniel Svozil, 1997).

2.3.3 Recurrent Neural Networks

Recurrent neural networks differ from the aforementioned FNN by having not only forward connections, but a hidden state to memorise the previous inputs and sequences. These hidden states, shown in the RNN's architecture in figure 4, enable the network to retain information from previous time steps and use it to process the current input data.

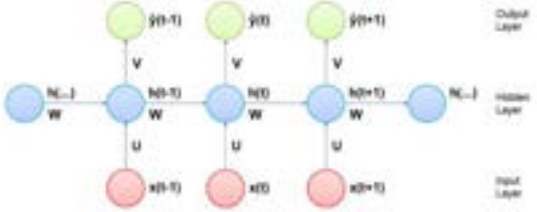


Figure 4: The general structure of an RNN, the vector x(t) is the input, y(t) is the output, and h(t) is the hidden state at time t which acts a memory for the network that is calculated based on both the current input and the previous time step's hidden state (Pra, 2020)

This aids the RNN in finding complex patterns within time series input data, however, not only are RNN's computational expensive to train, but they also suffer from a vanishing or exploding gradient problem, where the hidden layer parameters either do not change much or lead to numeric instability. RNN's also suffer from a weak memory and are unable take several past elements into further predictions (Pra, 2020).

2.3.4 Long Short-Term Memory (RNN Variant)

In order to combat the vanishing gradient problem with RNN's, the long short-term memory unit was developed to improving the gradient flow within the network by replacing a hidden layer. This is achieved by the LSTM's four unique features that are shown in figure 5. The first is a cell state *c(t)*, this represents the network's memory, and brings information along the entire sequence. The second is the forget gate, this will decide what elements from previous time steps should be kept or "forgotten". The third is an input gate, which will decide what information to add to the current time step to the information from that of the previous time step. Lastly, there is the output gate, this decides the value of the next hidden state and passes this and the new cell state to the next time step.

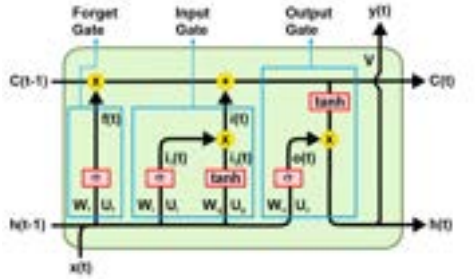


Figure 5: The general structure of an LSTM unit, the vector x(t) is the input, y(t) is the output, h(t) is the hidden state and c(t) is the cell state (Pra, 2020)

The LSTM is the most capable of all RNN types in regard to retaining long term dependencies in data, however, as such, it is also the most computationally intense RNN variant (Pra, 2020).

2.3.5 Gated Recurrent Units (RNN Variant)

The gated recurrent unit (GRU's) are a relatively new RNN variant similar to the LSTM in that it replaces a hidden layer in order to solve the vanishing gradient problem. This unit is more simplistic, having two gates to the LSTM's three, and is less effective at retaining long-term dependencies within data. This being said, the GRU is simpler and faster to both train and run than the traditional RNN and its variants in this paper.

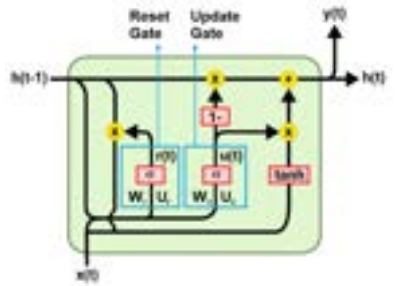


Figure 6: The general structure of a GRU unit, the vector x(t) is the input, y(t) is the output, h(t) is the (Pra, 2020)

Figure 6 illustrates the GRU and its three comprising features, the first being a reset gate, this decides how to combine the new input with the previous memory by deciding the amount of the previous time steps information which can be forgotten. The second feature is an update gate which allows the model to determine how much information from the previous time steps to pass onto the future. The last feature is the memory *h(t)*, which brings information along the sequence and passes to the next time step.

2.3.6 1D Convoluted Neural Networks

1D (1 dimensional) convolutional neural networks (CNN's) are able to process sequential data, such as time series data. They are composed of multiple layers of interconnected neurons and only one spatial dimension (i.e., width) and no height or depth. They are typically comprised of three elements, the first of which is a convolutional layer that applies convolutional filtering to extract potential features and patterns from the data it is analysing. The second is a pooling layer which serves to reduce the size of a data series while ensuring that the important patterns and characteristics of the data identified by the convolutional layer are preserved. The final element is a fully connected area that lies at the end of the network which maps the features extracted by the layers before it into coherent values as an output. Figure 7 displays this architecture and elements.

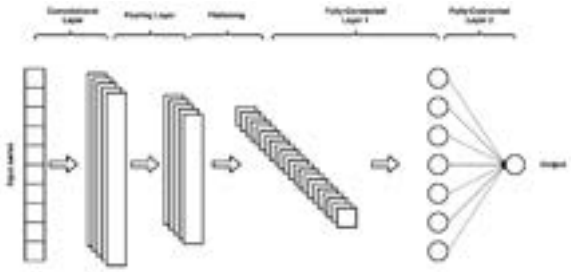


Figure 7: The general architecture of a 1D CNN (Lewinson, 2020)

CNN's are able to identify patterns independent of the time component and are considered noise-resistant. Additionally, CNN's are computationally less expensive than RNN's and in some cases can perform better. However, by design CNN's can struggle with long term dependencies, which are patterns that span over long time periods and are essential to predicting a time series (Lewinson, 2020).

3. Methodology

3.1 Introduction to the Method

The utilisation of dynamical systems with known analytical solutions representable as ODE's is a conscious choice designed to compare the

predictions generated by each neural network architecture trained upon noisy and imperfect data to the "real" system, thereby allowing for a measurement of the true resilience of the deep learning method. Both dynamical systems will be analysed in the same way, with the same noise applied at the same defined point for each neural network. Each part of the methodology applies to each combination of neural network and dynamical system, creating 10 total pairings, and will serve to achieve the objective goal of understanding the resilience of deep learning methods.

3.2 Data Generation and Hyperparameter tuning

The ODE's for the dynamical system being used were integrated across a time period which would allow enough data to be generated for a pattern to emerge (in the case of the MSS), or for the reaction of the system to go to completion (in the case of the biohydrogen production problem). This yielded deterministic datasets, upon which hyperparameter tuning was conducted to define the models to be used for the prediction of noisy data.

The generated data of length n was then processed to be formed of input windows of length 5, and dimensionality 0 which corresponded to a single output, the next "input".

- E.g., if the dataset was [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10] and the window size was 3, X_1 would be [0, 1, 2] and y_1 would be [3]. X_2 would be [1, 2, 3].

Table 1 below shows the hyperparameters used for the MSS:

Algorithm	Hyperparameters/Model Summary
Simple FNN	Dense (U = 32, ID=5, ReLU) Dense (U = 16, ReLU)
Simple RNN	Input Layer, (window size = 5, inputs = 1) SimpleRNN(U = 5), Dense(U = 5, ReLU)
LSTM	Input Layer, (window size = 5, inputs = 1) LSTM(U = 15), Dense(U = 5, ReLU)
GRU	Input Layer, (window size = 5, inputs = 1) GRU(U = 5), Dense(U = 5, ReLU)
1D CNN	Input Layer, (window size = 5, inputs = 1) CNN1D(U = 5), Flatten Dense(U = 5, ReLU)

Table 1

Where *U* represents the number of units in the layer and *ID* gives the input dimension. Since the MSS was treated as a univariate system, in all cases, the input consisted of a sequential window of 5 inputs and the output layer was a 1-unit Dense layer with a linear activation function. The learning rate was 0.01 and 7 epochs were run. The solver used was Adam.

For the biohydrogen production system, the principles for hyperparameter tuning stayed are the same. The increased complexity of the system demanded that NNs were deeper and had more units in each layer. The number of data points and input window size also had to be increased as a result. Furthermore, since there were 3 measured (and output) variables, the input layer gained one dimension to allow multiple inputs, and the linear output layer was changed to have three outputs. Table 2 shows the hyperparameters used.

Algorithm	Hyperparameters/Model Summary
Simple FNN	*It was found that predicting multivariate systems with multivariate outputs was impractical with the FNN*
Simple RNN	Input Layer (window size = 100, inputs = 3) SimpleRNN(U = 32), Dense(U = 64, ReLU) * 3
LSTM	Input Layer, (window size = 100, inputs = 3) LSTM(U = 64), Dense(U = 64, ReLU) * 3
GRU	Input Layer (window size = 100, inputs = 3) GRU(U = 32), Dense(U = 64, ReLU) * 3
1D CNN	Input Layer (window size = 100, inputs = 3) CNN1D(U = 64), Flatten Dense(U = 64, ReLU) * 3

Table 2

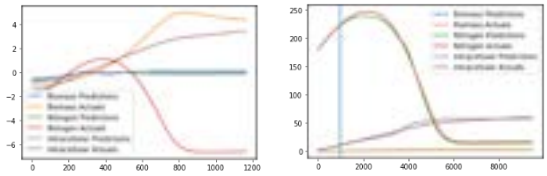


Figure 8: The same model, before and after tuning hyperparameters.

3.3 Addition and testing of noise

Noise was added in two steps. The magnitude of random process disturbance was taken as its variance. The disturbance was included in the integration step, in the form of a spontaneous change in variable (normally distributed with mean equal to the point of disturbance and the variance being the quantifying metric of the noise). The integration was then continued with the "disturbed" value being the new initial condition for the solver. This was then looped to introduce multiple points of process disturbance to the system, at random times. This was done to decrease the likelihood of the algorithm "expecting" disturbance at a certain timestamp after training. Testing was started with a variance equal to ~1/2 the test set's mean and increased until the model broke. The model could be considered a failure if:

- The root mean-squared error was deemed too high,
- The model predicted something impractical for application,
- The model was underfitted/overfitted, or any combination thereof.

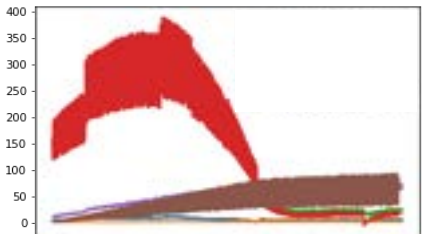


Figure 9: Example dataset for the multivariate system with both process disturbances and measurement noise

At the level of stochasticity *prior* to model failure, additional **measurement noise** was added through giving each data point a new value which was uniformly distributed. The key metric here was the range. For the MSS, a constant range was used. In the case of the multivariate system, due to the varying orders of magnitude of the different variables, a proportional range was deployed. The same principles for system failure were applied in the tests for system resilience against measurement noise.

4. Results

Results were found which were indicative of the speed vs accuracy of the different algorithms, their adaptability to process disturbances and resilience to measurement noise.

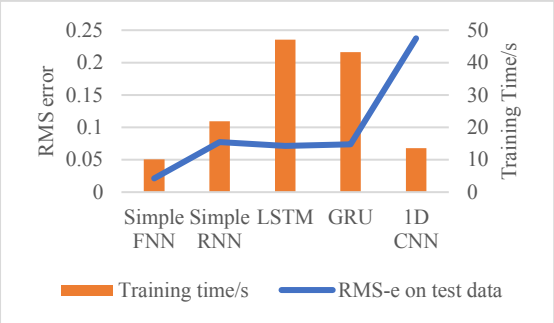


Figure 10: Graph showing the average training time vs the root mean square error for each algorithm

First, data was collected about the nature of each model's adaptability towards process disturbances.

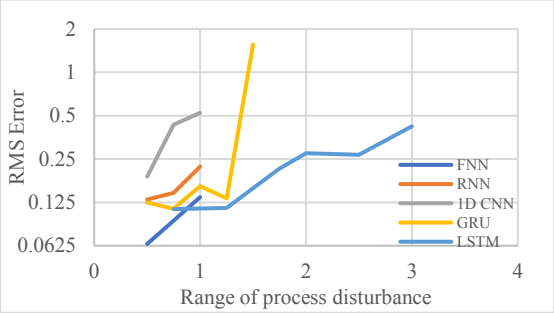


Figure 11: Graph showing the increasing RMS-e as a function of increasing process disturbance for MSS

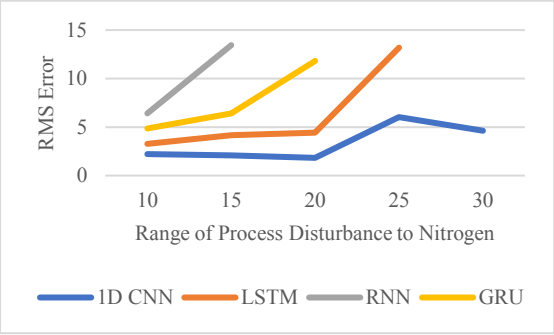


Figure 12: Graph showing the increasing RMS-e as a function of increasing process disturbance for biohydrogen

It is important to note that although it may seem that some algorithms appear to be better than others from the raw error data, this is only one factor in the judgement of the suitability of a model. Some good examples of models which yield low RMS-e values

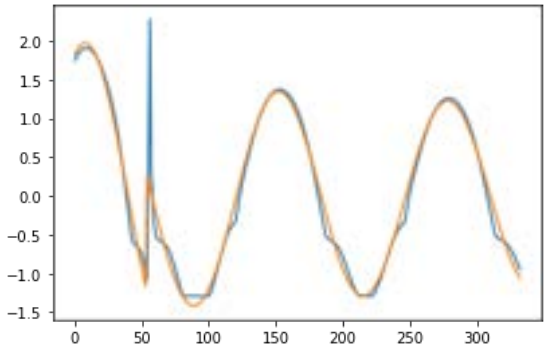


Figure 13:2 Simple RNN, PD = 1. RMS-e = 0.2213 Unacceptable level of overshoot and data is underfitted despite pleasing MSE scores

but are unsuitable include Figure 13, which shows substantial overshoot in the prediction set in the event of process disturbance, and Figure 14, which predicts values of biomass concentration orders of magnitude away from the real value.

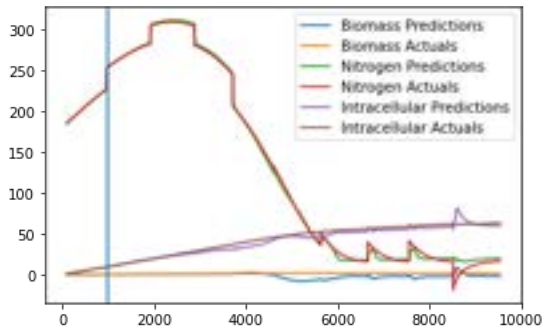


Figure 14:3 1D CNN, PD = 30, RMS-e = 4.6312. Overall trend seems strong but is let down by negative values of Biomass. (Nitrogen is also negative. However, this is a result of the method of adding process disturbances and not a symptom of a failing algorithm)

After this, data was collected on the breaking point of models with measurement noise on top of process disturbances.

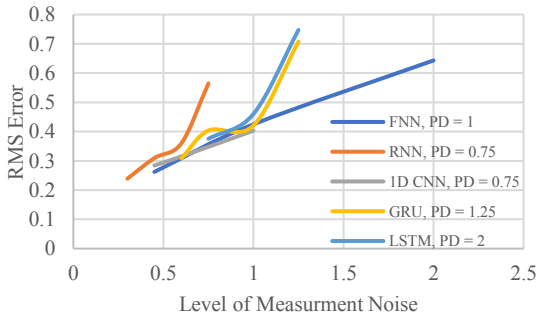


Figure 15: Univariate with measurement noise

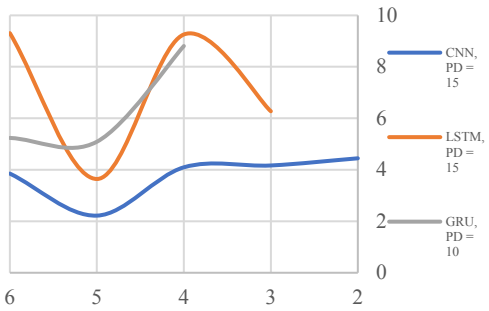


Figure 16: Multivariate with measurement noise (inverse)

Some interesting plots from this series included:

Univariate FNN

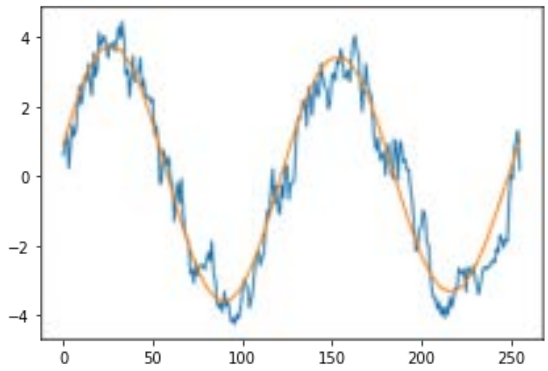


Figure 17: FNN, PD = 1, MN = 2: Very robust algorithm, ideal for solving linear systems

Simple RNN

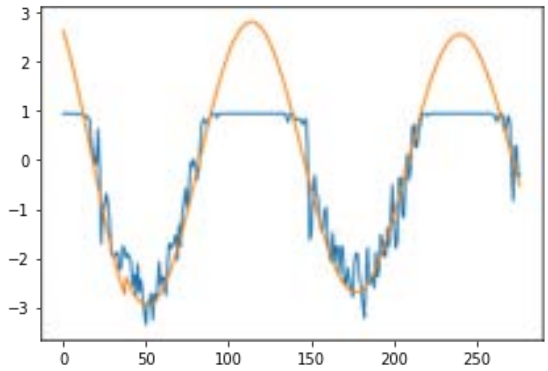


Figure 18: Simple RNN, PD = 0.75, MN = 0.75. Data is underfitted

CNN

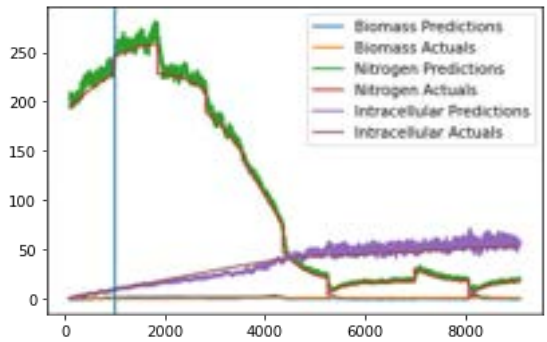


Figure 19: 1D CNN, PD = 15, MN = 1/2: Excellent prediction of multivariate systems considering the level of noise added to the system. (Figure 9 for a dataset used in training this model)

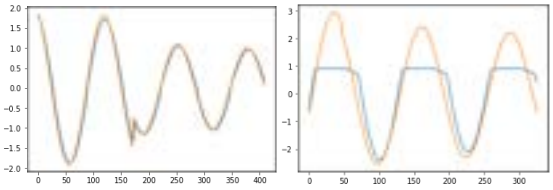


Figure 20: 1D CNN, PD = 1. The same level of noise can also cause the system to break. Underfitted data like this was generated on the second and 5th run

LSTM

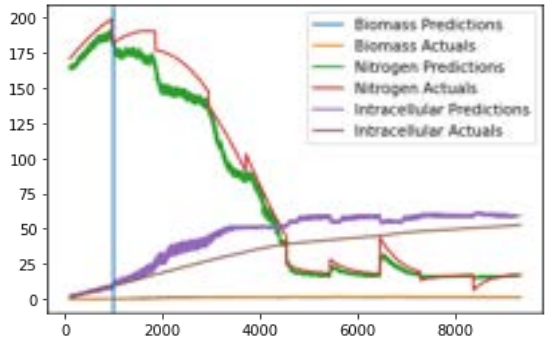


Figure 21: LSTM, PD = 15, MN = 1/4. More gradual failing of system as noise increases as opposed to other algorithms

GRU

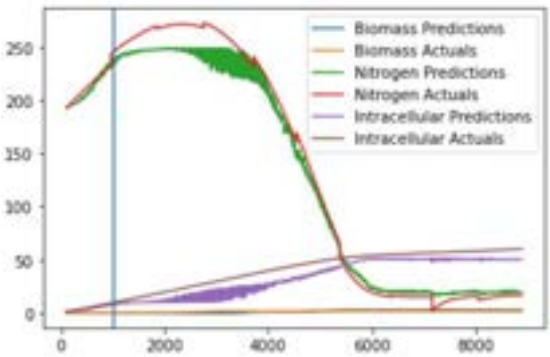


Figure 22: GRU, PD = 10, MN = 1/4. Comparably underfitted compared to LSTM, even with less stochastic noise

5. Discussion

As can be seen, the results show that the FNN is not only the least computationally expensive neural network, but also produces good results when applied to univariate systems, even when the dataset is subjected to noise. It cannot be used on multivariate systems however, due to its input limit. The CNN performed well the univariate system, however, due to unknown reasons, had the tendency to be unreliable at random points. Therefore, although it produced some of the best fitting models for the system, more research into the exact mechanisms of the CNN may be required before applying it to this purpose. In regard to the multivariate, of all the neural networks tested on this system, the CNN's model worked best. Not only did it produce the best fitting predictions, but it was also the most resilient to both process disturbances and measurement noise.

To now focus on the RNN's, as expected, the advanced LSTM was the most reliable of the three, followed by the GRU and then the simplistic RNN model. This was expected due to the nature of each variant discussed in the background. However, with increasingly better results came increasingly expensive computational requirements, the LSTM was especially taxing, which was made very clear during the data collection process, as it took over 4 times as long to train and test when compared to the CNN. This long run time and the fact that the FNN and CNN were able to produce competing results within a fraction of the time indicate that the strength of the LSTM lies within the more reliable prediction of complex systems. It can also be stipulated that the fact that LSTM is the only algorithm which is able to store long-term patterns means that it is actually not suited for application to batch systems and may find more success in a continuous process, in the aid of control systems.

6. Conclusions

To conclude, the overall objective of applying neural network-based data-driven deep learning methods to investigate neural networks has been successful. Through rigorous testing, it can be said with confidence that the best method for time series prediction across both univariate and multivariate applications is the LSTM. Though computationally expensive, it has the capability to accurately predict time series data with a good resilience to noise. It is superior to its other RNN variants and offers an overall improvement on simple RNN's and GRU's in terms of accuracy. Despite the fact that LSTM did not produce as close-fitting data as CNN, it was shown to have a much steadier decline in accuracy, as opposed to other algorithms which would suddenly fail at a threshold noise level.

Due to the nature of the LSTM however, the time to run the models is often long and it may be an issue for less powerful computers to use. This study found that FNN's and CNN's provide a less computationally expensive predictive method for univariate and multivariate systems respectively. Although these sacrifice reliability for speed, for less complex systems or cases where accuracy is less important, these can provide a good predictive deep learning model.

7. Outlook

Though the results were indicative of certain neural networks, namely CNN's, performing better than others, achieving the overall objective, it is worth noting that due to both the time constraints of the project and breadth of the investigation undertaken, that there are certain key areas that should be expanded upon from this study. The focus on exploring individual neural network architectures and their ability to predict time series predictions combined with the limitations of the technology used resulted in a lack of emphasis on hyperparameter tuning and experimenting. This process can prove computationally taxing, however, the use of, for example, different activations such as Sigmoid rather than ReLU could have provided different, and potentially better, predictive results, as observed during the initial setup of each model.

It is further discussed in literature that combinations of the neural networks' architectures can be useful to refine and improve prediction. In addition to combing together the neural network architectures discussed and investigated here, adding a Bayesian neural network layer – which adds a probabilistic predictive method to the predictions, and adding a transformer neural network – which can input sequences in parallel respectively and assist in identifying long term system trends, can further improve models for predicting complex dynamical systems (Brownlee, Deep Learning for Time Series Forecasting, 2018).

Additionally, further studies into the Kalman filter algorithm could prove useful to achieving the objective. This filter utilises system estimate covariance when predicting complex, noisy data and can handle systems with many dimensions, uncertain or incomplete information, and can adapt to changes to the system over time. Applying this to systems and combining it with the models explored in this paper is a robust solution to not only the objective, but even more complex systems due to the Kalman filter's intrinsic resilience to high levels of noise and missing data (Maitra, 2019). However, although this could be used for real examples, as the stochasticity and measurement noise used in this study was gaussian by default and known, the Kalman filter may be inappropriate to use in combination with the methods used here.

8. Acknowledgements

Thank you to Haiting Wang, whose guidance was both valuable and appreciated, to Professor Serafim Kalliadasis, who introduced us to deep learning in 2020, and to the staff and supervisors of the Chemical Engineering and Centre for Process Systems Engineering.

9. References

Brownlee, J. (2016, November 7). *Machine Learning Mastery*. Retrieved from How to Code a Neural Network with Backpropagation In Python (from scratch): https://machinelearningmastery.com/implement-backpropagation-algorithm-scratch-python/

Brownlee, J. (2018). *Deep Learning for Time Series Forecasting.* Machine Learning Mastery.

Daniel Svozil, V. K. (1997). Introduction to multi-layer feed-forward neural networks. *Chemometrics and Intelligent Laboratory Systems, Volume 39, Issue 1*, 43-62.

Del Rio-Chanona, E. A.-G. (2015). Optimal Operation Strategy for Biohydrogen Production. *Industrial and Engineering Chemistry Research*, 6334-6343.

Facundo Bre, J. M. (2017, November). *Prediction of wind pressure coefficients on building surfaces using Artificial Neural Networks.* Retrieved from ResearchGate: https://www.researchgate.net/publication/321259051_Prediction_of_wind_pressure_coefficients_on_building_surfaces_using_Artificial_Neural_Networks

H2O.ai. (2022). *Neural Network Architectures*. Retrieved from H2O.ai Wiki: https://h2o.ai/wiki/neural-network-architectures/#:~:text=The%20architecture%20of%20neural%20networks,power%20of%20a%20human%20brain.

Hanyu Gao, L.-T. Z.-H.-M. (2022). Machine Learning and Data Science in Chemical Engineering. *Industrial & Engineering Chemistry Research Volume 61, Issue 24*, 8357-8594.

Lewinson, E. (2020). *Python for Finance Cookbook.* Packt.

Maitra, S. (2019, December 2019). *Towards Data Science*. Retrieved from State Space Model and Kalman Filter for Time-Series Prediction: https://towardsdatascience.com/state-space-model-and-kalman-filter-for-time-series-prediction-basic-structural-dynamic-linear-2421d7b49fa6

Mills, C. (2022, December 20). *So, What Actually Is Chat GPT-3? And Can It Replace Us?* Retrieved from Student Edge: https://studentedge.org/article/chat-gpt-3-explained

Mussmann, T. (2021). *Data Driven Learning of Dynamical Systems Using Neural Networks.* The Ohio State University.

Nykamp, D. Q. (2008). *Dynamical system definition*. Retrieved from Math Insight: https://mathinsight.org/definition/dynamical_system

Pra, M. D. (2020, November 2). *Time Series Forecasting with Deep Learning and Attention Mechanism*. Retrieved from Towards Data Science: https://towardsdatascience.com/time-series-forecasting-with-deep-learning-and-attention-mechanism-2d001fc871fc

Shah, N. V. (2020, November 12). *Artificial Neural Network and Data-driven techniques: Scientific Computing in the era of emerging technologies*. Retrieved from Reduced Order Modelling, Simulation and Optimization of Coupled Systems: https://www.romsoc.eu/artificial-neural-network-and-data-driven-techniques-scientific-computing-in-the-era-of-emerging-technologies/#:~:text=The%20aim%20of%20data%2Ddriven,the%20field%20of%20numerical%20analysis.

ShareTechnote. (2015, May 12). *Engineering Math - Differential Equation*. Retrieved from ShareTechnote : http://www.sharetechnote.com/html/DE_Modeling.html

Urquhart, J. (2010, December 15). *Biohydrogen produced in air*. Retrieved from Royal Society of Chemistry : https://www.chemistryworld.com/news/biohydrogen-produced-in-air/3000547.article

Wahlström, M. B., & Dernsjö, A. (2021). *Data-Driven Learning for Approximating Dynamical Systems Using Deep Neural Networks.* Stockholm : KTH Royal Institute of Technology.

Author Index

Linked to paper number

Supervisor Index

Linked to paper number