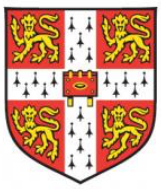
MACHINE-ASSISTED SYNTHESIS AND DEVELOPMENT IN PHARMACEUTICAL INDUSTRY

Perman Jorayev

Wolfson College

Department of Chemical Engineering and Biotechnology



UNIVERSITY OF CAMBRIDGE

This dissertation is submitted for the degree of *Doctor of Philosophy*.

September 2022

Preface

I hereby declare that this thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text.

This dissertation is not substantially the same as any work that has already been submitted before for any degree or other qualification except as declared in the preface and specified in the text.

As prescribed by the Department of Engineering Degree Committee, this dissertation does not exceed 65,000 words, including appendices, footnotes, tables, and equations, but excluding the bibliography, and 150 figures.

Perman Jorayev

September 2022

ABSTRACT

Process development of novel chemical transformations is often a laborious and complex task. This is mostly due to the difficulties in identifying the underlying reaction mechanism(s), selection of chemical (intrinsic) and physical (process) parameters that affect the process objective(s), quantifying the nonlinear interactions between them, and lack of data. Reducing the cost and time for development of robust processes from novel chemical transformations, therefore, requires more efficient solutions to address the individual challenges. In this project, we present several new workflows to tackle some of these challenges.

First, we explored use of black-box Bayesian optimisation algorithm TS-EMO for optimisation of complex reaction networks, such as the bio-waste crude sulphate turpentine conversion to functional molecules, with no prior mechanistic information. Using Gaussian processes as surrogate models and sampling from the reaction space based on the highest expected hypervolume improvement, algorithm-guided optimisation of eight continuous variables allowed for identification of the experimental Pareto front to account for the trade-offs between the reaction objectives conversion and yield. Then, expanding to the discrete variable space, we developed a solvent recommendation workflow based on similarity and data fusion techniques, and sustainability guides. Based on two solvent libraries, we demonstrated the use molecular informatics-driven workflow on various chemical transformations, and identified a significant overlap with solvent selection tools developed by AstraZeneca and Syngenta.

Next, considering all continuous and discrete variables in the reaction, a holistic modelling of Buchwald-Hartwig amine synthesis using DFT-based parameterisation techniques and a dozen machine learning algorithms led to development of highly predictive reaction models. Through several Explainable AI tools, reaction specific descriptors were identified, and their transferability was validated in the laboratory on a similar reaction. Finally, in order to develop a robust process of a sensitive photoredox amine synthesis reaction, we generated *a priori* knowledge in the form of solubility predictions and measurements, and quantification of absorbance and photon flux. Using a recently developed NEMO algorithm, we demonstrated simultaneous optimisation of continuous and discrete variables for reaction objectives yield and cost. The workflows developed in this project, as demonstrated on multiple case studies, validated their efficiency and use in process development.

ACKNOWLEDGEMENTS

First of all, I would like to express my sincere gratitude to Prof. Alexei Lapkin for his continuous guidance and support from start to the finish of this project. Doing an interdisciplinary research was both exciting and exhausting, and Prof. Lapkin was always supportive and available whenever I needed guidance. He trusted me to be independent and to explore my own ideas, and was always patient if my ideas did not result in anything significant and made sure that I was on a right path. I also want to thank Prof. Matthew Gaunt as my advisor and his students for productive discussions on chemistry. I am thankful to Prof. Saif Khan for having me as his student during my time in Singapore, his continuous support with my projects, and his students for such an amazing collaboration.

I am thankful to my industrial collaborators Dr. Paul Deutsch, Dr. Alexandre Barthelme, and Dr. Patrick Pasau from UCB Pharma. Their feedback during the monthly project update meetings helped stress test our ideas and allowed for development of projects for real-life applications. I am grateful to UCB Pharma and the National Research Foundation, Prime Minister's Office, Singapore for providing me with PhD scholarships and funding that made it possible to carry out this project.

It was an absolute pleasure to work with Dr. Simon Sung, Dr. Mohammed Jeraal, and Dr. Danilo Russo, and I deeply appreciate all the fruitful collaborations and their project mentorship. I learned a lot, and for that, I am forever grateful. I want to thank Dr. Akihiro Takada for collaborating on running Buchwald-Hartwig optimisation in-house and Mr. Kobi Felton for descriptor calculations using COSMO*therm*. Furthermore, I want to thank Dr. Artur Schweidtmann for our collaboration on extending the TSEMO algorithm for optimisation of functional molecules synthesis from bio-waste terpenes. To all the past and current SRE members, thank you! This journey was special because of all the group members that I had the pleasure of interacting with on a day-to-day basis. Thanks to all the friends from the Wolfson College, the College football team, Dept. of Chemical Engineering and Biotechnology, Dept. of Chemistry, and CARES Singapore.

I do not think I would have ever taken this path if it were not for my parents, who guided me and supported me as long as I can remember myself. Thanks to everyone in my family for being in my life and making this journey possible.

TABLE OF CONTENTS

LIST OF PRESENTATIONS AND PUBLICATIONS

Peer-reviewed publications

1. **Jorayev, P.**,* Russo, D.,* Tibbetts, J., Schweidtmann, A. M., Deutsch, P., Bull, S. D., & Lapkin, A. A. (2021). Multi-objective Bayesian optimisation of a two-step synthesis of p-cymene from crude sulphate turpentine. Chemical Engineering Science, 116938.

Manuscripts currently in preparation for submission

2. **Perman Jorayev**, Simon Sung, Mohammad Jeraal, Alexei Lapkin, *Machine learning-driven optimisation of photoredox amine synthesis in flow*, 2022
3. **Perman Jorayev**, Simon Sung, Akihiro Takada, Alexei Lapkin, *Prior knowledge generation through reaction modelling via supervised machine learning*, 2022
4. **Perman Jorayev**, Connor Taylor, Alexei Lapkin, *Molecular informatics-driven solvent recommendation*, 2022

* Contributed equally to the publication.

Conference presentations

1. Invited speaker at the RSC Process Chemistry and Technology Group, August 2022 (50 min webinar, oral)
2. ACS Green Chemistry conference, June 2022, VA, USA (invited, oral)
3. UCB industry presentations, October 2019 and October 2022 (oral)
4. CEB Research conference, Dept. of Chemical Engineering, Cambridge, June 2021 (oral)
5. 4th Annual ML/AI in Biochemical Engineering conference, Cambridge, July 2021 (co-organiser)
6. 3rd Annual ML/AI in Biochemical Engineering conference, Cambridge, July 2020, Cambridge, UK (poster)
7. Wolfson Research Event, Wolfson College, Cambridge, March 2019 (oral)

LIST OF TABLES

LIST OF FIGURES

LIST OF SCHEMES

Chapter 1

Introduction

Background and research context

Process development of novel chemical transformations is often a laborious and complex task. This is mostly due to the difficulties in identifying the underlying reaction mechanism(s), selection of chemical (intrinsic) and physical (process) parameters that affect the process objective(s), quantifying the nonlinear interactions between them (Figure 1), and lack of data. Reducing the cost and time for development of robust processes from novel chemical transformations, therefore, requires more efficient solutions to address the individual challenges.

With an increased use of algorithms and automation in chemical research, when optimising for continuous variables only (e.g., temperature, residence time, concentration), conventional and Bayesian optimisation-based Design of Experiments (DoE) have been demonstrated in the form of closed-loop self-optimisation systems.[1-5] However, most of the current literature is limited to optimising for a single objective over up to five continuous variables with the aim of identifying a few optimal conditions (e.g., local optima) as opposed to developing a robust predictive model.[3] Moreover, when expanded into discrete variables including solvents, reagents, or ligands, these approaches often struggle with the curse of dimensionality and inefficiencies (in predictive accuracy), often due to the sparsity of the chemical data and difficulty of selecting relevant molecular descriptors, of black-box optimisation algorithms.[6]

A common approach to overcome this problem has been demonstrated using molecular descriptors to map the discrete variables onto a continuous space.[7-9] However, existing approaches often: (1) utilise large number of variables / dimensions (15+) to parameterise discrete variables, (2) start the optimisation from scratch (i.e., selected descriptors are not confirmed to be the most relevant), and (3) do not demonstrate a holistic approach (i.e., keep other reaction variables fixed). Parameterisation of discrete variables with relevant descriptors could increase the predictive accuracy of surrogate models to accelerate development of optimal process conditions. Especially, reactions of high pharmaceutical relevance such as

photoredox amine synthesis are sensitive to reaction conditions and require robust process models to overcome scalability challenges. This necessitates significant amount of prior knowledge generation on the choice of discrete variables, their parameterisation, the choices and the ranges of continuous variables, light sources, and the reactor setup, making it a challenging case study for holistic and robust process development.

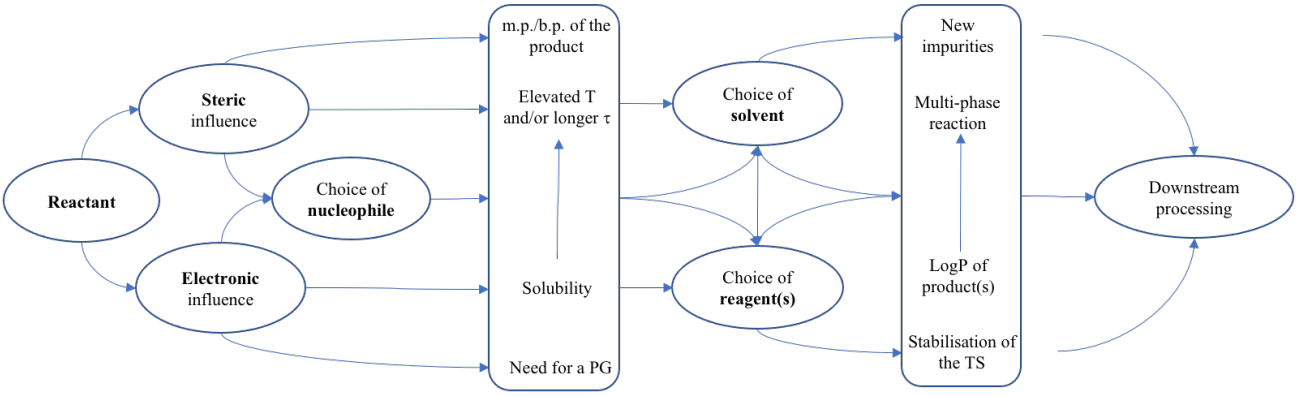


***Figure 1**. Some of the interactions in the process development pipeline. PG: protecting group, m.p.: melting point, b.p.: boiling point, and TS: transition state.*

Demonstrated on several case studies, the workflow (Figure 2) developed in this project addresses these challenges through a combination of cheminformatics tools (e.g., COSMO*therm*, Gaussian, RDKit), feature engineering techniques, supervised machine learning algorithms (e.g., Artificial Neural Networks, Boosting Algorithms), and Bayesian optimisation algorithms (e.g., TS-EMO, NEMO) to simultaneously optimise for discrete and continuous variables in complex chemical reactions (e.g., liquid-liquid, liquid-gas, photoredox) in batch and semi-automated continuous flow set-ups.

Outline and objectives

The project was started in collaboration with UCB Pharma. It aimed at addressing the challenge of incorporating *a priori* chemical information into automated process development based on robotic experiments. Specifically, black-box data-driven optimisation algorithms were to be used to guide through the search space and generate good process models incorporating discrete variables (e.g., solvents, ligands, reactants). Simultaneous optimisation of continuous and discrete variables for multiple competing objectives was aimed to be demonstrated on a

pharmaceutically relevant transformation such as photoredox synthesis of amines. The choice and parameterisation of discrete variables were to be selected using physically meaningful molecular descriptors in the form of *a priori* knowledge. The workflow developed in this project is given in Figure 2.

To achieve these goals, the following objectives were proposed:

1. Set-up experimental systems for robotic experiments with the chemistry of interest in low-pressure non-hazardous environment.
2. Implement DoE algorithm for the batch-sequential and flow-based workflows within the new robotic system.
3. Demonstrate optimisation of complex reactions using black-box Bayesian optimisation approach for continuous variables and multiple objectives, where identifying and developing a mechanistic kinetic model is specifically challenging.
4. Develop intuition and molecular informatics-driven DoE methodology for sampling from discrete variables space.
5. Develop a workflow for generation, selection, and validation of molecular descriptors in the form of *a priori* knowledge from literature data using cheminformatics tools and supervised ML.
6. Demonstrate the transfer of model validated results for reaction optimisation in the lab (in-house) on a separate case study.
7. Integrate prior knowledge generation, sampling from discrete variables space, and black-box optimisation of chemical (intrinsic) and physical (process) parameters to develop a robust process on air, water, and light sensitive photoredox amine synthesis in semi-automated continuous flow.

Thesis structure

The chapters in the thesis tackle various steps introduced in automated workflow developed in this work (Figure 2). Every chapter includes a summary of overall challenges, literature and scientific background, methodology, and results to answer the research question(s) addressed in the respective chapters. The thesis structure is as follows for each chapter.

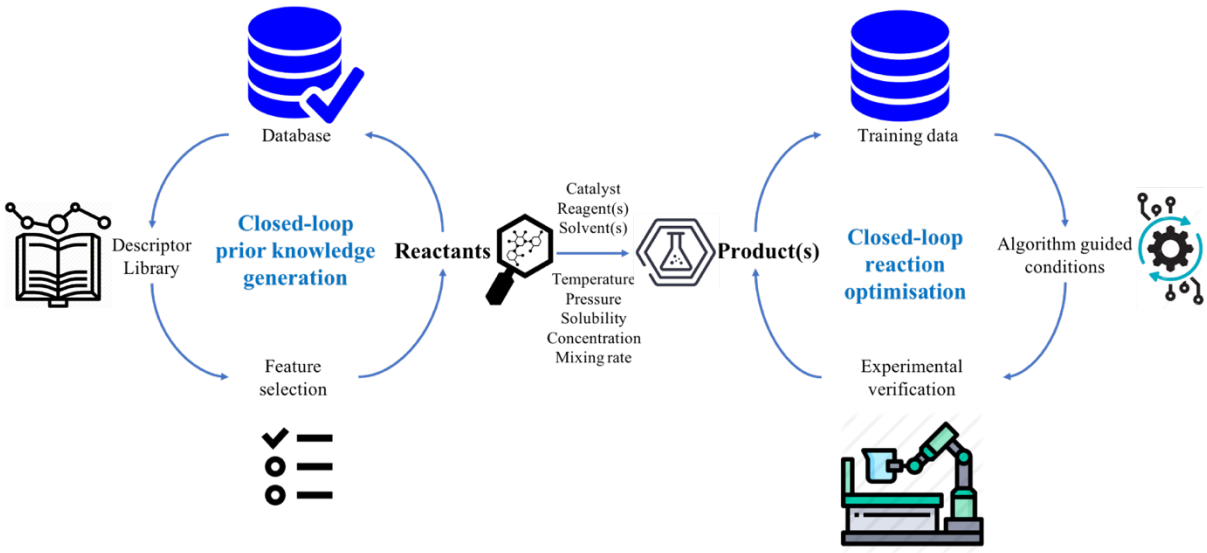


Figure 2. Schematic illustration of process development workflow developed in this project.

In Chapter 2, multi-objective Bayesian optimisation of a two-step synthesis of p-cymene from crude sulphate turpentine is presented. The first step (i.e., biphasic acid-catalysed ring opening of a mixture of terpenes) in this reaction is particularly challenging for traditional process development tools, since many species in the reaction interconvert into each other and are too similar for time-resolved sampling techniques for development of kinetic models. Both steps are optimised for conversion and yield over eight continuous variables using batch-sequential sampling, and the work has been published by us.[10] Moreover, model sensitivity analysis over input parameters space using *lengthscale* hyperparameters allowed for model interpretability.

In Chapter 3, an alternative approach to conventional DoE, such as the commonly used Latin Hypercube Sampling (LHS),[11, 12] to generate the training dataset before initiating an algorithm-guided optimisation is demonstrated on selection of solvent candidates. The workflow utilises different sustainability guides, molecular featurisation techniques, and similarity metrics to generate information-driven list of suggested solvents for a given reaction. The results are compared against the solvent selection guides by AstraZeneca[13] and Syngenta,[14] and the applicability of the workflow is demonstrated on the experimental solvent selection datasets

for various chemical transformations. The results demonstrate that the workflow could serve as a starting point for identifying and testing alternative reaction components (e.g., solvents) to those reported in literature.

In Chapter 4, I present a workflow to: (*i*) generate physically meaningful molecular descriptors, including a novel way to compute steric descriptors for complex structures such as ligands, (*ii*) automatically train and optimise ten ML algorithms, and (*iii*) extract physical insights from black-box models using explainable AI tools to validate selection of most relevant descriptors. The results from the first case study of Buchwald-Hartwig amination reaction are then utilised to model a separate Buchwald-Hartwig transformation optimised in-house and to benchmark against the results achieved when molecular descriptors were not used.

In Chapter 5, robust process development of sensitive photoredox tertiary amine synthesis in semi-automated continuous flow using a novel Bayesian optimisation algorithm is presented. Starting with an initial list of 115 solvents, solubility predictions and measurements, photon flux study, and UV-Vis studies are carried out to generate *a priori* knowledge. Optimisation results are also benchmarked using pool-based sampling technique to compare the learning efficiency of the algorithm with different dataset sizes.

Overall conclusions, and suggestions on future work are given in Chapter 6.

References

1. Echtermeyer, A.; Amar, Y.; Zakrzewski, J.; Lapkin, A., Self-optimisation and model-based design of experiments for developing a C–H activation flow process. *Beilstein Journal of Organic Chemistry* **2017,** *13*, 150-163.
2. Jeraal, M. I.; Holmes, N.; Akien, G. R.; Bourne, R. A., Enhanced process development using automated continuous reactors by self-optimisation algorithms and statistical empirical modelling. *Tetrahedron* **2018,** *74* (25), 3158-3164.
3. Mateos, C.; Nieves-Remacha, M. J.; Rincón, J. A., Automated platforms for reaction self-optimization in flow. *Reaction Chemistry & Engineering* **2019,** *4* (9), 1536-1544.
4. Clayton, A. D.; Manson, J. A.; Taylor, C. J.; Chamberlain, T. W.; Taylor, B. A.; Clemens, G.; Bourne, R. A., Algorithms for the self-optimisation of chemical reactions. *Reaction Chemistry & Engineering* **2019,** *4* (9), 1545-1554.
5. McMullen, J. P.; Stone, M. T.; Buchwald, S. L.; Jensen, K. F., An Integrated Microreactor System for Self-Optimization of a Heck Reaction: From Micro- to Mesoscale Flow Systems. *Angewandte Chemie International Edition* **2010,** *49* (39), 7076-7080.
6. Skoraczyński, G.; Dittwald, P.; Miasojedow, B.; Szymkuć, S.; Gajewska, E. P.; Grzybowski, B. A.; Gambin, A., Predicting the outcomes of organic reactions via machine learning: are current descriptors sufficient? *Scientific Reports* **2017,** *7* (1).
7. Amar, Y.; Schweidtmann, Artur M.; Deutsch, P.; Cao, L.; Lapkin, A., Machine learning and molecular descriptors enable rational solvent selection in asymmetric catalysis. *Chemical Science* **2019,** *10* (27), 6697-6706.
8. Zhang, C.; Amar, Y.; Cao, L.; Lapkin, A. A., Solvent Selection for Mitsunobu Reaction Driven by an Active Learning Surrogate Model. *Organic Process Research & Development* **2020,** *24* (12), 2864-2873.
9. Reizman, B. J.; Jensen, K. F., Feedback in Flow for Accelerated Reaction Development. *Accounts of Chemical Research* **2016,** *49* (9), 1786-1796.
10. Jorayev, P.; Russo, D.; Tibbetts, J. D.; Schweidtmann, A. M.; Deutsch, P.; Bull, S. D.; Lapkin, A. A., Multi-objective Bayesian optimisation of a two-step synthesis of p-cymene from crude sulphate turpentine. *Chemical Engineering Science* **2022,** *247*.
11. McKay, M. D.; Beckman, R. J.; Conover, W. J., A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. *Technometrics* **1979,** *21* (2).
12. Tang, B., Orthogonal Array-Based Latin Hypercubes. *Journal of the American Statistical Association* **1993,** *88* (424).
13. Diorazio, L. J.; Hose, D. R. J.; Adlington, N. K., Toward a More Holistic Framework for Solvent Selection. *Organic Process Research & Development* **2016,** *20* (4), 760-773.
14. Piccione, P. M.; Baumeister, J.; Salvesen, T.; Grosjean, C.; Flores, Y.; Groelly, E.; Murudi, V.; Shyadligeri, A.; Lobanova, O.; Lothschütz, C., Solvent Selection Methods and Tool. *Organic Process Research & Development* **2019,** *23* (5), 998-1016.

Chapter 2

Multi-objective Bayesian optimisation of a three-step synthesis of terephthalic acid from crude sulphate turpentine

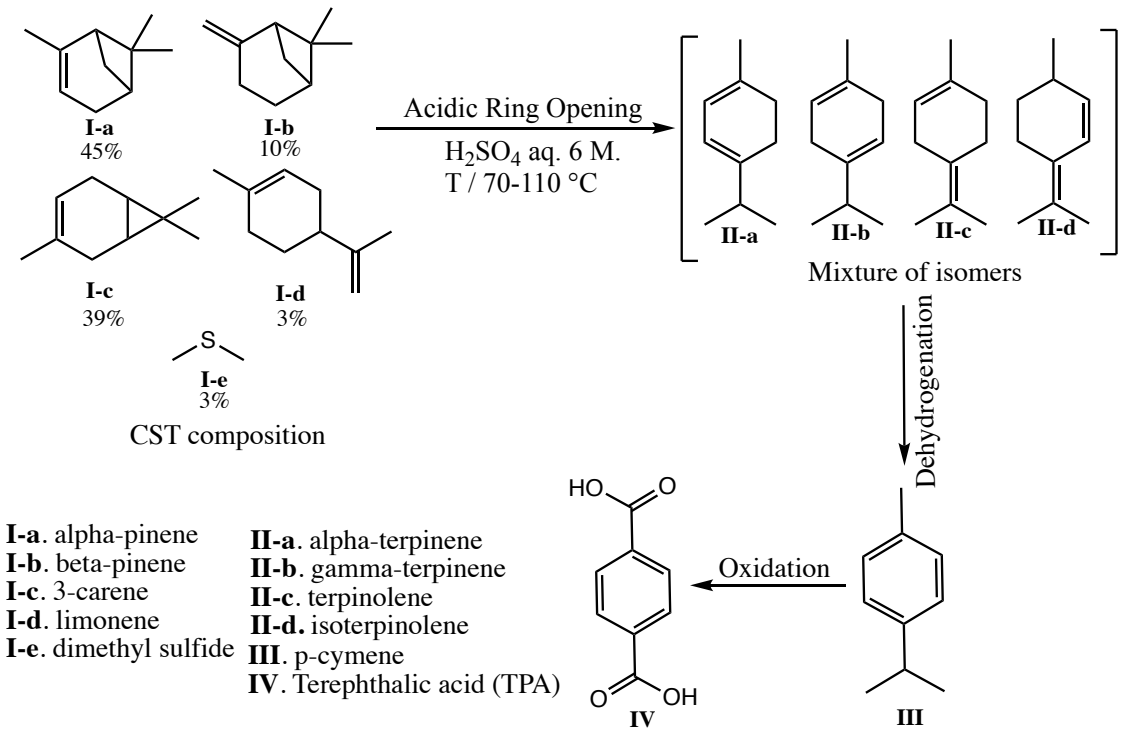
Introduction

Development of routes to functional molecules starting from bio feedstocks is one of the strategies to de-carbonise the chemical supply chain: replacement of fossil carbon sources with renewable carbon, preferably from bio-waste sources, allows to significantly reduce green-house gases emissions along the chemical supply chain.[1] Therefore, developing novel routes for valorisation of bio-waste materials which allow access to functional molecules, as drop-in replacements or with novel structures, with quantifiable importance in reaction networks is of high interest.[2-4] Most of such routes will start from feedstocks which are complex mixtures: products of de-polymerisation of lignin, of hydrolysis of cellulose, crude glycerol, mixtures of terpenes, etc., and which allow access to a very broad range of functionalised molecules.[1, 5]

(Bio)chemical conversion of such feedstocks requires either highly selective and robust chemistry, or prior purification. For example, crude sulphate turpentine (CST) is a waste by-product from the pulp and paper industry and can be used to produce *p*-cymene (Scheme 1).[6, 7] *p*-Cymene is used as a solvent for dyes and varnishes and it is the main precursor of p-cresol.[8] It shows a wide range of antioxidant and biological activity which makes it particularly suited for applications in the food and pharmaceutical industry. Other applications include the synthesis of fragrances and it has been proposed as an alternative precursor of terephthalic acid. The route could start from a purified single terpene, but in our work, the starting point is directly the CST.[9-11] In this work a mixture of alpha-pinene, beta-pinene, 3-carene, limonene, and dimethyl sulfide (DMS) were used to mimic CST.

The first reaction in the proposed route is an acid-catalysed ring opening of a mixture of terpenes **I**a-d to a mixture of isomers **II**a-c. This reaction exemplifies a highly promising approach to valorisation of bio-waste feedstocks *via* chemical routes that converge a starting

mixture of substrates to a single compound, in this case *p*-cymene (**III**), without significant loss of carbon. *p*-Cymene can then be converted to a range of functional molecules, as, for example, terephthalic acid (**IV**).[12]



Scheme 1. *A proposed route for conversion of crude sulphate turpentine (CST) to terephthalic acid.*

Within this reaction sequence, optimisation of the acid-catalysed ring opening is particularly challenging for traditional process development tools, since many species interconvert into each other and are too similar for most time-resolved sampling techniques. The mechanism of this reaction has been suggested.[6, 9, 11, 13] It is a multiphase reaction occurring through the reversible addition of DMS to the double bonds of monoterpenes that generate surfactant-like species improving mass transfer between the acidic and the organic phase. Previous studies have demonstrated the preliminary formation of mixtures rich in limonene and terpinolene, converting to α-terpinene, γ-terpinene and isoterpinolene through equilibrium protonation reactions.[9] Acid-catalysed polymerisations decrease the selectivity to the products of interest. As a result, the development of a mechanistic kinetic model for rational scale-up of this process is rather challenging. Model-free optimisation approaches are a promising alternative to model-based methods.

Design of Experiments

Traditionally, one way to explore optimal (i.e., often to find a few good points) conditions is through one variable at a time (OVAT) approach.[14] Combining domain knowledge with a preliminary data, different reaction conditions are explored by varying single reaction parameter (e.g., temperature) at a time. However, use of OVAT often requires more experiments, does not always lead to optimal conditions, and does not account for interactions between reaction parameters.[15] These drawbacks are addressed with use of statistical design of experiments (DoE), an efficient tool to screen reaction conditions over a constrained input space to identify the interactions between reaction parameters that affect the target(s) by fitting a polynomial model.[16-19] A few examples for different types of designs are given in Figure 3. However, a major shortcoming of classical DoE is their limited scalability over large number of reaction parameters, leading to an expansion in number of required experiments for optimisation, making this approach time and resource intensive.[20, 21] Besides the limited use in screening discrete variables, this approach also does not lead to any physical insights about the reaction.

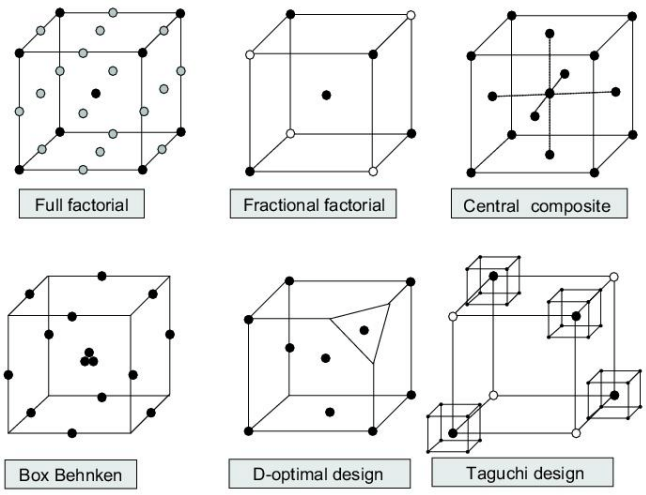


***Figure 3**. Different types of conventional design of experiments adopted from ref.[22]*

Self-optimisation algorithms

In recent years, a number of optimisation algorithms has been used for the design of experiments (DoE) in the 'self-optimisation' mode of experiments.[23, 24] In self-optimisation, automated experimental systems perform optimisation of a reaction, and potentially also separation, without intervention of a human.[25, 26] Several algorithms such as Nelder-Mead-Simplex,[27] SNOBFIT,[21] steepest descent,[28] MOAL[29] and TS-EMO[25] were used for optimising reactions ranging from nanoparticle synthesis[30] to heterogeneous catalytic reactions.[31] These

methods do not require any prior model of the chemical system and focus on the relationships between input and the output variables or objectives.[32]

Out of the algorithms mentioned above, Nelder-Mead-Simplex is a local search algorithm that scales poorly to larger systems, and is known to have limited convergence guarantees even for convex problems.[33] Steepest descent is an established algorithm but not suitable for experimental systems because derivatives need to be estimated through finite differences, which is expensive and inaccurate for noisy systems. SNOBFIT, is a global search algorithm that has been successfully used in many self-optimisation studies but has also shown rather slow convergence.[34, 35] Bayesian optimisation approaches, such as MOAL,[29] TS-EMO,[36] MVMOO,[37] and Google Vizier[38] construct non-parametric statistical models, Gaussian processes, using a sequential active learning approach (re-training a model when new experimental observations become available). Bayesian optimisation approaches utilise all available data to build statistical models and are thus data efficient. The models provide quantification of uncertainty that is used to solve the inherent exploration-exploitation trade-off within the derivative-free optimisation. The methods can be used to either develop an accurate model that is valid on the whole input domain or they can be used for efficient optimisation. There exist algorithm extensions to multi-objective optimisation that have been used successfully in self-optimisation and solvent selection.[25, 39-41]

Gaussian Processes

Gaussian processes are models of functions such that any finite set of function values $f(x_1), f(x_2), \dots, f(x_N)$ assumes to follow a joint Gaussian distribution.[42] Gaussian processes are often used as a surrogate to approximate a black-box model that aims to map a set of inputs x to their observed target values y

$$y = g(x) + \epsilon \tag{1}$$

where the associated noise term ϵ is assumed to follow a normal distribution (equation 2).

$$\epsilon \sim N(0, \sigma^2) \tag{2}$$

Before conditioning on a data, a prior distribution of GP models could be expressed using a mean function, $\mu(x)$ and a covariance function, also called the *kernel*, $\kappa(x, x')$ as given in equation 3.

$$g(x) \sim GP\ \{\mu(x), \kappa(x, x')\} \tag{3}$$

$$\mathbb{E}\,[f(x)] = \mu(x) \tag{4}$$

$$Cov\,[f(x), f(x')] = \kappa(x, x') \tag{5}$$

Assuming a Gaussian distribution over the training and the test data, the kernel influences how different data points relate to each other, ultimately affecting model generalisability and extrapolation to new data. Gaussian distributions are iteratively updated as more data points are observed, reducing the associated predicted uncertainty over the optimisation. Figure 4 shows an objective function (dashed line) approximation using a GP model (blue line) with confidence interval (shaded region) of a sampled GP (black line) over observed data points.

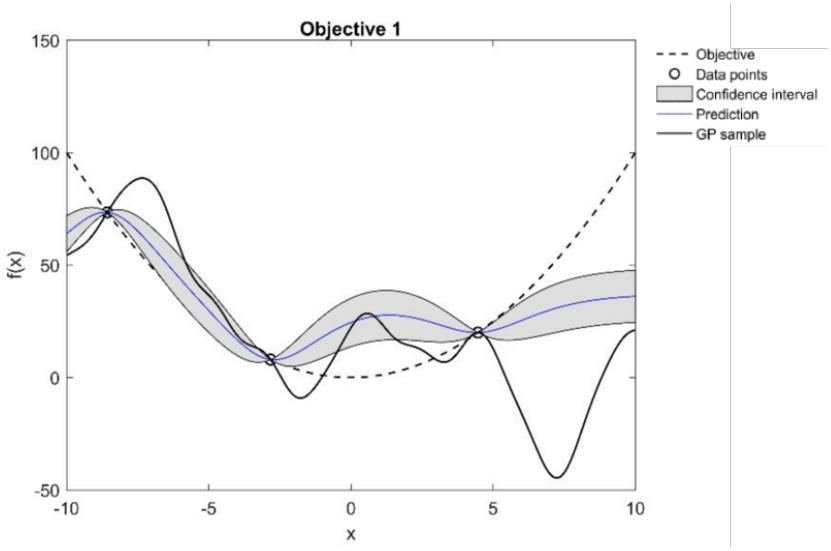


Figure 4. *An illustration of a Gaussian process model from ref.*[43]

Algorithm development

Thompson sampling efficient multi-objective optimisation (TS-EMO) algorithm was developed internally in our group as a solution to replace computationally expensive models using Gaussian surrogate models for multiple, often competing, objectives. The algorithm was demonstrated to be efficient in approximating the true Pareto front of the objectives,[25, 39-41] where one objective cannot be improved without a negative effect on the other. Within the algorithm, individual GP surrogate models of the objectives, conversion and yield in this work, are trained on the collected dataset independently to approximate their response surfaces.[42] The TS-EMO algorithm then draws random samples from these GPs using spectral sampling, which accounts for the inherent exploration-exploitation trade off during the optimisation. Then, a multi-objective genetic algorithm is called within TS-EMO to identify the Pareto front of the random samples.[25] Finally, TS-EMO identifies a set of experiments from that Pareto front (of

the random GP samples), which aim to improve the hypervolume (i.e., expected hypervolume improvement – EHI) of the actual Pareto front (of the experiments conducted).

TS-EMO algorithm is described in detail in the provided references.[36] Briefly, after the first training of the algorithm, at each iteration, new experimental conditions were suggested, tested in the laboratory, and the results were fed back to the algorithm, until satisfactory outputs were obtained (Figure 5). For better clarity and replicability, the reader can refer to pseudocode reported in Figure 16 in the Appendix. Selected parameters for the algorithm are: maximum number of function evaluation = 35; number of spectral sampling points for each objective = 1000; Matern type = 1; function evaluations by direct algorithm per input dimension for each objective = 200. At each iteration, the computational time required for new suggestions was of the order of magnitude of hundreds of seconds, and in any case, completely negligible compared to the time required to run the experiments.

The original version of TS-EMO was developed for batch-sequential experimental protocol, suggesting a number of experiments to perform which have the highest predicted hypervolume improvement. However, this may result in repeating the same experiment if the algorithm suggested to collect data at the same reaction conditions, but at a different reaction time. To avoid this, we implemented a separate sampling for reaction time. Instead of running experiments at the suggested sample points, objectives functions at these points were evaluated and the predicted values were added back to the dataset. The algorithm was run again to sample multiple points (with predicted hypervolume improvement), but this time keeping all the reaction variables constant except the reaction time. In other words, the algorithm was run twice with the second sampling exclusively suggesting points for the reaction time under the same values of the other variables. This improved the data efficiency by allowing multiple sampling per reaction and generating more data using the same amount of reagents (e.g., sampling over six intervals during the same reaction).

In this work we extend the previously developed TS-EMO algorithm[25, 36,44] to improve its efficiency with respect to the experimental budget for optimisation of reaction time, and use it for the optimisation of the ring-opening reaction, step 1, and dehydrogenation, step 2 in Scheme 1. Most of the existing studies on self-optimisation published to date are limited to optimising up to five reaction variables for a single objective.[24] Using the modified version of TS-EMO, this work demonstrates a model-free approach to building an accurate statistical model for a

complex reaction by optimising eight continuous variables for multiple objectives. Successful generation of such accurate models for complex reactions paves the way for establishing model-free hierarchical optimisation of multi-step processes, facilitated by the use of Gaussian processes surrogate model, allowing a certain degree of interpretability.

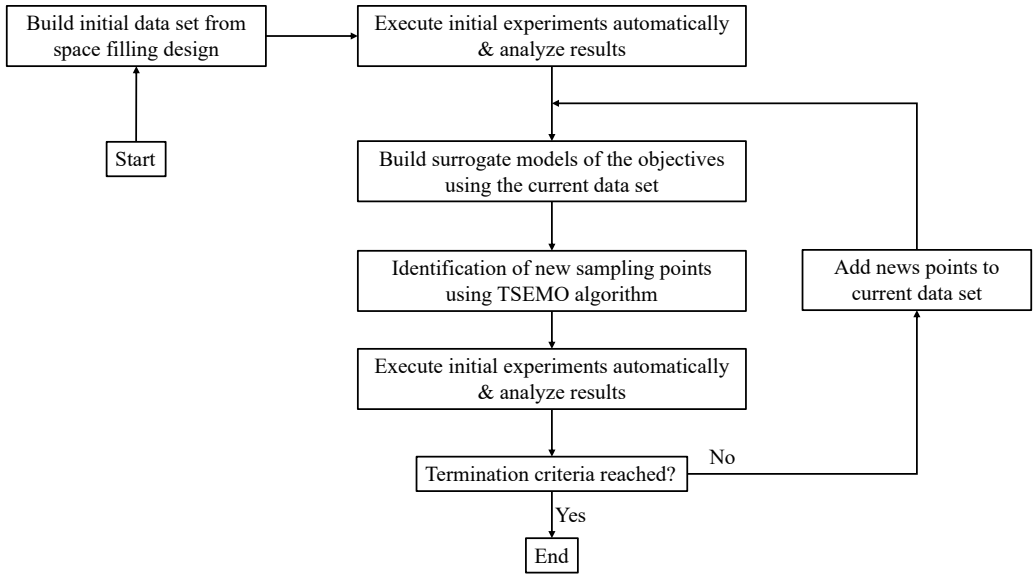


***Figure 5**. Self-optimisation methodology implemented in this chapter.*

Contributions

Dr. Artur Schweidtmann modified the original version of TS-EMO algorithm to allow for batch-sequential sampling of four experiments per iteration, and independent sampling for reaction time per experiment. Except for the solvent selection and other handful of experiments, which were done by me, optimisation experiments and analysis of step 2 were carried out by Dr. Danilo Russo. All the experiments and analysis of step 1 were carried out by me.

Methods and Materials

Materials

Reagents: 3-carene, 90%, stabilized; (1S)-(-)-beta-pinene, 98%; (1S)-(-)-alpha-pinene, 98%; dimethyl sulfide, 99+%, extra pure; alpha-terpinene, 90% tech.; gamma-terpinene, 97%, stabilized; ethyl acetate, 99%; cobalt(II) nitrate hexahydrate, 99%, pure; manganese(II) bromide, 99%, anhydrous were purchased from Acros Organics and used as received. (R)-(+)-Limonene, 97% was purchased from Alfa Aesar and used as received. Reagents 1,2,4,5-

tetramethylbenzene, 98%; sulfuric acid, ACS reagent, 95.0-98.0%; terpinolene, ≥90% (GC); *p*-cymene 99%; di-tert-butyl peroxide 98%, 1-methyl-2-pyrrolidinone (NMP), ACS reagent, ≥99.0%; acetic acid, reagent grade, ≥99%, di-tert-butyl peroxide, 98%, silicone oil, chloroform-d, were all purchased from Sigma Aldrich and used as received. De-ionised water was used to prepare acid solutions.

Procedures

Experiments for the first step of CST conversion were carried out in batch mode using 100 mL glass bottles immersed in a silicone oil bath (Figure 17 in the Appendix). The reaction mixture was magnetically stirred (650 rpm). A glass thermometer was immersed into the reaction medium through a hole in the plastic cap to monitor the reaction temperature. The caps were not sealed and no pressure build up was observed. The temperature was kept constant using an IKA magnetic hot plate equipped with Pt sensor and a feedback controller, for precise temperature control. To start the reaction, first, the calculated relative amounts of all five components of crude-sulphate turpentine (CST) were weighed and added into the reaction bottle. Then, 0.032 mol (0.5 M) of 1,2,4,5-tetramethylbenzene (durene) was added as an internal standard. The reaction was then stirred until the mixture was homogeneous and heated to the given reaction temperature (explored range 70 – 110 °C). The volume of the organic phase – five CST components and the internal standard in the reaction mixture was fixed to 50 mL in every experiment, roughly three times the one previously reported in the literature.[9] The sulfuric acid solution with the given concentration (4.0 to 7.0 mol L^{-1}, depending on the experiment) and volume (7.5 to 17.5 mL, depending on the experiment) was added to the organic phase; the time of addition was taken as t=0. Sampling of the reaction mixture for ^{1}H NMR analysis was at different time intervals, as suggested by the algorithm, over the approx. 4.5 h of reaction. Collected samples were decanted in the fridge to quench the reaction and allow for phase separation. The lighter organic phase was then dissolved in chloroform-d in NMR tubes. Details of the reaction set-up (Figure 17) and analysis (Figure 14) are given in the Appendix. Repeated experiments under certain conditions allowed to estimate an average RMSE of 1.41 and 1.83, for conversion and yield, respectively.

The second reaction step was carried out feeding compressed air and the liquid phase containing organic substrate to a continuous-flow stainless steel compact reactor with recycle, packed with glass inert spheres to increase the contact surface between the phases. A full

description of the device can be found elsewhere.[45] The organic mixture was prepared using *p*-xylene as a solvent, and tert-butyl hydroperoxide was added as a radical initiator to increase the concentration of radicals at the start of the reaction. A 9 mL stirred reservoir was used for the liquid phase, equipped with a condenser at -18 °C to reflux the evaporated compounds. Liquid phase was pumped through the continuous flow reactor using a Vapourtec HPLC pump module and continuously recirculated to the reservoir. Air was directly fed to the reactor using a MFC (SmartTrack 100 Sierra). Samples were collected at different reaction times from the reservoir using a sampling port connected to a syringe. Samples were rapidly diluted in acetonitrile (30 μL in 1 mL) and analysed by GC/MS. Detailed reaction set-up is given in Figure 18 in the Appendix. This reaction was inspired by the one reported in a previous paper, but with three fundamental differences: (*i*) no DMS was used in the starting mixture, (*ii*) p-xylene was used as solvent, and (*iii*) the experimental set-up for changed from batch to continuous flow with recycle, which does mimic the overall batch behaviour. Repeated experiments under certain conditions allowed to estimate an average RMSE of 2.27 and 1.27, for conversion and yield, respectively.

Analytical methods

^{1}H NMR spectra were recorded using a Bruker AVANCE III 400 MHz spectrometer with Bruker QNP Cryoprobe. The experimental condition was 32 scans with the receiver gain set to 4, d1 = 2 s, and the total acquisition time of 2.94 s. An example of an analysis spectrum is given in Figure 14 in the Appendix.

For the isoaromatisation, the product composition was analysed using GC-MS (Agilent Technologies 7890B GC, 5977A MSD, CTC PAL autosampler; HP-Innowax Agilent column 19091N-133, 30 m x 0.25 mm, 0.25 μm; the system was built and supplied by JSB UK and Ireland Ltd). The inlet condition for the sample injection was 300 °C at the septum purge flow of 3 mL min^{-1} with the split ratio of 100:1 and split flow of 300 mL min^{-1}, column flow 3 mL min^{-1}. The initial oven temperature was held at 60 °C for 1 min and then ramped to 85 °C and then to 180 °C at 2 °C min^{-1} and 100 °C min^{-1}, respectively.

Results and Discussion

Optimisation of the acid catalysed ring opening

Initial exploration of the decision space

As the automated reaction optimisation requires well-defined bounds of the optimisation variables, we conducted an initial exploration of the variables. For this, we designed and carried out 11 batch reactions. The final ranges of the optimisation variables are shown in Table 1. Upper and lower bounds for the molar fractions of single compounds in CST were chosen based on the variability in turpentine composition depending on its geographical origin.

The starting materials were found to polymerise at elevated temperatures, resulting in lower yields (Figure 17). Based on this, and to avoid boiling off of the aqueous layer, the upper boundary for temperature was set to 110 °C. Conversion reached 100% in under 5.5 h for the medium set of reaction conditions (middle of the ranges of temperature and starting concentrations, rows 1-4 in Table 2). Although sulfuric acid is the catalyst and a higher concentration of acidic hydrogen increases conversion, it also leads to an increase in polymerisation and to a formation of insoluble by-products. Results of the study of the effect of the individual reaction parameters on the reactions outcomes are given in Table 2.

***Table 1**. Lower and upper bounds for the input variables in the optimisation. The numbers given for the starting materials indicate their mole fraction. Limonene molar fraction is kept constant at 0.04, which is the commonly observed concentration of limonene in this bio-waste feedstock. Mole fraction ranges for the other starting materials were chosen according to the composition of industrial produced CST waste.[46]*

Range	Temp / °C	H_2SO_4 / mol L^{-1}	Aq./org. ratio	α-pinene	3-carene	β-pinene	DMS	Time / min
Lower	70	4.0	0.15	0.40	0.00	0.05	0.009	1
Upper	110	7.0	0.35	0.80	0.35	0.4	0.037	270

The first set of data (rows 1-4, Table 2) shows that the reaction reaches completion within 5.5 h, depending on the adopted conditions. The second set of data (rows 5-7, Table 2) shows the effect of temperature, *i.e.,* conversion increases with the increase in temperature. For the

aq./org. phase ratio, we observed higher conversion and yield with the lower amount of the aqueous phase (rows 8,9). Since the reaction is acid catalysed, higher conversions and yields were observed with higher concentrations of sulfuric acid (6 *vs* 5 M), rows 10,11 in Table 2. In terms of time, longer duration of the reaction gives higher conversion. However, the combination of high temperature and high acid concentration (e.g., 6.92 M H_2SO_4) leads to the decrease in yield over time due to polymerisation reactions (rows 12-14, Table 2). Thus, there exists a trade-off between yield and conversion motivating the use of multi-objective optimisation methods.[9]

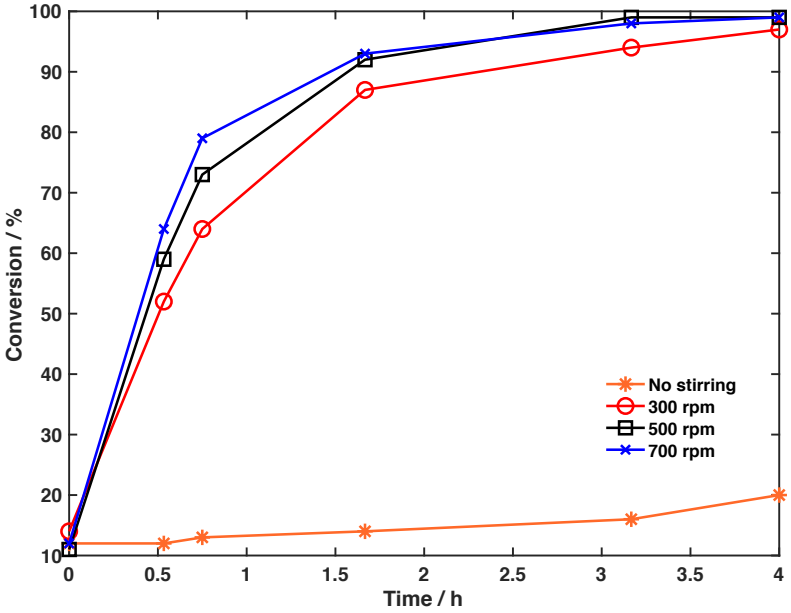
Table 2*. The effects of the individual reaction parameters on the reaction outcomes.*

T / °C	H_2SO_4 / M	Phase ratio (aq./org.)	α-pinene	3-carene	β-pinene	DMS	Time / min	Conversion / %	Yield / %
90	6	0.2	0.45	0.35	0.12	0.04	**60**	81	66
90	6	0.2	0.45	0.35	0.12	0.04	**120**	93	65
90	6	0.2	0.45	0.35	0.12	0.04	**195**	97	63
90	6	0.2	0.45	0.35	0.12	0.04	**330**	100	56
70	6	0.2	0.45	0.35	0.12	0.04	60	53	41
90	6	0.2	0.45	0.35	0.12	0.04	60	81	66
110	6	0.2	0.45	0.35	0.12	0.04	60	86	65
90	6	**0.2**	0.45	0.35	0.12	0.04	60	81	66
90	6	**0.33**	0.45	0.35	0.12	0.04	60	71	55
90	**6**	0.2	0.45	0.35	0.12	0.04	60	81	66
90	**5**	0.2	0.45	0.35	0.12	0.04	60	25	15
109	**6.92**	0.263	0.45	0.35	0.12	0.04	65	100	**38**
109	**6.92**	0.263	0.45	0.35	0.12	0.04	105	100	**5**
109	**6.92**	0.263	0.45	0.35	0.12	0.04	180	100	**0**

The effect of stirring rate

Several experiments were conducted to ensure that the reaction was not mass transfer limited. Reactions were run at 0, 300, 500, and 700 rpm with identical other reaction conditions. It was observed that conversion and yield are independent of the stirring rate between 500 and 700 rpm (Figure 6).

a)



b)

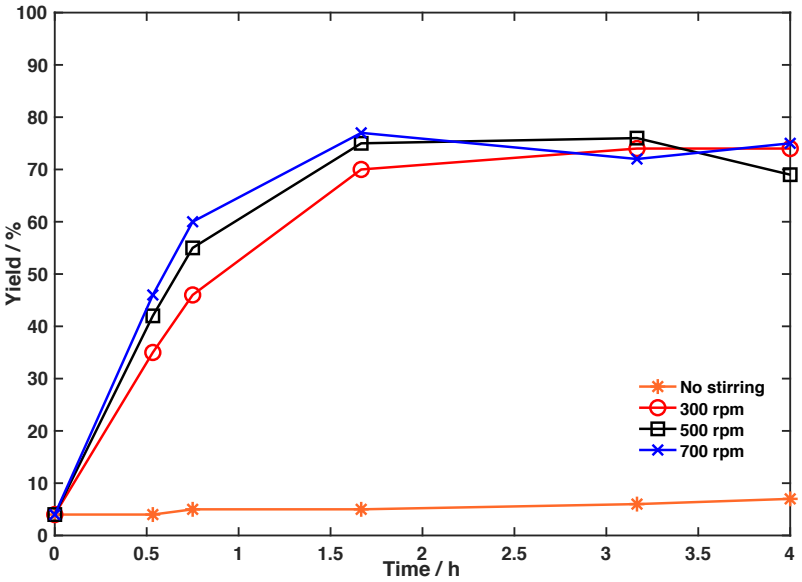


***Figure 6**. The effect of stirring rate on (a) conversion and (b) yield. One can see that conversion and yield are independent of the stirring rate between 500 rpm and 700 rpm.*

The effect of DMS

Another factor that affects the reaction outcome is dimethyl sulfide (DMS) concentration, which is part of the CST in the industrially produced waste. Three experiments were conducted with different quantities of DMS, whilst keeping all other reaction parameters constant. It was found that higher amounts of DMS significantly increase reaction rate in the investigated range 3.2 – 17.0 mol % DMS, whilst there was no significant difference in conversion and yield at the end of the reactions with different DMS quantities (Figure 7).

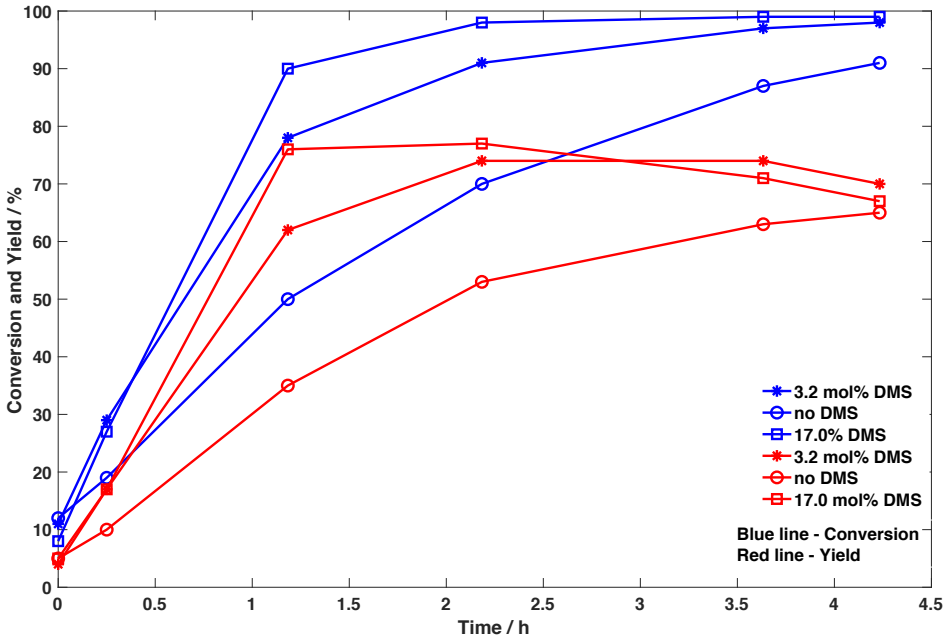


Figure 7*. The effect of different DMS concentrations on conversion (blue line) and yield (red line). High amounts of DMS increase the reaction rate, whilst there was no significant difference in the conversion and yield at the end of the reaction.*

Initial dataset collection and an algorithm-guided reaction optimisation

Following the initial exploratory experiments to set the optimisation variables bounds, the next set of experiments was performed using a space-filling Latin Hypercube sampling (LHS), which provides an initial dataset to initialise the TS-EMO algorithm.[47,48] 15 reactions were carried out, collecting an average of 6 samples at different times for each experiment. The dataset that combines the data collected in experiments used for exploration of the ranges of input variables, together with those designed by LHS, was then used to train the TS-EMO algorithm. The number of initial data points was chosen based on the order of magnitude reported by similar studies,[25] and the availability of experimental resources. The optimal

number of training data to select depending on the specifics of a chemical system, or more generally – on the shape of function to be discovered – is a research question without a definitive answer as of today. In some cases, it may be feasible to start optimisation with almost no training data, which would result in the significantly larger exploration requirement and the risk of experimental failures. In order to speed up data collection and to run experiments in parallel, batch-sequential sampling was implemented in the TS-EMO algorithm. This means that the algorithm designs four new reaction conditions in each run and these can be conducted in parallel.[36]

As one can see in Figure 8, the initial training dataset includes conditions with both high and low outputs for both objectives. The algorithm suggested conditions, which start from experiment 26, also include conditions resulting in both high and low objectives. This illustrates the behaviour of TS-EMO algorithm, balancing the trade-off between exploration (developing a good model, specifically targeting experiments to reduce uncertainty of the model) and exploitation (finding conditions that give optimal objective values – high yield and conversion in this case). The exploration of the algorithm can be observed mainly in the beginning (experiments from 26 to 54, Figure 8), where also regions with comparable low objective values are explored. This is the result of the intrinsic behaviour of the adopted algorithm, aiming to maximise objectives and, at the same time, reduce the uncertainty of GP model predictions by exploring areas of the input space distant from the found local optima. After a certain amount of exploration, the algorithm was mainly suggesting conditions that achieved high objective values. It is important to point out that the suggested conditions significantly differ from each other in terms of the process parameters that lead to high objectives, confirming that the algorithm is not stuck at a local. For instance, very similar outcomes were observed for significantly different values of temperature, e.g., 77 *vs* 105 °C (rows 3, 5, Table 3). Also note that the algorithm is never relying on pure exploration or exploitation but is always solving an exploration-exploitation trade-off.

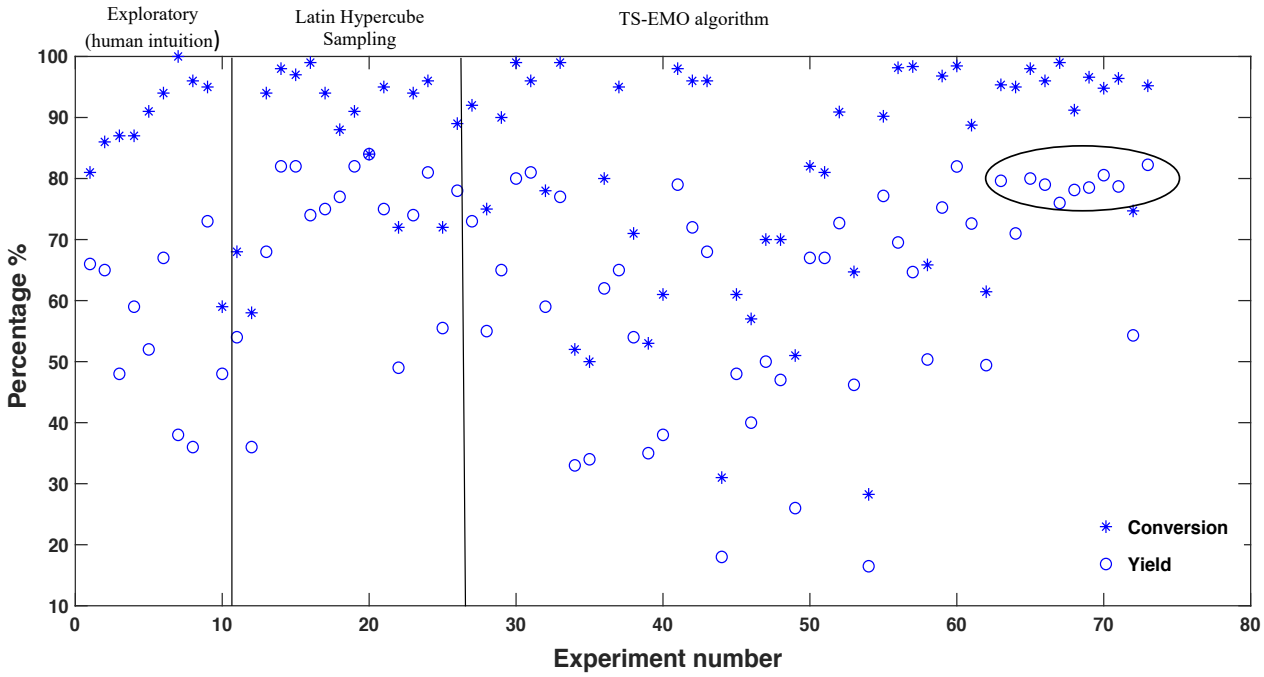


Figure 8. *Results of the optimisation driven by a statistical algorithm in the absence of a physical process model. Dataset is split in a way as human intuition vs LHS (initial dataset) vs algorithm generated reaction conditions by TS-EMO.*

Table 3. *Experimentally validated reaction conditions and composition for the best solutions selected from the final Pareto front.*

T / °C	H_2SO_4 / M	Phase ratio (aq./org.)	α-pinene	3-carene	β-pinene	Time / min	Conversion / %	Yield / %
89	6.7	0.21	0.75	0.06	0.13	61	97	82
83	6.1	0.24	0.54	0.11	0.29	134	94	82
77	5.5	0.31	0.53	0.05	0.36	140	96	81
101	5.3	0.22	0.48	0.11	0.35	102	96	81
105	6.9	0.34	0.63	0.006	0.29	30	99	80

A commonly used rule of thumb to decide the number of experiments to reach the termination criteria in an optimisation process is to set a target for the objectives (e.g., 99% for conversion and 80% for yield) and stop when the target is reached.[27] Another approach is to see if the newly achieved objectives differ by less than a pre-defined values, say by 3%, from the previously obtained result.[27] As shown in Table 3 and Figure 8, 32 experiments were enough

to surpass both of these criteria. Loeppky *et al*. reported a study on choosing the sample size for a computer experiment and used the $10 \times d$ rule, d being the number of variables involved in a reaction.[49] In another study, based on TS-EMO algorithm, 68 and 78 experiments were performed to optimise four continuous variables in an S_NAr and N-benzylation reactions, respectively.[25] In this work, we see that after 60 experiments the TS-EMO algorithm suggests a dozen of consecutive optimal solutions, Figure 8. A further stopping criterium is to reach a stable plateau in the hypervolume change as a function of the number of experiments, as highlighted in Figure 12 in the Appendix for both steps of reaction considered in this work.

In the earlier studies we have used TS-EMO to optimise automated reaction and separation systems in continuous flow[25] or batch experiments with a fixed batch time.[39] In this study, the batch time is considered as a continuous degree of freedom. However, the batch time needs to be handled individually during optimisation because it is cheap to withdraw several samples from the same batch at different batch times. Thus, we extend the TS-EMO algorithm accordingly to a two-step optimisation procedure: First, we run the algorithm on the full set of optimisation variables (including batch time). This gives us a suggested experiment, i.e., reaction conditions, and a suggested batch time. Second, we fix the reaction conditions and re-run the algorithm to suggest additional batch times. This gives us a suggested experiment with a set of batch times. The number of suggested batch times can be adapted through a batch sequential approach that has been described in our previous work[36] and is available as an option in the open-source Matlab implementation.[44]

The results of the optimisation of acid catalysed ring opening step are given in Figure 9. The results include a range of high and low values for the objectives, which are useful to efficiently explore the input variable space and reduce the model uncertainty. Although the objectives increase simultaneously for most of the sample points, one can see that maximum conversion could be achieved at the lowest value for yield. This indicates a non-obvious correlation between the objectives. The blue asterisks are experimentally identified Pareto points. One can see a cluster of optimal solutions explored by the TS-EMO algorithm leading up to the Pareto points. In Figure 9, we also highlighted the Pareto points of the initial training data set, comprising the expert-guided experiments and the LHS points. As one can see, two of these points can be also found in the Pareto front of the final optimisation (97% conversion and 82% yield, and 84% conversion and 84% yield). However, the rest of the points found at the end of the optimisation procedure was significantly higher than the ones only based on the training

dataset and, most importantly, the outcomes were obtained under a variety of conditions, some of which are particularly relevant from an industrial point of view, i.e., lower temperature and lower acid concentration.

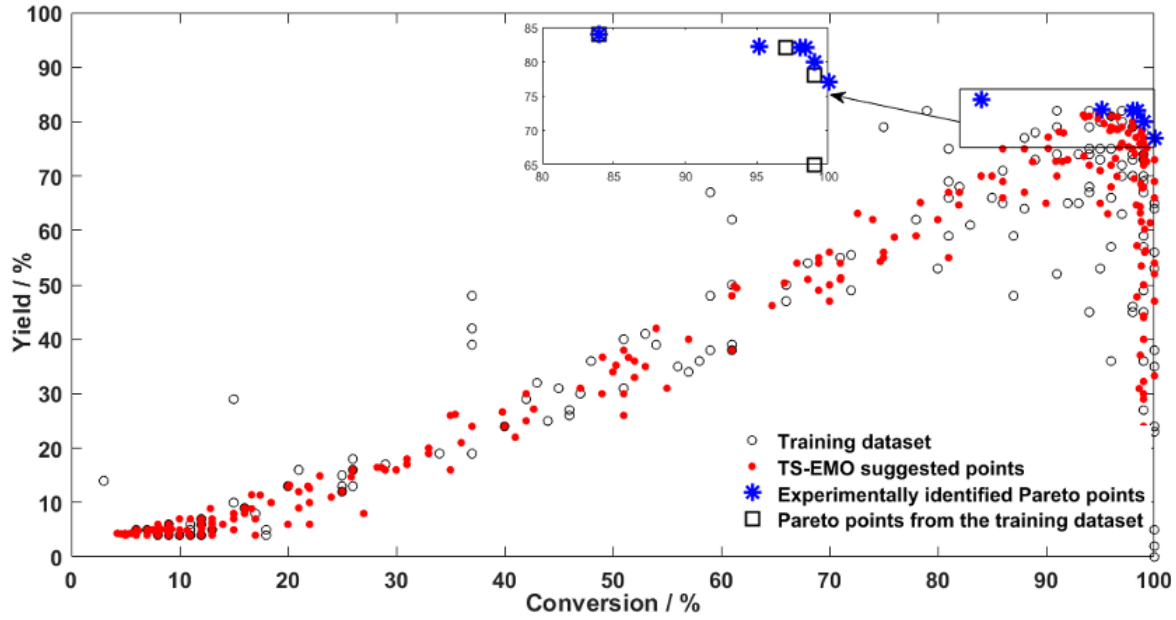


***Figure 9**. All the sample points for conversion and yield in acid catalysed ring opening reaction. Empty circles indicate the training dataset, red full circles indicate the result from TS-EMO suggested conditions, and the blue asterisks indicate Pareto points, including top five solutions given in Table 3; Empty squares indicate the Pareto point of the original training dataset.*

Finally, it is worth stressing that the optimised solutions given in Table 3 are in agreement with general chemistry-based observations previously reported in the literature.[6, 9, 11] In particular, low 3-carene content is crucial to obtain fast conversion and low occurrence of polymerisation. This supports the previously reported suggestion of blending turpentine feedstocks or distilling carene away from pinenes mixtures to ensure good yields. Also, with more forcing conditions, i.e., higher acid concentration and temperature, the reaction can be stopped sooner and still achieve good yields compared to milder conditions which take longer and would favour polymerisation of the alkene bonds. The best solutions summarised in Table 3 were generally higher than the ones previously reported for similar systems, i.e., ~ 70%. Higher isolated yields were reported (~ 90%) for systems with no 3-carene and higher contents of DMS, even though the results are not directly comparable because of the different upper limit for DMS concentration of 3.7 % used in this work.[9] It is important to highlight that high outputs values

were also obtained under a variety of conditions that were previously unexplored, with reaction times reduces down to 30 min for systems with a volume three times larger the one previously reported in the literature.[9] A final remark concerns the attainment of high values of yield and conversion, 99% and 72%, respectively, even in the presence of high contents of 3-carene, i.e., 0.35 molar fraction (exp. 64, Table 9 in the Appendix). In fact, depending on its origin, CST can contain high amounts of 3-carene, whose separation adds significant costs to the process. Although yield maximisation in the presence of high 3-carene content was not the aim of our optimisation, the results of explorative experiments, guided by the surrogate model-based approach, suggested new interesting conditions that will be further investigated in future research.

Optimisation of the radical dehydrogenation reaction

A similar procedure was followed for the second reaction step in order to maximise conversion of terpinene isomers **II**a-c and the selectivity to *p*-cymene. However, four important differences must be highlighted: (*i*) a high accuracy analytical method was used to determine the product concentration, (*ii*) a sequential sampling was adopted, since the experimental setup only allowed to carry out one reaction at a time, (*iii*) constraints of input variables were only based on human intuition and limitations of the experimental setup, and (*iv*) the algorithm was only trained using a set of experiments suggested by LHS.

Eight input variables were selected: temperature, flow rates of the air and the liquid phases, volumetric fractions of the reactants and the radical initiator in the liquid phase, and time. The ranges are summarised in Table 4. The experiment was performed in a flow system with a recirculation, effectively resulting in a batch-like observed temporal response. However, using the compact reactor with internal structure and embedded heat exchangers allows to operate at a very broad range of temperatures and pressures, whilst not compromising on safety. The internal packing of the reactor is an efficient radical scavenger, whereas the micro-heat exchanger has previously been shown to be highly efficient in elevated temperature selective oxidation reactions.[45] We hypothesised that the addition of a radical initiator may enhance the reaction rate, especially at short residence times; this is based on our previous work on aerobic oxidation in the liquid phase.[50]

Table 4. *Lower and upper bounds for the input variables in the optimisation of the second reaction. The numbers given for the starting materials indicate their volumetric fraction.*

Range	Temp / °C	Q_{liq} / mL min^{-1}	Q_{gas} / mL min^{-1}	α-terpinene	γ-terpinene	Terpinolene	Tert-butyl hydroperoxide	Time / min
Lower	80	0.1	5	0.00	0.00	0.00	0.00	0
Upper	150	5.0	120	0.22	0.22	0.22	0.22	240

The results with the highest yields, and conversions above 96% are reported in Table 5, whereas the whole sequence of experiments is reported in Figure 10.

Table 5. *Experimentally validated reaction conditions and composition for best solutions; the first line row represents the single Pareto solution.*

Temp / °C	Q_{liq} / mL min^{-1}	Q_{gas} / mL min^{-1}	α-terpinene	γ-terpinene	Terpinolene	Tert-butyl hydroperoxide	Time / min	Conversion (%)	Yield (%)
138	2.63	98.5	0.059	0.142	0.0068	0.220	234	100	62.5
132	3.48	119.0	0.073	0.175	0.0220	0.146	240	96.6	58.3
136	2.85	105.0	0.096	0.205	0.0296	0.180	230	99.2	54.5
138	4.48	45.6	0.054	0.166	0.0198	0.119	235	97.7	51.2
136	4.70	75.6	0.106	0.165	0.0222	0.098	210	98.0	51.1

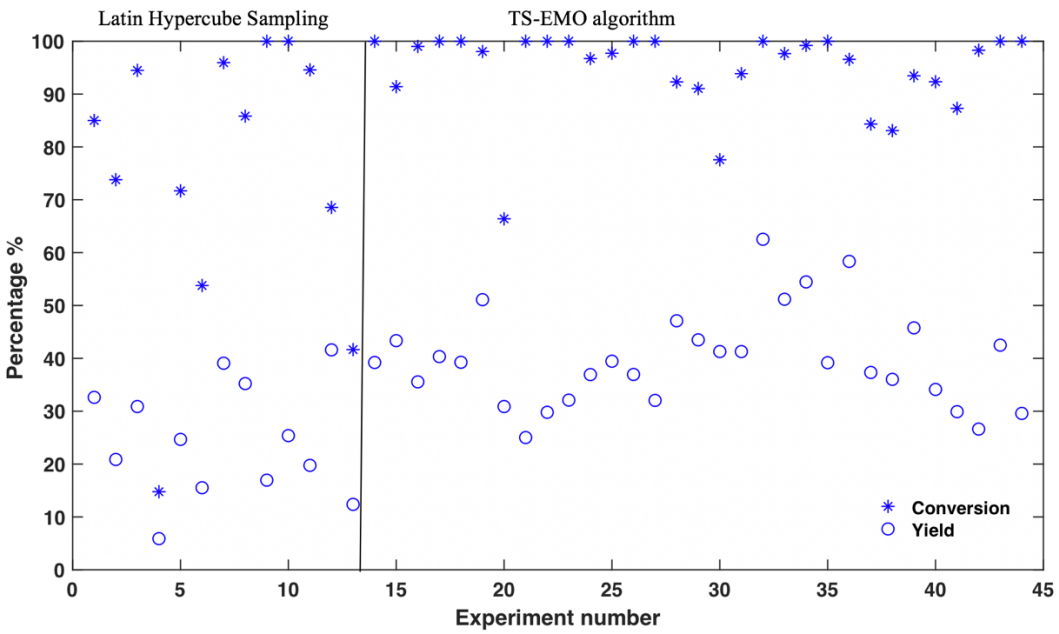


***Figure 10**. Results of the optimisation of reaction of p-cymene synthesis driven by TS-EMO algorithm. The data is split into the experiments suggested by LHS (initial dataset), and the algorithm-generated reaction conditions.*

It is worth highlighting several differences in this optimisation with respect to the previous case. For the second step of the overall route to terephthalic acid, the optimum was found in 44 experiments and the best results were all obtained under similar conditions, e.g., temperature and reaction time are in a relatively narrow range, 132 – 138 °C and 210 – 240 min, respectively. Any attempt of exploration outside of this narrow area of input variables resulted in a significant decrease in the target outputs. This indicates a much simpler solution space with a single optimum, compared to multiple optimal solutions in the first reaction. The entire dataset in the output space is reported in Figure 11, highlighting the existence of a single Pareto point in this specific case. As for the first step of reaction, also in this case we reported the Pareto points of the training data set, demonstrating that the optimization procedure led to a significant improvement of the desired targets.

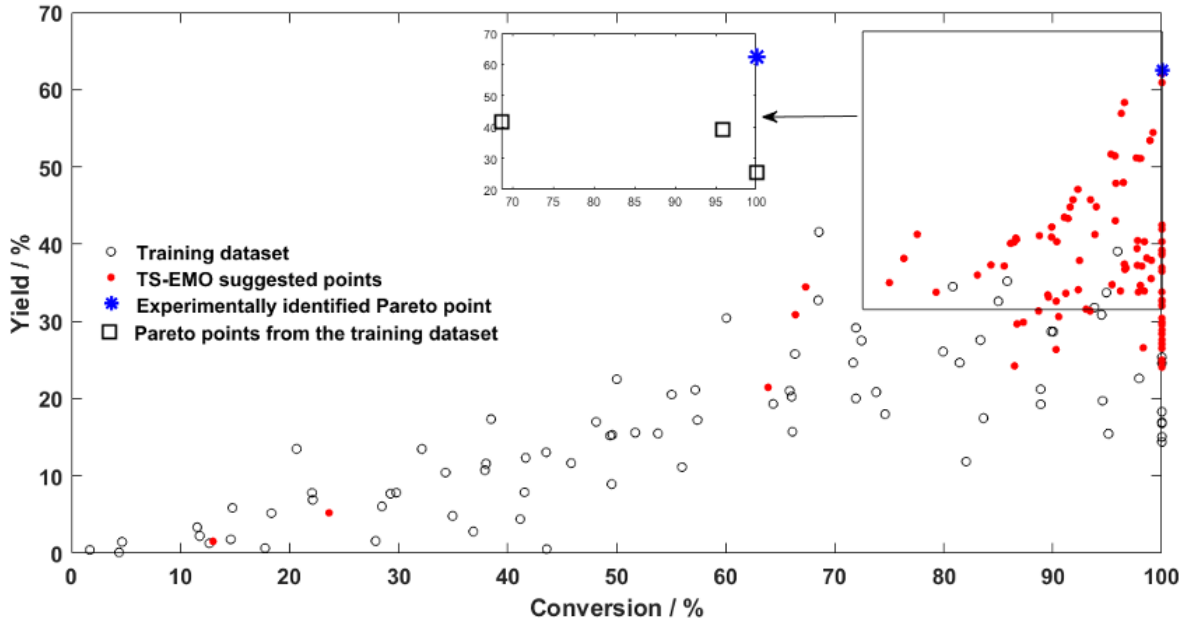


***Figure 11**. All the sample points for conversion and yield in p-cymene synthesis. Empty circles indicate the training dataset, red full circles indicate the result from TS-EMO suggested conditions, and the blue asterisk indicate the identified Pareto point; empty squares indicate the Pareto front of the initial training dataset. TS-EMO suggested cluster of optimal solutions leading up to the Pareto point.*

We also show that in this case the sequential (as opposed to batch-sequential) optimisation strategy results in a more efficient knowledge acquisition from the algorithmic point of view. In this case, each suggested experiment takes into account the results of the previous ones, whereas in a batch-sequential optimisation a certain number of experiments is suggested and some of the suggested experiments may not be as informative as each of the experiments in the sequential optimisation. Batch-sequential optimisation may suit better the problems where both discrete and continuous reaction variables need to be optimised simultaneously, as well as when balancing algorithmic efficiency and experimental budget (for example, when it is cheap to realise batch experiments).

Some of the best obtained results are in line with the observations reported elsewhere.[51-54] In particular, the highest yields reported in the literature are obtained when γ-terpinene is the major component.[54] Thus, it could be possible to further optimise the overall reaction sequence by modifying the feed into the second reaction step (Scheme 1) after the first reaction. In this case direct comparison with previously reported yields is not possible, since the investigated

system is substantially different for the following reasons: (*i*) no DMS was used in the starting mixture, (*ii*) p-xylene was used as solvent, and (*iii*) the experimental set-up was changed from batch to continuous flow with recycle.

Sensitivity of the developed statistical models to input variables

Within the TS-EMO algorithm, Gaussian process surrogate modes are trained on the experimental data. The Gaussian processes contain a number of hyperparameters that can be used to analyse sensitivity of model to input variables, which also implies the relevance or significance of the specific input variables to the objectives of the optimisation. In particular, *lengthscale* hyperparameters describe the influence of input variables on the reaction output, a concept known as *automatic relevance determination*.[42] A lower value of a lengthscale, θ_i, indicates a greater contribution of the corresponding input variable to the objective. In our study of the first step of CST conversion, we can see that temperature, reaction time and acid concentration have strong influence on conversion and yield, Table 6. It is important to point out that the acid concentration has a relatively stronger contribution to conversion than it does to yield. Temperature, on the other hand, has a relatively stronger contribution to yield than it does to conversion. This shows that a higher acid concentration has stronger impact on formation of by-products (e.g., polymers) than temperature. Compared to α-pinene, 3-carene and β-pinene concentrations have smaller contributions to conversion, which matches with the reaction composition where α-pinene exists in a larger quantity and reacts significantly faster than 3-carene leading to a greater contribution to the objectives (Table 8 in the Appendix). Even though β-pinene is quickly consumed, its lower quantity in the starting CST sample, compared to α-pinene, decreases its contribution to conversion and yield.

Besides the lengthscales, hyperparameters σ_f and σ_{noise} correspond to the output variance and noise hyperparameter, respectively. Low output variance, σ_f, indicates that the model is not sensitive to small fluctuations (noise) in the input data and can be used to make accurate predictions on a new dataset. High output variance, on the other hand, implies that the model does not perform well on the dataset it is not trained on, indicating overfitting of the input data. Overfitting happens when noise in the input data interferes with the signal and causes the algorithm to model noise when it is trained on a noisy dataset. Low values for hyperparameters σ_{noise} and σ_f indicate the quality of the input data and model accuracy to make good predictions on a new dataset, respectively.

Hyperparameters for the model of the second reaction step are listed in Table 7. In this case, the lowest hyperparameter values are associated with reaction time for both objectives. This is in agreement with the experimental data. The decrease in yield cannot be ascribed to consecutive reactions involving the desired product, since yield monotonously increases with reaction time. High gas flow rates appear to be associated with the higher conversion, suggesting an influence of mass transfer limitations on the reaction. Concentration of radical initiator does not affect the rate of the reaction significantly, but it is highly correlated with selectivity to the product of interest.

***Table 6**. Hyperparameters of GP surrogate models for the acid catalysed ring opening reaction.*

Hyperparameter	GP1 (Conversion)	GP2 (Yield)
$\theta_{temperature}$	3.27	2.29
$\theta_{phase\ ratio}$	11.69	10.98
$\theta_{conc.\ H2SO4}$	1.46	2.23
$\theta_{alpha\text{-}pinene}$	3.85	7.00
$\theta_{3\text{-}carene}$	22.74	14.00
$\theta_{beta\text{-}pinene}$	31.62	4.04
θ_{time}	1.48	2.13
σ_f	1.57	2.443
σ_{noise}	$6.14 \cdot 10^{-6}$	$6.14 \cdot 10^{-6}$

Table 7. *Hyperparameters of GP surrogate models for the synthesis of p-cymene.*

Hyperparameter	GP1 (Conversion)	GP2 (Yield)
$\theta_{temperature}$	5.11	31.49
θ_{Qliq}	18.73	11.91
θ_{Qgas}	2.42	7.53
$\theta_{alpha\text{-}terpinene}$	14.72	5.75
$\theta_{gamma\text{-}terpinene}$	23.21	11.77
$\theta_{terpinolene}$	7.05	4.37
θ_{tBuOOH}	31.62	2.21
θ_{time}	2.30	3.95
σ_f	1.08	1.79
σ_{noise}	$1.43 \cdot 10^{-2}$	$6.80 \cdot 10^{-6}$

Conclusions

We have demonstrated multi-objective reaction development on previously reported crude sulphate turpentine (CST) conversion to *p*-cymene,[9, 10] in the absence of any prior kinetic model using an extended version of Bayesian optimisation algorithm TS-EMO. Eight continuous variables were optimised for acid-catalysed ring opening reaction (step 1) in batch, and radical dehydrogenation reaction (step 2) in flow to maximise conversion and yield, without including any prior physico-chemical mechanistic information. Optimisation results showed that the algorithm suggested experimental points that include low and high values for the objectives. This was necessary to reduce the model uncertainty by balancing the exploration-exploitation trade-off.

For step 1, the algorithm was able to suggest a group of optimal conditions after 60 experiments. Optimal solutions were achieved under very different range of input variables values, indicating that the algorithm was not stuck at a local optima. However, this was not the case for step 2, where the optimal conditions were achieved in 44 experiments under similar values for temperature and reaction time. The narrow range for these input variables indicates simpler solution space for step 2 with a single optimum. In both reactions, the model suggested

cluster of optimal conditions around the experimentally identified Pareto points, which revealed the trade-off between the objectives.

The hyperparameters of the models revealed further information about model sensitivity to input variables. For instance, in step 1, the low values for hyperparameters of temperature, acid concentration, and reaction time indicated strong contribution to conversion and yield than other input variables. Low values for hyperparameters σ_{noise} and σ_f indicate the quality of the input data and model accuracy to make good predictions on a new dataset, respectively.

In summary, Bayesian optimisation algorithms are data efficient tools to develop accurate reaction models where *a priori* knowledge is not available, the number of input variables is large, and the objectives are competing. The developed models for individual steps could be used for potential process design and scale-up.

This is particularly relevant for the development of bio-waste routes to functional molecules, considering the chemical complexity and the high number of variables usually associated with such organic matrices. As a test case, in this work, we show that the application of hybrid approaches for the optimisation of reactions allows to speed up the identification of the best conditions to obtain aromatic components of highly valuable products, like p-cymene, from bio-based side-stream of traditional processes, representing a further step in the de-carbonisation of the chemical supply chain.

References

1. Guo, Z.; Yan, N.; Lapkin, A. A., Towards circular economy: integration of bio-waste into chemical supply chain. *Current Opinion in Chemical Engineering* **2019,** *26*, 148-156.

2. Jacob, P.-M.; Yamin, P.; Perez-Storey, C.; Hopgood, M.; Lapkin, A. A., Towards automation of chemical process route selection based on data mining. *Green Chem.* **2017,** *19*, 140-152.

3. Lapkin, A. A.; Heer, P. K.; Jacob, P.-M.; Hutchby, M.; Cunningham, W.; Bull, S.; Davidson, M. G., Automation of route identification and optimisation based on datamining and chemical intuition. *Farad. Discuss.* **2017,** *202*, 483-496.

4. Weber, J. M.; Lió, P.; Lapkin, A. A., Identification of strategic molecules for future circular supply chains using large reaction networks. *Reaction Chemistry & Engineering* **2019**.

5. Corma, A.; Iborra, S.; Velty, A., Chemical routes for the transformation of biomass into chemicals. *Chem. Rev.* **2007,** *107*, 2411-2502.

6. Linnekoski, J. A.; Asikainen, M.; Heikkinen, H.; Kaila, R. K.; Räsänen, J.; Laitinen, A.; Harlin, A., Production of p-Cymene from Crude Sulphate Turpentine with Commercial Zeolite Catalyst Using a Continuous Fixed Bed Reactor. *Organic Process Research & Development* **2014,** *18* (11), 1468-1475.

7. Zou, J.-J.; Chang, N.; Zhang, X.; Wang, L., Isomerization and Dimerization of Pinene using Al-Incorporated MCM-41 Mesoporous Materials. *ChemCatChem* **2012,** *4* (9), 1289-1297.

8. Eggersdorfer, M., Terpenes. *Ullmann's Encyclopedia of Industrial Chemistry* **2000**.

9. Tibbetts, J. D.; Bull, S. D., Dimethyl sulfide facilitates acid catalysed ring opening of the bicyclic monoterpenes in crude sulfate turpentine to afford p-menthadienes in good yield. *Green Chemistry* **2021,** *23* (1), 597-610.

10. Tibbetts, J. D.; Bull, S. D., p-Menthadienes as Biorenewable Feedstocks for a Monoterpene-Based Biorefinery. *Advanced Sustainable Systems* **2021**.

11. Williams, C. M.; Whittaker, D., Evidence for intimate ion-pair formation in the addition of acids to olefins. *Journal of the Chemical Society D: Chemical Communications* **1970,** (15).

12. Tibbetts, J. D.; Russo, D.; Lapkin, A. A.; Bull, S. D., Efficient Syntheses of Biobased Terephthalic Acid, p-Toluic Acid, and p-Methylacetophenone via One-Pot Catalytic Aerobic Oxidation of Monoterpene Derived Bio-p-cymene. *ACS Sustainable Chemistry & Engineering* **2021,** *9* (25), 8642-8652.

13. Holmen AB,. WO/2015/023225/A1, 2015.

14. Razavi, S.; Gupta, H. V., What do we mean by sensitivity analysis? The need for comprehensive characterization of "global" sensitivity in Earth and Environmental systems models. *Water Resources Research* **2015,** *51* (5), 3070-3092.

15. Rasheed, M.; Wirth, T., Intelligent Microflow: Development of Self-Optimizing Reaction Systems. *Angewandte Chemie International Edition* **2011,** *50* (2), 357-358.

16. Weissman, S. A.; Anderson, N. G., Design of Experiments (DoE) and Process Optimization. A Review of Recent Publications. *Organic Process Research & Development* **2014,** *19* (11), 1605-1633.

17. Pilipauskas, D. R., *Using Factorial Experiments in the Development of Process Chemistry*. Dekker: New York: 1999.

18. Box, G. E. P.; Hunter, W. G.; Hunter, J. S., *Statistics for experimenters : an introduction to design, data analysis, and model building*. Wiley: New York, 1978; p xviii, 653 p.

19. Gujral, G.; Kapoor, D.; Jaimini, M., An Updated Review on Design of Experiment (Doe) in Pharmaceuticals. *Journal of Drug Delivery and Therapeutics* **2018,** *8* (3).

20. Lendrem, D.; Owen, M.; Godbert, S., DOE (Design of Experiments) in Development Chemistry: Potential Obstacles. *Organic Process Research & Development* **2001,** *5* (3), 324-327.

21. Jeraal, M. I.; Holmes, N.; Akien, G. R.; Bourne, R. A., Enhanced process development using automated continuous reactors by self-optimisation algorithms and statistical empirical modelling. *Tetrahedron* **2018,** *74* (25), 3158-3164.

22. Veldhuis, C.; Gornicz, T.; Scholcz, T. P., Ship optimization using viscous flow computations in combination with generic shape variations and Design of Experiments. . In *PRAGS*, 2016.

23. Clayton, A. D.; Manson, J. A.; Taylor, C. J.; Chamberlain, T. W.; Taylor, B. A.; Clemens, G.; Bourne, R. A., Algorithms for the self-optimisation of chemical reactions. *Reaction Chemistry & Engineering* **2019,** *4* (9), 1545-1554.

24. Mateos, C.; Nieves-Remacha, M. J.; Rincón, J. A., Automated platforms for reaction self-optimization in flow. *Reaction Chemistry & Engineering* **2019,** *4* (9), 1536-1544.

25. Schweidtmann, A. M.; Clayton, A. D.; Holmes, N.; Bradford, E.; Bourne, R. A.; Lapkin, A. A., Machine learning meets continuous flow chemistry: Automated optimization towards the Pareto front of multiple objectives. *Chemical Engineering Journal* **2018,** *352*, 277-282.

26. Fabry, D. C.; Sugiono, E.; Rueping, M., Online monitoring and analysis for autonomous continuous flow self-optimizing reactor systems. *Reaction Chemistry & Engineering* **2016,** *1* (2), 129-133.

27. McMullen, J. P.; Stone, M. T.; Buchwald, S. L.; Jensen, K. F., An Integrated Microreactor System for Self-Optimization of a Heck Reaction: From Micro- to Mesoscale Flow Systems. *Angewandte Chemie International Edition* **2010,** *49* (39), 7076-7080.

28. Moore, J. S.; Jensen, K. F., Automated Multitrajectory Method for Reaction Optimization in a Microfluidic System using Online IR Analysis. *Organic Process Research & Development* **2012,** *16* (8), 1409-1415.

29. Echtermeyer, A.; Amar, Y.; Zakrzewski, J.; Lapkin, A., Self-optimisation and model-based design of experiments for developing a C–H activation flow process. *Beilstein Journal of Organic Chemistry* **2017,** *13*, 150-163.

30. Krishnadasan, S.; Brown, R. J. C.; deMello, A. J.; deMello, J. C., Intelligent routes to the controlled synthesis of nanoparticles. *Lab on a Chip* **2007,** *7* (11).

31. Ley, S. V.; Fitzpatrick, D. E.; Ingham, R. J.; Myers, R. M., Organic Synthesis: March of the Machines. *Angewandte Chemie International Edition* **2015,** *54* (11), 3449-3464.

32. Reizman, B. J.; Jensen, K. F., Feedback in Flow for Accelerated Reaction Development. *Accounts of Chemical Research* **2016,** *49* (9), 1786-1796.

33. Lagarias, J. C.; Reeds, J. A.; Wright, M. H.; Wright, P. E., Convergence Properties of the Nelder--Mead Simplex Method in Low Dimensions. *SIAM Journal on Optimization* **1998,** *9* (1), 112-147.

34. McMullen, J. P.; Jensen, K. F., An Automated Microfluidic System for Online Optimization in Chemical Synthesis. *Organic Process Research & Development* **2010,** *14* (5), 1169-1176.

35. Holmes, N.; Akien, G. R.; Savage, R. J. D.; Stanetty, C.; Baxendale, I. R.; Blacker, A. J.; Taylor, B. A.; Woodward, R. L.; Meadows, R. E.; Bourne, R. A., Online quantitative mass spectrometry for the rapid adaptive optimisation of automated flow reactors. *Reaction Chemistry & Engineering* **2016,** *1* (1), 96-100.

36. Bradford, E.; Schweidtmann, A. M.; Lapkin, A., Efficient multiobjective optimization employing Gaussian processes, spectral sampling and a genetic algorithm. *Journal of Global Optimization* **2018,** *71* (2), 407-438.

37. Manson, J. A.; Chamberlain, T. W.; Bourne, R. A., MVMOO: Mixed variable multi-objective optimisation. *Journal of Global Optimization* **2021,** *80* (4), 865-886.

38. Golovin, D.; Solnik, B.; Moitra, S.; Kochanski, G.; Karro, J.; Sculley, D., Google Vizier. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '17*, 2017; pp 1487-1495.

39. Amar, Y.; Schweidtmann, Artur M.; Deutsch, P.; Cao, L.; Lapkin, A., Machine learning and molecular descriptors enable rational solvent selection in asymmetric catalysis. *Chemical Science* **2019,** *10* (27), 6697-6706.

40. Jeraal, M. I.; Sung, S.; Lapkin, A. A., A Machine Learning-Enabled Autonomous Flow Chemistry Platform for Process Optimization of Multiple Reaction Metrics. *Chemistry–Methods* **2020,** *1* (1), 71-77.

41. Zhang, C.; Amar, Y.; Cao, L.; Lapkin, A. A., Solvent Selection for Mitsunobu Reaction Driven by an Active Learning Surrogate Model. *Organic Process Research & Development* **2020,** *24* (12), 2864-2873.

42. Williams, C. K.; Rasmussen, C. E., *Gaussian processes for machine learning*. MIT press Cambridge, MA: 2006; Vol. 2.

43. Amar, Y. Accelerating process development of complex chemical reactions. University of Cambridge, 2019.

44. Bradford, E.; Schweidtmann, A. M. TS-EMO GitHub. https://github.com/Eric-Bradford/TS-EMO.

45. Bavykin, D. V.; Lapkin, A. A.; Kolaczkowski, S. T.; Plucinski, P. K., Selective oxidation of alcohols in a continuous multifunctional reactor: ruthenium oxide catalysed oxidation of benzyl alcohol. *Appl. Catal. A: Gen.* **2005,** *288*, 165-174.

46. Helmdach, D.; Yaseneva, P.; Heer, P. K.; Schweidtmann, A. M.; Lapkin, A. A., A Multiobjective Optimization Including Results of Life Cycle Assessment in Developing Biorenewables-Based Processes. *ChemSusChem* **2017,** *10* (18), 3632-3643.

47. McKay, M. D.; Beckman, R. J.; Conover, W. J., A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. *Technometrics* **1979,** *21* (2).

48. Tang, B., Orthogonal Array-Based Latin Hypercubes. *Journal of the American Statistical Association* **1993,** *88* (424).

49. Loeppky, J. L.; Sacks, J.; Welch, W. J., Choosing the Sample Size of a Computer Experiment: A Practical Guide. *Technometrics* **2009,** *51* (4), 366-376.

50. Aworinde, S. M.; Wang, K.; Lapkin, A. A., Borate-assisted liquid-phase selective oxidation of n-pentane. *Applied Catalysis A: General* **2018,** *563*, 28-42.

51. Bi, L. W.; Zhang, Q. G.; Wang, P.; Zhao, Z. D.; Li, D. W.; Chen, Y. X.; Li, D. M.; Gu, Y.; Wang, J.; Liu, X. Z., Study on Gas-Phase Catalytic Conversion of Turpentine-Based Dipentene (TBDP) by Pd/C Catalysts. *Advanced Materials Research* **2011,** *236-238*, 27-34.

52. Colonna, M.; Berti, C.; Fiorini, M.; Binassi, E.; Mazzacurati, M.; Vannini, M.; Karanam, S., Synthesis and radiocarbon evidence of terephthalate polyesters completely prepared from renewable resources. *Green Chemistry* **2011,** *13* (9).

53. Iwamuro, H.; Ohshio, T.; Matsubara, Y., *Nippon Kagaku Kaishi* **1978,** (6), 909-911.

54. Asikainen, M.; Jauhiainen, O.; Aaltonen, O.; Harlin, A., Continuous catalyst-free aromatization of γ-terpinene using air as an oxidant. *Green Chem.* **2013,** *15* (11), 3230-3235.

Appendix

Table 8. *Reaction profiles for two different experiments to demonstrate the concentration change of species during the reaction.*

Time / min	α-pinene / M	3-carene / M	β-pinene / M	Limonene / M
0	2.298	0.531	1.781	0.194
2	2.263	0.536	1.717	0.202
87	0.620	0.408	0.034	0.303
119	0.212	0.330	0.012	0.231
200	0.005	0.157	0.007	0.103
258	0.000	0.049	0.000	0.005

Time / min	α-pinene / M	3-carene / M	β-pinene / M	Limonene / M
12	2.332	1.486	0.512	0.216
23	2.258	1.486	0.470	0.216
82	1.771	1.422	0.194	0.248
88	1.697	1.432	0.172	0.251
101	1.589	1.417	0.135	0.253
147	1.181	1.358	0.049	0.266
260	0.330	1.156	0.010	0.239

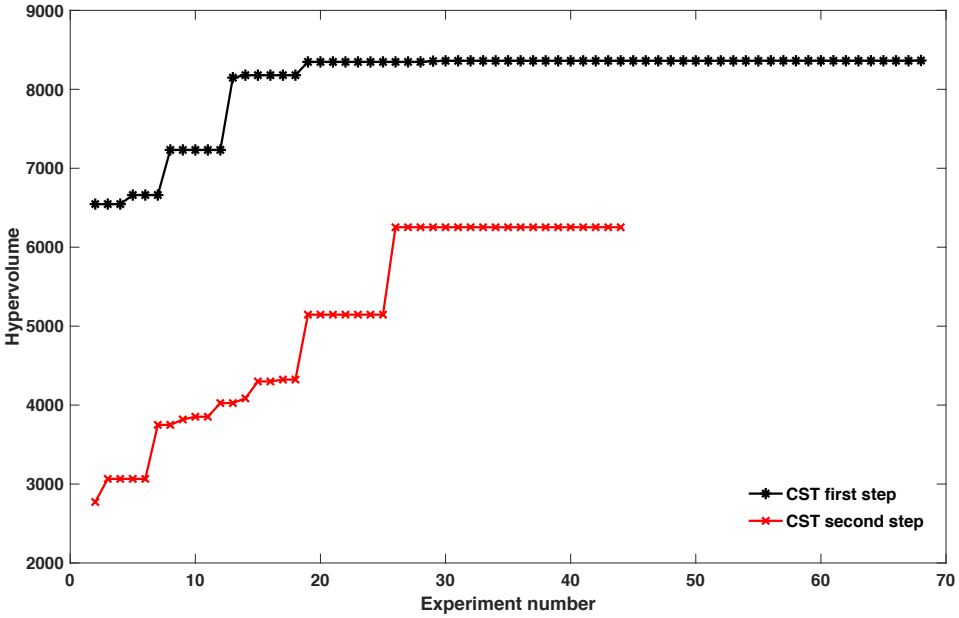


Figure 12. *Change in hypervolume for the first and the second steps over number of experiments. One could see that the model converges to a certain hypervolume for both of the reaction steps.*

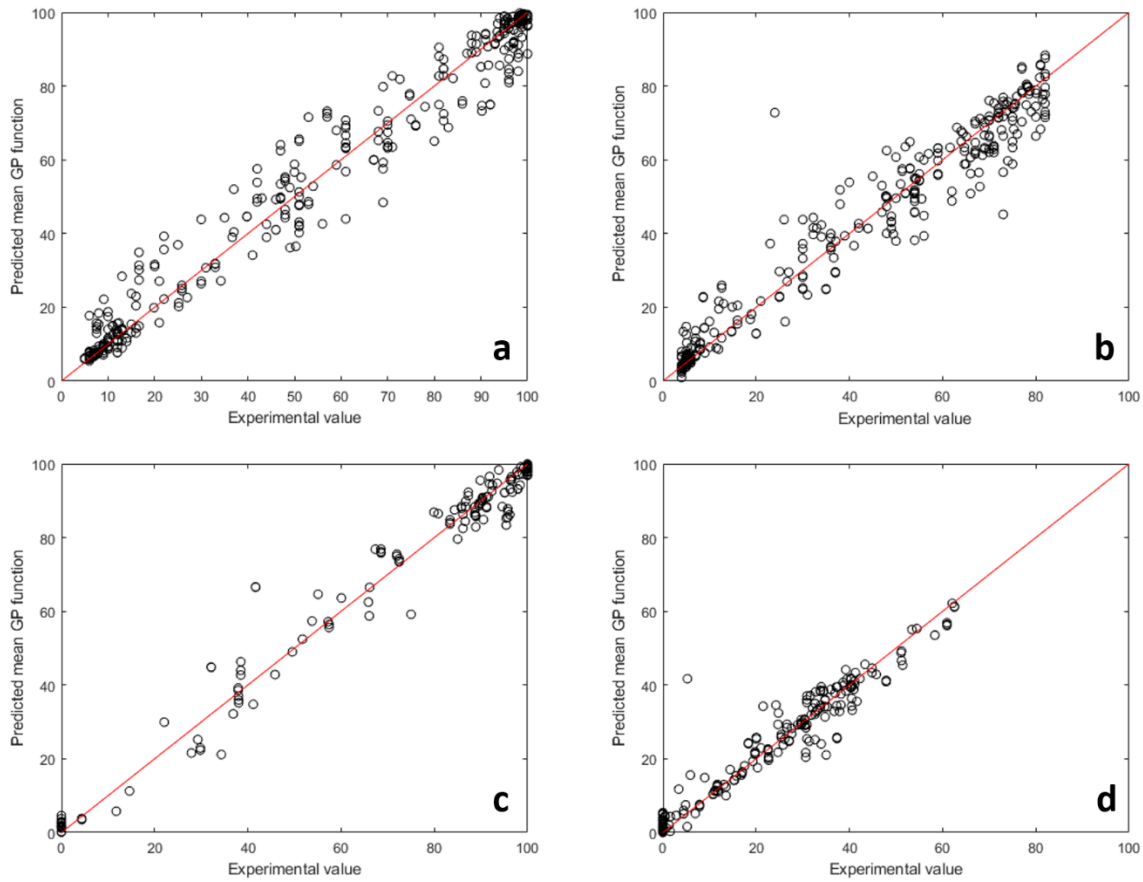


Figure 13. *10-fold cross validation of the GP models. Prediction vs. experimental value. a) 1st step of reaction, conversion; b) 1st step of reaction, yield; c) 2nd step of reaction, conversion; d) 2nd step of reaction, yield.*

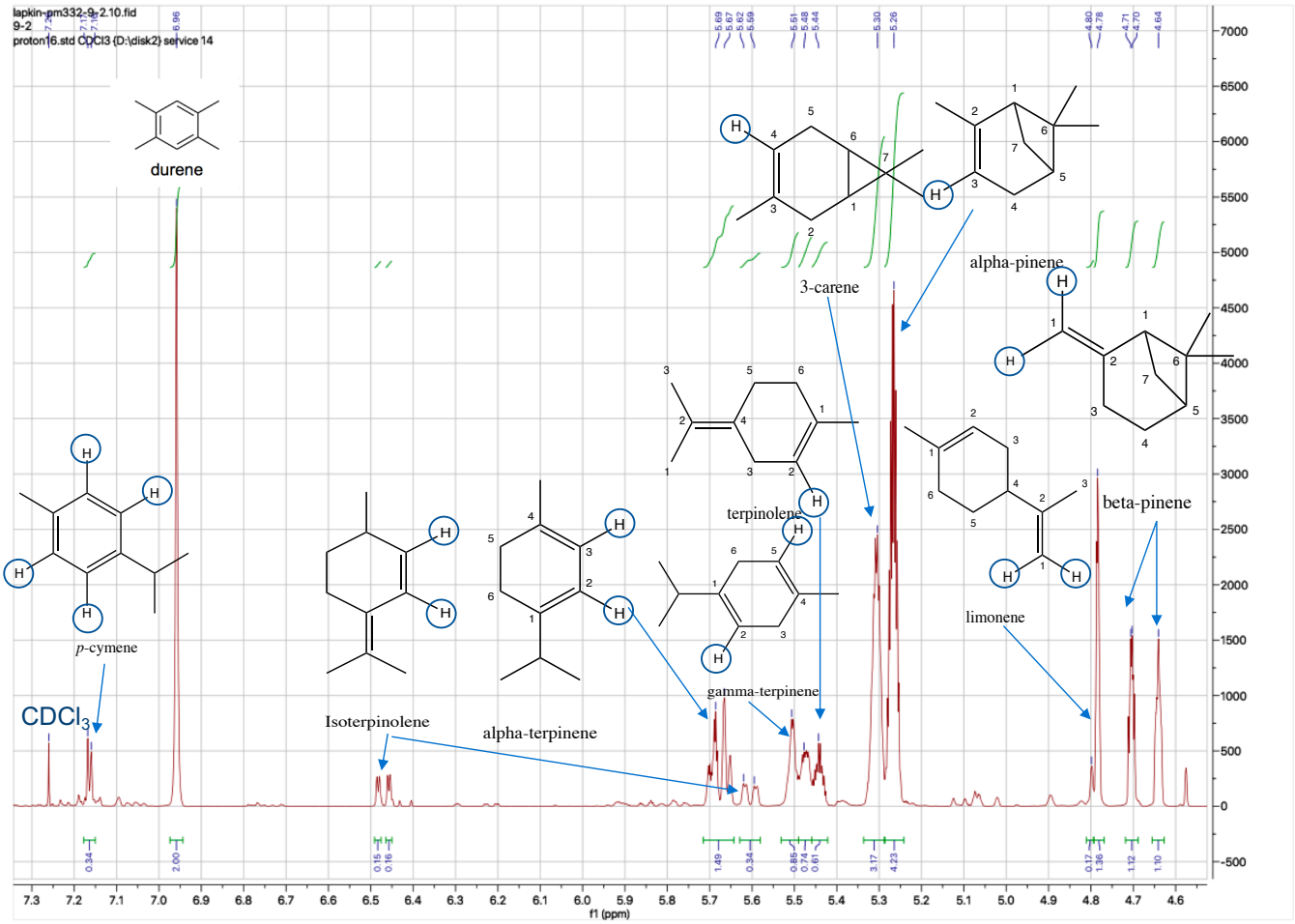


Figure 14. *^{1}H NMR analysis of the acid catalysed ring opening reaction. 1,2,4,5-tetramethylbenzene (0.032 mol) was used as an internal standard and the reaction species were quantified relative to the internal standard amount. The following peaks were integrated to calculate the amount of reaction species by ^{1}H NMR (400 MHz, $CDCl_3$) δ: 7.16 (4H, p-cymene), 6.48 (1H, isoterpinolene), 5.68 (2H, α-terpinene), 5.59 (1H, isoterpinolene), 5.51 (2H, γ-terpinene), 5.44 (1H, terpinolene), 5.30 (1H, 3-carene), 5.26 (1H, α-pinene), 4.78 (2H, limonene), 4.70 (1H, β-pinene), 4.64 (1H, β-pinene). For isoterpinolene, average of peaks at 6.48 and 5.59 were taken as the final peak integration value. Same approach was taken for two separate peaks for β-pinene at 4.70 and 4.64.*

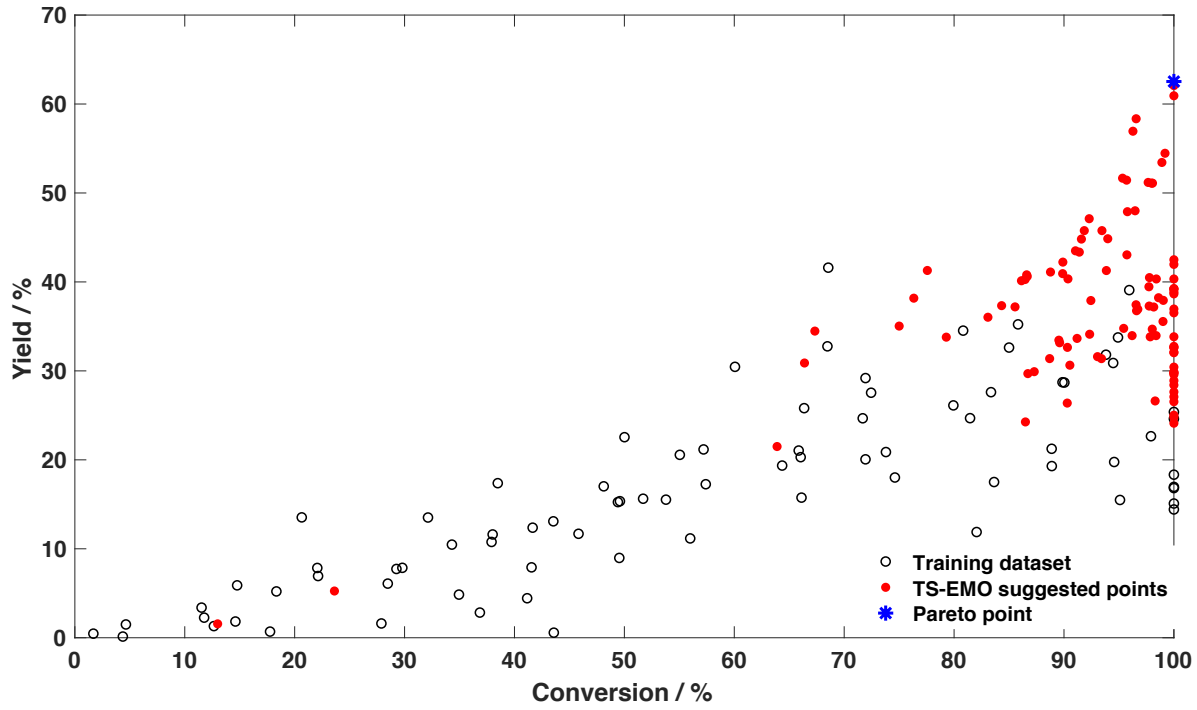


Figure 15. *All the sample points for conversion and yield in p-cymene synthesis. Empty circles indicate the training dataset, red circles indicate the result from TS-EMO suggested conditions, and the blue stars indicate experimentally identified Pareto points. The algorithm suggested conditions have significantly high objective values than the training dataset. The blue point corresponds to the experimentally identified Pareto point. TS-EMO suggested cluster of optimal solutions leading up to the Pareto point.*

***Figure 16**. Pseudocode for the modified TS-EMO algorithm used in this work.*

TS-EMO extended version: A pseudocode for the model-free Bayesian optimisation of continuous variables for multiple objectives using exploratory, Latin-Hypercube sampling (LHS), and surrogate Gaussian Process (GP) models for complex reaction optimisation.

START

1. Conversion, Yield ← objectives #identify reaction objectives
2. var_list ← list of variables # define list of reaction variables
3. dataset_1 ← exploratory data # run reactions based on the domain knowledge
4. LB_var_list ←list of lower bounds for the variables
5. UB_var_list ← list of upper bounds for the variables.

return dataset_1, LB_var_list, UB_var_list

LHS

1. var_limits ← array([LB, UB])
2. selected_conditions ←LHS(Var_limits, number_of_reactions)
3. dataset_2 ← experimental values for the selected conditions
4. training_dataset ← dataset_1 + dataset_2

return training_dataset

TS-EMO

Input: Black-box function g(x)
Type of covariance function of GP
Training dataset $X_0 = \{x_1^{(1)}, x_2^{(1)}, x_3^{(1)}, \dots, x_N^{(m)}\}$ # m as total number of variables
Corresponding set of observations $Y_0 = \{y_1^{(obj1)}, y_2^{(obj1)}, y_3^{(obj1)}, \dots, y_N^{(obj2)}\}$

for objective functions:

1. for i ← 0, …, N
2. Build GP models for each objective from current dataset X_i and Y_i
3. Approximately sample $f^{(i)}(\boldsymbol{x})$ from $GP(m, k|X_i, Y_i)$ using spectral sampling
4. Determine $x_{n+1} = argmin_{x\in\chi} f^{(i)}(x)$ to find Pareto front of sampled functions
5. Evaluate $y_{n+1} = g(x_{n+1})$ to select points with largest estimated improvement in hypervolume
6. For multiple sampling for time variable, repeat steps 3-4-5 using the predicted values for y_{n+1} whilst keeping all the variables constants, except time.
7. Carry out the reactions for experimental validation of y_{n+1}
8. Update the dataset $X_{i+1} = \left\{x_1^{(1)}, x_2^{(1)}, x_3^{(1)}, \dots, x_{n+1}^{(m)}\right\}$, $Y_{i+1} = \{y_1^{(obj1)}, y_2^{(obj1)}, y_3^{(obj1)}, \dots, y_{n+1}^{(obj2)}\}$
9. Update n ← n+1
10. **if** termination criteria reached:
    terminate
    else:
    repeat steps 2-9.

return X_n, Y_n, Pareto front, lengthscale parameters.

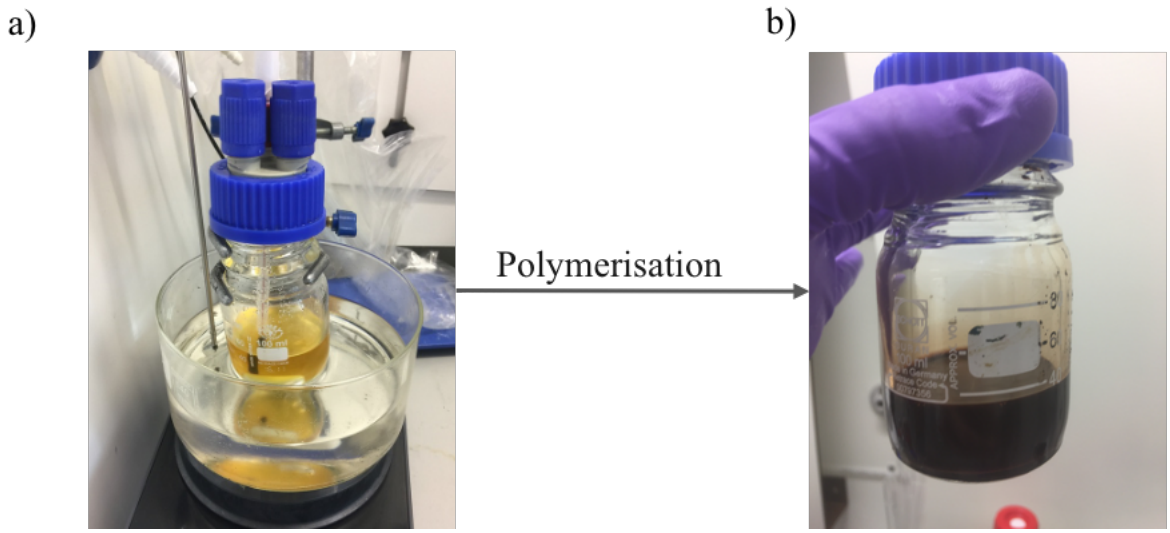


***Figure 17**. a) Experimental batch setup used for the first step of reaction: acid-catalysed ring opening and b) polymerisation product at high reaction temperature.*

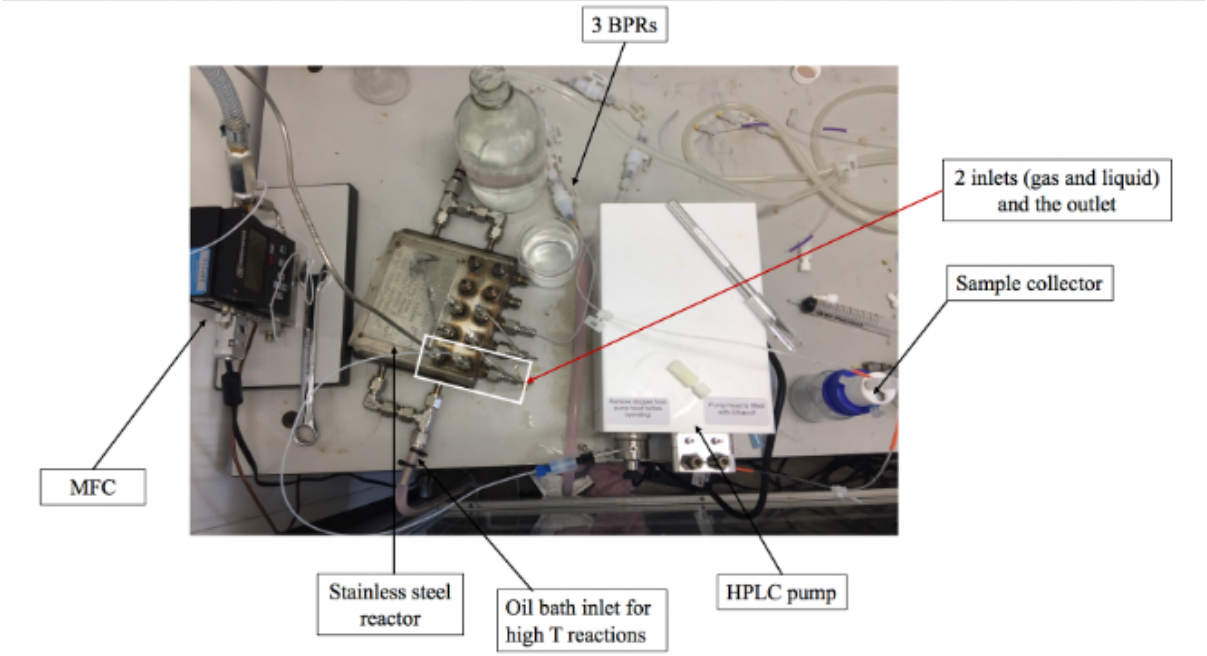


***Figure 18**. Experimental setup used for the second step of reaction: aerobic dehydrogenation of terpinenes. Detailed description of the continuous packed reactor can be found in Plucinski et al., 2005.*

Table 9. *The entire dataset for optimisation of acid catalysed ring opening reaction (step 1). Nonzero values for conversion and yield at time = 0 hr indicates fast reactivity as soon as the organic and the aqueous phases are mixed at the start of the reaction. The dataset is categorized according to their experiment number. The numbers given for the starting materials indicate their mole fraction. Limonene molar fraction is kept constant at 0.04, which is the commonly observed concentration of limonene in this bio-waste feedstock.*

Exp No:	T / °C	Phase ratio (aq./org.)	H_2SO_4 / M	α-pinene	3-carene	β-pinene	Time / min	Conversion / %	Yield / %
1	90	0.20	6.00	0.45	0.35	0.12	60	81.00	66.00
							120	93.00	65.00
							195	97.00	63.00
							330	100.00	56.00
2	110	0.20	6.00	0.45	0.35	0.12	30	57.00	34.00
							60	86.00	65.00
							90	96.00	57.00
							150	100.00	35.00
3	90	0.20	5.00	0.45	0.35	0.12	90	47.00	30.00
							120	61.00	38.00
							180	87.00	48.00
							240	94.00	45.00
4	90	0.33	6.00	0.45	0.35	0.12	60	71.00	55.00
							90	87.00	59.00
							120	95.00	53.00
							165	98.00	45.00
							199.8	100.00	23.00
5	70	0.26	7.00	0.45	0.35	0.12	30	42.00	29.00
							60	66.00	47.00
							90	80.00	53.00
							120	91.00	52.00
							180	100.00	24.00
6	96	0.17	5.10	0.45	0.35	0.12	73.02	83.00	61.00
							138	92.00	65.00
							181.8	94.00	67.00
							238.02	96.00	66.00
7	109	0.26	6.92	0.45	0.35	0.12	65	100.00	38.00
							105	100.00	5.00
							180	100.00	0.00
8	76	0.20	5.50	0.45	0.35	0.12	60	40.00	24.00
							80	56.00	35.00
							160	96.00	36.00

							230	100.00	2.00
9	70	0.20	6.00	0.45	0.35	0.12	60	53.00	41.00
							120	82.00	68.00
							195	95.00	73.00
							330	100.00	53.00
10	75	0.20	4.90	0.45	0.35	0.12	150	43.00	32.00
							230	59.00	48.00
11	77	0.20	4.90	0.65	0.15	0.1	160	48.00	36.00
							240	68.00	54.00
12	74.10	0.27	4.22	0.60	0.20	0.14	0	9.20	4.50
							12	12.80	4.68
							65	25.15	11.93
							117	36.65	18.88
							154	44.05	24.53
							198	51.25	30.50
							245	57.65	36.05
13	101.43	0.33	4.04	0.51	0.08	0.34	0	7.55	4.23
							30	26.40	12.65
							66	46.35	26.48
							102	58.60	37.93
							166	80.95	58.68
							197	87.75	64.05
							251	94.20	68.15
14	83.07	0.24	6.08	0.54	0.11	0.29	0	9.75	4.38
							16	51.10	40.23
							48	80.75	69.38
							134	94.25	81.60
							170	95.45	82.13
							213	96.50	82.45
							254	97.60	81.70
15	89.14	0.21	6.71	0.75	0.06	0.13	0	11.40	6.45
							24	91.40	79.40
							61	97.25	81.80
							121	99.10	78.40
							174	99.40	72.65
							208	99.40	68.70
							247	99.70	65.38
16	109.00	0.16	6.95	0.53	0.01	0.40	6	98.65	73.93
							50	98.40	46.35
							101	98.60	27.48
							160	98.70	23.10
17	85.56	0.30	6.25	0.34	0.25	0.33	0	16.00	8.65
							10	52.65	40.88

							65	86.00	71.10
							113	94.15	75.43
							157	97.50	73.13
							210	99.10	69.65
							250	99.50	64.48
18	94.89	0.25	4.75	0.45	0.32	0.15	157	80.75	74.88
							215	88.45	77.30
							264	93.40	74.33
19	96.95	0.26	6.59	0.49	0.36	0.09	77	91.20	82.40
							123	94.90	80.38
							168	97.70	74.30
							205	98.80	66.80
							266	98.80	49.10
20	87.29	0.29	4.85	0.43	0.27	0.24	32	28.70	14.90
							62	47.95	37.30
							115	67.03	59.00
							200	81.65	79.35
							235	84.13	84.30
21	92.13	0.17	5.07	0.60	0.15	0.18	0	11.00	4.45
							15	28.65	17.40
							71	77.75	62.25
							131	91.25	74.35
							171	94.70	75.20
							218	97.20	72.28
							254	98.20	69.58
22	72.26	0.23	4.46	0.74	0.03	0.15	0	8.95	4.18
							10	9.95	4.38
							66	24.60	12.58
							98	34.20	18.73
							141	45.75	27.23
							204	60.75	39.48
							255	71.80	49.40
23	104.79	0.34	5.27	0.66	0.14	0.14	0	11.90	5.68
							5	19.85	12.83
							84	89.00	73.15
							114	94.15	73.85
							155	97.30	70.00
							217	99.30	56.90
							244	99.20	45.23
24	76.56	0.31	5.52	0.53	0.05	0.36	0	12.40	4.30
							6	16.90	7.93
							71	84.60	66.40
							114	93.80	78.80

							140	95.80	80.78
							188	97.45	79.68
							234	97.90	78.75
25	78.85	0.18	5.72	0.48	0.27	0.19	0	7.65	4.68
							13	9.10	5.50
							70	25.75	15.53
							134	45.05	31.25
							170	53.95	38.95
							224	65.70	49.83
							264	71.95	55.50
26	104.57	0.20	5.84	0.63	0.14	0.17	0	12.35	6.63
							19	60.85	49.70
							64	89.35	77.53
							97	95.50	74.98
							140	98.15	70.23
							185	99.10	59.18
							268	98.90	36.20
27	79.00	0.24	4.60	0.46	0.10	0.33	1	8.00	4.00
							5	9.00	4.00
							71	40.00	24.00
							137	71.00	51.00
							203	86.00	66.00
							269	92.00	73.00
28	78.30	0.22	4.14	0.46	0.10	0.37	0	6.00	4.00
							2	6.00	4.00
							24	10.00	5.00
							85	30.00	16.00
							148	49.00	30.00
							233	69.00	49.00
							270	75.00	55.00
29	104.00	0.19	4.09	0.52	0.22	0.18	0	9.00	4.00
							2	10.00	4.00
							5	10.00	5.00
							70	37.00	24.00
							138	68.00	51.00
							205	90.00	65.00
							270	99.00	50.00
30	105.00	0.34	6.94	0.63	0.01	0.29	0	11.00	4.00
							20	97.00	79.00
							30	99.00	80.00
							60	100.00	73.00
							98	100.00	66.00
							140	100.00	54.00

31	101.00	0.22	5.33	0.48	0.11	0.35	0	8.00	4.00
							51	86.00	75.00
							102	96.00	81.00
							153	100.00	77.00
							204	100.00	69.00
							257	98.00	52.00
							269	99.00	47.00
32	78.00	0.28	4.25	0.41	0.13	0.39	0	7.00	4.00
							10	8.00	4.00
							20	10.00	4.00
							25	11.00	4.00
							35	13.00	5.00
							250	75.00	56.00
							269	78.00	59.00
33	80.00	0.34	6.58	0.41	0.13	0.40	0	10.00	4.00
							1.5	13.00	4.00
							45	94.00	81.00
							60	96.00	81.00
							160	99.00	77.00
34	72.00	0.24	4.15	0.43	0.17	0.36	0	6.00	4.00
							10	7.00	4.00
							75	16.00	8.00
							140	30.00	16.00
							205	42.00	25.00
							270	52.00	33.00
35	76.00	0.27	4.38	0.55	0.32	0.08	0	7.00	5.00
							5	7.00	5.00
							35	10.00	5.00
							55	12.00	7.00
							180	37.00	24.00
							235	50.00	34.00
36	84.00	0.23	4.16	0.43	0.12	0.40	0	7.00	4.00
							5	9.00	4.00
							15	16.00	9.00
							35	33.00	20.00
							60	52.00	36.00
							150	80.00	62.00
37	89.00	0.26	4.00	0.41	0.13	0.39	0	17.00	4.00
							1.5	17.00	4.00
							3	17.00	4.00
							10	22.00	6.00
							20	27.00	8.00
							60	55.00	31.00

							120	81.00	55.00
							180	95.00	65.00
38	81.00	0.19	4.00	0.60	0.26	0.07	0	9.00	4.00
							30	13.00	6.00
							85	29.00	16.00
							95	33.00	19.00
							105	36.00	21.00
							250	71.00	54.00
39	85.00	0.18	4.05	0.56	0.26	0.10	0	9.00	4.00
							30	12.00	4.00
							38	12.00	5.00
							68	15.00	7.00
							120	25.00	12.00
							230	53.00	35.00
40	70.00	0.16	4.12	0.43	0.14	0.36	0	6.00	4.00
							19	8.00	4.00
							30	9.00	4.00
							90	21.00	9.00
							180	41.00	22.00
							270	61.00	38.00
41	109.00	0.29	6.00	0.62	0.27	0.05	1	8.00	5.00
							25	67.00	54.00
							39	86.00	69.00
							80	94.00	72.00
							109	96.00	72.00
							197	99.00	44.00
							240	99.00	30.00
42	106.00	0.18	7.00	0.60	0.28	0.05	1	12.00	5.00
							15	69.00	54.00
							39	88.00	67.00
							80	96.00	68.00
							152	99.00	40.00
							190	99.00	30.00
43	74.00	0.15	5.20	0.63	0.23	0.09	23	8.00	5.00
							31	8.00	6.00
							44	10.00	7.00
							265	61.00	48.00
							270	61.00	48.00
44	81.00	0.19	4.30	0.62	0.24	0.08	3	7.00	5.00
							29	9.00	6.00
							41	11.00	7.00
							54	15.00	8.00
							120	26.00	16.00

							260	57.00	40.00
45	80.00	0.23	4.70	0.47	0.34	0.11	12	12.00	5.00
							23	14.00	6.00
							82	31.00	17.00
							88	33.00	19.00
							101	36.00	21.00
							147	47.00	31.00
							260	70.00	50.00
46	77.00	0.20	4.50	0.42	0.12	0.39	0	10.00	4.00
							4	11.00	4.00
							15	12.00	4.00
							27	13.00	5.00
							41	14.00	6.00
							62	17.00	7.00
							87	22.00	10.00
							96	24.00	11.00
							180	51.00	30.00
							260	70.00	47.00
47	71.00	0.20	4.50	0.42	0.12	0.39	0	12.00	4.00
							53	15.00	5.00
							91	20.00	6.00
							180	35.00	16.00
							260	51.00	26.00
48	89.00	0.20	4.10	0.55	0.17	0.22	20	8.00	6.00
							60	20.00	13.00
							120	42.00	30.00
							269	82.00	67.00
49	80.00	0.23	5.30	0.50	0.05	0.40	4	5.00	4.00
							16	6.00	5.00
							120	35.00	26.00
							242	74.00	62.00
							269	81.00	67.00
50	90.00	0.18	4.00	0.42	0.15	0.36	0	8.00	4.00
							2	8.90	4.58
							14	12.05	5.33
							18	12.05	5.33
							44	25.85	14.73
							49	25.85	14.73
							54	28.65	16.43
							180	81.95	64.68
							260	90.90	72.70
51	83.00	0.25	4.50	0.45	0.09	0.39	0	7.00	4.00
							37	13.00	6.48

							56	16.65	8.88
							70	18.45	10.00
							88	22.00	12.60
							154	42.70	27.18
							230	64.70	46.20
52	72.00	0.19	5.48	0.41	0.12	0.40	0	7.00	5.00
							2	7.70	4.80
							4	7.65	4.80
							150	76.00	58.75
							270	90.20	77.15
53	109.00	0.34	6.90	0.41	0.14	0.40	0	11.00	7.00
							1	16.65	11.43
							88	98.15	69.53
							98	98.75	68.55
							175	99.60	61.40
54	105.00	0.32	7.00	0.51	0.18	0.25	0	14.00	5.00
							7	50.30	35.23
							87	95.70	63.05
							100	98.35	64.68
							115	98.70	63.23
							269	100.00	33.30
55	77.00	0.17	4.80	0.47	0.25	0.23	0	7.00	5.00
							1	7.55	5.03
							72	21.85	13.00
							136	39.80	26.65
							179	51.45	36.65
							270	65.85	50.35
56	105.00	0.34	4.75	0.43	0.13	0.38	0	8.00	5.00
							14	35.45	26.23
							157	91.50	72.68
							178	93.50	73.65
							220	96.40	73.23
							230	96.80	75.25
57	72.00	0.17	7.00	0.44	0.17	0.34	0	8.00	5.00
							1	8.85	5.08
							18	49.05	36.70
							103	93.45	81.25
							120	94.85	81.48
							205	98.45	81.98
							269	98.85	78.33
58	109.00	0.18	5.00	0.47	0.11	0.36	0	6.00	4.00
							2	8.20	5.13
							87	78.40	65.13

							119	88.75	72.63
							200	96.55	69.95
							258	99.00	32.25
59	109.00	0.18	5.30	0.40	0.13	0.40	0	7.00	5.00
							2	9.25	5.10
							33	61.45	49.43
							223	98.40	47.80
							244	98.70	37.08
							260	98.60	30.95
60	89.00	0.15	6.80	0.53	0.11	0.30	0	10.00	4.00
							1	10.70	4.88
							54	95.35	79.63
							120	98.65	75.93
							230	99.10	60.23
							249	99.10	55.98
61	107.00	0.16	5.80	0.40	0.28	0.27	0	10.00	5.00
							1.5	13.00	7.00
							89	91.00	70.00
							120	95.00	71.00
							250	99.00	29.00
							270	99.00	24.00
62	88.00	0.15	6.30	0.41	0.13	0.40	0	10.00	5.00
							2	12.00	6.00
							8	21.00	12.00
							32	84.00	70.00
							120	98.00	80.00
							190	99.00	73.00
							230	99.00	68.00
63	85.00	0.15	5.80	0.41	0.12	0.40	0	9.00	4.00
							32	54.00	42.00
							45	67.00	54.00
							100	88.00	75.00
							190	96.00	79.00
							240	98.00	75.00
64	72.00	0.35	6.80	0.41	0.35	0.18	0	13.00	7.00
							1.5	13.00	7.00
							42	70.00	56.00
							149	97.00	76.00
							204	99.00	76.00
							213	99.00	74.00
							245	99.00	72.00
65	97.00	0.30	6.60	0.57	0.05	0.34	0	4.00	4.00
							26	72.60	63.13

							45	91.20	78.13
							190	99.20	56.30
							196	98.80	53.48
							260	99.00	44.30
66	107.00	0.35	5.70	0.44	0.11	0.39	0	7.00	5.00
							54	91.60	77.93
							91	95.90	78.35
							105	96.60	78.53
							140	97.60	75.40
							155	98.20	74.05
							200	98.90	67.80
67	76.00	0.15	6.40	0.42	0.11	0.40	0	8.00	4.00
							120	94.80	80.55
							180	98.80	77.95
							240	98.70	64.35
							250	98.80	61.60
							270	98.40	57.20
68	70.00	0.35	6.60	0.44	0.11	0.40	0	7.00	4.00
							100	90.20	75.05
							170	96.40	78.70
							218	97.60	77.93
							240	98.30	77.10
							269	98.70	75.63
69	90.00	0.34	4.00	0.74	0.03	0.17	0	5.00	4.00
							45	12.85	8.93
							80	17.45	11.35
							90	20.15	13.20
							110	22.95	14.90
							260	71.05	51.33
							270	74.70	54.30
70	95.00	0.24	6.00	0.53	0.04	0.37	0	6.00	4.00
							40	61.20	49.75
							90	93.60	80.83
							104	95.20	82.25
							143	96.60	80.83
							245	99.30	75.15
							267	99.30	72.65

Table 10. *The entire dataset for p-cymene synthesis via radical dehydrogenation (step 2).*

Exp # 1 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	100	0	0.00	0.00
Q_{liq} / mL min^{-1}	0.5	60	20.64	13.53
Q_{gas} / mL min^{-1}	40	150	71.94	20.06
α-terpinene / mL	1.463	180	71.94	29.19
γ-terpinene / mL	0.623	240	85.00	32.62
Terpinolene / mL	0.934			
Tert-butyl hydroperoxide / mL	0.726			
Exp # 2 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	113	0	0.00	0.00
Q_{liq} / mL min^{-1}	0.86	14	17.76	0.69
Q_{gas} / mL min^{-1}	15.6	54	34.95	4.85
α-terpinene / mL	1.326	100	45.83	11.69
γ-terpinene / mL	0.856	133	51.70	15.63
Terpinolene / mL	1.209	156	57.41	17.26
Tert-butyl hydroperoxide / mL	0.397	180	64.36	19.36
		231	73.80	20.87
Exp # 3 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	122	0	0.00	0.00
Q_{liq} / mL min^{-1}	4.44	25	29.80	7.86
Q_{gas} / mL min^{-1}	33.2	52	49.60	15.35
α-terpinene / mL	1.525	83	65.86	21.05
γ-terpinene / mL	0.249	106	72.45	27.54
Terpinolene / mL	0.328	140	83.36	27.61
Tert-butyl hydroperoxide / mL	01.483	187	89.86	28.72
		224	94.48	30.89
Exp # 4 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	145	0	0.00	0.00
Q_{liq} / mL min^{-1}	0.215	60	1.68	0.46
Q_{gas} / mL min^{-1}	12.2	180	11.53	3.39
α-terpinene / mL	1.117	240	14.77	5.89

γ-terpinene / mL	1.08			
Terpinolene / mL	1.526			
Tert-butyl hydroperoxide / mL	1.115			
Exp # 5 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	86	0	0.00	0.00
Q_{liq} / mL min^{-1}	1.453	7	4.36	0.13
Q_{gas} / mL min^{-1}	110	50	27.89	1.61
α-terpinene / mL	0.872	84	22.13	6.94
γ-terpinene / mL	1.631	127	49.54	8.97
Terpinolene / mL	1.775	159	55.05	20.56
Tert-butyl hydroperoxide / mL	1.345	201	71.69	24.67
Exp # 6 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	84	0	0.00	0.00
Q_{liq} / mL min^{-1}	0.744	1.5	4.64	1.49
Q_{gas} / mL min^{-1}	25.9	66	36.85	2.83
α-terpinene / mL	0.404	84	41.15	4.44
γ-terpinene / mL	1.361	130	37.91	10.77
Terpinolene / mL	1.884	154	43.54	13.09
Tert-butyl hydroperoxide / mL	1.774	182	49.41	15.25
		230	53.78	15.52
Exp # 7 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	126	0	0.00	0.00
Q_{liq} / mL min^{-1}	3.739	6	28.47	6.09
Q_{gas} / mL min^{-1}	85.5	44	66.04	20.30
α-terpinene / mL	1.972	76	79.94	26.12
γ-terpinene / mL	0.413	109	90.03	28.69
Terpinolene / mL	0.121	150	93.82	31.82
Tert-butyl hydroperoxide / mL	0.703	176	94.91	33.76
		211	95.94	39.08
Exp # 8 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	111	0	0.00	0.00
Q_{liq} / mL min^{-1}	4	6	11.76	2.26

Q_{gas} / mL min^{-1}	53.5	60	48.12	17.02
α-terpinene / mL	0.596	80	57.21	21.17
γ-terpinene / mL	0.939	103	66.36	25.81
Terpinolene / mL	0.511	157	81.45	24.68
Tert-butyl hydroperoxide / mL	0.046	186	80.82	34.53
		218	85.82	35.23
Exp # 9 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	141	0	0.00	0.00
Q_{liq} / mL min^{-1}	2.923	7	18.34	5.21
Q_{gas} / mL min^{-1}	97.5	47	82.05	11.88
α-terpinene / mL	0.172	90	95.10	15.49
γ-terpinene / mL	0.116	131	100.00	14.42
Terpinolene / mL	0.837	159	100.00	16.82
Tert-butyl hydroperoxide / mL	1.015	182	100.00	15.09
		219	100.00	16.96
Exp # 10 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	130	0	0.00	0.00
Q_{liq} / mL min^{-1}	2.6	16	29.26	7.73
Q_{gas} / mL min^{-1}	79.2	45	74.61	18.01
α-terpinene / mL	1.686	73	88.88	21.24
γ-terpinene / mL	0.721	114	97.91	22.65
Terpinolene / mL	0.946	145	100.00	18.33
Tert-butyl hydroperoxide / mL	1.664	192	100.00	24.61
		234	100.00	25.38
Exp # 11 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	91	0	0.00	0.00
Q_{liq} / mL min^{-1}	1.971	14	14.61	1.83
Q_{gas} / mL min^{-1}	112	56	41.55	7.91
α-terpinene / mL	1.717	81	55.99	11.17
γ-terpinene / mL	0.548	123	66.12	15.75
Terpinolene / mL	1.38	165	83.64	17.50
Tert-butyl hydroperoxide / mL	0.472	193	88.89	19.30
		225	94.57	19.76

Exp # 12 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	105	0	0.00	0.00
Q_{liq} / mL min^{-1}	2.082	11	12.65	1.31
Q_{gas} / mL min^{-1}	60.2	42	22.06	7.85
α-terpinene / mL	0.048	80	38.48	17.38
γ-terpinene / mL	1.783	107	50.01	22.55
Terpinolene / mL	0.683	152	60.06	30.46
Tert-butyl hydroperoxide / mL	1.868	190	68.48	32.76
		224	68.55	41.60

Exp # 13 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	98	0	0.00	0.00
Q_{liq} / mL min^{-1}	3.294	4	43.58	0.57
Q_{gas} / mL min^{-1}	41.6	47	34.31	10.47
α-terpinene / mL	0.656	83	38.01	11.61
γ-terpinene / mL	1.891	106	32.13	13.52
Terpinolene / mL	1.554	165	41.65	12.37
Tert-butyl hydroperoxide / mL	0.218			

Exp # 14 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	145	0	0.00	0.00
Q_{liq} / mL min^{-1}	4.572	110	96.60	36.76
Q_{gas} / mL min^{-1}	81	132	98.16	37.18
α-terpinene / mL	1.309	177	100.00	38.86
γ-terpinene / mL	1.251	200	100.00	39.22
Terpinolene / mL	0.124			
Tert-butyl hydroperoxide / mL	1.518			

Exp # 15 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	133	0	0.00	0.00
Q_{liq} / mL min^{-1}	4.288	145	86.13	40.13
Q_{gas} / mL min^{-1}	80	150	86.46	40.27
α-terpinene / mL	0.498	193	88.77	41.11
γ-terpinene / mL	1.579	204	90.36	40.34
Terpinolene / mL	0.107	240	91.39	43.35

Tert-butyl hydroperoxide / mL	0.29			
Exp # 16 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	130	0	0.00	0.00
Q_{liq} / mL min^{-1}	4.482	110	93.41	31.38
Q_{gas} / mL min^{-1}	43.1	179	97.85	33.83
α-terpinene / mL	1.545	188	98.03	34.69
γ-terpinene / mL	1.991	240	99.01	35.55
Terpinolene / mL	0.508			
Tert-butyl hydroperoxide / mL	1.377			
Exp # 17 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	127	0	0.00	0.00
Q_{liq} / mL min^{-1}	3.988	152	97.78	40.49
Q_{gas} / mL min^{-1}	90.6	182	98.40	40.33
α-terpinene / mL	1.431	228	100.00	41.97
γ-terpinene / mL	1.574	240	100.00	40.33
Terpinolene / mL	0.435			
Tert-butyl hydroperoxide / mL	1.641			
Exp # 18 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	131	0	0.00	0.00
Q_{liq} / mL min^{-1}	3.554	140	97.75	37.29
Q_{gas} / mL min^{-1}	106	188	98.60	38.22
α-terpinene / mL	1.299	240	100.00	39.26
γ-terpinene / mL	1.782			
Terpinolene / mL	0.489			
Tert-butyl hydroperoxide / mL	0.726			
Exp # 19 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	136	0	0.00	0.00
Q_{liq} / mL min^{-1}	4.698	0.5	12.99	1.56
Q_{gas} / mL min^{-1}	75.6	46	75.00	35.04
α-terpinene / mL	0.95	86	86.65	40.61

γ-terpinene / mL	1.488	113	93.98	44.85
Terpinolene / mL	0.2	140	95.33	51.66
Tert-butyl hydroperoxide / mL	0.884	210	97.97	51.10
		231	98.05	51.10

Exp # 20 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	139	0	0.00	0.00
Q_{liq} / mL min^{-1}	4.236	145	100.00	28.90
Q_{gas} / mL min^{-1}	76.7	230	100.00	29.89
α-terpinene / mL	1.562	233	100.00	29.79
γ-terpinene / mL	1.093			
Terpinolene / mL	0.735			
Tert-butyl hydroperoxide / mL	1.486			

Exp # 21 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	137	0	0.00	0.00
Q_{liq} / mL min^{-1}	4.033	143	100.00	30.42
Q_{gas} / mL min^{-1}	79.5	190	100.00	32.60
α-terpinene / mL	1.639	198	100.00	32.76
γ-terpinene / mL	1.278	240	100.00	32.10
Terpinolene / mL	0.532			
Tert-butyl hydroperoxide / mL	0.879			

Exp # 22 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	124.5	0	0.00	0.00
Q_{liq} / mL min^{-1}	3.149	99	90.53	30.64
Q_{gas} / mL min^{-1}	80.2	190	95.43	34.78
α-terpinene / mL	0.715	240	96.72	36.93
γ-terpinene / mL	1.027			
Terpinolene / mL	0.397			
Tert-butyl hydroperoxide / mL	0.646			

Exp # 23 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	115	0	0.00	0.00
Q_{liq} / mL min^{-1}	4.453	145	93.05	31.60

Q_{gas} / mL min^{-1}	112	155	92.44	37.91
α-terpinene / mL	1.302	198	96.56	37.43
γ-terpinene / mL	1.282	240	97.73	39.46
Terpinolene / mL	0.4			
Tert-butyl hydroperoxide / mL	1.455			

Exp # 24 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	136.6	0	0.00	0.00
Q_{liq} / mL min^{-1}	4.982	137	100.00	32.65
Q_{gas} / mL min^{-1}	81.9	170	100.00	33.84
α-terpinene / mL	1.679	209	100.00	36.52
γ-terpinene / mL	1.525	240	100.00	36.96
Terpinolene / mL	0.328			
Tert-butyl hydroperoxide / mL	0.831			

Exp # 25 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	104	0	0.00	0.00
Q_{liq} / mL min^{-1}	2.516	168	89.60	33.16
Q_{gas} / mL min^{-1}	104.9	170	89.53	33.45
α-terpinene / mL	1.483	180	90.31	32.64
γ-terpinene / mL	1.41	240	93.86	41.28
Terpinolene / mL	0.296			
Tert-butyl hydroperoxide / mL	1.411			

Exp # 26 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	137.7	0	0.00	0.00
Q_{liq} / mL min^{-1}	2.627	206	100.00	60.93
Q_{gas} / mL min^{-1}	98.5	228	100.00	62.07
α-terpinene / mL	0.529	234	100.00	62.53
γ-terpinene / mL	1.278			
Terpinolene / mL	0.061			
Tert-butyl hydroperoxide / mL	1.984			

Exp # 27 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	137.5	0	0.00	0.00
Q_{liq} / mL min^{-1}	4.478	170	95.77	47.90
Q_{gas} / mL min^{-1}	45.6	180	96.47	48.00
α-terpinene / mL	0.484	235	97.66	51.18
γ-terpinene / mL	1.496			
Terpinolene / mL	0.178			
Tert-butyl hydroperoxide / mL	1.071			

Exp # 28 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	135.9	0	0.00	0.00
Q_{liq} / mL min^{-1}	2.853	180	95.70	51.44
Q_{gas} / mL min^{-1}	105	220	98.91	53.43
α-terpinene / mL	0.864	230	99.19	54.46
γ-terpinene / mL	1.844			
Terpinolene / mL	0.266			
Tert-butyl hydroperoxide / mL	1.624			

Exp # 29 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	142.7	0	0.00	0.00
Q_{liq} / mL min^{-1}	3.321	180	99.03	37.93
Q_{gas} / mL min^{-1}	72.9	215	100.00	39.18
α-terpinene / mL	1.041	235	100.00	38.66
γ-terpinene / mL	0.849			
Terpinolene / mL	0.255			
Tert-butyl hydroperoxide / mL	1.527			

Exp # 30 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	132	0	0.00	0.00
Q_{liq} / mL min^{-1}	3.475	200	96.27	56.94
Q_{gas} / mL min^{-1}	119	240	96.57	58.34
α-terpinene / mL	0.657			
γ-terpinene / mL	1.577			
Terpinolene / mL	0.198			

Tert-butyl hydroperoxide / mL	1.312			
Exp # 31 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	138	0	0.00	0.00
Q_{liq} / mL min^{-1}	2.738	185	100.00	28.41
Q_{gas} / mL min^{-1}	39	191	100.00	29.76
α-terpinene / mL	1.614	214	100.00	32.04
γ-terpinene / mL	0.7			
Terpinolene / mL	0.283			
Tert-butyl hydroperoxide / mL	0.788			
Exp # 32 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	132.5	0	0.00	0.00
Q_{liq} / mL min^{-1}	0.798	162	89.88	40.93
Q_{gas} / mL min^{-1}	109	170	89.90	42.23
α-terpinene / mL	0.356	203	91.59	44.82
γ-terpinene / mL	1.453	215	92.30	47.11
Terpinolene / mL	0.152			
Tert-butyl hydroperoxide / mL	1.22			
Exp # 33 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	137.9	0	0.00	0.00
Q_{liq} / mL min^{-1}	4.17	148	23.62	5.25
Q_{gas} / mL min^{-1}	68.9	180	66.39	30.89
α-terpinene / mL	0.738			
γ-terpinene / mL	1.463			
Terpinolene / mL	0.273			
Tert-butyl hydroperoxide / mL	1.15			
Exp # 34 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	149	0	0.00	0.00
Q_{liq} / mL min^{-1}	4.611	193	100.00	25.02
Q_{gas} / mL min^{-1}	115	240	100.00	24.12
α-terpinene / mL	0.779			

γ-terpinene / mL	0.905
Terpinolene / mL	1.391
Tert-butyl hydroperoxide / mL	1.086

Exp # 35 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	118	0	0.00	0.00
Q_{liq} / mL min^{-1}	0.969	150	85.55	37.19
Q_{gas} / mL min^{-1}	54.6	155	86.61	40.80
α-terpinene / mL	0.968	240	91.05	43.50
γ-terpinene / mL	1.585			
Terpinolene / mL	0.17			
Tert-butyl hydroperoxide / mL	0.092			

Exp # 36 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	132.2	0	0.00	0.00
Q_{liq} / mL min^{-1}	2.749	150	63.89	21.50
Q_{gas} / mL min^{-1}	88.1	217	77.57	41.29
α-terpinene / mL	0.843			
γ-terpinene / mL	1.922			
Terpinolene / mL	0.549			
Tert-butyl hydroperoxide / mL	1.398			

Exp # 37 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	98	0	0.00	0.00
Q_{liq} / mL min^{-1}	2.688	177	86.49	24.26
Q_{gas} / mL min^{-1}	63.8	198	84.33	37.33
α-terpinene / mL	1.667	240	90.30	26.38
γ-terpinene / mL	0.352			
Terpinolene / mL	0.307			
Tert-butyl hydroperoxide / mL	1.594			

Exp # 38 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	114	0	0.00	0.00
Q_{liq} / mL min^{-1}	1.834	154	67.33	34.47

Q_{gas} / mL min^{-1}	90.6	169	76.33	38.17
α-terpinene / mL	0.683	188	79.29	33.79
γ-terpinene / mL	1.761	222	83.08	36.03
Terpinolene / mL	0.193			
Tert-butyl hydroperoxide / mL	0.595			

Exp # 39 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	129.5	0	0.00	0.00
Q_{liq} / mL min^{-1}	3.389	187	91.85	45.77
Q_{gas} / mL min^{-1}	112	207	93.45	45.77
α-terpinene / mL	0.388	240	95.72	43.04
γ-terpinene / mL	0.804			
Terpinolene / mL	0.23			
Tert-butyl hydroperoxide / mL	1.17			

Exp # 40 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	113.3	0	0.00	0.00
Q_{liq} / mL min^{-1}	3.908	169	88.69	31.38
Q_{gas} / mL min^{-1}	52.3	202	91.19	33.65
α-terpinene / mL	1.04	240	92.33	34.12
γ-terpinene / mL	1.768			
Terpinolene / mL	0.198			
Tert-butyl hydroperoxide / mL	0.406			

Exp # 41 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	127.7	0	0.00	0.00
Q_{liq} / mL min^{-1}	0.543	220	86.71	29.69
Q_{gas} / mL min^{-1}	86.3	235	87.29	29.91
α-terpinene / mL	0.736			
γ-terpinene / mL	0.202			
Terpinolene / mL	0.122			
Tert-butyl hydroperoxide / mL	1.108			

Exp # 42 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	118	0	0.00	0.00
Q_{liq} / mL min^{-1}	3.976	189	98.29	26.62
Q_{gas} / mL min^{-1}	60.2	240	100.00	24.78
α-terpinene / mL	1.822			
γ-terpinene / mL	0.789			
Terpinolene / mL	0.447			
Tert-butyl hydroperoxide / mL	0.675			

Exp # 43 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	127	0	0.00	0.00
Q_{liq} / mL min^{-1}	4.366	120	96.20	33.96
Q_{gas} / mL min^{-1}	98	218	98.38	33.96
α-terpinene / mL	0.364	238	100.00	42.49
γ-terpinene / mL	1.725			
Terpinolene / mL	0.965			
Tert-butyl hydroperoxide / mL	0.621			

Exp # 44 initial condition		Time / min	Conversion (%)	Yield (%)
Temp / °C	149	0	0.00	0.00
Q_{liq} / mL min^{-1}	3.18	98	100.00	27.08
Q_{gas} / mL min^{-1}	107.4	116	100.00	27.63
α-terpinene / mL	1.813	125	100.00	26.54
γ-terpinene / mL	1.261	217	100.00	29.59
Terpinolene / mL	0.812			
Tert-butyl hydroperoxide / mL	0.774			

Chapter 3

Molecular informatics-driven solvent recommendation

Introduction

Solvents often exist in an excess in a solution or are present in the largest amount in a reaction.[1] This is indeed reflected in solvent consumption in industry where 80-90% of the material usage[2] and ~70% of the overall waste in pharmaceutical industry is attributed to solvent use.[3, 4] Given the significant material and energy consumption, ACS Green Chemistry Institute (GCI) Pharmaceutical Roundtable (PR) identified solvent selection as one of the top five priorities for green engineering research.[5] Moreover, new regulations such as the EU REACH (Registration, Evaluation, Authorisation, and Restriction of Chemicals) highlighted issues with commonly used solvents (e.g., *N*-methyl-2-pyrrolidone, 1,2-dichloroethane), pushing for finding greener alternatives.[6] As a result, several guidelines by pharmaceutical companies such as GSK,[7-9] Pfizer,[10] AZ,[4] Sanofi,[11] and Syngenta[12] have been developed for assessing solvents based on safety, health, and environmental (SHE) impact (Table 11).

Motivation to select an ideal solvent originates from the 15-18th centuries where alchemists searched for a "universal solvent" that could dissolve every other substance.[13] Whilst the primary role of a solvent could be to dissolve other reaction components (e.g., reactants, catalyst, etc.) and serve as a heat and mass transfer medium,[14, 15] choice of solvent could significantly enhance reaction rates and selectivities,[16] or even alter the reaction mechanism.[17] For instance, solvolysis of 2-chloro-2-methylpropane in water was found to be 335,000 times faster than in ethanol, highlighting the influence of solvent selection on reaction rate.[18, 19] Similarly, Campbell *et al.* reported 2,800 times increase in reaction rate in nitrobenzene compared to carbon tetrachloride for the reaction of 2,4-dinitrobenzenesulphenyl bromide with cyclohexene.[20] Cox *et al.* reported a 6-order of magnitude increase in rate constant of a nucleophilic aromatic substitution reaction of 4-fluoronitrobenzene with azide anion in hexamethylphosphorotriamide (HMPT) compared to water.[21]

Despite the significant role of solvents in optimising for process parameters and minimising for SHE impacts given the heavy use in industry, there is no universally accepted framework

for solvent selection. For a given reaction, chemists often select suitable solvents based on domain knowledge, experience, and heuristics.[22] A commonly employed approach is to screen list of solvents, often under fixed reaction conditions, to select a solvent that meets reaction objectives, such as high yield. A drawback of this approach is that this mostly limits the solvent choice to what's available in the lab, which could miss out on "non-obvious" solvents that could potentially achieve better process and environmental metrics, and does not account for the relationship of continuous variables (i.e., other reaction variables) and the solvent. Whilst a "good" solvent could be found this way, selecting the most optimal solvent is often not considered.[2]

Whilst SHE-based solvent selection guides are comprehensive, the challenge behind developing universal framework for solvent selection is threefold: (*i*) the difficulty to find reaction specific descriptors of solvents,[23-25] (*ii*) which is caused by the difficulty of capturing complexity of reactions, and accounting for small variations that could significantly affect the outcomes,[26] and (*iii*) lack of standardised similarity metrics (e.g., Euclidean, Tanimoto) and data pre-treatment techniques (e.g., interval scaling, principal component analysis), which affect what would be considered a "similar" solvent (to find an alternative) for a given solvent.[27-30] As opposed to designing the solvent space from scratch when developing new processes, there exists a need to efficiently generate initial list of solvents to study for a given reaction based on different parameterisation techniques, similarity metrics, and SHE impact. A look up table similar to NMR chemical shifts of trace impurities and commonly used laboratory solvents,[31] but for finding (greener) alternatives to given set of solvents would significantly accelerate development of new processes whilst accounting for both the process parameters and SHE impact.

In Chapter 2, we demonstrated multi-objective reaction optimisation based on continuous variables. In this chapter, we address selection of solvents with no or minimal available information for a given reaction. We explore various data treatment methods, similarity metrics, and descriptor sets for suggestion of alternative solvents for a target solvent. The results are compared against the solvent selection guides by AstraZeneca[4] and Syngenta,[12] and applied to identify solvent influence on the experimental solvent selection datasets of various transformations. Relation of this chapter to Chapter 4 and 5 are that they address holistic reaction development, one based on purely computational methods, and the other with experimental optimisation, respectively.

Contributions

Contributions to the chapter is as follows. Mr. Kobi Felton calculated the five sigma moments (i.e., molecular descriptors calculated using COSMO*therm*) used for the commonly used 120 solvents. Datasets collection, implementation of similarity metrics, generation of solvent recommendations per solvent candidate in the datasets, case study selections and applicability analysis of the workflow were done by me.

Solvent selection guides – sustainability

One of the earlier surveys comparing solvent guides based on SHE impact was reported by Sanofi and GSK in 2014,[32] with an updated version together with Pfizer in 2015.[33] Using solvent guides from five pharmaceutical companies, the authors evaluated 51 solvents to categorise into one of the four groups: recommended, problematic, hazardous, and highly hazardous (Table 11 and Table 12). When scoring, sum of the scores for each of safety, health, and environmental impact was used, taking the worst score from each category (e.g., if a solvent had 5, 8, and 2 for air, water, and waste based on GCI-PR guide, then 8 was used as the overall score for environmental impact). Out of 51 solvents, 34 could be categorised in one of four groups as there is a high overlap between different guides, whilst 17 solvents cannot be ranked due to the differences in weighing and approaches of scoring solvent SHE scores. For instance, AstraZeneca considers seven environmental factors whilst GSK and ACS GCI consider three environmental factors, which the latter two also differ from each other (Table 11 and Table 12). In terms of grouping, Sanofi categorised solvents into four groups whilst Pfizer grouped solvents into three categories. Considering only commonly used (i.e., classical) 51 solvents limits the applicability of such report, and as highlighted by the authors, the report was not developed as a universal solvent selection guide, but rather a comparison of different solvent selection guides.

***Table 11**. An overview of solvent selection guides.*

Source	Dataset size	Criteria
Pfizer[10]	39	Preferred, usable, and undesirable
AstraZeneca[4]	272	Green, yellow, and red based on 2 safety (flammability and resistivity, 1 health, 7 environmental (impact on air, VOC, impact on water, potential bio-treatment plant load, recycling, incineration, and life cycle) values with assigned scores from 1-10
GSK[7-9]	154	2 safety (flammability/explosion and reactivity/stability), 1 health, 3 environmental (waste, environmental impact, life cycle).
ACS GCI[34]	63	Green, yellow, and red based on 1 safety, 1 health, and 3 environmental (air, water, and waste) values with assigned scores 1-10
Sanofi[11]	96	Recommended, substitution advisable, substitution requested, and banned
Syngenta[12]	209	Internal tool used at Syngenta for solvent selection based on solvent properties and descriptors, and SHE impact based on GSK's guide

Table 12. *Comparison of different solvent selection guides for commonly used solvents, reproduced from ref.*[32] [a]*Subst. adv.: substitution advisable; Subst. req.: substitution requested.* [b]*TBC: to be confirmed; HH: highly hazardous.*

Family	Solvent	AZ	GCI-PR	GSK	Pfizer	Sanofi[a]	Issues	Overall[b]
Water	Water	—	—	24	Preferred	Recommended	—	Recommended
Alcohols	MeOH	19	14	14	Preferred	Recommended	—	TBC
	EtOH	16	13	17	Preferred	Recommended	—	Recommended
	i-PrOH	16	16	17	Preferred	Recommended	—	Recommended
	n-BuOH	17	13	18	Preferred	Recommended	—	Recommended
	t-BuOH	20	15	15	Preferred	Subst. adv.	—	TBC
	Benzyl alcohol	—	11	20	—	Subst. adv.	—	TBC
	Ethylene glycol	—	13	21	Usable	Subst. adv.	—	TBC
Ketones	Acetone	21	15	15	Preferred	Recommended	—	TBC
	MEK	21	16	15	Preferred	Recommended	—	TBC
	MIBK	22	17	15	—	Recommended	—	TBC
	Cyclohexanone	—	14	20	—	Subst. adv.	—	TBC
Esters	Methyl acetate	—	14	14	—	Subst. adv.	—	TBC
	Ethyl acetate	18	15	16	Preferred	Recommended	—	Recommended
	i-PrOAc	18	13	18	Preferred	Recommended	—	Recommended
	n-BuOAc	13	14	21	—	Recommended	—	Recommended
Ethers	Diethyl ether	27	21	3	Undesirable	Banned	H224	HH
	Diisopropyl ether	—	—	4	Undesirable	Subst. adv.	Perox.	Hazardous
	MTBE	24	21	4	Usable	Subst. adv.	—	TBC
	THF	23	16	4	Usable	Subst. adv.	H351	TBC
	Me-THF	24	15	11	Usable	Recommended	—	Problematic
	1,4-Dioxane	28	21	11	Undesirable	Subst. req.	—	Hazardous
	Anisole	18	13	18	—	Recommended	—	Recommended
	DME	21	23	2	Undesirable	Subst. req.	H360	Hazardous
Hydrocarbons	Pentane	—	—	7	Undesirable	Banned	H224	Hazardous
	Hexane	26	21	1	Undesirable	Subst. req.	—	Hazardous
	Heptane	21	17	14	Usable	Subst. adv.	—	Problematic
	Cyclohexane	25	18	14	Usable	Subst. adv.	—	TBC
	Me-cyclohexane	—	17	16	Usable	Subst. adv.	—	Problematic
	Benzene	—	21	1	Undesirable	Banned	H350	HH
	Toluene	22	18	11	Usable	Subst. adv.	H351	Problematic
	Xylenes	19	15	13	Usable	Subst. adv.	—	Problematic
Halogenated	DCM	20	18	5	Undesirable	Subst. adv.	H351	TBC
	Chloroform	—	18	4	Undesirable	Banned	—	HH
	CCl_4	—	19	3	Undesirable	Banned	H420	HH
	DCE	—	19	4	Undesirable	Banned	H350	HH
	Chlorobenzene	25	16	18	—	Subst. adv.	—	Problematic
Aprotic polar	Acetonitrile	24	14	14	Usable	Recommended	—	Problematic
	DMF	20	17	7	Undesirable	Subst. req.	H360	Hazardous
	DMAc	20	16	4	Undesirable	Subst. req.	H360	Hazardous
	NMP	18	16	7	Undesirable	Subst. req.	H360	Hazardous
	DMPU	—	—	14	—	Subst. adv.	—	Problematic
	DMSO	8	15	14	Usable	Subst. adv.	—	Problematic
	Sulfolane	9	13	21	—	Subst. adv.	—	Recommended
	Nitromethane	—	—	1	—	Banned	Explo.	HH
Miscellaneous	Methoxy-ethanol	21	20	3	—	Subst. req.	H360	Hazardous
Acids	Formic acid	20	15	—	—	Subst. req.	—	TBC
	Acetic acid	17	15	17	Usable	Subst. adv.	—	TBC
	Ac_2O	—	16	15	—	Subst. adv.	—	TBC
Amines	Pyridine	26	16	5	Undesirable	Subst. adv.	—	TBC
	TEA	23	18	3	—	Subst. req.	—	Hazardous

Whilst considering SHE impact is important in developing new processes, selecting the "greenest" solvent could potentially compromise process performance. Therefore, selection of an alternative solvent should behave similarly to the initial choice of solvent. One of the initial attempts to map solvents based on their physicochemical properties was demonstrated by Martin *et al*., where authors used principal component analysis (PCA),[35] a statistical technique to map high dimensional data onto lower dimensions without significant information loss, to classify 18 solvents using 18 physicochemical properties.[36] The result of this approach was similar to chemical intuition-based classification of solvents reported by Parker *et al*.[37] Expanding on their 46 solvents excel spreadsheet, AstraZeneca developed an interactive tool

for selection and screening of 272 solvents.[4] The compiled dataset contains both experimental properties (e.g., boiling point, density) and computational descriptors (e.g., logP, H-Bond acceptor and donor strengths, area) calculated using Gaussian and COSMO*therm*, for solvents, expanding the initial 46x11 spreadsheet to 272x100+ database. The interactive tool was built based on applying PCA on the original 31 experimental properties for solvents to reduce the solvent space to 6 principal components (PCs) whilst retaining 87.90% of the original information. Practical applicability of the tool was demonstrated on identifying certain patterns in three principal component space for a handful of reactions, as elaborated and benchmarked against our approach in Results and Discussions section. Although useful, the tool only allows using one of the three principal components, which accounts for 70.30% of the original physical properties, and such approach requires extensive collection and calculation of solvent properties and descriptors.

Considering both SHE impact and physicochemical properties, an interactive tool developed by Syngenta for internal use allows for more flexibility in solvent design.[12] For instance, a user can define list of properties (e.g., density, acidity), descriptors (e.g., Kamlet-Taft parameters, Abraham's parameters), and SHE constraints (e.g., GSK guide: Waste) to generate a list of recommended solvents. Based on a dataset of 209 solvents, the tool allows for generation of initial group of solvents based on dissimilarity of solvents for screening, selection of an intermediate solvent (e.g., identify solvents intermediate to 1,4-dioxane and diethyl ether on the PCA map), and suggestion of alternative solvents. The results from alternative solvent recommendation are benchmarked against our results in the Results and Discussions section. Similar to the solvent selection tool by AstraZeneca, this approach also requires extensive collection of solvent properties and descriptors, was not validated for applicability in various reactions, and is limited to internal use only. Finally, neither of these tools evaluate different similarity metrics (e.g., Euclidean, Tanimoto, Cosine) when selecting alternative solvents, which could affect the choice of recommended solvent depending on the implemented similarity metric.

Solvent selection methods – reaction optimisation

When selecting solvents for reaction optimisation, existing methods could be grouped into four categories: intuition-based selection using a reported data in literature for a similar transformation, screening a handful of manually selected solvents,[38-40] Computer-Aided Molecular Design (CAMD),[14, 22, 41, 42] and machine learning-driven optimisation.[43, 44] The first two approaches depend heavily on domain expertise, limits the solvent choices to commonly used solvents, and does not explore "non-obvious" solvents, hence potentially missing out on identifying the most optimal solvent. CAMD, on the other hand, is based on developing quantitative relationships, known as solvatochromatic equations,[22] between physiochemical descriptors of solvents (e.g., Kamlet-Taft parameters,[45] Abraham descriptors[46]) and reaction targets such as logarithms of rate and equilibrium constants. In order to maximise the reaction rate of Menschutkin reaction, Adjiman and colleagues evaluated 1,341 solvents *in silico* based on a quantum-mechanical calculation of rate constants in nine solvents.[41] Model suggested solvents were evaluated *in situ* and the identified solvent improved the rate constant by 40%. However, most of the studies have been focused on reaction kinetics, and do not account for other discrete and continuous variables in the reaction. Moreover, these approaches do not demonstrate a holistic reaction optimisation and are not optimised for multiple objectives. ML-driven reaction optimisations, as demonstrated by our group,[43, 44] tackle some of these challenges using sequential Bayesian optimisation algorithms and by mapping the discrete variables such as solvents onto continuous variable space using descriptors. As explained in depth in Chapter 5, which demonstrates ML-driven holistic reaction optimisation, one of the outstanding challenges in the field is identifying reaction relevant descriptors for solvents. For instance, Amar *et al*. compared six models based on different descriptors for solvents in asymmetric hydrogenation of α-β unsaturated γ-lactam to produce Brivaracetam,[47] for objectives conversion and diastereoselectivity (d.e.). Compared to information rich model built on 17 descriptors, the authors found that the model built on five screening charge densities achieved higher accuracy in predicting low *vs* high d.e., whilst model performances were comparable for predicting conversion.[43] Similarly, we also demonstrate in Chapter 5 that describing solvents in five dimensions in the form of sigma moments allowed for efficient optimisation of photoredox amine synthesis.

Molecular descriptors - Theory

Solvent parameterisation is a critical step in describing solvation phenomena in the form of quantitative structure-property relationships (QSPR),[48, 49] linear free energy relationships (LFER),[46, 50-52] computer-aided molecular design (CAMD),[22, 41, 53] and reaction optimisation.[43, 44] In the form of solvatochromatic equations (equation 1), it has been demonstrated feasible to quantify, correlate, and predict solvent influence on various physicochemical properties and reactivity parameters. Earlier work on parameterising solvents was demonstrated by Kamlet and colleagues, where the authors reported a general form of solvatochromatic equation (equation 1) and parameters π*, α, and β, also known as Kamlet-Taft parameters, where π* describes solvent dipolarity/polarizability, α provides a measure of HBD (hydrogen bond donor) acidities, and β provides a measure of HBA (hydrogen bond acceptor) basicities (i.e., solvent's ability to accept a proton).[50] The general form of solvatochromatic equation developed by Kamlet and Taft is as following:

$$XYZ = XYZ_0 + s(\pi^* + d\delta) + a\alpha + b\beta + h\delta_H + e\xi \quad (1)$$

Where δ is a polarisability correction term (e.g., 0.0 for nonchlorinated aliphatic solvents), δ_H is the Hildebrand solubility parameters, and ξ is a coordinate covalency. The model coefficients $s, d, a, b, h,$ and e measure the relative dependency of XYZ on the indicated solvent properties. Later, Abraham and colleagues reported a set of five descriptors, also known as Abraham descriptors, to describe the solvation phenomena (equation 2).[46, 51, 52] Abraham descriptors correspond to E (excess molar refraction), S (dipolarity/polarisability), A (hydrogen bond acidity), B (hydrogen bond basicity), and V (McGowan characteristic volume). The general form of the equation using Abraham descriptors is as following:

$$LogSP = c + eE + sS + aA + bB + vV \quad (2)$$

Where the model coefficients are determined using multiple linear regression. Later, Abraham, Klamt, and co-workers reported a comparison of experimental Abraham descriptors with the five computational COSMOments of Klamt's COSMO-RS for 470 compounds.[54] Using five COSMOments, which will be referred to as sigma moments in the rest of the chapter, namely sig2 (electrostatic polarity), sig3 (asymmetry of sigma profile), sig0 (CSA - molecular surface area), Hbacc3 (hydrogen bond acceptor strength), and Hbdon3 (hydrogen bond donor strength),

authors demonstrate a high information overlap between experimental Abraham descriptors and computational Klampt's sigma moments. For instance, Abraham parameter A (hydrogen bond acidity) and V (McGowan molecular volume) could be calculated using sigma moments in the following forms for the 470 compounds studied.

$$A = 0.030 - 0.006Sig3 + 0.085Hbdon3 + 0.074Hbacc3 \quad (3)$$

$$V = -0.262 - 0.001Sig2 + 0.001Sig3 + 0.01Hbdon3 + 0.008CSA \quad (4)$$

Comparing the three most commonly used descriptors for describing the solvation phenomena, Klamt's five sigma moments were selected in this study due to high information overlap with other two descriptor sets, time and cost efficiency of computationally calculated descriptors, and the ability to parameterise solvents in small number of descriptors as opposed to sampling from a large descriptor space. Moreover, given the success of using information rich descriptors libraries as explained in the previous section, descriptors used by Amar *et al*. were also used as a comparison.[43] Finally, due to the popularity of using fingerprints in similarity search in medicinal chemistry, and its ease of calculation, solvent fingerprints were also included in this study. Bender *et al*. reported that there is no significant difference in the performances of commonly used fingerprints.[28] Therefore, Morgan fingerprints with radius=2 were calculated. Table 13 gives the summary of types and sources of descriptors used for solvent parameterisation.

COSMOtherm - descriptor calculations

COSMOtherm, which is a program on COSMO-RS, was used to calculate the sigma moments used in this study.[55] COSMO-RS is a COSMO, the Conductor-like Screening Model, theory for "real solvents".[56, 57] COSMO is a class of continuum solvation models (CSMs) based on the theory of interacting molecular surface charges calculated using quantum chemical methods (QM).[58] The theory is based on the claim that if a solvent has opposite surface charge densities at all faces of a solute, then the solvent is considered to screen the solute as good as a conductor. In other words, COSMO calculates a discrete surface around a solute embedded in a virtual conductor (i.e., solvent), where each segment *i* is characterised by its area a_i and the screening charge density (SCD) on this segment. All interactions are considered to be local pair wise interactions of surface segments. COSMO-RS calculates liquid interactions based on three terms: electrostatic interactions ($E_{misfit} \sim (\sigma + \sigma')^2$), hydrogen bond interactions

($E_{H-bond}\sim(\sigma * \sigma')^2$), and Van der Waals interactions ($E_{V\ dW}\sim Area$).[58] Besides the local pair wise surface interactions, COSMO-RS equations are based on the following assumptions: the liquid state is incompressible, the surfaces are in close interactions and all parts of the molecular surfaces could interact with each other, and the 3D geometry may be neglected.

Representation of molecular interactions in COSMO-RS are mainly based on two parameters - σ-profiles, which describes the probability of finding a mean screening charge density on a given segment of the surface, and chemical potentials ($\mu_s(\sigma)$), which describes the affinity of the system S to a surface of polarity σ.[56] The chemical potential, for a given pure liquid compound or a liquid mixture S, could be calculated based on σ as following:

$$\mu_s(\sigma) = \sum_{l}^{m} c_S^l\, M_l^X \quad with\ M_l^X = \int p^X(\sigma)\sigma^l d\sigma \tag{5}$$

Where c_S^l are σ-moment coefficients (SMCs) describing the liquid system S, and M_l^X are the σ-moments of solute X. Total of 7 sigma moments could be calculated (l_{max}= 6) using COSMO*therm*, where, as described above, M_0^X (i.e., Sig0) corresponds to the molecular area of the compound or system.[54, 57] Moreover, the hydrogen bond donor and acceptors strengths can be defined as following:

$$M_{1,HB}^X = \int p^X(\sigma) f_{1,HB}(\sigma) d\sigma \ \ with\ f_{1,HB}(\sigma) = \begin{cases} 0 & if\ \pm\sigma \le \sigma_{HB} \\ \sigma \pm \sigma_{HB} & if\ \pm\sigma \ge \sigma_{HB} \end{cases} \tag{6}$$

Where (HB) describes either a hydrogen bond donor or hydrogen bond acceptor, and σ_{HB} defines the COSMOtherm's hydrogen bonding threshold.

Similarity metrics

A framework to calculate similarity amongst a group of molecules plays an important role in cheminformatics. Whilst the choice of parameterisation could significantly affect the end result of what is considered to be "similar",[59] similarity calculations strongly depend on the choice of the similarity coefficients. However, similarity is a subjective metric when considered based on absolute magnitude of similarity values, and there is no golden standard of choosing a similarity coefficient. Broadly, the commonly used similarity metrics could be categorised in

three: distance (e.g., Euclidean, Manhattan), association (e.g., Dice, Tanimoto), and correlation coefficients (e.g., Kendall's τ and Spearman's ρ).[29] Distance coefficients measure the degree of difference between two variables (i.e., vector of solvent descriptors in our study), and are suitable with integer-valued and real-valued variables. Association coefficients, on the other hand, express whether a certain fingerprint or a descriptor is present in two variables, and is more suitable to work on binary datasets. Finally, correlation coefficients calculate the correlation or dependence between two variables.

On the search of the "best" similarity metric,[27] several studies have been carried out comparing 8,[27] 16,[60] 22,[29] and 51[30] similarity coefficients. Benchmarking 16 metrics on high-content screening (HCS) dataset of cellular phenotypic responses, Reisen *et al*. found that non-linear correlation-based similarity metrics such as Kendall's τ and Spearman's ρ outperformed the commonly used similarity metrics such as Euclidean distance.[60] Bajusz and co-workers quantified the differences between various similarity metrics using sum of ranking differences (SRD) technique over different dataset types (fragment, leadlike, druglike) and sizes (Figure 19).[27] SRD values for a given metric refers to sum of absolute values of differences of an entry (i.e., a molecule) in that metric and an entry in the reference column (i.e., average of all similarity metrics used in the calculation). Smaller values of SRD for a given metric indicates a lower variation from the reference. As given in Figure 19, Cosine coefficient had the lowest SRD variation, followed by Tanimoto index, Dice index, and Soergel distance, which were identified to be the best (and in some sense equivalent) metrics for similarity calculations. On the other hand, distance metrics such as Euclidean and Manhattan distances were found to be far from being optimal and are not recommended as a standalone metric. Moreover, the authors highlighted the influence of size of molecules and dataset selection methods on certain metrics, especially on the rankings of Euclidean and Manhattan distances. Finally, whilst the Tanimoto coefficient is often considered as a superior metric as supported by several studies,[30, 61, 62] the authors highlighted the size dependence, favouring larger molecules in similarity and smaller molecules in diversity selections,[28] of the Tanimoto coefficient.[27, 63]

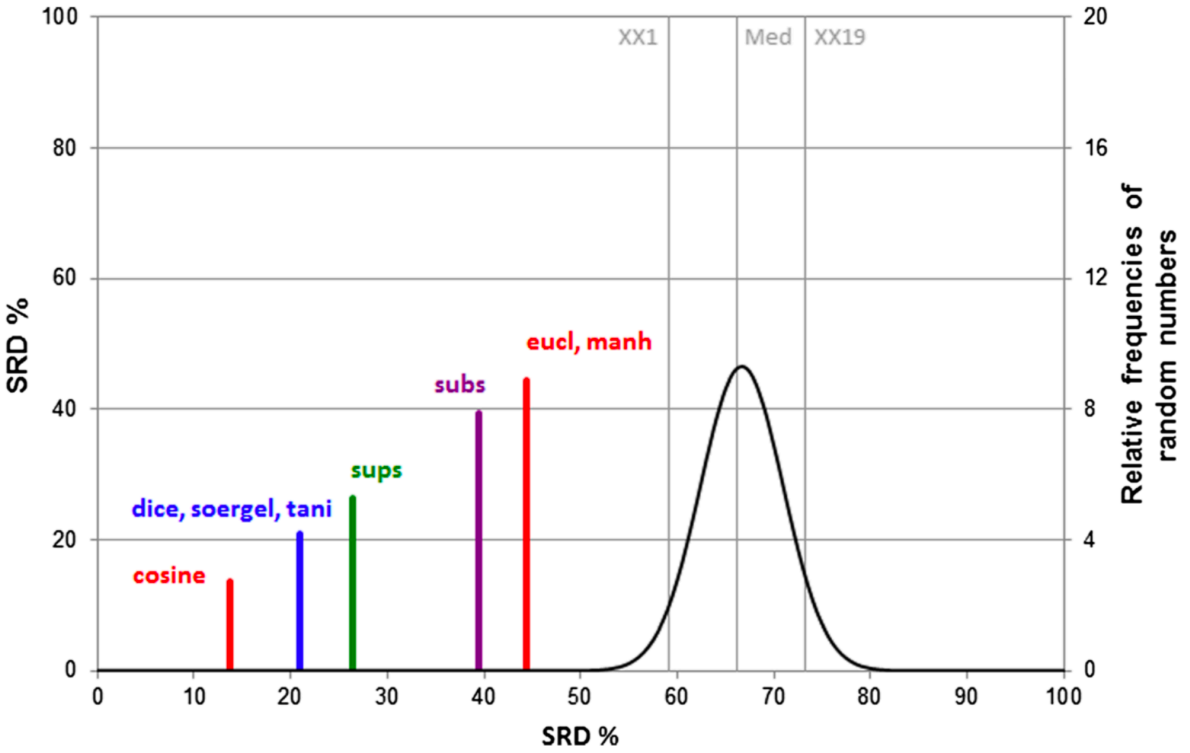


Figure 19. *Visualisation of SRD ranking and grouping of similarity metrics. X-axis represents scaled SRD values and y-axis shows the relative frequencies for the fitted Gauss curve on random numbers (black) (XX1 = 5% error limit, med = median, XX19 = 95% limit). If a similarity metric overlaps with the Gaussian curve, it is not distinguishable from random ranking. The figure was reproduced from ref.*[27]

In this work, due to the variations in the choice of similarity coefficients, and a superior performance when several metrics are used in combination (i.e., similarity fusion),[29, 64] the following 11 similarity metrics were implemented in the solvent recommendation workflow: Euclidean, Cosine, Manhattan, squared Euclidean, Dice, Correlation, Jaccard, Kendall's τ, Spearman's ρ, Pearson's r, and Tanimoto coefficient. Depending on the case study given in Results and Discussions, five or ten closest solvents were recommended per similarity metric for a given solvent, and solvents were ranked based on the frequency of selection by different metrics.

Methods and Materials

Datasets

In terms of solvent libraries, two datasets were used. The first dataset contains 120 solvents from the proprietary UCB Pharma solvent library. The main reason behind this dataset selection was to explore solvent recommendations based on practicality (i.e., selecting from solvents commonly used in the pharmaceutical industry). The second dataset contains a library of 457 solvents, a more comprehensive library to explore "non-obvious" solvents that might not be commonly selected by a synthetic chemist in a lab, but perhaps provides a better process and environmental parameters. This library of 457 solvents were previously used for Mitsunobu esterification and asymmetric hydrogenation reactions for optimising for yield and conversion & d.e., respectively.[43, 44]

Solvent parameterisation

As explained in the Molecular descriptors section above, choice of reaction or study (e.g., QSAR) specific molecular descriptors are important in achieving a highly predictive model with a small amount of data. However, previous studies led to inconclusive results in terms of size and choices of molecular descriptors.[23, 26] Moreover, it is important to note that meaningful physicochemical descriptors could allow for interpretation of suggested results using domain knowledge. Therefore, for solvent selection studies, the parameterisation techniques given in Table 13 were compared. Specifically, fingerprints were selected as they are often used in molecular similarity search. Sigma moments were selected as they were demonstrated to explain the solute-solvent behaviour in five dimensions by Klampt *et al*., where authored reproduced experimentally measured Abraham parameters using sigma moments.[54] Five sigma moments were calculated using COSMO*therm* and solvent structures were loaded from the COSMO*therm* database with BP-TZVP-FINE parameterisation. 17 descriptors library was selected as they were demonstrated to be successful in asymmetric hydrogenation reaction. In the same study, comparing six different models based on different parameterisation of solvents, the best model was selected to be based on sigma profiles of solvents as opposed to the 17 descriptors.[43, 65] Therefore, sigma profiles were included in the study as a standalone descriptor set and also as part of the 17 descriptors.

***Table 13**. List of solvent descriptors used in this study. *indicates the descriptors calculated and used by Amar et al. in asymmetric hydrogenation reaction.[43] More details were provided elsewhere.[65]*

Descriptors	Source
Morgan fingerprints (radius = 2)	RDKit[66]
Sigma moments Sig0: molecular surface area Sig2: electrostatic polarity Sig3: asymmetry of sigma profile Hbacc3: hydrogen bond acceptor strength Hbdon3: hydrogen bond donor strength	COSMO*therm*[55]
Molecular weight (g/mol) Density (g/mL) Molar volume (mL/mol) Refractive index (-) Molecular refractive power (mL/mol) Dielectric constant Dipole moment (D) Melting point (°C) Boiling point (°C) Viscosity (cP) lnP (partition coefficient) Vapour pressure (mbar)	Amar *et al*.*
Sigma profile Partitioned into five regions (Sig1-5)	Amar *et al*.*

Similarity calculations

Similarity calculations between two solvent descriptors vectors were performed using scipy.distance and scipy.stats functions.[67] For distance metrics, similarities were ranked based on increasing values, with lower values indicating a higher similarity. For correlation metrics, high values indicate higher similarity, so the rankings were performed on a decreasing order. Closest five or ten solvents were returned per a given solvent in the dataset depending on the

study (Results and Discussions). Tanimoto similarities were calculated using simplified molecular-input line-entry system (SMILES)[68] representations of solvents, which were computed using custom Cas_No_to_SMILES and Names_to_SMILES functions. All descriptors were scaled to be in the same range using StandardScaler function in Sklearn. Finally, for a given solvent, all recommended solvent alternatives by different metrics were combined in the form of "voting" and most similar solvents were ranked according to their selection frequency by different metrics.

Results and Discussion

First, this section compares (theoretical) results based on the influence of using dimensionality reduction techniques, choices and sizes of parameterisation, and the choices of similarity metrics in suggesting similar solvents. This is demonstrated using both 120 and 457 solvent libraries, with a recommended output example given for dichloromethane (DCM) in Table 14. Second, the recommendation results using our approach is compared against the solvent selection tool by AstraZeneca on two case studies – keto-enol tautomerisation and nucleophilic substitution reactions, and the solvent selection tool developed by and used internally at Syngenta. Then, recommended solvent alternatives are analysed on two case studies (e.g., Menschutkin reaction, *t*-butyl chloride solvolysis). Finally, the section ends with list of recommended greener alternatives for industrially identified hazardous and highly hazardous solvents based on guidelines by GSK, AZ, and others.[33]

Example of an output of alternative solvents for dichloromethane (DCM)

An example of an output for DCM alternatives is given in Table 14. Solvents are ranked based on their distance (i.e., similarity) to the solvent candidate, the first being the "closest" (i.e., most similar). As a sanity check, the solvent itself is intentionally returned as the closest solvent to itself, which is useful for confirming the similarity metric accuracy and robustness. For instance, Dice index (similarity coefficient) works on binary data, and this is why DCM is only selected as the closest solvent to itself when parameterised using 207 nonzero bits of Morgan fingerprints for the solvents (row 4, Table 14). Similarly, similarity coefficients Kendall's tau and Spearman's ρ did not select (i.e., rank) DCM as the closest to itself (rows 8 and 10, Table 14). The reason for such similarity rankings is, probably, because the selected similarity values for Kendall's tau and Spearman's ρ correspond to one of the values in [1.0, 0.33, 0.00, -0.33, -1.0] and [1.0, 0.5, 0.0, -0.5, -1.0], respectively. Therefore, especially when solvents are

described in a small set of descriptors (e.g., 3 principal components), dozens of solvents are ranked to be similar with a same similarity value (e.g., 1.0). Whilst the list includes the original solvent DCM as similar with 1.0 similarity value, the function is not able to differentiate the similarities any further, and does not include the original solvent itself in top five closest alternatives.

Looking at the output example in Table 14, most metrics return similar solvents (e.g., $CHCl_3$, 1,2-dichloroethane, benzene, chlorobenzene), solvents that would normally be selected based on domain knowledge, along with some alternatives such as anisole, nitromethane, benzyl alcohol, carbon disulfide (CS_2), and (trifluoromethyl)benzene – solvents that might not be selected as an obvious alternative to DCM. It is important to highlight that the solvents are recommended based on various parameterisation techniques, and the objective is to have a list of suggestions that could achieve similar process performance, but perhaps with better SHE impact or more process specific physical properties. When expanded to suggesting the closest ten solvents from the library of 457 solvents, suggested alternatives include 1,1-dichloroethane, *cis*- and *trans*-1,2-dichloroethene, 1,1,2-trichloroethane, 1,2- and 1,2-dichloropropane, ethanethiol, acetyl chloride, and others (Table 22 and 23, Appendix). Although the expanded list may not include solvents with significantly better SHE scores, the list includes solvents that might not normally be selected. The reason for choosing DCM as an example was due to relatively easier identification of its homologous (i.e., structurally similar) solvents, making it easier to validate the suggested alternatives. Main advantages of using multiple similarity metrics and parameterisation techniques are explained in the case studies in Results and Discussion.

***Table 14**. An example of an output for solvent alternatives for DCM using various similarity metrics and parameterisation techniques. DCM: Dichloromethane, DCE: Dichloroethane, BnOH: Benzyl alcohol, TFA: Trifluoroacetic acid, TFMA: Trifluoromethansulfonic acid, PhCF3: (trifluoromethyl)benzene, TMG: 1,1,3,3-tetramethylguanidine, NMP: N-methyl-2-pyrrolidinone, HMPA: hexamethylphosphoramide, DMI: dimethyl isosorbide. PEG: polyethylene glycol. PCs: principal components, FPs: Morgan fingerprints.*

Metrics	5 sigma moments	PCA (n=3 PCs)	207 nonzero FPs	PCA (n=60 PCs)
Euclidean	DCM 1,2-DCE $CHCl_3$ Benzene Chlorobenzene	DCM $CHCl_3$ Benzene 1,2-DCE Chlorobenzene	DCM 1,2-DCE H_2O Benzene PEG	DCM 1,2-DCE H_2O 1-Chlorobutane PEG
Cosine	DCM 1,2-DCE $CHCl_3$ Benzene Chlorobenzene	DCM 1,2-DCE $CHCl_3$ Nitromethane Benzene	DCM 1,2-DCE 1-Chlorobutane Glycol $CHCl_3$	DCM 1,2-DCE 1-Chlorobutane $CHCl_3$ H_2O
Correlation	DCM $CHCl_3$ TFMA TFA Benzyl alcohol	DCM CS_2 $CHCl_3$ TFMA Benzyl alcohol	DCM 1,2-DCE 1-Chlorobutane Glycol $CHCl_3$	DCM 1,2-DCE 1-Chlorobutane $CHCl_3$ H_2O
Dice	*DMI,* *Ethyl oleate* *Glycerol* *1,3-Propanediol,* *Diglyme*	*DMI* *HMPA* *Dimethyl adipate,* *Diglyme,* *Ethyl succinate*	DCM, 1,2-DCE, 1-Chlorobutane, Glycol, $CHCl_3$	*Ethyl oleate,* *Aceticacid-2-ethylhexylester,* *Dipentene,* *Methyl acetate*
SEuclidean	DCM 1,2-DCE $CHCl_3$ Benzene Chlorobenzene	DCM $CHCl_3$ Benzene 1,2-DCE Chlorobenzene	DCM 1,2-DCE H_2O Benzene PEG	DCM 1,2-DCE H_2O 1-Chlorobutane PEG
Manhattan	DCM 1,2-DCE Benzene $CHCl_3$ Chlorobenzene	DCM $CHCl_3$ Benzene 1,2-DCE Cyclohexane	DCM 1,2-DCE H_2O Benzene PEG	DCM 1,2-DCE H_2O 1-Chlorobutane Glycol
Jaccard	DCM Hexane CS_2 Benzene Pentane	DCM TMG n-Pentyl acetate N-Methylmorpholine N-Methyl-2-pyrrolidinone	DCM 1,2-DCE 1-Chlorobutane Glycol $CHCl_3$	DCM TMG n-Pentyla cetate N-Methylmorpholine N-Methyl-2-pyrrolidinone
Kendalltau	DCM Chlorobenzene $CHCl_3$ Anisole 1,2-DCE	*Methyl formate,* *Benzyl alcohol,* *Methanesulfonic acid,* *Isopentanol,* *Isobutanol,*	DCM 1,2-DCE 1-Chlorobutane Glycol $CHCl_3$	DCM 1,2-DCE 1-Chlorobutane Glycol CS_2
Pearsonr	DCM $CHCl_3$ TFMA TFA Benzyl alcohol	DCM CS_2 $CHCl_3$ TFMA Benzyl alcohol	DCM 1,2-DCE 1-Chlorobutane $CHCl_3$ Glycol	DCM 1,2-DCE 1-Chlorobutane $CHCl_3$ H_2O
Spearmanr	DCM $CHCl_3$ Anisole DCE PhCF3	*Methyl formate,* *Benzyl alcohol,* *Methanesulfonic acid,* *Isopentanol,* *Isobutanol,*	DCM 1,2-DCE 1-Chlorobutane glycol $CHCl_3$	DCM 1,2-DCE 1-Chlorobutane $CHCl_3$ CS_2
Tanimoto	DCM 1,2-DCE 1-Chlorobutane $CHCl_3$ Glycol	DCM 1,2-DCE 1-Chlorobutane $CHCl_3$ Glycol	DCM 1,2-DCE 1-Chlorobutane $CHCl_3$ Glycol	DCM 1,2-DCE 1-Chlorobutane $CHCl_3$ Glycol

Overlap between different similarity metrics

Even though dozens of similarity metrics were studied for similarity search for larger molecules in medicinal chemistry,[27, 29] to the best of our knowledge, their influence on solvent selection has not been studied. Table 15 contains the overlap of closest five suggested solvents for ten similarity metrics in two solvent libraries. First, Dice index was removed as it only works on binary data. Second, Squared Euclidean and Euclidean, both distance coefficients, returned the same list of suggestions as expected. Similarly, Correlation coefficient and Pearson's r returned the same list of suggestions. To avoid selection bias amongst the metrics, average overlap of metrics was calculated after removing Squared Euclidean and Pearson's r metrics (row Average(n), Table 15). Kendall's τ and Spearman's ρ, which are considered as general correlation coefficients,[69] resulted in a similar overlap percentage. Pearson's r, same as the Correlation coefficient, is robust to non-linear correlations and resulted in a larger overlap (63.23% in 120 dataset and 72.02% in 457 datasets) compared to Spearman's ρ, which is used to capture linear correlations. The metrics Euclidean, Cosine, and Manhattan all resulted in a similar overlap, with Jaccard and Tanimoto having the lowest overlap of suggested solvents with other metrics. Moreover, when larger set of descriptors are used (e.g., 17 descriptors versus five sigma moments), the overlap between different metrics increased, suggesting that the choice of similarity metric might have a relatively less influence compared to sampling from the solvent library parameterised using a smaller set of descriptors.

As demonstrated, there is a significant difference in solvent suggestions based on the choices of similarity metrics and parameterisation techniques. This is why combination (i.e., similarity fusion) of different similarity metrics was used in this study. In addition to Tanimoto index, which captures the chemical substructure overlap based on solvent fingerprints, one metric per similarity coefficient class – distance (Euclidean), association (Cosine), and correlation (Correlation) was selected to define the final list of similarity metrics. In the following sections, only solvents suggested using combination of these four metrics will be used.

***Table 15**. Average overlap % between different similarity metrics based on the closest five recommended solvents for every solvent in the 120 solvents library, parameterised using five sigma moments, and every solvent in the 457 solvents library, parameterised using 17 descriptors. Eucl: Euclidean, Cos: Cosine, Corr: Correlation, SEucl: Squared Euclidean, Manh: Manhattan, Tan: Tanimoto.*

	120 solvents									
	Eucl.	Cos.	Corr.	SEucl.	Manh.	Jaccard	Kendalltau	Pearsonr	Spearmanr	Tan.
Eucl.	100									
Cos.	84.31	100								
Corr.	73.75	79.30	100							
SEucl.	100.00	84.31	73.75	100						
Manh.	91.53	82.92	72.36	91.53	100					
Jaccard	38.05	38.61	38.33	38.05	38.47	100				
Kendalltau	61.11	63.19	65.00	61.11	60.97	35.55	100			
Pearsonr	73.75	79.30	100.00	73.75	72.36	38.33	65.00	100		
Spearmanr	61.53	63.19	66.52	61.53	61.25	36.52	83.75	66.53	100	
Tan.	49.03	48.05	47.36	49.03	50.42	36.66	43.75	47.36	45.14	100
Average	70.34	69.24	68.49	70.34	69.09	37.62	59.94	68.49	60.66	46.31
Average(n)	65.62	65.66	63.23	-	65.42	37.46	59.05	-	59.70	45.77
	457 solvents									
	Eucl.	Cos.	Corr.	SEucl.	Manh.	Jaccard	Kendalltau	Pearsonr	Spearmanr	Tan.
Eucl.	100									
Cos.	81.51	100								
Corr.	77.53	85.27	100							
SEucl.	100.00	81.51	77.53	100						
Manh.	86.21	76.18	73.30	86.21	100					
Jaccard	40.88	40.19	39.90	40.88	42.34	100				
Kendalltau	86.10	89.57	90.70	86.10	86.51	44.53	100			
Pearsonr	92.52	98.18	100.00	92.52	86.91	43.36	73.45	100		
Spearmanr	84.97	88.95	91.32	84.97	82.90	43.25	87.89	74.16	100	
Tan.	44.53	45.11	46.10	44.53	44.53	35.59	28.21	29.06	28.17	100
Average	77.14	76.27	75.74	77.14	73.90	41.21	74.78	76.68	74.06	38.43
Average(n)	71.68	72.40	72.02	-	70.28	40.95	73.36	-	72.49	38.89

Overlap between different parameterisation techniques

This section compares the overlap of suggested solvents based on different descriptor sets. First, only approximately half of the suggested solvents are the same based on sigma moments versus molecular fingerprints in both solvent libraries (Table 16 and 17), highlighting a significant difference in two parameterisation techniques. Second, whilst PCA is commonly used to reduce the dimensions of the sampling space,[35] it could introduce variations in the selection of closest alternatives as given by the overlap between three principal components (PCs) and five sigma moments (80.14%, Table 16) and ten PCs and 17 descriptors (90.76%, Table 17). Indeed, it was observed in solvent selection for photoredox amine synthesis in Chapter 5 that surrogate reaction models (e.g., Neural Networks, XGBoost) achieved a better prediction performance of reaction yield when five sigma moments were used instead of three PCs. This suggests that using PCA to reduce the solvent parameters dimension might not

always achieve a better reaction performance, as demonstrated by the variations in suggested solvents. However, as observed in the choice of similarity metrics, using larger dimensions was also observed to increase the overlap on average (Table 16 versus Table 17).

Table 16*. Overlap % between list of top five (i.e., most similar or closest 5) solvents based on 120 solvents library. PCs: Principal Components, FPs: Fingerprints. 3 PCs account for 94.03% of the original information in five sigma moments. 60 PCs retain 95.27% of original information in 207 nonzero fingerprints data.*

All 10 similarity metrics				
	3 PCs from sigma moments	5 sigma moments	60 PCs from nonzero FPs	207 nonzero FPs
3 PCs from sigma moments	100			
5 sigma moments	69.40	100		
60 PCs from FPs	52.93	54.86	100	
207 nonzero FPs	46.73	52.76	84.43	100
Only 4 similarity metrics				
	3 PCs from sigma moments	5 sigma moments	60 PCs from nonzero FPs	207 nonzero FPs
3 PCs from sigma moments	100			
5 sigma moments	80.14	100		
60 PCs from FPs	58.51	61.15	100	
207 nonzero FPs	58.65	61.39	92.95	100

Table 17*. Overlap between list of top five solvents based on 457 solvents library. Results compare the overlap between use vs no use of PCA on 17 descriptors library. PCs: Principal Components. 10 PCs account for 96.30% of the original information in 17 descriptors*

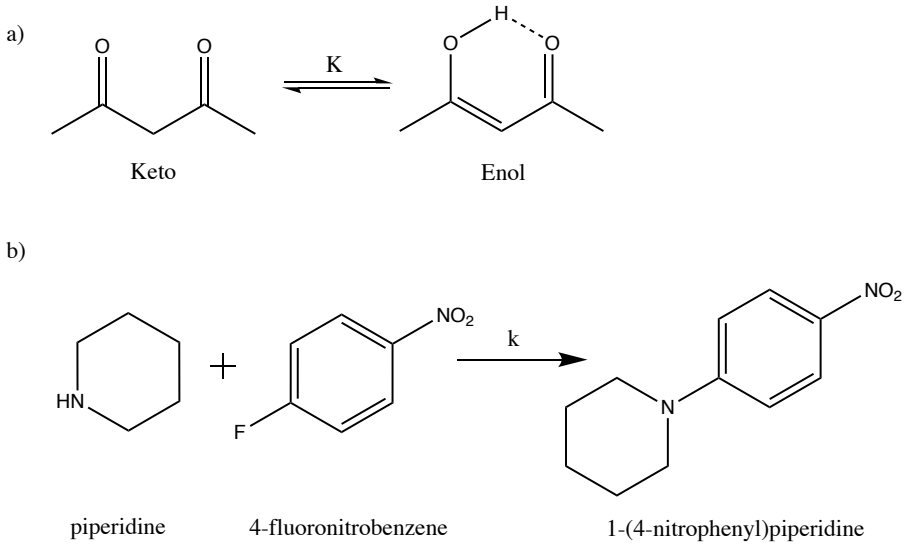
All 10 similarity metrics			
	10 PCs from 17 descriptors	17 descriptors	Sigma profile 1-5
10 PCs from 17 descriptors	100		
17 descriptors	79.55	100	
Sigma profile 1-5	57.08	58.16	100
Only 4 similarity metrics			
	10 PCs from 17 descriptors	17 descriptors	Sigma profile 1-5
10 PCs from 17 descriptors	100		
17 descriptors	90.76	100	
Sigma profile 1-5	66.82	67.93	100

Comparison to AstraZeneca's (AZ) solvent selection guide

AZ case study 1: keto-enol tautomerisation

Available as an interactive tool to sample from 272 solvents database, parameterised using both experimental and computational descriptors, the authors validated the applicability of the solvent selection guide on a few case studies.[4] For instance, in tautomerisation of acetylacetone (Scheme 2a), dependence of % enol on the solvent choice was observed as a clear trend in the first principal component (PC1) versus the second principal component (PC2) plot on the PCA map. Larger values for PC1 indicated a higher percentage of the enol form. List of solvents included in the study is given in Table 18. For a given solvent, the closest (i.e., most similar) solvents from the list of 120 commonly used solvents list are selected. For example, carbon disulfide (CS_2), hexane, and benzene are recommended to be similar to each other, and all three solvents have a comparable amount of enol form in the equilibrium, suggesting a similar reaction performance for similar solvents. Diethyl ether was not recommended to be similar to benzene, hexane or CS_2, which could be due to the availability of other closer solvents than diethyl ether. However, hexane was a recommended alternative for diethyl ether, indicating a similar reaction performance. Solvents in the parenthesis (e.g., DCM) are included as mutually similar to the solvents included in the study. For instance, DCM was selected to be similar to CS_2, benzene, and chloroform – solvents with 87-94% enol form. This implies that DCM could lie somewhere mid-point on the distance map to the given solvents. Similarly, solvents with

relative lower amount of the enol form (62-82%) were recommended to be alternatives to each other, suggesting a trend in the equilibrium composition in various solvents.



***Scheme 2**. a) keto-enol tautomerisation of acetylacetone[70] and b) nucleophilic substitution reaction between 4-fluoronitrobenzene and piperidine.[71]*

***Table 18**. List of solvents and enol % in tautomerisation of acetylacetone, reproduced from Rogers et al.,[70] alongside recommended "similar" solvents based on 120 solvents library. Carbon tetrachloride (CCl4) was removed from the list since it was not available in the 120 solvents library. Order of solvents are given based on the similarity (i.e., distance) to the target solvent. THF: tetrahydrofuran, DCM: dichloromethane, DMSO: dimethylsulfoxide.*

Solvent	% enol	Suggested solvent alternatives
Hexane	95	Benzene, carbon disulfide, chloroform
Diethyl ether	95	Hexane (THF)
Carbon disulfide	94	Benzene, hexane (DCM, acetonitrile)
Benzene	89	Carbon disulfide, hexane, chloroform (DCM)
Chloroform	87	Benzene, carbon disulfide (DCM)
Dioxane	82	(THF)
Ethanol (absolute)	82	Methanol
Methanol	74	Ethanol, acetonitrile (formic acid, water)
Acetic acid	67	Ethanol (water, formic acid, acetone)
Acetonitrile	62	DMSO, ethanol, methanol (water, acetone)
DMSO	62	Dioxane (acetone)

AZ case study 2: nucleophilic substitution reaction

Schmid *et al*. studied the relationship between solvent properties (e.g., donor number, acceptor number, and dielectric constant) and second order reaction rate for S_N2 reaction of p-fluoronitrobenzene and piperidine (Scheme 2b).[71] Solvent rankings based on the experimental -*lnk* value is given in Table 19. Based on this reported data, in the solvent selection guide by AstraZeneca, authors simply highlight a change in the solvent pattern on the solvent PCA map when the nucleophile choice was changed from anionic azide ion to piperidine, suggesting a slight difference in the underlying mechanisms of the two reactions. No clear patterns (e.g., linear trend, clustering) of solvents were observed. In our solvent recommendation guide, based on the list of selected solvent alternatives, a clear pattern is observed in the solvent ranking. For instance, first three solvents DMSO, DMF, and DMA are recommended to be alternatives to each other, with propylene carbonate as a common solvent alternative for all the three. Similarly, solvents from acetonitrile to nitroethane have mutual recommended solvents, suggesting a comparable reaction rate in these solvents. A further trend could be observed for solvents from butanone to ethyl acetate (entries 9-15), and for solvents anisole and benzene. Although dibutyl ether was recommended as an alternative to 1,2-dimethoxyethane (entry 14), none of the solvents in the Table 19 was recommended to be similar to dibutyl ether, indicating a limitation for some of the solvents. However, considering a clear pattern in the rankings of solvents and recommended solvent alternatives, our workflow demonstrates a certain degree of generalisability of solvent selection for the reaction in study.

***Table 19**. List of solvents studied in a nucleophilic substitution reaction between piperidine and 4-fluoronitrobenzene, alongside solvent ranking based on the experimental -lnk values of the reaction based on the work by Schmid et al.[71] *indicates that solvents were sampled from the 457 solvents library as they were not available in the library of commonly used 120 solvents. Order of solvents are given based on the similarity to the target solvent. DME: 1,2-dimethoxyethane, DMSO: dimethylsulfoxide, DMF: N,N-dimethylformamide, DMA: N,N-dimethylacetamide, MeCN: acetonitrile, MeOAc: methyl acetate, EtOAc: ethyl acetate, PC: propylene carbonate.*

Solvent	Rank	Suggested solvent alternatives
DMSO	1	DMF, nitromethane, acetone (PC)
DMF	2	DMA, DMSO (propionitrile, PC)
DMA	3	DMF, DMSO, Acetone, (PC)
MeCN	4	Nitromethane, Acetone, DMA, DMSO
Nitromethane	5	Nitroethane, Acetone
Benzonitrile	6	DMF, DMA (hexanitrile)
Acetone	7	MeCN, EtOAc
Nitroethane	8	Nitromethane, MeCN (propionitrile, methyl formate)
Butanone	9	Acetone, 3-pentanone, EtOAc (propionitrile)
3-pentanone*	10	3,3-dimethylbutanone, 2-butanone, acetone, DME
3,3-dimethylbutanone*	11	3-pentanone, 2-butanone, MeOAc, acetone, DMA
THF	12	Butanone, acetone, DME (diethyl ether)
MeOAc	13	EtOAc, butanone, acetone, acetonitrile, DMA, DMSO (methyl formate)
DME*	14	*EtOAc, MeOAc, 3-pentanone, dibutyl ether (dimethoxymethane)
EtOAc	15	MeOAc, butanone, DME, acetone
Anisole	16	Benzene, DME (toluene, chlorobenzene, isopropylbenzene)
Benzene	17	Anisole (toluene, chlorobenzene, isopropylbenzene)
Dibutyl ether*	18	-

AZ case study 3: selection of primary alcohols

On the PCA map of solvents by AstraZeneca, the first seven primary alcohols (i.e., methanol to 1-heptanol) follow a clear line.[4] In our work, ethanol, 1-propanol, and 1-butanol were selected to be alternatives to methanol, alongside other recommendations such as 2-propanol, 1,2-ethanediol, and acetone. The reason for the lack of C5-C7 primary alcohols in the recommendation list is mainly because only the closest 10 solvents to methanol were selected and the library of 457 solvents contains other solvents than C5-C7 primary alcohols that are more "similar" to methanol. Looking at the alternatives to 1-butanol, we have identified 1-pentanol, 1-hexanol, 1-heptanol, and 1-octanol, alongside tert-butanol, isopentanol, and cyclohexanol. Considering the list of recommended solvents for methanol and 1-butanol, this suggests a significant overlap between the AstraZeneca's solvent selection guide and our workflow.

Overall, compared to the AZ solvent selection guide, which requires extensive list of both experimental and computational parameters, all three benchmarked case studies demonstrate a relatively better performance (i.e., clear pattern) in solvent selection in our approach. For a given solvent, a list of alternatives is provided based on their similarity score as opposed to selecting solvents from a PCA map. The three principal components on the PCA map accounts for 70.3% of the original solvent descriptors, which indicates a loss of certain amount of information regarding the solvents. Moreover, even though the authors suggest a clear pattern on the PCA map for a given list of solvents, which was not the case for the nucleophilic substitution reaction, a list of solvents ranked based on their similarity provides a clear guide for selection of alternative solvents.

Comparison to the internal tool developed by Syngenta

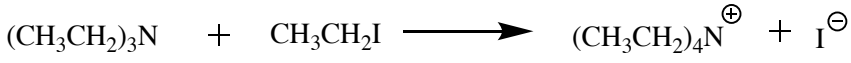
The interactive solvent selection tool by Syngenta provides more flexibility in selecting solvents from the dataset of 209 solvents.[12] Herein, closest ten solvent alternatives to diethyl ether, using Syngenta's tool as reported by the authors, were compared to the alternatives using our approach. Solvents *t*-butylmethyl ether, cyclopentyl methyl ether, 2-methyltetrahydrofuran, diethyl carbonate, and diisopropyl ether were all recommended as an alternative to diethyl ether in both selection guides. Compared to butyl acetate and isopropyl acetate suggested by Syngenta, our list includes ethyl acetate and ethyl formate, which indicates that both approaches selected similar esters as an alternative to diethyl ether.

Additionally, compared to tributylamine suggestion, our list includes triethylamine, again, both approaches suggesting a similar solvent from a different solvent "class". Whilst Syngenta's tool provides more functionalities, these examples demonstrate a significant overlap in selecting solvent alternatives using only a handful of computational descriptors in our approach. Moreover, Syngenta's tool does not evaluate different similarity coefficients (e.g., Euclidean, Tanimoto, Cosine), requires extensive list of solvent parameters, and is limited to internal use only.

Validation using experimental Menschutkin reaction data

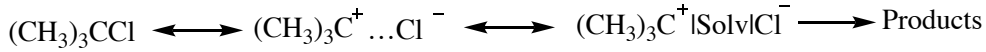
The first study of solvent influence on reaction rates was reported by Nikolai Menschutkin reaction in 1890 for the reaction of triethylamine and ethyl iodide (Scheme 3a) at 100 °C in 22 solvents.[72] It was observed that the reaction rate in benzyl alcohol was ~800 faster than in n-hexane, which led to the generalisable trend of increasing reaction rate in hydrocarbons < ethers < esters < alcohols / ketones < ketones / alcohols.[73] Since then, Menschutkin reaction was studied both computationally and experimentally as the reaction rate is known to be influenced significantly by the reaction medium.[22, 41, 73, 74]

a)



Triethylamine Ethyl iodide

b)



t-Butyl chloride

***Scheme 3**. a) Menschutkin reaction of triethylamine and ethyl iodide at 298.15 K and b) solvolysis of t-butyl chloride at 298.15 K.*

Table 20 shows list of 30 solvents and their rankings based on experimental reaction rate. Similar observations to the “class” based rankings reported by Menschutkin are observed in the table, but with more specific solvent examples and their alternatives. For instance, except for diethyl ether alternatives, every solvent in the first ten entries includes a suggested solvent that’s next to the target solvent in the ranking. Similar patterns are observed for the entries in the rest of the table as well, except for acetophenone (entry 21), where no neighbour solvent from the rankings was selected to be similar. However, the recommended alternatives benzaldehyde, 2-nitrotoluene, 2- and 4-nitroanisole are all recommended to be similar to benzonitrile (entry 22) and nitrobenzene (entry 24), suggesting a significant overlap in similarity. Similarly, all solvents from entry 25 to 30 are suggested (i.e., neighbouring solvents in the ranking) as an alternative to each other, except for 1,1,2,2-tetrachloroethane. As demonstrated in this example, recommended “similar” solvents show a similar reaction performance, based on the suggestions from the top ten closest solvents list, validating the similarity recommendations. A comparable reaction performance in recommended list allows for wider selection of solvents based on different criteria (e.g., SHE impacts, cost) without sacrificing on the process performance.

***Table 20**. List of solvents and their rankings based on experimental reaction rate constant for the Menschutkin reaction of triethylamine and ethyl iodide at 298.15 K. The data was taken from ref.[73, 74] Solvents cyclohexyl chloride and cyclohexyl bromide were removed from the list since they were not included in the solvent library. Alternative solvents were selected from the library of 457 solvents and listed based on their similarity to the target solvent. * indicates solvents that were selected from the library of 120 commonly used solvents. DCE: dichloroethane, DCM: dichloromethane, TCE: trichloroethane, THF: tetrahydrofuran, DMSO: dimethylsulfoxide, DMF: N,N-dimethylformamide, DMA: N,N-dimethylacetamide, PC: propylene carbonate, CS_2: carbon disulfide, CCl4: carbon tetrachloride.*

Solvent	Rank	Suggested solvent alternatives
n-Hexane*	1	Cyclohexane, toluene, benzene (mecyclohexane,)
Cyclohexane*	2	n- Hexane, toluene, THF (CS_2, mecyclohexane)
Diethyl ether*	3	THF, n-hexane, ethyl acetate, (2-MeTHF)
CCl4	4	1,1,1-TCE (2,2-dichloropropane)
1,1,1-TCE	5	CCl4 (1,1,2-TCE, 2,2-dichloropropane)
Toluene*	6	Benzene, chlorobenzene, n-hexane, cyclohexane (methoxybenzene)
Benzene*	7	Toluene, chlorobenzene, chloroform, 1,2-DCE, cyclohexane
Ethyl acetate*	8	Dioxane, 2-butanone, acetone
Dioxane*	9	2-Butanone, THF, acetone, ethyl acetate (diethyl ether)
THF*	10	Dioxane, 2-butanone, acetone, cyclohexane (diethyl ether)
Ethyl benzoate	11	Acetophenone (ethoxybenzene)
Chlorobenzene	12	Bromobenzene (1,2-, 1,3-, and 1,4-dichlorobenzene, fluorobenzene, methoxybenzene)
Bromobenzene	13	Chlorobenzene, iodobenzene (1,2- and 1,3-dichlorobenzene)
1,1-DCE	14	1,2-DCE, DCM, chloroform, 1,1,2,2-tetrachloroethane
Chloroform	15	DCM, 1,1,2,2-tetrachloroethane, 1,1-DCE, 1,1,2- and 1,1,1-TCE
2-Butanone	16	Acetone, propionitrile, ethyl acetate (butyronitrile, 3-pentanone)

Iodobenzene	17	Bromobenzene, chlorobenzene (fluorobenzene, 1,2-, 1,3-, and 1,4-dichlorobenzene)
Acetone	18	2-Butanone, propionitrile (2-pentanone, butyronitrile)
1,2-DCE	19	DCM, 1,1-DCE, nitromethane (1,1,2-TCE)
DCM	20	1,2-DCE, 1,1-DCE, chloroform (1,1,2-TCE)
Acetophenone	21	(2-nitrotoluene, methyl benzoate, 2- and 4-nitroanisole, benzaldehyde)
Benzonitrile	22	Nitrobenzene, acetophenone, bromobenzene, (benzaldehyde, 2- and 4-nitroanisole)
Propionitrile	23	Acetonitrile, acetone, 2-butanone, DMF (butyronitrile)
Nitrobenzene	24	Benzonitrile, acetophenone, iodobenzene, chlorobenzene (2-nitrotoluene, 2- and 4-nitroanisole, benzaldehyde)
DMF	25	Propionitrile, PC, DMSO (DMA, butyronitrile)
1,1,2,2-Tetrachloroethane	26	Chloroform, DCM, 1,1-DCE (1,1,2-TCE)
Acetonitrile	27	Propionitrile, nitromethane, DMF, Acetone
Nitromethane	28	Acetonitrile, propionitrile (2-nitrotoluene)
PC	29	DMSO, DMF, (DMA, sulfalone)
DMSO	30	PC, nitromethane, DMF, acetone (DMA, sulfalone)

Solvolysis of t-butyl chloride

Based on solvatochromatic equations, Zhou *et al*. studied the influence of 136 commonly used solvents on reaction rate of various reactions.[75] Using 12 segments of solvents' sigma potential curves, reduced down to four principal components using PCA, authors successfully modelled the dependence of experimental reaction rate (log k) for solvolysis of t-butyl chloride on solvent properties (Scheme 3b). Similar to case studies in previous sections, list of solvents and rankings are given in Table 21. Considering the experimental data collected from Zhou *et al*.'s work is based on commonly selected solvents, solvent alternatives were selected from the library of 120 solvents as opposed to the larger library of 457 solvents, unless stated otherwise. For simplicity, only five alternatives are provided for a given solvent, and certain ranges (e.g., entry 1-5, 6-15) are provided for an easier analysis to the reader. Overall, good pattern is observed. In the list of 30 solvents, only acetophenone (entry 32, Table 21) does not include a suggested alternative with a similar ranking. However, acetophenone alternatives include 2-methylbenzonitrile and 2-nitroanisole, which are also recommended to be similar to benzonitrile (entry 31), both solvents containing a several mutually similar solvents. It is important to highlight that a recommended solvent alternative could be relatively away on the rankings, which could be due to the alternative solvent being relatively far away on the descriptors space, as would be validated based on the calculated similarity value (e.g., Euclidean distance).

Table 21. *List of solvents and rankings based on the experimental data (log k_{exp}) for t-butyl chloride solvolysis at 298.15 K. Solvents alternatives were selected from the library of 120 solvents, except for solvents labelled with *, which were selected from the library of 457 solvents. DMSO: dimethylsulfoxide, IPA: 2-propanol, PC: propylene carbonate, DMF: N,N-dimethylformamide, NMP: N-methylpyrrolidine, DMA: N,N-dimethylacetamide, DCM: dichloromethane, DCE: 1,2-dichloroethane, THF: tetrahydrofuran, EtOAc: ethyl acetate, DME: 1,2-dimethoxyethane, NMF: N-methylformamide, NMA: N-methylacetamide, NMP: N-methylpyrrolidine, DEC: diethyl carbonate, THF: tetrahydrofuran.*

Solvent	Rank	Suggested solvent alternatives
Water*	1	Formic acid, formamide, ethane-1,2-diol, ethanol, acetic acid
Formic acid*	2	Formamide, water, ethane-1,2-diol, acetic acid, ethanol
Propane-1,2,3-triol*	3	Ethane-1,2-diol, formamide, water, formic acid, ethanol
Formamide*	4	Formic acid, water, ethane-1,2-diol, propane-1,2,3-triol, NMF
Ethane-1,2-diol*	5	Propane-1,2,3-triol, acetic acid, formic acid, formamide, water
Methanol*	6	Ethanol, 1-propanol, IPA, 1-butanol, water
NMF *	7	NMA, formamide, water, formic acid, propane-1,2,3-triol
Acetic acid*	8	Ethane-1,2-diol, formic acid, NMA, water, 2-methyl-2-propanol
Ethanol	9	1-propanol, 1-butanol, IPA, ethane-1,2-diol, water
1-butanol	10	1-propanol, IPA, ethanol, water, methanol
1-propanol	11	1-butanol, ethanol, 2-propanol, methanol, water, 1-octanol
NMA*	12	NMF, formamide, PC, DMSO, DMA, acetic acid
1-hexanol*	13	1-octanol, 1-butanol, cyclohexanol (2-methyl-2-butanol)
1-octanol	14	Cyclohexanol, 1-propanol (2-methyl-2-butanol)
IPA	15	1-propanol, 1-butanol, 2-methyl-2-propanol, cyclohexanol (2-methyl-2-butanol)
DMSO	16	DMF, DMA, NMP, propanone, nitromethane
Sulfalone*	17	DMSO, PC, NMP, DMF, DMA
Cyclohexanol	18	2-methyl-2-propanol, IPA, 1-propanol, 1-butanol, 2-methyl-2-butanol
Nitromethane	19	Acetonitrile, PC, DMSO, DMF, chloroform,
PC	20	EtOAc, NMP, acetonitrile
2-methyl-2-propanol	21	2-methyl-2-butanol, IPA, 1-butanol, cyclohexanol, 1-propanol
DMF	22	DMA, NMP, DMSO, propanone

Acetonitrile	23	Nitromethane, PC, propanone, 2-methyl-2-propanol, DMSO
2-methyl-2-butanol	24	2-methyl-2-propanol, cyclohexanol, IPA, 1-butanol, 1-propanol
NMP	25	DMA, DMF, DMSO, cyclohexanone, PC
DMA	26	DMF, NMP, DMSO, propanone, diethyl ether
Propanone	27	Cyclohexanone, DMSO, DMA, NMP, EtOAc
Cyclohexanone	28	DMF, EtOAc, NMP, diethyl ether
Chloroform	29	DCM, DCE, chlorobenzene, benzene, IPA (toluene)
Nitrobenzene*	30	Benzonitrile, acetophenone, chlorobenzene (2-nitroanisole, toluene)
Benzonitrile*	31	Nitrobenzene, acetophenone, (2-methylbenzonitrile, 2-nitroanisole, toluene)
Acetophenone*	32	(2-methylbenzonitrile, 2-nitroanisole)
Diethyl oxalate*	33	EtOAc (DEC)
Diethyl ether	34	THF, EtOAc, pentane (DEC, 2-MeTHF)
DCM	35	DCE, chloroform, chlorobenzene, benzene, nitromethane,
DCE	36	DCM, chloroform, chlorobenzene, benzene, pentane
THF	37	Diethyl ether, propanone, cyclohexanone, (2-MeTHF, DME, cyclohexane)
Chlorobenzene	38	Benzene, DCE, pentane (toluene, cyclohexane)
EtOAc	39	Cyclohexanone, diethyl ether (DEC, methyl acetate, DME)
Benzene	40	Chlorobenzene, DCM, DCE, chloroform, pentane (toluene, hexane)
Heptane	41	Pentane (hexane, cyclohexane)
Pentane	42	Heptane, benzene, chlorobenzene, (cyclohexane, toluene)

Greener alternatives to (highly) hazardous and problematic solvents

As mentioned earlier, it is important to balance for SHE impacts and process performance. Skowerski *et al*. screened alternative solvents for olefin metathesis in order to replace the commonly used solvents such as DCM, 1,2-DCE, toluene, and benzene that have major regulatory issues.[76] Indeed, in our workflow, all three solvents are recommended to be similar to DCM, suggesting that solvents demonstrate similar process performance. Studied alternative solvents, for achieving high reaction yield using ten most frequently used ruthenium (Ru) catalysts, included methanol, 2-propanol (IPA), ethyl acetate (EtOAc), dimethyl carbonate (DMC), cyclopentyl methyl ether (CPME), and 2-methyl-tetrahydrofuran (2-MeTHF). Overall reaction performance in solvents could be categorised as following: EtOAc ~ DMC > CPME, 2-MeTHF > IPA, methanol. Looking at the suggested alternatives using our workflow for EtOAc, DMC is indeed recommended to be similar to EtOAc. Moreover, CPMA and 2-MeTHF are suggested to be alternatives for each other, whilst they are not selected to be similar to EtOAc. Similar pattern was observed for methanol and IPA (i.e., neither of the solvents is recommended as an alternative to the others, whilst they are suggested as an alternative to each other). Finally, even though benzyl alcohol and anisole were not included in the study by Skowerski *et al.*, both of these solvents have significantly better SHE scores compared to the four targeted hazardous solvents in this study, and are recommended to be potential replacements based on our workflow.

A similar study was carried out by McGonagle *et al*. in order to find alternatives to DCM, 1,2-DCE, tetrahydrofuran (THF), and *N,N*-dimethylformamide (DMF) in a direct aldehyde reductive amination reaction.[38] Authors reported a comparable reaction yield (78-82%) in *t*-butyl methyl ether (TBME), cyclopentylmethyl ether (CPME), dimethyl carbonate (DMC), (EtOAc), and 2-MeTHF. Looking at the alternatives for THF in our work, CPME, TBME, and 2-MeTHF are all recommended to be similar, whilst achieving a relative better SHE scores. DMC is selected as an alternative to DMF, which in return, suggests EtOAc as an alternative. The same group reported a similar conclusion of suggesting DMC, EtOAc, and 2-MeTHF as potential replacements to DCM and DMF for various types of amide coupling reactions.[77] Considering both EtOAc and 2-MeTHF are recommended to be similar to DMC in our study, this demonstrates a significant potential to explore alternative solvents with better SHE scores without sacrificing on the reaction performance. An additional example to using benzyl alcohol and anisole as greener alternative[78] to benzene and toluene, dimethyl isosorbide, which is a

biobased green solvent,[79] is selected as an alternative to carcinogenic hexamethylphosphoramide (HMPU) and 1,2-dimethyltetrahydropyrimidin-2(1h)-one (DMPU) in our workflow. A summary of list of alternatives to hazardous and highly hazardous solvents are given in Table 24 in the Appendix.

Conclusions

In conclusion, we have developed a workflow to explore alternative solvents from a library of commonly used 120 solvents and a larger library of 457 solvents. The workflow includes various parameterisation techniques (e.g., sigma moments, screening charge densities, fingerprints, larger list of solvent descriptors), similarity metrics (e.g., Cosine, Jaccard, Kendall's tau), and data pre-treatment methods (e.g., standard scaling, principal component analysis). Based on their influence on solvent selection, a list of alternative solvents was provided for each solvent in the 120 and 457 solvents library. For the commonly used solvents, closest ten solvents were selected using five sigma moments and three principal components of sigma moments based on similarity metrics Euclidean distance, Cosine similarity, Correlation coefficient, and Tanimoto index. Similar list was provided for the solvents in the library of 457 solvents using a longer list of information rich descriptors and five screening charge densities.

Applicability of this workflow was compared with solvent selection tools developed by AstraZeneca and Syngenta. Despite using relatively small number of computationally generated descriptors, a significant overlap was observed with list of alternative solvents to diethyl ether selected by the internal tool at Syngenta. Analysed on three case studies, using closest ten solvents selected from the solvent libraries, a better ranking (i.e., pattern identification) of solvents was demonstrated compared to the case studies by AstraZeneca. Moreover, the library of 457 solvents used in this work allows for selection of wider range of solvents compared to 272 and 209 solvents reported by AstraZeneca and Syngenta, respectively. Moreover, based on the experimental data on dozens of solvents in various transformations, suggested solvent alternatives in our workflow demonstrated a similar process performance to target solvents. Finally, selection of greener solvent alternatives to solvents with major regulatory issues was demonstrated.

Whilst most of solvent selection guides focus on the SHE impacts of solvents, it is important to highlight that solvent alternatives should be selected based on the process performance first, as demonstrated in our approach, and solvents with better SHE scores could then be selected from the wider list of suggested alternatives without sacrificing on the process performance. It is also important to reiterate that this approach is unlikely to generalise to all transformation types, since roles of solvent differ from one reaction to another, and requires different solvent parameters to optimise for reaction objectives. It is unlikely that a general set of descriptors can capture the complexity of chemical reactions. Therefore, our workflow serves a starting point and a guide to select initial list of solvents, based combination of various solvent descriptors, to design new reactions and suggest non-obvious solvents, and serves as a tool to be combined with domain knowledge and expertise.

References

1. Buncel, E.; Stairs, R. A.; Wilson, H., *The role of the solvent in chemical reactions*. Oxford University Press: Oxford ; New York, 2003; p ix, 159 p.
2. Constable, D. J. C.; Jimenez-Gonzalez, C.; Henderson, R. K., Perspective on Solvent Use in the Pharmaceutical Industry. *Organic Process Research & Development* **2006,** *11* (1), 133-137.
3. Jimenez-Gonzalez, C.; Ponder, C. S.; Broxterman, Q. B.; Manley, J. B., Using the Right Green Yardstick: Why Process Mass Intensity Is Used in the Pharmaceutical Industry To Drive More Sustainable Processes. *Organic Process Research & Development* **2011,** *15* (4), 912-917.
4. Diorazio, L. J.; Hose, D. R. J.; Adlington, N. K., Toward a More Holistic Framework for Solvent Selection. *Organic Process Research & Development* **2016,** *20* (4), 760-773.
5. Jiménez-González, C.; Poechlauer, P.; Broxterman, Q. B.; Yang, B.-S.; am Ende, D.; Baird, J.; Bertsch, C.; Hannah, R. E.; Dell'Orco, P.; Noorman, H.; Yee, S.; Reintjens, R.; Wells, A.; Massonneau, V.; Manley, J., Key Green Engineering Research Areas for Sustainable Manufacturing: A Perspective from Pharmaceutical and Fine Chemicals Manufacturers. *Organic Process Research & Development* **2011,** *15* (4), 900-911.
6. ECHA, REACH Web page. https://echa.europa.eu/regulations/reach. (accessed September 26, **2022**).
7. Alder, C. M.; Hayler, J. D.; Henderson, R. K.; Redman, A. M.; Shukla, L.; Shuster, L. E.; Sneddon, H. F., Updating and further expanding GSK's solvent sustainability guide. *Green Chemistry* **2016,** *18* (13), 3879-3890.
8. Curzons, A. D.; Mortimer, D. N.; Constable, D. J. C.; Cunningham, V. L., So you think your process is green, how do you know? — Using principles of sustainability to determine what is green – a corporate perspective. *Green Chemistry* **2001,** *3* (1), 1-6.
9. Henderson, R. K.; Jiménez-González, C.; Constable, D. J. C.; Alston, S. R.; Inglis, G. G. A.; Fisher, G.; Sherwood, J.; Binks, S. P.; Curzons, A. D., Expanding GSK's solvent selection guide – embedding sustainability into solvent selection starting at medicinal chemistry. *Green Chemistry* **2011,** *13* (4).
10. Alfonsi, K.; Colberg, J.; Dunn, P. J.; Fevig, T.; Jennings, S.; Johnson, T. A.; Kleine, H. P.; Knight, C.; Nagy, M. A.; Perry, D. A.; Stefaniak, M., Green chemistry tools to influence a medicinal chemistry and research chemistry based organisation. *Green Chem.* **2008,** *10* (1), 31-36.
11. Prat, D.; Pardigon, O.; Flemming, H.-W.; Letestu, S.; Ducandas, V.; Isnard, P.; Guntrum, E.; Senac, T.; Ruisseau, S.; Cruciani, P.; Hosek, P., Sanofi's Solvent Selection Guide: A Step Toward More Sustainable Processes. *Organic Process Research & Development* **2013,** *17* (12), 1517-1525.
12. Piccione, P. M.; Baumeister, J.; Salvesen, T.; Grosjean, C.; Flores, Y.; Groelly, E.; Murudi, V.; Shyadligeri, A.; Lobanova, O.; Lothschütz, C., Solvent Selection Methods and Tool. *Organic Process Research & Development* **2019,** *23* (5), 998-1016.
13. Reichardt, C.; Welton, T., *Solvents and solvent effects in organic chemistry*. 4th, updated and enl. ed.; Wiley-VCH: Weinheim, Germany, 2011; p xxvi, 692 p.

14. Gani, R.; Jiménez-González, C.; Constable, D. J. C., Method for selection of solvents for promotion of organic reactions. *Computers & Chemical Engineering* **2005,** *29* (7), 1661-1676.
15. Sheldon, T. J.; Folić, M.; Adjiman, C. S., Solvent Design Using a Quantum Mechanical Continuum Solvation Model. *Industrial & Engineering Chemistry Research* **2006,** *45* (3), 1128-1140.
16. Chauvin, Y.; Mussmann, L.; Olivier, H., A Novel Class of Versatile Solvents for Two-Phase Catalysis: Hydrogenation, Isomerization, and Hydroformylation of Alkenes Catalyzed by Rhodium Complexes in Liquid 1,3-Dialkylimidazolium Salts. *Angewandte Chemie International Edition in English* **1996,** *34* (2324), 2698-2700.
17. Burgess, S. A.; Appel, A. M.; Linehan, J. C.; Wiedner, E. S., Changing the Mechanism for CO2 Hydrogenation Using Solvent-Dependent Thermodynamics. *Angewandte Chemie International Edition* **2017,** *56* (47), 15002-15005.
18. Fainberg, A. H.; Winstein, S., Correlation of Solvolysis Rates. III.1 t-Butyl Chloride in a Wide Range of Solvent Mixtures2. *Journal of the American Chemical Society* **1956,** *78* (12), 2770-2777.
19. Winstein, S.; Fainberg, A. H., Correlation of Solvolysis Rates. IV.1 Solvent Effects on Enthalpy and Entropy of Activation for Solvolysis of t-Butyl Chloride2. *Journal of the American Chemical Society* **1956,** *79* (22), 5937-5950.
20. Campbell, D. S.; Hogg, D. R., Electrophilic additions to alkenes. Part IV. Kinetics of the reaction of 2,4-dinitrobenzenesulphenyl bromide with cyclohexene in benzene and in chloroform solution. *Journal of the Chemical Society B: Physical Organic* **1967**.
21. Cox, B. G.; Parker, A. J., Solvation of ions. XVIII. Protic-dipolar aprotic solvent effects on the free energies, enthalpies, and entropies of activation of an SNAr reaction. *Journal of the American Chemical Society* **2002,** *95* (2), 408-410.
22. Folić, M.; Adjiman, C. S.; Pistikopoulos, E. N., Design of solvents for optimal reaction rate constants. *AIChE Journal* **2007,** *53* (5), 1240-1256.
23. Wigh, D. S.; Goodman, J. M.; Lapkin, A. A., A review of molecular representation in the age of machine learning. *WIREs Computational Molecular Science* **2022,** *12* (5).
24. Pattanaik, L.; Coley, C. W., Molecular Representation: Going Long on Fingerprints. *Chem* **2020,** *6* (6), 1204-1207.
25. David, L.; Thakkar, A.; Mercado, R.; Engkvist, O., Molecular representations in AI-driven drug discovery: a review and practical guide. *Journal of Cheminformatics* **2020,** *12* (1).
26. Pomberger, A.; Pedrina McCarthy, A. A.; Khan, A.; Sung, S.; Taylor, C. J.; Gaunt, M. J.; Colwell, L.; Walz, D.; Lapkin, A. A., The effect of chemical representation on active machine learning towards closed-loop optimization. *Reaction Chemistry & Engineering* **2022,** *7* (6), 1368-1379.
27. Bajusz, D.; Rácz, A.; Héberger, K., Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics* **2015,** *7* (1).
28. Bender, A.; Jenkins, J. L.; Scheiber, J.; Sukuru, S. C. K.; Glick, M.; Davies, J. W., How Similar Are Similarity Searching Methods? A Principal Component Analysis of Molecular Descriptor Space. *Journal of Chemical Information and Modeling* **2009,** *49* (1), 108-119.

29. Holliday, J. D.; Hu, C.-Y.; Willet, W., Grouping of Coefficients for the Calculation of Inter-Molecular Similarity and Dissimilarity using 2D Fragment Bit-Strings. *Combinatorial Chemistry & High Throughput Screening* **2002,** *5* (2).
30. Todeschini, R.; Consonni, V.; Xiang, H.; Holliday, J.; Buscema, M.; Willett, P., Similarity Coefficients for Binary Chemoinformatics Data: Overview and Extended Comparison Using Simulated and Real Data Sets. *Journal of Chemical Information and Modeling* **2012,** *52* (11), 2884-2901.
31. Fulmer, G. R.; Miller, A. J. M.; Sherden, N. H.; Gottlieb, H. E.; Nudelman, A.; Stoltz, B. M.; Bercaw, J. E.; Goldberg, K. I., NMR Chemical Shifts of Trace Impurities: Common Laboratory Solvents, Organics, and Gases in Deuterated Solvents Relevant to the Organometallic Chemist. *Organometallics* **2010,** *29* (9), 2176-2179.
32. Prat, D.; Hayler, J.; Wells, A., A survey of solvent selection guides. *Green Chem.* **2014,** *16* (10), 4546-4551.
33. Prat, D.; Wells, A.; Hayler, J.; Sneddon, H.; McElroy, C. R.; Abou-Shehada, S.; Dunn, P. J., CHEM21 selection guide of classical- and less classical-solvents. *Green Chemistry* **2016,** *18* (1), 288-296.
34. ACS Green Chemistry Institute Pharmaceutical Roundtable (GCI-PR), https://www.acsgcipr.org. (accessed September 26, **2022**).
35. Pearson, K., LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **1901,** *2* (11), 559-572.
36. Bohle, M.; Kollecker, W.; Martin, D., Application of Factorial Analysis in Organic-Chemistry. *Z Chem* **1977,** *17* (5), 161-168.
37. Parker, A. J., The effects of solvation on the properties of anions in dipolar aprotic solvents. *Quarterly Reviews, Chemical Society* **1962,** *16* (2).
38. McGonagle, F. I.; MacMillan, D. S.; Murray, J.; Sneddon, H. F.; Jamieson, C.; Watson, A. J. B., Development of a solvent selection guide for aldehyde-based direct reductive amination processes. *Green Chemistry* **2013,** *15* (5).
39. Reizman, B. J.; Jensen, K. F., Simultaneous solvent screening and reaction optimization in microliter slugs. *Chemical Communications* **2015,** *51* (68), 13290-13293.
40. Reizman, B. J.; Jensen, K. F., Feedback in Flow for Accelerated Reaction Development. *Accounts of Chemical Research* **2016,** *49* (9), 1786-1796.
41. Strubing, H.; Karamertzanis, P. G.; Pistikopoulos, E. N.; Galindo, A.; Adjiman, C. S., Solvent design for a Menschutkin reaction by using CAMD and DFT calculations. *Comput-Aided Chem En* **2010,** *28*, 1291-1296.
42. Grant, E.; Pan, Y.; Richardson, J.; Martinelli, J. R.; Armstrong, A.; Galindo, A.; Adjiman, C. S., Multi-Objective Computer-Aided Solvent Design for Selectivity and Rate in Reactions. In *13th International Symposium on Process Systems Engineering (PSE 2018)*, 2018; pp 2437-2442.
43. Amar, Y.; Schweidtmann, Artur M.; Deutsch, P.; Cao, L.; Lapkin, A., Machine learning and molecular descriptors enable rational solvent selection in asymmetric catalysis. *Chemical Science* **2019,** *10* (27), 6697-6706.

44. Zhang, C.; Amar, Y.; Cao, L.; Lapkin, A. A., Solvent Selection for Mitsunobu Reaction Driven by an Active Learning Surrogate Model. *Organic Process Research & Development* **2020,** *24* (12), 2864-2873.
45. Kamlet, M. J.; Abboud, J. L. M.; Abraham, M. H.; Taft, R. W., Linear solvation energy relationships. 23. A comprehensive collection of the solvatochromic parameters, .pi.*, .alpha., and .beta., and some methods for simplifying the generalized solvatochromic equation. *The Journal of Organic Chemistry* **2002,** *48* (17), 2877-2887.
46. Abraham, M. H., Scales of solute hydrogen-bonding: their construction and application to physicochemical and biochemical processes. *Chemical Society Reviews* **1993,** *22* (2).
47. Van Paesschen, W.; Hirsch, E.; Johnson, M.; Falter, U.; von Rosenstiel, P., Efficacy and tolerability of adjunctive brivaracetam in adults with uncontrolled partial-onset seizures: A phase IIb, randomized, controlled trial. *Epilepsia* **2013,** *54* (1), 89-97.
48. Karelson, M.; Maran, U.; Wang, Y.; Katritzky, A. R., QSPR and QSAR Models Derived Using Large Molecular Descriptor Spaces. A Review of CODESSA Applications. *Collection of Czechoslovak Chemical Communications* **1999,** *64* (10), 1551-1571.
49. Karelson, M.; Lobanov, V. S.; Katritzky, A. R., Quantum-Chemical Descriptors in QSAR/QSPR Studies. *Chemical Reviews* **1996,** *96* (3), 1027-1044.
50. Kamlet, M. J.; Abboud, J. L. M.; Abraham, M. H.; Taft, R. W., Linear solvation energy relationships. 23. A comprehensive collection of the solvatochromic parameters, .pi.*, .alpha., and .beta., and some methods for simplifying the generalized solvatochromic equation. *The Journal of Organic Chemistry* **1983,** *48* (17), 2877-2887.
51. Abraham, M. H.; Doherty, R. M.; Kamlet, M. J.; Harris, J. M.; Taft, R. W., Linear Solvation Energy Relationships .37. An Analysis of Contributions of Dipolarity Polarizability, Nucleophilic Assistance, Electrophilic Assistance, and Cavity Terms to Solvent Effects on Tert-Butyl Halide Solvolysis Rates. *J Chem Soc Perk T 2* **1987,** (7), 913-920.
52. Abraham, M. H.; Doherty, R. M.; Kamlet, M. J.; Harris, J. M.; Taft, R. W., Linear Solvation Energy Relationships .38. An Analysis of the Use of Solvent Parameters in the Correlation of Rate Constants, with Special Reference to the Solvolysis of Tert-Butyl Chloride. *J Chem Soc Perk T 2* **1987,** (8), 1097-1101.
53. Folić, M.; Gani, R.; Jiménez-González, C.; Constable, D. J. C., Systematic Selection of Green Solvents for Organic Reacting Systems. *Chinese Journal of Chemical Engineering* **2008,** *16* (3), 376-383.
54. Zissimos, A. M.; Abraham, M. H.; Klamt, A.; Eckert, F.; Wood, J., A Comparison between the Two General Sets of Linear Free Energy Descriptors of Abraham and Klamt. *Journal of Chemical Information and Computer Sciences* **2002,** *42* (6), 1320-1331.
55. COSMO*therm*, Version C3.0, Release 17.01; COSMOlogic GmbH & Co. KG.
56. Klamt, A., Conductor-like Screening Model for Real Solvents: A New Approach to the Quantitative Calculation of Solvation Phenomena. *The Journal of Physical Chemistry* **1995,** *99* (7), 2224-2235.
57. Klamt, A.; Eckert, F.; Hornig, M., COSMO-RS: A novel view to physiological solvation and partition questions. *Journal of Computer-Aided Molecular Design* **2001,** *15* (4), 355-365.

58. Klamt, A.; Schüürmann, G., COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J. Chem. Soc., Perkin Trans. 2* **1993,** (5), 799-805.
59. Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J., Molecular Similarity in Medicinal Chemistry. *Journal of Medicinal Chemistry* **2013,** *57* (8), 3186-3204.
60. Reisen, F.; Zhang, X.; Gabriel, D.; Selzer, P., Benchmarking of Multivariate Similarity Measures for High-Content Screening Fingerprints in Phenotypic Drug Discovery. *SLAS Discovery* **2013,** *18* (10), 1284-1297.
61. Chen, X.; Reynolds, C. H., Performance of Similarity Measures in 2D Fragment-Based Similarity Searching: Comparison of Structural Descriptors and Similarity Coefficients. *Journal of Chemical Information and Computer Sciences* **2002,** *42* (6), 1407-1414.
62. Willett, P., Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006,** *11* (23-24), 1046-1053.
63. Dixon, S. L.; Koehler, R. T., The Hidden Component of Size in Two-Dimensional Fragment Descriptors: Side Effects on Sampling in Bioactive Libraries. *Journal of Medicinal Chemistry* **1999,** *42* (15), 2887-2900.
64. Salim, N.; Holliday, J.; Willett, P., Combination of Fingerprint-Based Similarity Coefficients Using Data Fusion. *Journal of Chemical Information and Computer Sciences* **2002,** *43* (2), 435-442.
65. Amar, Y. Accelerating process development of complex chemical reactions. University of Cambridge, 2019.
66. Landrum, G., RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum* **2013**.
67. Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat, İ.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P.; Vijaykumar, A.; Bardelli, A. P.; Rothberg, A.; Hilboll, A.; Kloeckner, A.; Scopatz, A.; Lee, A.; Rokem, A.; Woods, C. N.; Fulton, C.; Masson, C.; Häggström, C.; Fitzgerald, C.; Nicholson, D. A.; Hagen, D. R.; Pasechnik, D. V.; Olivetti, E.; Martin, E.; Wieser, E.; Silva, F.; Lenders, F.; Wilhelm, F.; Young, G.; Price, G. A.; Ingold, G.-L.; Allen, G. E.; Lee, G. R.; Audren, H.; Probst, I.; Dietrich, J. P.; Silterra, J.; Webber, J. T.; Slavič, J.; Nothman, J.; Buchner, J.; Kulick, J.; Schönberger, J. L.; de Miranda Cardoso, J. V.; Reimer, J.; Harrington, J.; Rodríguez, J. L. C.; Nunez-Iglesias, J.; Kuczynski, J.; Tritz, K.; Thoma, M.; Newville, M.; Kümmerer, M.; Bolingbroke, M.; Tartre, M.; Pak, M.; Smith, N. J.; Nowaczyk, N.; Shebanov, N.; Pavlyk, O.; Brodtkorb, P. A.; Lee, P.; McGibbon, R. T.; Feldbauer, R.; Lewis, S.; Tygier, S.; Sievert, S.; Vigna, S.; Peterson, S.; More, S.; Pudlik, T.; Oshima, T.; Pingel, T. J.; Robitaille, T. P.; Spura, T.; Jones, T. R.; Cera, T.; Leslie, T.; Zito, T.; Krauss, T.; Upadhyay, U.; Halchenko, Y. O.; Vázquez-Baeza, Y., SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **2020,** *17* (3), 261-272.

68. Weininger, D., SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling* **1988,** *28* (1), 31-36.
69. Kendall, M. G., Rank Correlation Methods (4 ed.). **1970**.
70. Burdett, J. L.; Rogers, M. T., Keto-Enol Tautomerism in β-Dicarbonyls Studied by Nuclear Magnetic Resonance Spectroscopy.1 I. Proton Chemical Shifts and Equilibrium Constants of Pure Compounds. *Journal of the American Chemical Society* **1964,** *86* (11), 2105-2109.
71. Schmid, R., Re-interpretation of the solvent dielectric constant in coordination chemical terms. *Journal of Solution Chemistry* **1983,** *12* (2), 135-152.
72. Menschutkin, N., Über die Affinitätskoeffizienten der Alkylhaloide und der Amine. *Zeitschrift für Physikalische Chemie* **1890,** *6U* (1), 41-57.
73. Ganase, Z. An experimental study on the effects of solvents on the rate and selectivity of organic reactions. Imperial College London, 2015.
74. Abraham, M. H.; Grellier, P. L., Substitution at saturated carbon. Part XX. The effect of 39 solvents on the free energy of Et3N, EtI, and the Et3N–EtI transition state. Comparison with solvent effects on the equilibria Et3N + EtI ⇌ Et4N+I–and Et3N + EtI ⇌ Et4N++ I–. *J. Chem. Soc., Perkin Trans. 2* **1976,** (14), 1735-1741.
75. Zhou, T.; Qi, Z.; Sundmacher, K., Model-based method for the screening of solvents for chemical reactions. *Chemical Engineering Science* **2014,** *115*, 177-185.
76. Skowerski, K.; Białecki, J.; Tracz, A.; Olszewski, T. K., An attempt to provide an environmentally friendly solvent selection guide for olefin metathesis. *Green Chem.* **2014,** *16* (3), 1125-1130.
77. MacMillan, D. S.; Murray, J.; Sneddon, H. F.; Jamieson, C.; Watson, A. J. B., Evaluation of alternative solvents in common amide coupling reactions: replacement of dichloromethane and N,N-dimethylformamide. *Green Chemistry* **2013,** *15* (3).
78. Delolo, F. G.; dos Santos, E. N.; Gusevskaya, E. V., Anisole: a further step to sustainable hydroformylation. *Green Chemistry* **2019,** *21* (5), 1091-1098.
79. Russo, F.; Galiano, F.; Pedace, F.; Aricò, F.; Figoli, A., Dimethyl Isosorbide As a Green Solvent for Sustainable Ultrafiltration and Microfiltration Membrane Preparation. *ACS Sustainable Chemistry & Engineering* **2019,** *8* (1), 659-668.

Appendix

Table 22. *DCM alternatives when expanded to 457 library when all ten similarity metrics were used, identifying alternatives that are not in the commonly used practical solvents. Numbers represent the confidence level. Numbers next to the solvent name indicates how many times they are selected to be similar (i.e., how many metrics).*

17 descriptors	10 PCs from 17 descriptors	Sigma profiles 1 to 5
dichloromethane, 10, *cis*-1,2-dichloroethene, 9 1,2-**dichloroethane**, 8 1,1-**dichloroethane**, 8 *trans*-1,2-dichloroethene, 6, 1,1-dichloroethene, 4 1,2-dichloropropane, 3, acetyl chloride, 3 1,1,2-**trichloroethane**, 2 **trichloromethane**, 2 1-bromo-2-chloroethane, 2 nitric acid, 2 **iodomethane**, 1 1,2,3-trichloropropane, 1 3-chloropropene, 1 1,3-dichloropropane, 1 1,4-dichlorobutane, 1 2,2,3-trimethylbutane, 1	dichloromethane, 10 1,1-dichloroethane, 8 1,2-dichloroethane, 7 trans-1,2-dichloroethene, 6 cis-1,2-dichloroethene, 5 acetyl chloride, 4 1,1-dichloroethene, 3 1,2-dichloropropane, 3 3-bromopropene, 3 ethanethiol, 2 trichloroethene, 2 1-chloropropane, 2 bromoethane, 1 trichloromethane, 1 3-chloropropene, 1 1,3-dichloropropane, 1 1,4-dichlorobutane, 1 (trichloromethyl)benzene, 1 hexylamine, 1 hexyl acetate, 1 hexanoic acid, 1 1-bromo-2-chloroethane, 1	dichloromethane, 10 dibromomethane, 8 cis-1,2-dichloroethene, 8 trans-1,2-dichloroethene, 6 diiodomethane, 5 phenylhydrazine, 3 aniline, 2 4-methoxyaniline, 2 acetyl chloride, 2 1,2-dichloroethane, 2 1,1,2-trichloroethane, 2 (trichloromethyl)benzene, 1 2,4-dimethylpentane, 1 propane, 1 phenyl isothiocyanate, 1 2,4-dimethylaniline, 1 ethyl chloroacetate, 1 4-methylaniline, 1 1,1-dichloroethene, 1 2-methylaniline, 1 1-nitropropane, 1 nitroethane, 1 1,2-dibromoethane, 1 1,3-dichloropropane, 1 1,4-dichlorobutane, 1 1-bromo-2-chloroethane, 1

Table 23. *DCM alternatives when expanded to 457 library when selected four similarity metrics were used, identifying alternatives that are not in the commonly used practical solvents.*

17 descriptors	10 PCs from 17 descriptors	Sigma profiles 1 to 5
dichloromethane, 4, trans-1,2-dichloroethene, 3, cis-1,2-dichloroethene, 3, 1,2-dichloroethane, 4, 1,1-dichloroethane, 3, 1,1-dichloroethene, 3, acetyl chloride, 3, trichloromethane, 3, 1,1,2-trichloroethane, 4, 2-propen-1-amine, 1, 1,2-dichloropropane, 1, 1,2,3-trichloropropane, 1, 1,3-dichloropropane, 1, 1,4-dichlorobutane, 1, 1-bromo-2-chloroethane, 1, 1-chloropropane, 1, 1,5-dichloropentane, 1, 1-chloro-2-methylpropane, 1, 3-chloropropene, 1	dichloromethane, 4, trans-1,2-dichloroethene, 3, cis-1,2-dichloroethene, 3, 1,1-dichloroethane, 3, 1,2-dichloroethane, 4, 1,1-dichloroethene, 3, acetyl chloride, 3, trichloromethane, 3, 3-bromopropene, 3, ethyl nitrate, 1, 1,2-dichloropropane, 2, 1,3-dichloropropane, 1, 1,4-dichlorobutane, 1, 1-bromo-2-chloroethane, 1, 1,1,2-trichloroethane, 1, 1-chloropropane, 1, 1,5-dichloropentane, 1, 1-chloro-2-methylpropane, 1, 3-chloropropene, 1	dichloromethane, 4, dibromomethane, 3, cis-1,2-dichloroethene, 3, trans-1,2-dichloroethene, 3, diiodomethane, 3, 1,1,2-trichloroethane, 4, 1,2-dichloroethane, 3, 1,2-dibromoethane, 2, 1-bromo-2-chloroethane, 3, nitromethane, 1, 1,1-dichloroethane, 1, phenylhydrazine, 1, 4-methylaniline, 1, 1,1-dichloroethene, 1, 2-methylaniline, 1, 1,3-dichloropropane, 1, 1,4-dichlorobutane, 1, 1-chloropropane, 1, 1,5-dichloropentane, 1, 1-chloro-2-methylpropane, 1, 3-chloropropene, 1

Alternatives to highly hazardous solvents

***Table 24**. List of alternative solvents for hazardous and highly hazardous solvents categorised by Prat et al., selected from the library of commonly used 120 solvents and the library of 457 solvents based on three different parameterisation techniques. Numbers next to the solvent implies how many times the solvent was selected to be similar out of four similarity metrics.*

Molecule	Five sigma moments (120 library)	17 descriptors (457 library)	Sigma profile 1-5 (457 library)
Diisopropyl ether	2-ethoxy-2-methyl-propane, 3, methyl-tert-amylether, 3, tetrahydro-2,2,5,5-tetramethylfuran, 3, triethylamine, 3, cyclopentyl-methyl-ether, 3, methyl-t-butylether, 3, diethylether, 3, 2-MeTHF, 1, "2,2-oxybis-butane" n-ethyldiisopropylamine, 1, 2,4,6-collidine, 1, 2-propanol, 1, 1-methoxy2-propanol, 1, isobutanol, 1, dimethylcarbonate, 1, 2-butanol, 1, propyleneglycol, 1, aceticacid-2-methylpropylester, 1, lacticacid, 1	diisopropylamine, 3, dipropyl ether, 3, 2-methoxy-2-methylpropane, 3, 1,1-diethoxyethane, 3, triethylamine, 3, 1,1-diethoxymethane, 1, tetramethylsilane, 3, 1-hexene, 2, Z-3-hexene, 2, 2,2-dimethylbutane, 2, hexane, 1, dibutyl ether, 1, 2,2-dimethylpentane, 1, 2-propanol, 1, 2-bromopropane, 1, 1-iodopropane, 2, 1,1-dichloroethane, 1, isopropylamine, 1, 2,2,3-trimethylbutane, 1, methylbutane, 1	dipropyl ether, 3, 1Z,5Z-cycloocta-1,5-diene, 3, diisopropylamine, 3, diisopropyl sulfide, 3, N-methylpiperidine, 3, dipropylamine, 2, 2-methoxy-2-methylpropane, 1, N,N-dimethylcyclohexylamine, 3, N-methylcyclohexylamine, 1, 1,3,3-trimethyl-2-oxabicyclo2.2.2octane, 2, 1-pentanethiol, 1, dibutyl ether, 1, dipentyl ether, 1, 2-propanol, 1, 2-bromopropane, 1, 1-iodopropane, 2, 1,1-dichloroethane, 1, isopropylamine, 1, 1,1-diethoxyethane, 1, 2,2,3-trimethylbutane, 1, methylbutane, 1
Dioxane	cyclohexanone, 2, cyclopentanone, 3, 1,2-dimethoxyethane, 3, propanone, 3, butanone, 3, n-methylmorpholine, 3, pyridine, 2, ethylacetate, 1, dihydro-5-methyl-23h-furanone, 1, dimethylformamide, 2, n,n-dimethylacetamide, 2, PEG, 1, n-methyl-2-pyrrolidinone, 1, dimethylsulfoxide, 1, THF, 1, 1,3-dioxolan-2-one, 1, 2-MeTHF, 1, tetrahydrofurfurylalcohol, 1, tetrahydro-2,2,5,5-tetramethylfuran, 1, cyclohexane, 1, dimethylisosorbide, 1	dimethyl carbonate, 3, pyrazine, 3, morpholine 1, 1,2-dimethoxyethane, 3, methyl 2-propenoate, 3, diethyl carbonate, 3, methyl propanoate, 2, ethyl acetate, 1, ethyl propanoate, 1, methyl acetate, 1, dimethoxymethane, 1, 2,4,6-trimethyl-1,3,5-trioxane, 1, 2,4-pentanedione, 1, ethyl 2-hydroxypropanoate, 1, oxirane, 1, THF, 1, oxane, 1, 1,3-dioxolan-2-one, 1, oxolan-2-one, 1, 2-MeTHF, 1, cyclohexane, 1, cyclopentane, 1	N,N-dimethylformamide, 3, N,N-dimethylacetamide, 3, ethyl acetate, 3, methyl acetate, 3, methyl propanoate, 3, dimethoxymethane, 3, 2-propanone, 3, N-methylpyrrolidon, 2, 2-butanone, 3, butanal, 1, oxirane, 1, THF, 1, oxane, 1, morpholine, 1, 1,3-dioxolan-2-one, 1, oxolan-2-one, 1, 2-MeTHF, 1, cyclohexane, 1, cyclopentane, 1

Hexane	methylcyclohexane, 3, n-heptane, 4, pentane, 4, 2,2,4-trimethylpentane, 3, cyclohexane, 3, toluene, 3, trifluoromethylbenzene, 2, isopropylbenzene, 2, 1-chlorobutane, 4, n-ethyldiisopropylamine, 1, dipentene, 1, 1-pentanol, 1, 1-octanol, 1, 1-heptanol, 1, 1-butanol, 1, n-butylacetate, 1, n-pentylacetate, 1	2,2-dimethylbutane, 3, 3-ethylpentane, 3, 2,4-dimethylpentane, 3, 3-methylhexane, 3, 2,2-dimethylpentane, 3, 2-methylhexane, 3, heptane, 4, tetramethylsilane, 2, 3,3-dimethylpentane, 2, 2-methylpentane, 1, 2,3-dimethylpentane, 1, undecane, 1, nonane, 1, octane, 1, dodecane, 1, hexadecane, 1, decane, 1, pentane, 1, 1-pentanethiol, 1	2-methylpentane, 3, 2,2,3-trimethylbutane, 3, 3,3-dimethylpentane, 3, cyclooctatetraene, 3, methylcyclohexane, 3, 3-ethylpentane, 3, 2,3-dimethylpentane, 3, 2,2-dimethylpentane, 2, 2,2-dimethylbutane, 1, triethylamine, 1, 1-dodecanol, 1, undecanol, 1, undecane, 1, nonane, 1, octane, 1, dodecane, 1, hexadecane, 1, heptane, 1, decane, 1, pentane, 1, 1-pentanethiol, 1
N-methylpyrrolidone	n,n-dimethylacetamide, 3, 1,3-dimethyltetrahydropyrimidin-21h-one 1,1,3,3-tetramethylguanidine, 3, 1,3-dimethyl-2-imidazolidinone, 4, dimethylformamide, 3, dimethylsulfoxide, 2, p-dimethylaminopyridine, 2, triethylenediamine, 1, PEG, 2, 2,3,4,6,7,8,9,10-octahydropyrimido1,2-aazepine, 2, dioxane, 1, butanone, 1, cyclopentanone, 2, cyclohexanone, 2, dihydro-5-methyl-23h-furanone, 1, n-methylmorpholine, 1, 1,3-dioxolan-2-one, 1, propylenecarbonate, 1	N,N-dimethylacetamide, 3, N,N,N,N-tetramethylurea, 3, 1,3-dimethyltetrahydropyrimidin-21H-one, 4, N,N-dimethylformamide, 3, oxolan-2-one, 4, cyclohexanone, 4, dimethylcyanamide, 3, cyclopentanone, 3, ethyl thiocyanate, 1, dimethyl sulfoxide, 1, sulfolane, 2, N-methylpiperidine, 1, 1,3-dioxolan-2-one, 1, 4-methyl-2-oxo-1,3-dioxolane, 1, 1,2,3-tribromopropane, 1, tetrahydronaphthalene, 1	N,N-dimethylacetamide, 3, ethyl acetate, 2, propyl acetate, 3, methyl propanoate, 2, ethyl propanoate, 3, 1,2-dimethoxyethane, 2, methyl butanoate, 3, cyclopentanone, 3, dioxane, 2, 2-pentanone, 2, pentanal, 1, 3-pentanone, 1, butanal, 1, 1,3-dimethyltetrahydropyrimidin-21H-one, 1, cyclohexanone, 1, N-methylpiperidine, 1, oxolan-2-one, 1, 1,3-dioxolan-2-one, 1, 4-methyl-2-oxo-1,3-dioxolane, 1, 1,2,3-tribromopropane, 1, tetrahydronaphthalene, 1
N,N-dimethylacetamide	n-methyl-2-pyrrolidinone, 3, dimethylformamide, 4, 1,3-dimethyl-2-imidazolidinone, 3, 1,3-dimethyltetrahydropyrimidin-21h-one, 3, 1,1,3,3-tetramethylguanidine, 3, PEG, 2, p-dimethylaminopyridine, 2, dimethylsulfoxide, 3, triethylenediamine, 2, dioxane, 1, butanone, 2, cyclopentanone, 1, 1,2-dimethoxyethane, 1, propanone, 1, aceticanhydride, 1, methylacetate, 1, 1,2-ethanedioldiacetate, 1, ethylacetate, 1, 4-methyl-2-pentanone, 1	N-methylpyrrolidon, 3, N,N-dimethylformamide, 3, dimethylcyanamide, 3, oxolan-2-one, 3, N,N,N,N-tetramethylurea, 4, cyclopentanone, 2, E-2-butenenitrile, 2, 2,4-pentanedione, 2, cyclohexanone, 2, dimethyl sulfoxide, 2, 4-methyl-2-oxo-1,3-dioxolane, 2, sulfolane, 1, 2-propanone, 1, acetic acid, 1, acetyl chloride, 1, 3-methyl-2-butanone, 1, 2-butanone, 1, N-methylacetamide, 1, acetic anhydride, 1	ethyl acetate, 3, methyl propanoate, 3, dioxane, 3, N-methylpyrrolidon, 3, N,N-dimethylformamide, 3, methyl acetate, 2, dimethoxymethane, 2, propyl acetate, 2, methyl 2-propenoate, 1, dimethylcyanamide, 1, methyl butanoate, 1, 2-butanone, 2, ethyl 2-hydroxypropanoate, 1, butanal, 1, N,N,N,N-tetramethylurea, 1, 2-propanone, 1, acetic acid, 1, acetyl chloride, 1, 3-methyl-2-butanone, 1, 2,4-pentanedione, 1, N-methylacetamide, 1, acetic anhydride, 1

N,N-dimethylformamide	n,n-dimethylacetamide, 4, n-methyl-2-pyrrolidinone, 3, 1,3-dimethyl-2-imidazolidinone, 3, PEG, 2, n-methylmorpholine, 1, dioxane, 3, p-dimethylaminopyridine, 2, 1,2-dimethoxyethane, 2, 1,3-dimethyltetrahydropyrimidin-21h-one, 2, dimethylsulfoxide, 3, 1,1,3,3-tetramethylguanidine, 2, propanone, 2, butanone, 1, cyclopentanone, 1, pyridine, 1, methylformate, 1, hexamethylphosphoramide, 1, formicacid, 1, ethylformate, 1	dimethylcyanamide, 3, N,N-dimethylacetamide, 4, oxolan-2-one, 3, E-2-butenenitrile, 2, N-methylpyrrolidon, 3, ethyl thiocyanate, 3, cyclopentanone, 3, propanenitrile, 3, butanenitrile, 1, dimethyl sulfoxide, 2, 4-methyl-2-oxo-1,3-dioxolane, 1, N-methylformamide, 1, methyl formate, 1, propanal, 1, hexamethylphosphoramide, 1, N,N,N,N-tetramethylurea, 1, formic acid, 1, formamide, 1, butanal, 1	methyl acetate, 3, dioxane, 3, 2-propanone, 3, N,N-dimethylacetamide, 3, ethyl acetate, 3, dimethylcyanamide, 3, methyl propanoate, 2, methyl 2-propenoate, 2, dimethoxymethane, 3, 2-butanone, 1, oxolan-2-one, 1, propanal, 2, N-methylformamide, 1, methyl formate, 1, hexamethylphosphoramide, 1, N,N,N,N-tetramethylurea, 1, formic acid, 1, formamide, 1, butanal, 1
1,2-Dimethoxyethane	cyclohexanone, 3, n-methylmorpholine, 3, dioxane, 2, PEG, 3, cyclopentanone, 1, tert-butylacetate, 1, 2,4,6-collidine, 1, 4-methyl-2-pentanone, 1, isopropylacetate, 1, triethylenediamine, 2, n,n-dimethylacetamide, 2, dimethylformamide, 1, n-methyl-2-pyrrolidinone, 2, p-dimethylaminopyridine, 2, 1,3-dimethyltetrahydropyrimidin-21h-one, 1, 1,1,3,3-tetramethylguanidine, 1, diglyme, 2, 2-methoxyethanol, 1, 1-methoxy2-propanol, 1, methyl-tert-amylether, 1, 1,2-ethanedioldiacetate, 1, methyl-t-butylether, 1, diethylether, 1, methylformate, 1	ethyl acetate, 3, methyl propanoate, 3, ethyl propanoate, 3, methyl butanoate, 3, propyl acetate, 3, propyl formate, 2, 2-pentanone, 3, 1,1-diethoxymethane, 1, 3-pentanone, 2, dimethoxymethane, 3, methyl acetate, 2, bis2-methoxyethyl ether, 1, 2-methoxyethanol, 1, trimethyl phosphate, 1, dipropyl ether, 1, 3,6,9-trioxaundecane, 1, dibutyl ether, 1, dimethyl propanedioate, 1, dipentyl ether, 1	cyclohexanone, 3, N-methylpyrrolidon, 3, 2-pentanone, 3, cyclopentanone, 3, 3,3-dimethyl-2-butanone, 3, 3-pentanone, 3, propyl acetate, 2, ethyl propanoate, 1, pentanal, 3, 4-methyl-2-pentanone, 1, N,N,N,N-tetramethylurea, 2, bis2-methoxyethyl ether, 1, 2-methoxyethanol, 1, dimethoxymethane, 1, trimethyl phosphate, 1, dipropyl ether, 1, 3,6,9-trioxaundecane, 1, dibutyl ether, 1, dimethyl propanedioate, 1, dipentyl ether, 1
Pentane	cyclohexane, 3, methylcyclohexane, 3, hexane toluene, 3, benzene, 3, n-heptane, 4, cs2, 3, 1-chlorobutane, 4, 2,2,4-trimethylpentane, 2, chlorobenzene, 1, 1-butanol, 1, 1-pentanol, 1, 1-heptanol, 1, 1-octanol, 1, 2-hexanone, 1, propanol, 1	methylbutane, 3, 2-methyl-1-butene, 3, 2,2-dimethylbutane, 3, 1-pentene, 2, 2-methylpentane, 3, Z-3-hexene, 3, methylcyclopentane, 2, 1-hexene, 3, tetramethylsilane, 2, butane, 2, hexane, 2, decane, 1, undecane, 1, heptane, 1, hexadecane, 1, dodecane, 1, octane, 1, nonane, 1, butanol, 1	methylcyclopentane, 3, methylbutane, 3, cyclohexane, 3, 2,2-dimethylpropane, 3, 2,2-dimethylbutane, 3, methylcyclohexane, 3, 2-methylpentane, 1, cyclohexene, 2, hexane, 2, Z-3-hexene, 1, 1-pentene, 1, decanol, 1, nonanol, 1, 2-methyl-1-butene, 1, decane, 1, undecane, 1, heptane, 1, hexadecane, 1, dodecane, 1, octane, 1, nonane, 1,

			butanol, 1
2-methoxyethanol	1-methoxy2-propanol, 4, tetrahydrofurfurylalcohol, 2, tert-butanol, 1, cyclohexanol, 1, 2-propanol, 2, 1-butanol, 2, 2-methyl-2-butanol, 1, 2-butanol, 1, 1,3-propanediol, 3, propyleneglycol, 2, glycol, 3, glycerol, 1, imidazole, 1, ethanol, 2, methanol, 1, 1,2-dimethoxyethane, 1, diglyme, 2, propanol, 1	2-ethoxyethanol, 4, 2-propanol, 3, 2-propen-1-ol, 3, butanol, 2, 2-butanol, 1, propyl formate, 1, ethyl formate, 2, propanol 2-methyl-1-propanol, 1, ethanol, 3, methyl formate, 1, 2-propanone, 1, 1,2-propanediol, 1, ethanediol, 2, diethylene glycol, 2, 1,2,3-propanetriol, 1, 1,2-dimethoxyethane, 1, bis2-methoxyethyl ether, 1, 2-butoxyethanol, 1, dimethoxymethane, 1	2-ethoxyethanol, 4, morpholine, 1, N-methylacetamide, 3, 2-methyl-2-propanol, 1, 2-propen-1-ol, 2, propyl formate, 1, ethyl formate, 1, 2-propanol, 1, 2-butanol, 1, N-methylformamide, 2, acetic acid, 2, ethanediol, 3, 1,2-propanediol, 2, diethylene glycol, 3, 1,4-butanediol, 1, 1,2,3-propanetriol, 1, ethanethiol, 1, 1,2-dimethoxyethane, 1, bis2-methoxyethyl ether, 1, 2-butoxyethanol, 1, propanol, 1, dimethoxymethane, 1, ethanol, 1
Triethylamine	methyl-tert-amylether, 3, cyclopentyl-methyl-ether, 3, 2-ethoxy-2-methyl-propane, 3, diisopropylether, 3, methyl-t-butylether, 3, diethylether, 4, tetrahydro-2,2,5,5- tetramethylfuran, 3, 2,2-oxybis-butane, 3, 2-MeTHF, 1, n-ethyldiisopropylamine, 3, ethanol, 1, pentane, 1, hexane, 1, propanol, 1, n-heptane, 1, propionicacid, 1, 2-methyl-2-butanol, 1	3-ethylpentane, 3, 3,3-dimethylpentane, 3, 1-hexene, 2, 2,3-dimethylpentane, 3, methylcyclohexane, 1, Z-3-hexene, 1, 3-methylhexane, 3, 2,2-dimethylpentane, 3, 2,4-dimethylpentane, 3, 2-methylhexane, 2, 2-methylpentane, 2, 2,2,4-trimethylpentane, 1, bromoethane, 1, propane, 1, tributylamine, 1, 1-bromopropane, 1, butane, 1, ethanethiol, 1, ethanol, 1, iodoethane, 1, 1-bromobutane, 1	tetramethylsilane, 3, 2,2,3-trimethylbutane, 2, 3,3-dimethylpentane, 2, 1-hexene, 3, 2,3-dimethylpentane, 2, 3-ethylpentane, 2, 2,2-dimethylpentane, 3, hexane, 2, Z-3-hexene, 1, cyclooctatetraene, 1, 2-methoxyaniline, 1, 3-methylhexane, 1, 2,4-dimethylpentane, 1, 2-methylhexane, 1, 2,2,4-trimethylpentane, 1, heptane, 1, bromoethane, 1, propane, 1, tributylamine, 1, 1-bromopropane, 1, butane, 1, ethanethiol, 1, ethanol, 1, iodoethane, 1, 1-bromobutane, 1
1,2-Dichloroethane	ch2cl2, 4, benzene, 2, chlorobenzene, 3, anisole, 3, toluene, 2, 1-chlorobutane, 3, nitromethane, 2, chcl3, 4, trifluoromethylbenzene, 3, methanesulfonicacid, 1, benzylalcohol, 1,	cis-1,2-dichloroethene, 3, 1,1-dichloroethane, 3, 1,1,2-trichloroethane, 3, 1-chloro-2,3-epoxypropane, 2, 1-bromo-2-chloroethane dichloromethane, 4, 1,2-dichloropropane, 1, trans-1,2-dichloroethene, 3, 1,4-difluorobenzene, 1, nitromethane, 2, acetyl chloride, 2,	1,2-dibromoethane, 3, 1-bromo-2-chloroethane, 4, diiodomethane, 3, dibromomethane, 2, acetyl chloride, 2, 1,4-difluorobenzene, 2, 1,2,3-trichloropropane, 3, dichloromethane, 3, 1,1-dichloroethane, 3, 1,1,2-trichloroethane, 1, phenylhydrazine, 1,

	trifluoromethanesulfonicacid, 1, trifluoroaceticacid, 1, glycol, 1, ethanol, 1, 1,3-propanediol, 2, pentane, 1, hexane, 1	1,2,3-trichloropropane, 1, 1,3-dichloropropane, 1, 1,4-dichlorobutane, 1, 1-chloropropane, 1, 1,5-dichloropentane, 1, bis2-chloroethyl ether, 1, 1-chlorobutane, 1, 1-chloro-3-methylbutane, 1	4-methoxyaniline, 1, 2-methylaniline, 1, 1,3-dichloropropane, 1, 1,4-dichlorobutane, 1, 1-chloropropane, 1, 1,5-dichloropentane, 1, bis2-chloroethyl ether, 1, 1-chlorobutane, 1, 1-chloro-3-methylbutane, 1
Trichloromethane	ch2cl2, 4, benzene, 3, chlorobenzene, 3, 1,2-dichloroethane, 4, toluene, 2, trifluoromethylbenzene, 2, 1-chlorobutane, 2, pentane, 2, cyclohexane, 1, cs2, 1, trifluoromethanesulfonicacid, 1, trifluoroaceticacid, 1, 1-heptanol, 1, benzylalcohol, 1, isopentanol, 1, 2-propanol, 1, diisopropylether, 1, glycerol, 2, isobutanol, 1, 2-butanol, 1, dimethylcarbonate, 1	trichloroethene, 3, sulfuryl chloride, 3, trans-1,2-dichloroethene, 3, 1,1-dichloroethene, 3, dichloromethane, 3, cis-1,2-dichloroethene, 2, 1,1,1-trichloroethane, 1, 1,1-dichloroethane, 2, 1-bromo-2-chloroethane, 1, nitric acid, 2, trifluoroacetic acid, 2, 1,1,2,2-tetrachloroethane, 3, dibromomethane, 1, pentachloroethane, 1, 1,1,2-trichloroethane, 1, 1,2,3-trichloropropane, 1, 2-chlorobutane, 1, 1,2-dichloropropane, 1, dichloromethylbenzene, 1, 1-chloro-2-methylpropane, 1	tribromomethane, 3, octanoic acid, 3, pentachloroethane, 3, 2-methyl-1-propanol, 2, 2-butanol, 1, butanol, 1, 2-methyl-2-butanol, 1, 2-methyl-2-propanol, 1, 2-propanol, 1, nitric acid, 2, 1,1,2,2-tetrachloroethane, 3, 3-methylaniline, 2, trifluoroacetic acid, 2, water, 1, 2-methyl-2-phenylpropane, 1, furan, 1, 3-methylphenol, 1, 1,1-dichloroethane, 1, 1,1,2-trichloroethane, 1, 1,2,3-trichloropropane, 1, 2-chlorobutane, 1, 1,2-dichloropropane, 1, dichloromethylbenzene, 1, 1-chloro-2-methylpropane, 1
Tetrachloromethane	Not available in the 120 solvents library	tetrachloroethene, 3, bromotrichloromethane, 4, hexafluorobenzene, 3, trichlorofluoromethane, 4, 1,1,2-trichloro-1,2,2-trifluoroethane, 4, trichloroethanenitrile, 3, 1,1,1-trichlorotrifluoroethane, 4, trichloroethene, 2, trichloronitromethane, 2, sulfuryl chloride, 1, tribromomethane, 1, 1,2,3-tribromopropane, 1, 1,1,1-trichloroethane, 1, 2,2-dichloropropane, 1, 2-chloro-2-methylpropane, 1, pentachloroethane, 1	bromotrichloromethane, 4, 1-chlorobutane, 3, 1-chloro-2-methylpropane, 3, trichlorofluoromethane, 3, 2-chlorobutane, 2, 1,3-cyclohexadiene, 2, 1-bromo-2-methylpropane, 2, 1-bromobutane, 2, tetrachloroethene, 2, trichloroethanenitrile, 3, 1,1,1-trichlorotrifluoroethane, 2, 1,1,2-trichloro-1,2,2-trifluoroethane, 2, 2E,4E-2,4-hexadiene, 1, chlorocyclohexane, 1, 1,1,1-trichloroethane, 1, 2,2-dichloropropane, 1, 2-chloro-2-methylpropane, 1, pentachloroethane, 1
Nitromethane	acetonitrile, 3, methylformate, 4, 1,2-dichloroethane, 3, ethylformate, 2, ch2cl2, 2, 1,3-dioxolan-2-one, 3, anisole, 2, benzene, 2, dimethylcarbonate, 1, chcl3, 1, propylenecarbonate, 1, methanesulfonicacid, 2, aceticanhydride, 1, 2-furanmethanol, 1,	2-propenenitrile, 3, acetonitrile, 3, nitroethane, 4, methyl thiocyanate, 3, 2-aminoethanol, 3, E-2-butenenitrile, 2, 1,2-diaminoethane, 2, 1-chloro-2,3-epoxypropane, 3, propanenitrile, 1, 1-nitropropane, 3, furfural, 1, 1,3-dioxolan-2-one, 1, 2-nitropropane, 1, nitric acid, 1,	1,2-diaminoethane, 3, 2-propenenitrile, 3, nitroethane, 4, 2-aminoethanol, 2, furfural, 3, acetonitrile, 2, dichloromethane, 1, 1,2-dichloroethane, 1, dibromomethane, 1, 1-nitropropane, 3, methyl thiocyanate, 2, 2-propen-1-amine, 2, phenylhydrazine, 1, 4-methoxyaniline, 1,

	lacticacid, 1, propanone, 1, dimethylsulfoxide, 1, isopropylacetate, 1, tert-butylacetate, 1, n- propylacetate, 1, n-butylacetate, 1, dimethylformamide, 1	tetranitromethane, 1, trichloronitromethane, 1, ethyl nitrate, 1, propyl nitrate, 1, 2-nitrotoluene, 1	2-nitropropane, 1, nitric acid, 1, tetranitromethane, 1, trichloronitromethane, 1, ethyl nitrate, 1, propyl nitrate, 1, 2-nitrotoluene, 1
Benzene	toluene, 4, 1-chlorobutane, 3, chlorobenzene, 4, ch2cl2, 2, 1,2-dichloroethane, 2, pentane, 3, trifluoromethylbenzene, 4, cyclohexane, 3, chcl3, 1, cs2, 1, isopropylbenzene, 2, hexane, 1, methylcyclohexane, 1, pyridine, 1, PEG, 1, anisole, 1, benzylalcohol, 1, 2-furanmethanol, 1	1,4-cyclohexadiene, 3, cycloheptatriene, 3, fluorobenzene, 3, 1,4-dimethylbenzene, 3, 1,4-difluorobenzene, 3, 1,2-dimethylbenzene, 3, piperidine, 2, 1,3-cyclohexadiene, 2, styrene, 3, naphthalene, 1, 1,4-dichlorobenzene, 1, ethynylbenzene, 1, pyridine, 1, biphenyl, 1, phenol, 1, chlorobenzene, 1, aniline, 1, toluene, 1, bromobenzene, 1, iodobenzene, 1	2,2-dichloropropane, 3, ethyl isothiocyanate, 3, cycloheptatriene, 2, fluorobenzene, 4, dimethyl disulfide, 3, toluene, 3, 2-bromopropane, 3, iodoethane, 2, isoprene, 1, 3-chloropropene, 1, 1,3-dichloropropane, 1, chlorobenzene, 2, bromobenzene, 2, 1,4-dichlorobutane, 1, pyridine, 1, biphenyl, 1, phenol, 1, aniline, 1, iodobenzene, 1
Diethyl ether	methyl-t-butylether, 3, cyclopentyl-methyl-ether, 3, 2-MeTHF, 3, methyl-tert-amylether, 3, THF, 3, 2-ethoxy-2-methyl-propane, 4, diisopropylether, 3, triethylamine, 3, pyridine, 2, tetrahydro-2,2,5,5- tetramethylfuran, 1, diethylcarbonate, 1, ethylformate, 1, ethylacetate, 1, ethanol, 1, diglyme, 2, hexane, 1, pentane, 1	2-methoxy-2-methylpropane, 3, 2-MeTHF, 3, 2-methyl-1-butene, 3, diethylamine, 3, THF, 2, 2-methyl-2-propanamine, 1, 2-methylpropylamine, 1, 1-pentene, 3, isopropylamine, 1, butane, 2, methylbutane, 2, diethyl sulfide, 1, diisopropyl ether, 1, 1,1-diethoxymethane, 2, 3,6,9-trioxaundecane, 1, 2-ethoxyethanol, 1, dipropyl ether, 1, triethoxymethane, 1, dibutyl ether, 1, propane, 1, 1,1-diethoxyethane, 1, diethyl carbonate, 1	oxane, 3, 2-MeTHF, 3, 2-methoxy-2-methylpropane, 3, THF, 3, thiane, 3, diethylamine, 3, N-methylpiperidine, 2, diisopropyl ether, 1, 2-methyl-2-propanethiol, 3, butanethiol, 1, 2-butoxyethanol, 1, 3-isothiocyanatopropene, 1, 1,1-diethoxymethane, 1, 3,6,9-trioxaundecane, 1, 2-ethoxyethanol, 1, dipropyl ether, 1, triethoxymethane, 1, dibutyl ether, 1, propane, 1, 1,1-diethoxyethane, 1, diethyl carbonate, 1

Chapter 4

Prior knowledge generation through reaction modelling *via* supervised machine learning

Introduction

The overall challenge

Predicting reaction outcomes, often the major product(s), reaction yield, or diastereomeric excess, has been a long-standing topic in research. Ability to accurately predict reaction outcomes would significantly reduce the time and material needed to develop robust processes. Discovery of new molecules and transformations, and optimisation of reactions are often carried out incrementally, based on the previously reported work, combined with years of experience and domain knowledge. This approach has served well as demonstrated by ever increasing number of newly reported work in literature. However, doing so often introduces biases, limits the scope of reaction components to what is available in the lab, does not quantify the influence of relevant reaction parameters or validate choices of descriptors used to parameterise discrete variables. More importantly, this approach fails to quantitatively transfer results from previously reported reactions (i.e., literature data) to unknown reactions conditions for development of new processes. There exists a need to develop a holistic workflow to quantify influence of reaction parameters based on available literature data and predict the outcome of new conditions for more efficient process development.

In addition to the significant increase in available data (e.g., 56M reported reactions in Reaxys as of 25th August 2021),[1] last decade has seen significant adoption of machine learning (ML) algorithms, cheminformatics software, and laboratory automation tools to either generate new datasets and/or utilise literature data to model, and predict suitable reaction conditions and reaction outcomes.[2, 3] To do so, three approaches have been commonly employed. First, big data-based approaches either (*i*) through natural language processing architectures such as molecular transformer models to predict reaction yield[4, 5] or (*ii*) reaction template-based methods to classify reactions based on the transformation types to recommend suitable reaction conditions have been successfully demonstrated.[6, 7] The big data approach, however, is costly

in terms of accessing the data, computationally expensive and time consuming to train large models, does not take into account all the variables (e.g., stoichiometry), and does not reveal physical insights about the reaction being studied. The second approach utilises high-throughput experimentation (HTE) tools to generate large amounts of data in a short period of time to either screen conditions to access several optimal (i.e., good) results and/or for reaction modelling,[8, 9] which is also limited to the availability of such tools and is often labour and material intensive. The final approach focuses on extracting the relevant data (i.e., based on the type of transformation) from literature, either through manual collection or automated data mining, and parameterisation of reaction components (e.g., ligands, bases, solvents) for holistic reaction modelling.[10] This method allows for quantification of nonlinear interactions between physical and chemical parameters to validate important reaction parameters in the form of *a priori* knowledge.

In this chapter, I attempt to provide answers to generation, selection, and validation of *a priori* knowledge in the form of physically meaningful molecular descriptors using literature data, density-functional theory (DFT) based descriptors, and supervised ML algorithms, and demonstrate transfer and utilisation of this knowledge in optimising a new reaction in the lab. The workflow developed in this chapter is given in Figure 20. The chapter structure is as follows. First, a review of various reaction modelling approaches using different dataset sizes and sources, various transformation types, molecular descriptors, and ML algorithms is provided. Then, the mathematics behind some of the implemented algorithms are explained. Third, two Buchwald-Hartwig case studies are introduced with a literature review on the influences of ligands, solvents, and bases. Fourth, methodology for optimising molecular structures and calculation of descriptors using a novel method are provided with experimental procedure, analytical techniques, and materials. Finally, reaction yield prediction results in the form of R^2 score, mean absolute error (MAE), and root-mean-square error (RMSE) from ten ML algorithms are provided. The results are benchmarked against dummy regressor and y-scrambling results to avoid random correlation, and the learning performance of DFT-based descriptors are compared against One-Hot Encoding (OHE) featurisation. It is important to note that the entire ML workflow is fully automated (i.e., as opposed to random or grid search for hyperparameter optimisation, all the hyperparameters for each algorithm are fully optimised for using HyperOpt[11]) and robust due to more sophisticated boosting algorithms that are found to outperform commonly used algorithms such as Random Forest Regressor (RFR) or Multiple Linear Regression (MLR) in predicting reaction outcomes. Moreover, results from Explainable

AI (XAI) tools[12] - model feature importance and permutation feature importance (PFI) are used to extract physical insights from the best models and compared against the domain knowledge.

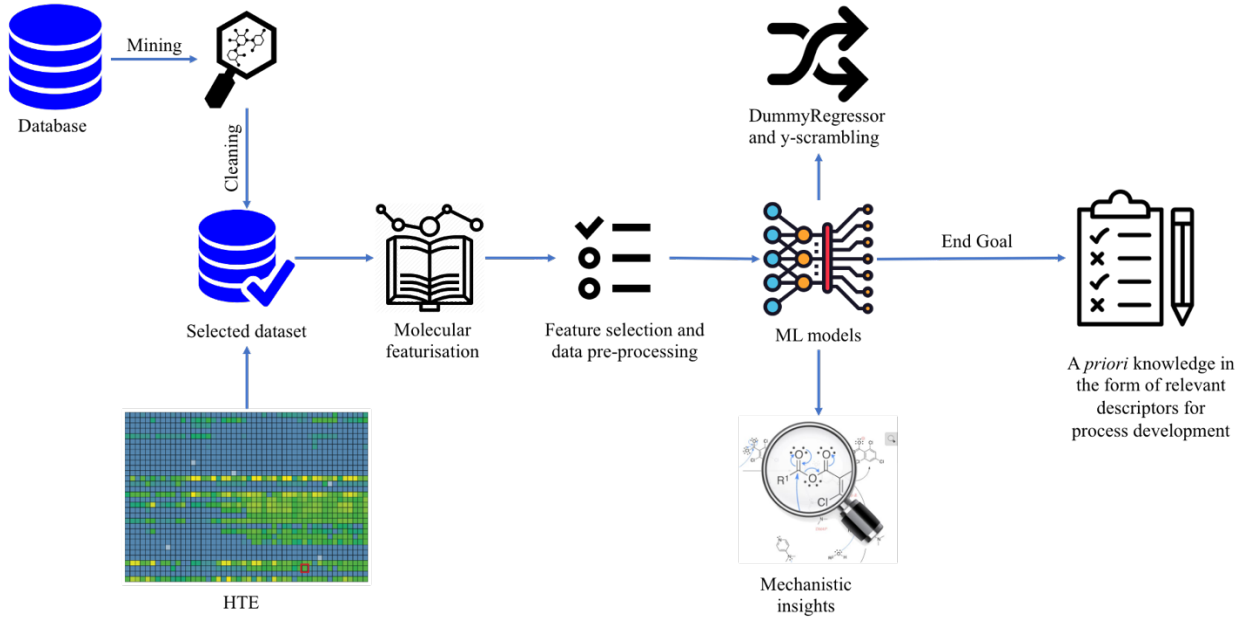


***Figure 20**. ML workflow developed for generating a priori knowledge from literature data.*

Contributions

The contributions to the work in this chapter is as follows: Dr. Simon Sung developed a workflow to add a reference H atom to a target atom (e.g., P atom in a ligand or N atom in a nucleophile to calculate the steric and electronic properties of) and optimise the structure to provide a point of reference for calculating steric descriptors (e.g., L, B1, B5, $\%V_{bur}$). Dr. Akihiro Takada performed the experiments in the optimisation of case study 2 using the automated flow set-up available in-house. Data collection, structure optimisations, descriptor calculations, implementation and optimisation of machine learning algorithms, implementation of Explainable AI tools, and project execution were carried out by me.

Background

Reaction modelling

Main objectives of reaction modelling are often to understand the influence of chemical and process parameters on the reaction outcome(s), quantify the (nonlinear) interactions between these parameters, extract physical insights, validate mechanistic models, and predict unforeseen conditions. With no available experimental data, *in silico* approaches in achieving some of these objectives could be used to explore plausible process models based on mechanistic DFT-level calculations.[13, 14] Cao *et al*. in our group optimised and analysed the structure of intermediates, based on the Hammond postulate that the lowest energy intermediate would be structurally and energetically similar to a transition state (TS), to predict the mechanistic pathway for Pd-catalysed C-H activation of heteroaromatic compounds.[14] Using relative stability of possible intermediates formed *via* proton-abstraction (PA) and electrophilic aromatic substitution (S_EAr), the two main mechanisms, as an approximate threshold, the authors validated the computational predictions against experimental data. Specifically, relative stability threshold of below 0 kcal mol^{-1} was observed for PA mechanism, and threshold of above five kcal mol^{-1} was observed for the S_EAr mechanism. Both mechanisms were considered plausible if the relative stability was between zero and five kcal mol^{-1}. The demonstrated mechanistic modelling could serve as a complementary approach to process development, but it requires significant domain knowledge in the form of mechanistic understanding, optimisation of all possible intermediates or transition states (TSs), does not account for continuous variables, and used to predict regioselectivity only.

An alternative to DFT-level mechanistic calculations is to use molecular descriptors for parameterisation of substrates, reagents, solvents, etc., in a reaction, and to map their influence on the reaction outcome. Sigman and colleagues hand crafted a collection of 367 reaction entries from 17 sources on nucleophilic addition reactions to imines in the presence of 1,1'-bi-2-naphthol (BINOL)-derived chiral phosphoric acids.[10] The authors used multiple-linear regression (MLR) model combined with DFT-calculated descriptors to predict enantioselectivity of out-of-sample reactions, and identification of key descriptors and interactions in asymmetric catalysis. For prediction of enantiomeric excess (% e.e.) using 4-fold leave-one-out (LOO) cross-validation, the authors achieved R^2 of 0.87 and identified the most important model parameters for reaction components (equation 1). For imine parameterisation, L_s (steric descriptor of the smallest imine substituent), NBO_N and NBO_C

(natural bond orbital parameters) were selected. On the other hand, parameterisation of solvent (*sol*), nucleophile (*H-X-CNu*: nucleophile angle measurement), and catalyst (L_{cat}: length of catalyst 2-substituent) were achieved using a single descriptor for each component. Using a similar workflow, the mechanistic transferability of the results was tested using leave-one-reaction-out (LORO) analysis. Reactions were categorised as "different" based on the individual publication. Leaving seven reactions out for the test dataset, the best result achieved R^2 of 0.85, suggesting that conditions from one reaction could be quantitatively transferred to predict outcomes of unknown reactions. Although the demonstrated MLR modelling allows for model interpretability, this approach ignores the nonlinear interactions between the parameters, requires significantly more hand-tuning of descriptor choices, and necessitates a deeper understanding of the mechanism to be able to choose relevant descriptors. Moreover, predicting reaction yield is often more difficult compared to regioselectivity that MLR methods have been mostly used for.[8]

$$Predicted\ \Delta\Delta G = 0.42 + 0.29sol - 0.90NBO_N - 0.75NBO_C + 0.33L_s + 0.63H - X - CNu + 0.20L_{cat} \quad (1)$$

In comparison to using literature data, Ahneman *et al*. employed automated ultra-high-throughput set-up (1,536-well plate) at Merck to conduct 4,608 reactions to study C-N cross-coupling.[8] The reaction scope included 15 aryl halides, 23 additives, 4 catalysts, and bases whilst keeping the continuous variables and 4-methylaniline starting material fixed. Compared to linear regression, k-nearest neighbors (kNN), support vector machines (SVM), and Bayes generalised linear model, the authors found that a single-layer neural network and random forest (RF) algorithms provided the best performance in predicting reaction yield. When trained on the entire dataset with 70/30 train/test split, RF model achieved R^2 value of 0.92 and root-mean-square error (RMSE) of 7.8% on the test data. In order to study the model generalisability on different additives, the algorithm was trained on a dataset containing 15 distinct additives out of 23. Out-of-sample prediction on the rest 8 additives had a model performance of 0.83 for R^2 value and 11.3% for RMSE. Out of 120 initial list of descriptors, 21 were selected to be important in the final model, Table 25.

The same group employed the HTE set-up with ML algorithms to predict the reaction yield of deoxyfluorination reaction of alcohols using sulfonyl fluorides.[15] Total of 640 distinct reactions were conducted to cover the reaction space of 32 alcohols, 5 sulfonyl fluorides, and 4 bases.

Using automatically extracted 43 descriptors, a random forest model achieved R^2 value of 0.89 and RMSE value of 9.3%, with significantly low out-of-sample prediction performance of 0.163 for R^2 value and RMSE of 24.6%. In comparison, manually selected 8 descriptors (Table 25) achieved slightly better performance with R^2 value of 0.93 and RMSE value of 7.4%, with out-of-sample prediction performance of 0.71 for R^2 and for 16.1% RMSE. In comparison to DFT-based descriptors, features with no physical information such as one-hot encoding (OHE) resulted in R^2 value of 0.86 and RMSE value of 10.5%, only 1.2% less compared to use of automatically extracted 43 information rich DFT-based descriptors. This suggests that for certain transformations, especially when the categorical variables have a significant distinction in reaction performance (i.e., such that presence or absence of certain reaction component in the training dataset), OHE or fingerprints (fps) based featurisation could achieve comparable results with DFT-based and manually selected descriptors. Similar performance metrics for different featurisation techniques (e.g., DFT, fps, OHE) was reported by Pomberger *et al*. on a HTE generated dataset for catalytic C(sp^3)-H activation of tertiary amines, suggesting that an increase in the dataset size was more significant in achieving higher prediction accuracy as opposed to the choices of featurisation.[16]

Table 25*. (a) Molecular, atomic, and vibrational descriptors based on optimised DFT structures of reaction components in C-N cross-coupling (top) and deoxyfluorination (bottom) reactions. The workflow does not include the process and validation of selected descriptors from the initial list.*

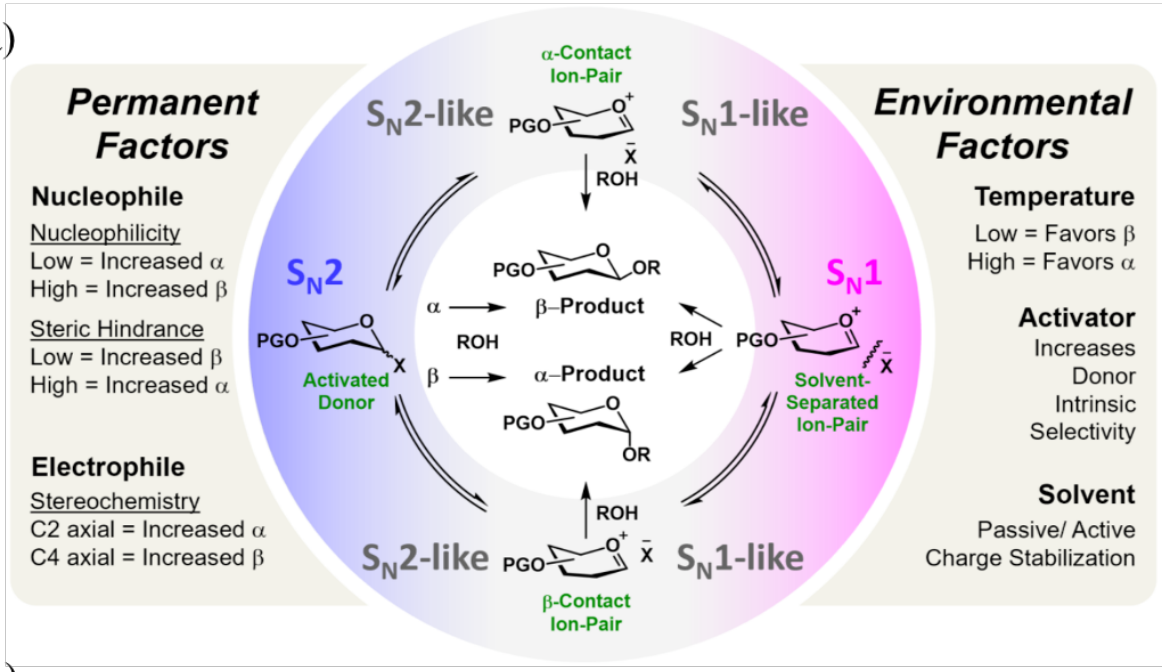
Molecule	Initial no. of descriptors	Selected no. of descriptors	Selected descriptors
	C-N coupling reaction		
Additive	19	9	C3 Electrostatic Charge, C5 NMR Shift, O1 Electrostatic Charge, V1 Frequency, V1 Intensity, Dipole Moment, Electronegativity, Hardness, Ovality
Aryl halide	27	7	C1 NMR Shift, C3 NMR Shift, C3 Electrostatic Charge, C4 NMR Shift, Dipole Moment, Electronegativity, Ovality
Base	10	2	E_{LUMO}, Ovality
Ligand	64	3	P1 electrostatic charge, V1 intensity, V6 frequency
	Deoxyfluorination reaction		
Alcohol	Not reported	4	C1 electrostatic charge, C1 exposed area, electronegativity and categorical descriptors (e.g., primary, tertiary, cyclic).
Base	Not reported	1	N1 exposed area
Sulfonyl fluoride	Not reported	3	S1 electrostatic charge, F1 electrostatic charge, O1 electrostatic charge

However, besides the high cost of setting up HTE tools, modelling reactions using a dataset that completely exhaust the combinatorial space for the reaction components poses an issue of ML algorithms exploiting the patterns in the experimental design instead of learning the underlying physical meaning of the descriptors. This issue was pointed out by Chuang *et al*. on the C-N cross-coupling reaction dataset of 4,608 different reactions.[17] Instead of DFT-based descriptors used in the original paper, the authors used randomised numbers to describe reaction components. For instance, ^{13}C NMR shift for 5C (carbon index 5) for parameterising additives was replaced by random numbers sampled from a standard normal distribution. Doing so, all five models used in the original paper achieved similar results for both physically meaningful descriptors and fully randomised numbers. Similarly, results for one-hot encoding (OHE) were almost same as the other two approaches. Furthermore, to validate the top 10 features selected using molecular descriptors, of which 8 were additive descriptors, authors compared the feature importance using a randomised data, in which the output (yield) was randomly shuffled to decorrelate from the features. In 100 trials using randomised-data set, additive features consistently occupied 9 of top 10 features, making it difficult to validate that models had learned the underlying physically meaningful descriptors. The authors emphasize that their findings do not conclude that the selected features in the original paper by Ahneman *et al*. are incorrect, nor the chemical features are unimportant. However, it raises concern whether the models actually learned the underlying meaning and correlation of molecular descriptors with the reaction output, given randomised parameterisation resulted in a similar model performance, raising a question around combinatorically exhausted experimental design for ML-driven modelling.

An alternative to high cost HTE set-up is the use of microreactors or droplet-based systems for efficient data generation. Gilmore and colleagues employed a standardised 78 μL silicon microreactor platform to predict the stereoselectivity (α- or β-product) of mechanistically complicated glycosylation reaction over 7 electrophiles (donor), 6 nucleophiles (acceptor), 4 acids (activator), 7 solvents and reaction temperature (Figure 21a).[18] A total of 268 data points were used as the training dataset. An initial list of 64 descriptors were generated to parameterise all reaction components. In order to select the most relevant features, maximum of 18 descriptors were selected per model per iteration to avoid overfitting, whilst keeping track of the most important features selected at every iteration. The selected features in the final model include 10 descriptors, a selection highly influenced by current understanding of the

glycosylation reaction, along with temperature (Figure 21b). The optimal model using CART (classification and regression tree) algorithm achieved overall RMSE of 6.8% for selectivity. It was found that stereoselectivity of glycosylation reactions is influenced more by the environmental parameters such as solvent (27%), temperature (19%), and acid activator (7%) than the coupling partners, donor (27%) and acceptor (20%), in the chemical space covered in this study, with the numbers in the parenthesis indicating the degree of influence.

a)



b)

Molecule	Initial number of potential descriptors	Selected number of descriptors	Selected descriptors
Electrophile (donor)	24	3	^{13}C NMR shift of C1, Dihedral angle, J coupling constant
Nucleophile (acceptor)	13	3	^{17}O NMR shift, exposed Area of O and αC
Acid (activator)	16	2	HOMO and Area of O- or N- in the conjugate base
Solvent	11	2	MaxElPot and MinElPot
Temperature	-	-	-

Figure 21*. a) Proposed pathways for glycosylation reaction, and the physical and chemical parameters affecting the selectivity; b) Initial list of descriptors and selected ones. Image a) has been reproduced from ref.*[18]

***Table 26**. Summary of various transformations, data sources, choices of molecular descriptors, and reaction objectives reported in the literature. MOLMAP: molecular map of atom-level properties, EED: electronic effect descriptors, ISIDA, CDK: Chemistry Development Kit.*

Transformation	Data source	Descriptors	Objectives
Asymmetric BINOL catalysis[10]	Literature	DFT	Enantioselectivity
Aza-Michael conjugate addition[19]	Literature	DFT	Enantioselectivity
BH C-N coupling[8]	HTE	DFT	Yield
Sulfonyl reaction[15]	HTE	DFT, OHE	Yield
C(sp3)-H activation[16]	HTE	DFT, OHE, FPs	Yield
Michael addition[20]	Literature	MOLMAPS, ISIDA, EED, CDK	Michael addition feasibility
BH C-N coupling[21]	Literature	Big data	Ligand/base recommendation
Glycosylation[18]	78 μL microreactor	DFT	Stereoselectivity

Supervised ML, implementation of algorithms, and hyperparameter optimisation.

Previous *Reaction Modelling* section highlights the use of supervised ML techniques in predicting reaction outcomes (yield, regioselectivity, enantioselectivity, etc.). Supervised ML is a branch of machine learning that learns patterns from labelled data, and maps a set of input features (i.e., independent variables) to an output (i.e., a dependent variable).[22] For a set of N training examples $[(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)]$, where x_i is a feature vector of the *i*-th incident in the training dataset and y_i is its label (i.e., class for classification tasks and a value for regression tasks), a machine learning algorithm would develop a function such that when fit to a test data (X_test), the predicted values ($\hat{y}$) would minimise the loss function (e.g., $error = \frac{1}{N}\sum_i L\,(y_i, \hat{y})$).[23]

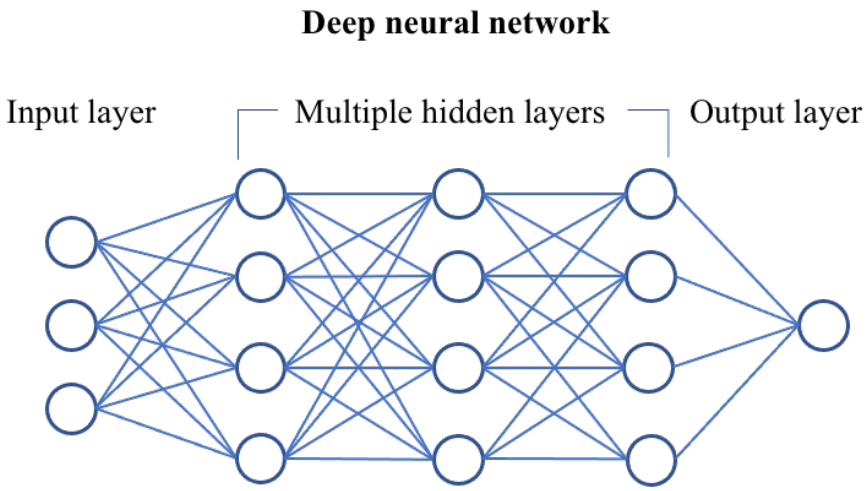


***Figure 22**. A representation of an artificial neural network.*

In order to develop a robust ML pipeline, several powerful algorithms such as artificial neural networks[24, 25] and gradient boosting[26-28] (e.g., ADA,[29] XGB[30]) algorithms were implemented. In this section, mathematics behind artificial neural networks is explained given their sophisticated architecture.

Artificial Neural Networks

For an input vector $x = [x_1, x_2, \dots, x_m]$ and target values $y = [y_1, y_2, \dots, y_n]$, and associated weights $w = [w_1, w_2, \dots, w_m]$ for each neuron in the input layer, output of the first neuron in the second layer (i.e., first hidden layer) or simply the output neuron is expressed as the dot product of x and w as given in equation 1, where b stands for bias.

$$Output\ z = x.w + b \tag{2}$$

$$\hat{y} = \sigma(z) = \frac{1}{1 + e^{-z}} \tag{3}$$

$$Loss\ (aka\ Cost) function = (argmin)\ MSE = \frac{1}{2m}\sum_{i=1}^{n}(y - \hat{y})^2 \tag{4}$$

where $\sigma(z)$ in equation 2 stands for the activation function, in this case, sigmoid function. Activation functions introduce a non-linearity into neural networks. Without use of activation functions over the estimated output, z, for a given neuron, the final prediction would have been a combination of linear functions. Common activation functions used in literature include Sigmoid, ReLu, Softmax, and Tanh.[31]

In order to minimise the cost function, gradient descent learning algorithm is implemented to find optimal weights and bias. First, using the chain rule, rate of change in the cost function is calculated in the form of partial derivate of the cost function with respect to the change in w_i.

$$\frac{\partial C}{\partial w_i} = \frac{\partial C}{\partial \hat{y}} \, x \frac{\partial \hat{y}}{\partial z} \, x \frac{\partial z}{\partial w_i} \tag{5}$$

Where each partial derivative is given as

$$\frac{\partial C}{\partial \hat{y}} = \frac{\partial}{\partial \hat{y}} \, x \frac{1}{2m} x \sum_{i=1}^{n} (y - \hat{y})^2 = \frac{1}{m} \sum_{i=1}^{n} (y - \hat{y}) \tag{6}$$

$$\frac{\partial \hat{y}}{\partial z} = \frac{\partial}{\partial z} \, \sigma(z) = \frac{\partial}{\partial z} \left(\frac{1}{1 + e^{-z}} \right) = \frac{e^{-z}}{(1 + e^{-z})^2} = \sigma(z) \, x \left(1 - \sigma(z) \right) \tag{7}$$

$$\frac{\partial z}{\partial w_i} = \frac{\partial}{\partial w_i} \, (x.w + b) = \frac{\partial}{\partial w_i} \sum_{i=1}^{n} (x_i w_i + b) = x_i \tag{8}$$

Combining equation 5-7, equation 4 can be written as

$$\frac{\partial C}{\partial w_i} = \frac{1}{m} \sum_{i=1}^{n} (y - \hat{y}) \cdot \sigma(z) \cdot \left(1 - \sigma(z) \right) \cdot x_i \tag{9}$$

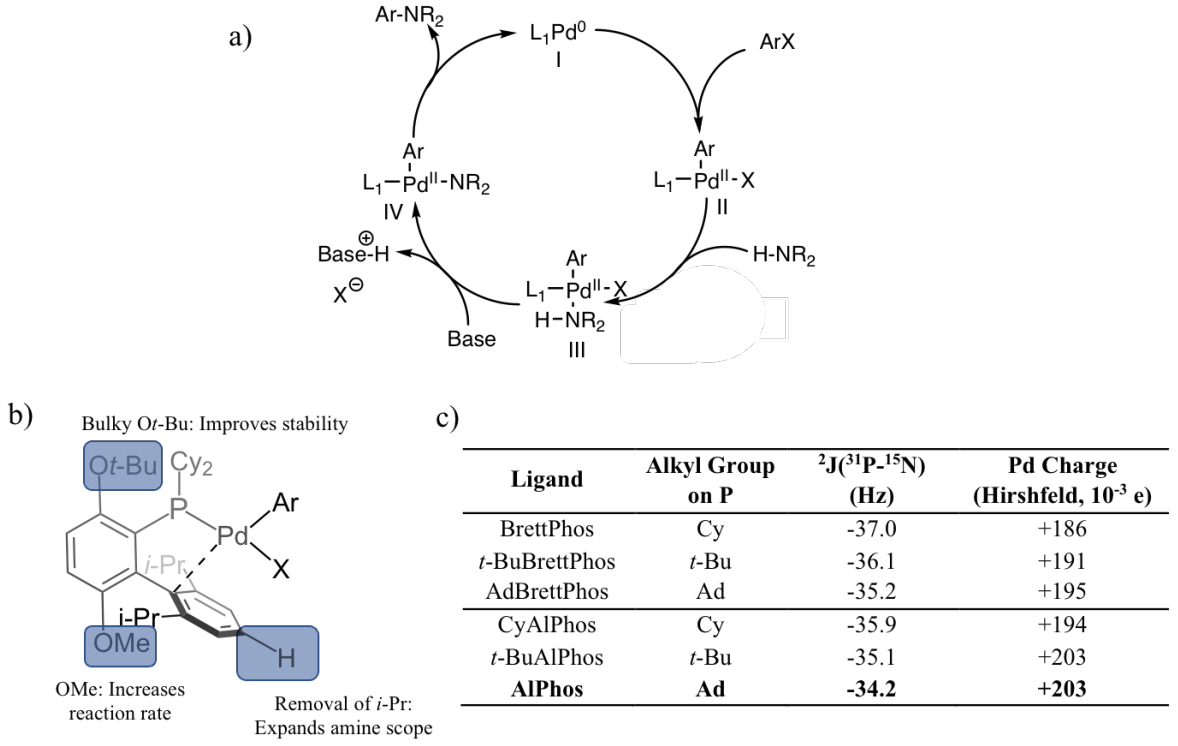
Therefore, a NN architecture can be optimised by iteratively updating w_i until convergence.

$$w_i = w_i - \left(\alpha \cdot \frac{\partial C}{\partial w_i} \right) \tag{10}$$

α hyperparameter in equation 10 stands for learning rate, which controls the change in weights and bias per iteration. Smaller values of α leads to slow convergence, whilst larger α values might fail to converge. Moreover, given the smaller values for $\frac{\partial C}{\partial w_i}$ as gradient descent optimisation approaches a local optima, learning rate does not need to be adjusted over the optimisation. The NN architecture used in our workflow and optimisation criteria are provided in *Method and Materials* section.

Buchwald-Hartwig reaction

Buchwald-Hartwig (BH) C-N cross-coupling plays a significant role in synthesis of complex molecules in pharmaceutical industry[32] with 5,000 new BH couplings reported every year.[21] Given its wide-spread use in discovery of novel molecules and applications in various areas, several "user's guide"[33] or "cheatsheet"[21] recommendations have been provided. Moreover, specific focus has been directed towards understanding the selection of optimal catalyst,[34] ligand,[35-37] base,[38, 39] and solvent[40-43] combinations based on screening of various nucleophiles and electrophiles.[9, 44, 45]



c)

Ligand	Alkyl Group on P	$^{2}J(^{31}P\text{-}^{15}N)$ (Hz)	Pd Charge (Hirshfeld, 10^{-3} e)
BrettPhos	Cy	-37.0	+186
t-BuBrettPhos	*t*-Bu	-36.1	+191
AdBrettPhos	Ad	-35.2	+195
CyAlPhos	Cy	-35.9	+194
t-BuAlPhos	*t*-Bu	-35.1	+203
AlPhos	**Ad**	**-34.2**	**+203**

Figure 23. *a) Proposed mechanism for Pd-catalysed C-N cross-coupling reaction,[38] b) role of substituents on various positions in a BH ligand,[34] and c) $^{31}P\text{-}^{15}N$ ^{2}J coupling constant for commonly used ligands.[38]*

Based on three decades of learnings of Pd-catalysed cross-coupling reactions, the user's guide by Buchwald and colleagues provides a firm starting point in directing synthetic chemists towards selection of ideal ligands and conditions based on the coupling substrates.[33] For instance, for the choice of α-branched primary amine as a nucleophile, the recommended ligands are (*t*-Bu)PhCPhos or (Cy)PhCPhos, selected based on kinetic analysis of arylation of hindered primary amines.[37] Slightly more quantified approach was reported by Fitzner *et al.*[21] as a cheatsheet based on analysis of 62,000 BH couplings gathered from CAS,[46] Reaxys,[1] and

the USPTO.[47] The entire dataset was categorised based on different nucleophile (e.g., amide, aryl, alkyl-aryl) and electrophile (e.g., Br_aryl, Cl_heteroaryl, other) classes. Based on this classification, the top recommended ligand/base combination for a coupling between a primary aniline (aryl) and a heteroaryl chloride (Cl_heteroaryl) would be XPhos and KO^{t}Bu with a literature-reported median yield of 90%. The authors also identified that the most popular ligand Xantphos displayed a relatively low average yield, probably due to its use for (screening of) wide substrate scope given its low cost. Also, as expected, stronger bases such as potassium hydroxide were used for arylation of weakly acidic alkylamines for easier deprotonation of Pd-coordinated amine during the catalytic cycle (Figure 23a).[33]

However, as highlighted by the authors, neither of the reports is comprehensive nor do they provide a predicted yield value. Moreover, the use of large datasets does not account for the scale the reactions were performed at, reaction stoichiometry, or the differences in the source of experimental errors. A more focused approach, on particular transformations, to study specific roles of bases and ligands have been demonstrated for more efficient design of reactions. For instance, to tackle the lack of stability of certain ligands (e.g., BrettPhos-based ligands) at room temperature, McCann *et al*. performed kinetic studies on the influence of substituent type and position on certain ligands in aryl amination reaction (Figure 23b). Driven by the hypothesis that primary amines and N-heteroaromatic substrates can displace the ligand in the Pd-complex, leading to formation of an off-cycle complex, which could require heating for reactivation, the catalyst turnover number is significantly reduced, especially when excess nucleophile is used. Therefore, it was found that a presence of a bulky substituent (e.g., *t*-BuO-) at the ortho position to the dialkylphospino group improved the catalyst stability to reduce off-cycle complex formation whilst a presence of -OMe substituent at the meta position improved the reaction rate, especially when heated to higher temperatures. Finally, removal of *i*-Pr- substituent from the para position to the non-phosphorous-containing ring led to accommodation of a wider substrate scope such as large α-tertiary primary amine nucleophiles (Figure 23b). Combining all three substituent modifications, this new class of GPhos-based ligands proved to perform the functions of three ligand classes: BrettPhos, CPhos, and EPhos.

In terms of optimal base selection, most of initial C-N coupling reactions employed strong, inorganic, and insoluble bases that are incompatible with base-sensitive functional groups, difficult to stir and/or dispose to automate in HTE design (e.g., using Mosquito to dispose bases

in Merck nanomole BH screening[9]), and are not reproducible due to variations in particle size. In order to break this base barrier, Dennis *et al*. studied the use of electron deficient Pd catalysts and weak and soluble organic bases in BH coupling under mild reaction conditions since organic bases usually require harsh conditions (e.g., microwave heating to 150 °C when DBU was used with XPhos ligand). It was found that higher acidity of amine-bound Pd-L complex makes it easier to deprotonate using weaker bases. Acidity of the amine-bound complex was influenced by the electron donation of the phosphine ligand (dependent on substituents) to Pd, which was inversely correlated with ^{31}P-^{15}N ^{2}J coupling constant (Figure 23c). AlPhos ligand has the smallest ^{2}J coupling constant, making the amine-bound Pd complex most cationic and active, to accommodate weak organic bases.

Simultaneous screening of ligands, organic bases, and continuous variables on fixed substrates was demonstrated by Buchwald and colleagues using a droplet platform (Scheme 4).[44] Four nucleophile types – aniline, primary amide, primary aliphatic amine, and secondary aliphatic amine was each optimised separately in 2-MeTHF, except for primary amide for which DMSO was used. Using one-hot encoding (OHE) for discrete variables, and three continuous reaction variables, D-optimal design was used with 5-level full factorial design to optimise the reactions. The list of bases and ligands are given in Scheme 5.

Results from the opimisation revealed further physical insights and patterns regarding the choice of a base and the concentration, and ligand per nucleophile class. As an example, coupling between aryl triflate and aniline resulted in low yield when triethylamine (TEA) or 1,1,3,3-tetramethylguanidine (TMG) was used, suggesting that bases are not strong enough. However, despite having similar pK_{BH^+} for TMG and 2-tert-butyl-1,1,3,3-tetramethylguanidine (BTMG), BTMG performed better, probably due to the presence of a bulky *t*-Bu substituent on the basic TMG nitrogen to avoid unfavourable bindings to Pd to prevent side reactions. Similarly, ligands with bulkier substituents (Cy < *t*-Bu < Ad) form an electron-deficient Pd-complex, which acidifies the bound amine and promotes deprotonation by weaker bases. Therefore, higher yield was achieved using AlPhos than *t*-BuBrettPhos, which was better than *t*-BuXPhos. Moreover, increasing base concentration did not increase the reactivity with mechanistic studies showing an inverse dependence on the base concentration. In cross-coupling of primary amides, DMSO was used to overcome the solubility issues instead of 2-MeTHF, and similar observations were made for bases with sterically hindered bases such as BTMG and MTBD resulted in high yields under mild conditions. Some other observations

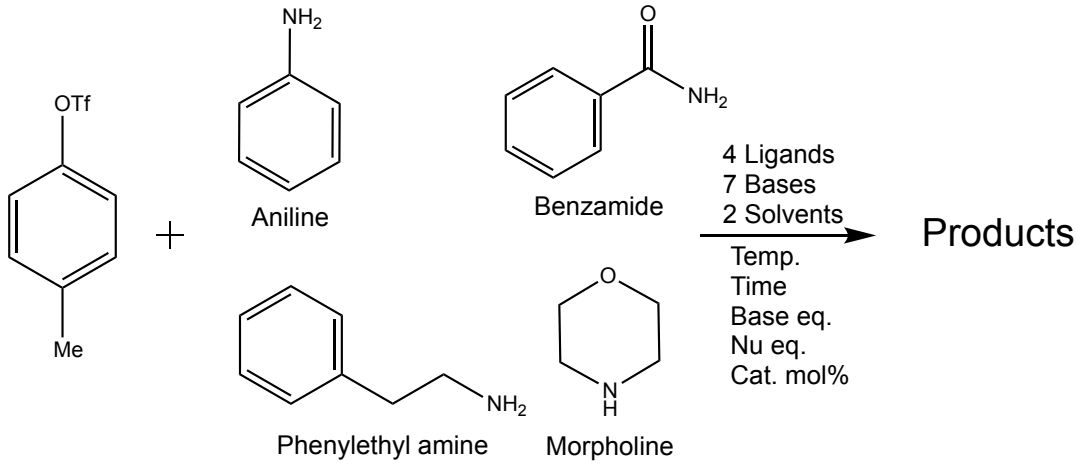
included use of BTTP for primary aliphatic amines, and number of side products (e.g., p-cresol, p-tolyl ether) in the coupling of secondary aliphatic amine morpholine, probably due to presence of water in the reaction mixture and slow Pd-N bond forming step.

Despite the qualitative guideline provided in the paper, it is not clear how the authors decided which ligands or bases to screen per the nucleophile type (e.g., use of BTTP for aliphatic amines, but not for aniline or primary amide) in the D-optimal design. Moreover, optimisation has been done using fixed substrates, limiting the generalisability across different substrates, and with not so data efficient optimisation algorithm (i.e., Bayesian optimisation algorithms).

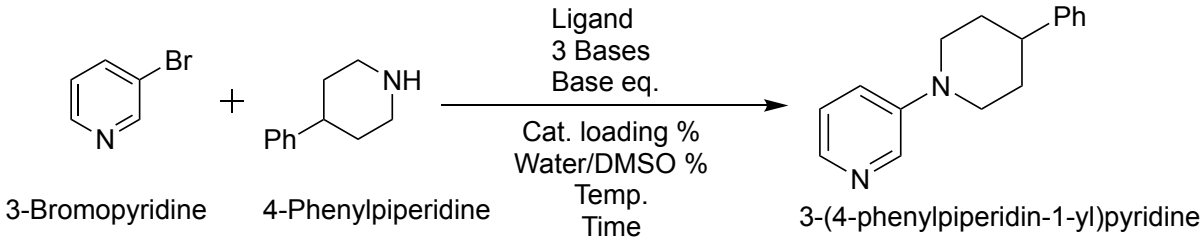
Introduction to the case studies

Given the different approaches presented above to better design a BH coupling reaction, there exists a need for holistic and quantified approach to develop a reaction model for BH couplings to predict the reaction outcome. Herein, we demonstrate a holistic approach to model two different BH coupling reactions using physically meaningful molecular descriptors. The workflow includes training and automated optimisation of ten ML algorithms (Table 29 in the Appendix), all benchmarked against DummyRegressor and scrambled target values (i.e., y-scrambled) to avoid random correlations. The first case study is based on the data from Buchwald and co-workers generated using a droplet platform explained above. The reaction scope covers four nucleophile classes, four ligands, seven bases, two solvents, and six continuous variables (Scheme 4a). The second case study is based on the optimisation of 4-phenylpiperidine and 3-bromopyridine coupling reaction for yield over five continuous variables and three choices of base (DBU, MTBD, and BTMG), using OHE to parameterise the base choice. All the experiments and analysis of case study 2 were performed by Dr. Akihiro Takado in i-DMT using a Vapurtec-based automated flow set-up, described elsewhere in detail[48] and in Chapter 5, and TS-EMO[49] algorithm described in Chapter 2. The full dataset generated in the optimisation is given in Table 36 in the Appendix.

a)

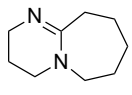


b)



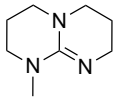
Scheme 4. *Chemistry of two case studies of interest. a) Chemistry from optimisation of different nucleophile classes using a droplet system by Baumgartner et al.,*[44] *whilst a holistic reaction modelling was performed in this project. b) Chemistry of case study 2 optimised in-house by Dr. Akihiro Takada.*

a)



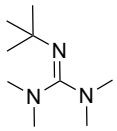
DBU

1,8-Diazabicyclo[5.4.0]undec-7-ene



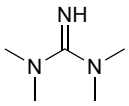
MTBD

7-Methyl-1,5,7-triazabicyclo[4.4.0]dec-5-ene



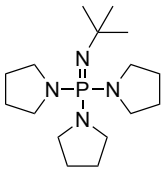
BTMG

2-tert-Butyl-1,1,3,3-tetramethylguanidine



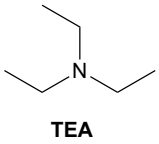
TMG

N,N,N′,N′-Tetramethylguanidine



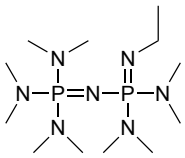
BTTP

Phosphazene base P1-t-Bu-tris(tetramethylene)



TEA

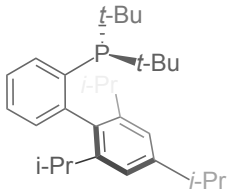
Triethylamine



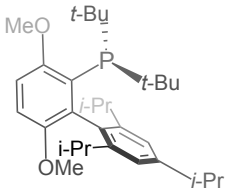
P_2Et

Tetramethyl(tris(dimethylamino)phosphoranylidene)phosphorictriamid-Et-imin

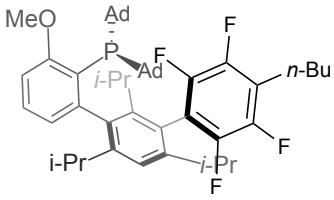
b)



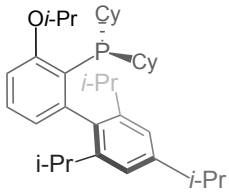
t-BuXPhos



t-BuBrettPhos



AlPhos



EPhos

Scheme 5*. List of a) bases and b) ligands used in modelling two different Buchwald-Hartwig reactions.*

Materials and methods

Experimental procedure

Case study 1

The dataset for case study 1 was collected from Electronic Supplementary Information of the study by Buchwald and co-workers using a droplet platform and design of experiments to identify patterns in base and ligand selection for different nucleophile classes.[44]

Case study 2

Stock solutions of each reaction components were prepared and placed in Gilson Liquid handler rack to be loaded according to respective reaction condition. *Stock solution 1*: 3-Bromopyridine (838 mg, 1.0 eq.) and 1-fluoronaphthalene (140 mg, 0.18 eq.) were dissolved in 30%/15%/6% H_2O/DMSO (34.3 mL). *Stock solution 2*: DBU (1.22 g, 1.5 eq.), MTBD (1.23 g, 1.5 eq.), BTMG (1.39 g, 1.5 eq.) and 4-Phenylpiperidine (1.12 g, 1.3 eq.) were dissolved in DMSO (33.6 mL). *Stock solution 3*: *t*-BuXPhos Pd G3 (37.7 mg/0.02 eq., 94.2 mg/0.05 eq., 188 mg/0.10 eq.) was dissolved in DMSO (15 mL). Each stock solution was transferred to 1 mL sample loop *via* Gilson 271 Liquid Handler. Experiments were carried out in flow using Vapourtec R2 Series in 10 mL R2 reactor and 10 bar BPR for the given residence time and temperature. 60 nL of crude reaction mixture was injected to Shimadzu HPLC *via* VICI switching valve for online analysis. All the experiments for case study 2 were performed by Dr. Akihiro Takada in the i-DMT centre using PyOptimiser[48].

Materials

Reagents 3-Bromopyridine, 98%; 1,8-Diazabicyclo[5.4.0]undec-7-ene (DBU), 98%; 7-Methyl-1,5,7-triazabicyclo[4.4.0]dec-5-ene (MTBD), 98%; 2-tert-Butyl-1,1,3,3-tetramethylguanidine (BTMG), 97%; tBuXPhos Pd G3, 95% were purchased from Sigma Aldrich and used as received. 1-Fluoronaphthalene, 98% was purchased from Alfa Aesar. 4-Phenylpiperidine, 97% was purchased from ACROS Organics. DMSO, 99.7% was purchased from Fisher Scientific.

Analysis (HPLC)

Reaction composition was analysed using HPLC (Shimadzu LC-20A, D2 lamp with PDA detector; Eclipse Plus C18, 95Å, 3.0 x 100 mm, 3.5 μm column). Injection volume of 60 nL (online direct injection), oven temperature of 30 °C, and total flowrate of 1.0 mL/min with acetonitrile/water (5%/95%) was found to be optimal. Acetonitrile concentration was increased to 95% in 9 mins, held for 1 min, then reducing it to 5% and holding it for another 1 min.

DFT optimisation methods

All substrates, bases, and ligand structures (Scheme 4 and 5) were optimised using Gaussian 16 using B3LYP functional and 6-31G(d) basis set. Ultrafine integration grid (Int=UltraFine) and Opt = (CalcFC, MaxCycle=250, NoEigenTest) with NoRaman frequency configuration was used. Optimised structures (i.e., log output files) were used to calculate the steric and electronic descriptors as explained below. An example of output log file, optimised structure and coordinates, calculated electronic and steric descriptors are all provided in the Appendix. Table 33-35 in the Appendix provides the summary of electronic and steric descriptors calculated for parameterisation of nucleophiles, bases, and ligands.

Electronic descriptors: charges and natural bond orbital (NBO) analysis

Charges - Merz-Singh-Kollman (M-K) scheme

Atomic charges are commonly used to describe the electron density distribution on individual atoms in the molecule.[50] Commonly used approaches to calculate atomic charges include Mulliken population analysis, two electrostatic potential fitting methods (M-K and ChelpG), the natural population analysis (NPA), the Hirshfeld population analysis, the minimal basis set (MBS) procedure, and charge model 5 (CM5).[51] In our approach, we used Merz-Singh-Kollman (M-K) scheme to calculate the atomic charges.[52] Optimised structures using Gaussian 16 were used to calculate the atomic charges through electrostatic potential (ESP) fitting at the following setting: HF/STO-3G pop=(ReadRadii, MK) scf=(direct, tight). Values are calculated for all atoms, but only the values for reaction centre in the nucleophiles or the most basic N atom in the bases were used alongside maximum and minimum values for the whole molecule.

Natural Bond Orbital (NBO) analysis

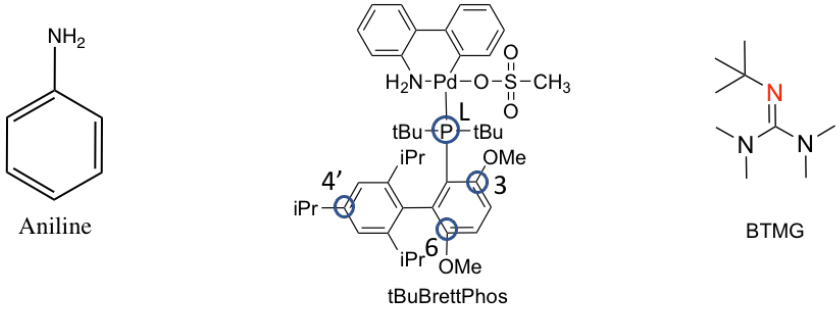
Distribution of electron density in atoms and in bonds could be calculated using natural (localised) orbitals, of which, natural bond orbitals (NBOs) includes the highest possible percentage of electron density.[53] Similar procedure to M-K calculations was followed for NBO analysis at the setting: HF/STO-3G scf=tight pop=nbo. Values are calculated for all atoms, but only the values for reaction centre in the nucleophiles or the most basic N atom in the bases were used.

Steric descriptors: sterimol parameters, percent buried volume, and molecular volume

Steric influences express the nonbonding interactions that are known to affect molecular reactivity. Knowles *et al*. demonstrated that enantioselectivity in asymmetric catalysis is highly affected by steric influences.[54] Specifically, unidimensional steric parameters such as A-values,[55] Charton,[56] and Taft parameters[57] often fail in quantifying the relationship between structure and enantioselectivity.[58] Sterimol parameters, on the other hand, are multidimensional descriptors that quantify the sterics influence across multiple principal axes and are found to be effective in multivariate modelling for enantioselectivity prediction.[10, 59, 60] Sterimol subparameter L represents the total distance (i.e., length) across the primary axis (e.g., Pd→P, H→N). Subparameters B1 and B5 represent the shortest and the longest distance perpendicular to the primary axis, respectively. In other words, B1 and B5 represent the minimum and maximum widths of a substituent, influenced by the extent of branching. The percent buried volume ($\%V_{bur}$), represents the percent of the volume the ligand occupies in a sphere with the metal atom positioned at the centre.[61] In our case study, this represents the percent of the volume occupied by the substituents on the P atom in the ligands and N atom in the nucleophiles and bases.

Optimised structures of ligands, bases, and nucleophiles were used to calculate the steric descriptors. Calculation of sterimol parameters requires a point of reference (e.g., Pd as a point of reference in calculating sterimol parameters of the P atom in the ligand using Pd → P bond axis). In order to simulate the sterics during the reaction (i.e., sterics around P after ligand addition to Pd or around N in the nucleophile and base after the substitution), a workflow was developed internally in the group by Dr. Simon Sung to add a hydrogen atom to the target atom (P or N) and optimise the structure to use H atom as a point of reference. In our calculations, the distance for point of reference for P was 2.28 Å to represent the Pd-P distance and 1 Å for

point of reference to N basic atom to represent N-H bond.[62] A pseudocode for the proposed workflow is given in Figure 26 in the Appendix. The final structure, along with a reference point, was used to calculate the sterimol parameters, percent buried volume, and molecular volume using Paton's code.[63] Summary of selected reaction components and parameters, types and choices of descriptors are given in Figure 24.



Ligand	Nucleophile	Base	Solvent	Continuous variables
Ligand_ESP	Nu_ESP	Base_ESP	2-MeTHF	Nu eq.
Ligand_NBO	Nu_ESP_min	Base_NBO	DMSO	Pre-catalyst mol%
Ligand_Vbur	Nu_ESP_max	Base_L		Aryl Triflate conc.
Ligand - L	Nu_NBO	Base_B1		Base eq.
Ligand - B1	Nu_NBO_min	Base_B5		T/ C
Ligand - B5	Nu_NBO_max	Base_Vbur		Res time / min
Ligand-3-L	Nu-Vbur			
Ligand-3-B5	Nu-MolVol			
Ligand-6-L	Nu-L			
Ligand-6-B5	Nu-B1			
Ligand-4'-L	Nu-B5			
Ligand-4'-B5				

Figure 24*. Choice of descriptors calculated for each reaction component and the continuous variables used for modelling of case study 1. The same six base descriptors were also used to model the case study 2 (as the base is the only discrete variable).*

Supervised ML algorithms

In the developed ML pipeline for this project, the following algorithms were implemented and optimised for: Dummy Regressor (DR) as a baseline comparison, Artificial Neural Networks (NN), Gradient Boosting Regressor (GBR), Extreme Gradient Boosting Regressor (XGB), Natural Gradient Boosting (NGB), Ridge Regression (RR), Extra Trees Regressor (ETR), Ada Boosting Regressor (ADA), Random Forest Regressor (RFR), K-Neighbors Regressor (KNR), and Support Vector Regressor (SVR). Hyperparameters for each algorithm were automatically optimised for using HyperOpt[11] with 1000 evaluations to minimise RMSE of predicted values. The choices of hyperparameters and their ranges are provided in Table 29 in the Appendix.

Results and Discussions

Model prediction results

Case study 1 dataset has 363 reaction entries and contains four nucleophile types, four ligands, seven bases, two solvents, temperature, time, nucleophile equivalence, base equivalence, and catalyst mol% (Scheme 4 and 5). Performance of three different feature sets - *full_features*, *no_corr_features* (i.e., correlated features are removed during the data pre-processing step), and *OHE_features* for predicting the reaction yield were compared. Full feature set contains 37 features whilst 27 features were left after removing correlated features (Figure 25c). Third featurisation was based on OHE of discrete variables and contained 22 features.

When comparing performance of different models, RMSE score is often considered as a better metric. In this case, average yield (i.e., target) is 40%, suggesting that data is not skewed towards low or high yield region, where a random guess would perform reasonably well. This is also reflected in the model performance of DummyRegressor (Figure 25a), where average R^2 score, MAE, and RMSE values are -0.03, 37% and 40%, respectively. Similarly, when the target yield values are randomly shuffled (i.e., y-scrambling), performance of all models were comparable at ~37% for MAE and ~41% for RMSE.

When full dataset was utilised for modelling, all three featurisation techniques resulted in a comparable performance (Table 30, Appendix), with *no_corr_features* achieving the best overall performance (Figure 25a) when comparing R^2 score, MAE, and RMSE values over predicted yield. Specifically, four models – XGBoost Regressor, Extra Trees Regressor, Gradient Boosting Regressor, and Artificial Neural Networks achieved ~0.90 R^2 score (Figure 25a-b). These ensemble learning-based algorithms, except for NNs, outperformed commonly selected models such as Random Forest Regressor or Support Vector Regressor in chemical yield prediction,[8, 18, 64] suggesting a higher learning efficiency. Using RMSE as the evaluation metric, XGBoost Regressor and GradientBoosting Regressor achieved the best results with RMSE values of 11.77 and 11.78, respectively (Figure 25a, 25f). More interpretable metric is MAE, to evaluate how many % points are the predicted values off from the actual value for yield. The experimental datasheet for case study 1 does not report the experimental error associated with 363 reaction points. However, an entry in the dataset contains 104% yield, meaning there is minimum of 4% experimental error for that specific entry (ESI of ref[44]). Considering a potential 3-5% experimental error, the lowest MAE value of 8.12 validates the predictive accuracy of the models (Figure 25a, 25d).

a)

Model	R^2 score	MAE	RMSE
DummyRegressor	-0.03	37.17	40.14
XGBRegressor	**0.91**	**8.25**	**11.77**
NGBRegressor	0.81	13.65	17.37
ExtraTreesRegressor	**0.90**	**8.12**	**12.30**
GradientBoostingRegressor	**0.91**	**8.13**	**11.78**
RandomForestRegressor	0.85	10.86	15.23
AdaBoostRegressor	0.78	15.86	18.54
KNeighborsRegressor	0.72	13.57	20.65
SupportVectorRegressor	0.75	14.49	19.63
RidgeRegressor	0.71	17.84	21.42
Neural Networks	**0.89**	**8.68**	**13.02**

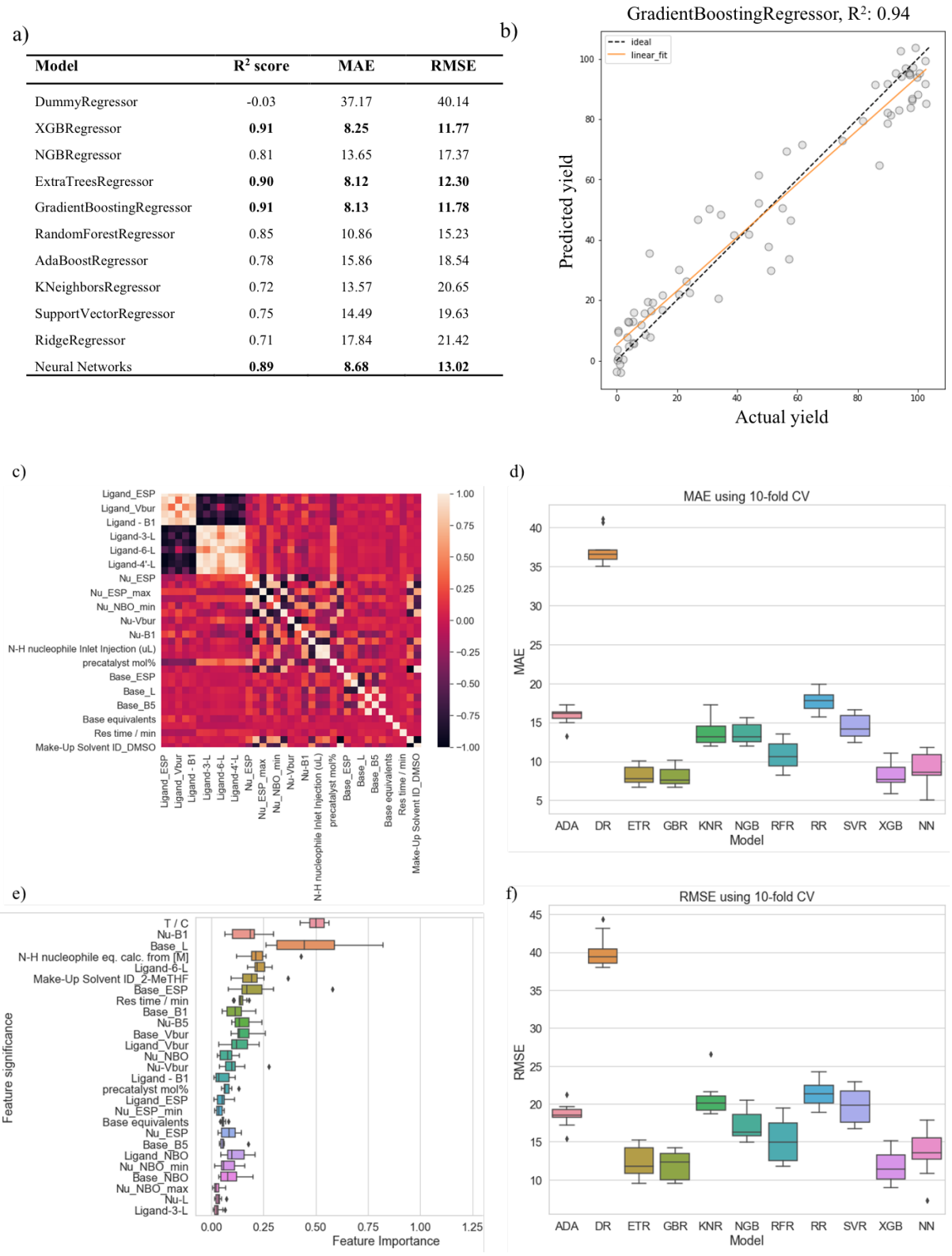


Figure 25. *Algorithms were trained using 80/20 train and test split with 10-fold randomised Monte Carlo cross-validation (i.e., repeated random sub-sampling validation). a) Summary of model metrics when trained on no_corr_features set (363, 28) with b) showing a predicted vs actual yield plot. c) Correlation plot for the features and the correlation numbers are provided*

in the Appendix. d) and f) show MAE and RMSE plots for different models. e) Representation of a permutation feature importance plot when trained with NNs, with PFI selected features ranking provided in Figure 27 in the Appendix.

Considering a dataset of 363 reaction entries, learning performance of *full_features*, *no_corr_features*, and *OHE_features* were evaluated on different dataset sizes. Sampling every third entry in the dataset, hence ensuring enough sampling of data points per nucleophile classes, model performances were compared using only 121 reaction entries (Table 30 in the Appendix). For *no_corr_features* dataset (121, 28), best model was still selected as XGBoost Regressor similar to using full dataset of 363 reaction points, with expectedly reduced R^2 score from 0.91 to 0.86 and increased RMSE value from 12.61 to 14.26, only 1.65 difference despite using one third of the dataset. Comparing the performance of molecular descriptors to OHE, best model resulted in R^2 score of 0.75 as opposed to 0.86 with descriptors, and RMSE value of 18.87, which is 4.61 difference using the best performing model. Whilst OHE based featurisation achieved a comparable performance on a large dataset (in this case, 363 data points), use of molecular descriptors allowed for more efficient learning even with a relatively smaller dataset size of 121 data points for a holistic modelling of ligands, bases, nucleophiles classes, solvents, and continuous reaction variables. Besides, use of molecular descriptors leads to physically meaningful insights as described in *Explainable AI* section and allows for more generalisability.

With the goal of being able to transfer learned prior knowledge from a publicly available dataset to reaction optimisation in-house, the model performances using molecular descriptors were evaluated on the dataset from case study 2. It is important to highlight that, unlike the case study 1 where the average yield is 40%, the average yield for case study 2 is 22.01%, with the maximum yield of 67.42. Hence, evaluating model performances based on MAE and RMSE might introduce a bias. Therefore, model performances were evaluated based on the fitting accuracy R^2 score (Table 27). Using the full dataset of 94 reaction points, the best model using molecular descriptors (94, 12) achieved R^2 score of 0.85 compared to R^2 score of 0.75 when OHE (94, 6) was used, highlighting a difference in predictive accuracy of models when molecular descriptors were used. To compare the metrics on a smaller dataset, algorithms were trained on a different dataset size (Table 27), with and without molecular descriptors. With molecular descriptors, there was no significant decrease in the model performances (i.e., comparable R^2 values) whilst the model performances decreased significantly with OHE-based

parameterisation. Another observation was that the R^2 scores of models using molecular descriptors, on half of the dataset (47,12), were still considerably higher than when using OHE on the full dataset (94,6), highlighting a similar learning efficiency difference as in case study 1. Amongst the models, Gradient Boosting Regressor (GBR) was found to be more efficient in low-data regimes such as half of data use actually improved the model. NNs performed the worst in the second case study due to low data regime.

Table 27. *Modelling results from the case study 2. Optimisation was performed using OHE for bases (three choices) and five continuous variables with total of 94 reactions.*

	12 features					
	94 training points			47 training points		
Model	**R^2 score**	**MAE**	**RMSE**	**R^2 score**	**MAE**	**RMSE**
DummyRegressor	-0.21	19.75	22.64	-0.15	21.01	22.95
XGBRegressor	**0.85**	**5.95**	**7.68**	**0.80**	**5.63**	**7.64**
NGBRegressor	0.73	7.73	10.34	0.74	7.20	9.54
ExtraTreesRegressor	**0.82**	**6.65**	**8.73**	**0.73**	**6.64**	**9.08**
GradientBoostingRegressor	**0.85**	**5.75**	**7.81**	**0.86**	**4.89**	**6.74**
RandomForestRegressor	0.73	7.52	10.51	0.68	8.02	10.45
AdaBoostRegressor	0.76	6.85	9.70	0.66	6.47	9.82
KNeighborsRregressor	0.64	8.03	11.79	0.32	9.91	15.22
SupportVectorRegressor	0.71	7.79	10.86	0.63	8.82	11.68
RidgeRegressor	0.58	10.08	13.16	0.50	11.10	14.32
Neural Networks	**0.43**	**11.57**	**16.07**	**0.26**	**13.38**	**17.88**
	One-Hot Encoding (OHE)					
	94 training points			47 training points		
Model	**R^2 score**	**MAE**	**RMSE**	**R^2 score**	**MAE**	**RMSE**
DummyRegressor	-0.24	22.05	24.30	-0.47	21.04	23.07
XGBRegressor	**0.69**	**7.68**	**10.55**	**0.55**	**9.36**	**12.25**
NGBRegressor	0.53	10.81	13.46	0.38	12.21	14.38
ExtraTreesRegressor	**0.75**	**7.73**	**10.28**	**0.42**	**10.43**	**14.01**
GradientBoostingRegressor	**0.66**	**7.96**	**10.92**	**0.58**	**9.50**	**12.03**
RandomForestRegressor	0.55	9.56	12.52	0.31	11.47	15.24
AdaBoostRegressor	0.45	9.59	14.85	0.22	10.50	15.76
KNeighborsRregressor	0.17	9.89	14.28	-0.19	13.60	19.14
SupportVectorRegressor	0.65	10.14	12.36	0.38	12.53	14.62
RidgeRegressor	0.27	13.97	16.69	0.11	14.41	17.44
Neural Networks	**0.48**	**11.26**	**15.69**	**0.45**	**12.96**	**16.31**

Explainable AI (XAI)

One of the drawbacks of black-box ML models is their lack of interpretability. In order to bridge the gap between black-box ML models and physical models, and extract further insights about the reaction, there has been significant focus on identifying feature importance in chemical reaction modelling.[12, 65] Therefore, two different feature importance metrics – model selected feature importance (i.e., which features are selected to be most significant by the trained model) and permutation feature importance (i.e., significance of a feature is calculated based on the decrease in the model performance when that specific feature is randomly shuffled) were tracked for every model at every iteration. Rankings and coefficients of selected features could potentially reveal physical insights about the reaction. To avoid bias in feature selection by removing correlated features, full feature set of 37 features (and 363 reaction points) were tracked over ten trainings of each model. Expectedly, DummyRegressor does not have feature importance, and hence was not included. Moreover, in order to further improve the real learning of the models as opposed to focusing on a single best performing model to select important features, a "voting" system was created that tracks how many times each feature was selected in top 20 feature list by each model. Table 28 summarises the list of features selected by both model feature importance and permutation feature importance (PFI) alongside the frequency of their selection as an important feature. For ranking of important features when correlated features were removed for modelling (i.e., *no_corr_features* dataset), one can refer to Table 31 in the Appendix.

First, there is a significant overlap between top features selected per feature category by the two approaches (Table 28). Starting with the base descriptors, top three features selected by both of the methods are $\%V_{bur}$, B1, and L steric descriptors followed by the two electronic descriptors NBO and ESP charges. For a comparable electronic density on the basic atom N, sterics around N decides the nucleophilicity of the base, serving as the differentiating factor amongst bases choices. This observation is in alignment with observed increase in yield when bulkier bases with similar pK_{BH+} values were used.[44] Base equivalence was not selected to be too important, which also partly aligns with kinetic observations, at least for the case of aniline where increasing base concentration did not improve the reaction yield.[44] Similarly, B1 descriptor was selected to be the most important for the base choice in the case study 2 (Table 32, Appendix), whilst ESP and NBO descriptors were selected to be the least important. A reason behind the insignificance of electronic descriptors is probably due to the extremely

similar ESP (-0.64, -0.65, and -0.62) and NBO (-0.36, -0.33, and -0.38) values for the base choices BTMG, DBU, and MTBD, respectively. As they do not provide a differentiating information, they were not selected to be significant by the models.

For ligands, sterimol subparameter L for position 6 was selected to be most important, which was observed to influence reaction rate.[34] Then, $\%V_{bur}$ and electronic density (Ligand_ESP, Ligand_NBO) of the P atom in the ligand were selected, suggesting that the free volume around P influences Pd-L complex formation. This is supported by formation of electron-deficient Pd-complex (i.e., based on the electron donation of the phosphine ligand represented using ESP and NBO), which acidifies the bound amine and promotes deprotonation by weaker bases.[38] Next, selected features represent the sterics around position 3 (Ligand-3-L) and 4' (Ligand-4'-B5), which were found to influence the stability of the catalyst and expansion of amine scope, respectively. Given the dataset contains four different nucleophile classes, selection of Ligand-4'-B5 is similar to reported literature that smaller steric values at position 4' expands the amine scope.[34] Selection of certain number of descriptors per position to be important, instead of choosing the descriptors only for specific positions such as P atom or position 6, suggests that models learned the influence of individual substituent positions.

Finally, continuous variables such as the reaction temperature, residence time, starting material concentration, and pre-catalyst mol% were selected as more important than the discrete variable descriptors. One of the possible reasons for this observation is that the authors Baumgartner *et al*. that carried out optimisation of different nucleophiles in case study 1 probably conducted pre-screening for the choices of ligands and bases per nucleophile type, as they do not explain why specific bases and ligands were used for certain nucleophiles, but not for others (e.g., use of *t*-BuXPhos for all nucleophiles except for morpholine or use of BTTP base for morpholine and phenylethyl amine couplings, but not for other nucleophiles). Selection of process parameters to be most important in reaction optimisation is also observed in literature.

***Table 28**. Selected features ranking based on model feature importance and permutation feature importance. Numbers indicate selection frequency by different models over 10 training cycles. DummyRegressor does not track for model importance.*

PFI selected features							
Nucleophile		**Base**		**Ligand**		**Reaction Variables**	
Nu_ESP_max	60	Base_Vbur	66	Ligand-6-L	69	T / C	80
Nu-Vbur	55	Base_B1	64	Ligand_ESP	51	N-H eq.	71
Nu_ESP	54	Base_L	62	Ligand_NBO	45	Res time / min	68
Nu-B1	50	Base_NBO	51	Ligand_Vbur	44	SM conc. / [M]	66
Nu_NBO	49	Base_ESP	50	Ligand - L	42	precatalyst mol%	54
Nu-MolVol	49	Base_B5	41	Ligand-4'-B5	39		
Nu_ESP_min	47	Base eq.	31	Ligand-6-B5	38		
Nu-L	44			Ligand - B1	38		
Nu-B5	41			Ligand-3-L	37		
Nu_NBO_min	39			Ligand - B5	35		
Nu_NBO_max	36			Ligand-3-B5	34		
				Ligand-4'-L	31		
Model selected features							
Nucleophile		**Base**		**Ligand**		**Reaction Variables**	
Nu-B1	64	Base_Vbur	79	Ligand-6-L	70	T / C	80
Nu-Vbur	62	Base_B1	75	Ligand_Vbur	50	N-H eq.	80
Nu-B5	58	Base_L	70	Ligand-4'-B5	49	Res time / min	79
Nu-MolVol	37	Base_NBO	70	Ligand-3-L	46	precatalyst mol%	76
Nu_ESP_max	32	Base_ESP	66	Ligand_NBO	33	SM conc. [M]	75
Nu_ESP_min	31	Base eq.	57	Ligand_ESP	32		
Nu-L	27	Base_B5	55	Ligand - B1	19		
Nu_NBO	23			Ligand - B5	14		
Nu_ESP	20			Ligand-3-B5	12		
Nu_NBO_min	10			Ligand-6-B5	9		
Nu_NBO_max	5			Ligand - L	8		
				Ligand-4'-L	5		

Conclusions

In conclusion, we have developed a workflow that utilises publicly available data and physically meaningful descriptors to generate *a priori* knowledge using a robust ML pipeline. Results from the second case study serves as a validation of transferable results of learnings from literature data to laboratory validation, ultimately reducing the time and material cost required for process development. The results also provide a good comparison between OHE and physically meaningful descriptors despite the "curse of dimensionality" that comes with using a handful of descriptors as opposed to a single numeric value. Main observations were around the learning efficiency with smaller datasets, extraction of physical insights, and model generalisability when using descriptors. As opposed to MLR approaches, where nonlinear interactions are not considered, tuning of models and descriptor selection requires deep understanding of the reaction in study, our approach uses more efficient learning algorithms (e.g., boosting algorithms) that were shown to outperform commonly used algorithms for reaction yield prediction. Using XAI tools such as model feature importance and permutation feature importance, selected features by models through a "voting" system were observed to be in alignment with physical knowledge reported in literature for the choice and substituents of bases and ligands. Also, instead of focusing on a single nucleophile class as was the case with optimisation of case study one by Baumgartner *et al.*, this porojects demonstrates a holistic approach tackling influence of every chemical and physical parameter choice in the reaction.

References

1. Reaxys, reaxys.com (accessed on September 27th, **2022**).
2. Coley, C. W.; Eyke, N. S.; Jensen, K. F., Autonomous Discovery in the Chemical Sciences Part I: Progress. *Angewandte Chemie International Edition* **2020,** *59* (51), 22858-22893.
3. Coley, C. W.; Eyke, N. S.; Jensen, K. F., Autonomous Discovery in the Chemical Sciences Part II: Outlook. *Angewandte Chemie International Edition* **2020,** *59* (52), 23414-23436.
4. Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A., Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Central Science* **2019,** *5* (9), 1572-1583.
5. Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J.-L., Prediction of chemical reaction yields using deep learning. *Machine Learning: Science and Technology* **2021,** *2* (1).
6. Gao, H.; Struble, T. J.; Coley, C. W.; Wang, Y.; Green, W. H.; Jensen, K. F., Using Machine Learning To Predict Suitable Conditions for Organic Reactions. *ACS Central Science* **2018,** *4* (11), 1465-1476.
7. Coley, C. W.; Green, W. H.; Jensen, K. F., Machine Learning in Computer-Aided Synthesis Planning. *Accounts of Chemical Research* **2018,** *51* (5), 1281-1289.
8. Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G., Predicting reaction performance in C–N cross-coupling using machine learning. *Science* **2018,** *360* (6385), 186-190.
9. Buitrago Santanilla, A.; Regalado, E. L.; Pereira, T.; Shevlin, M.; Bateman, K.; Campeau, L.-C.; Schneeweis, J.; Berritt, S.; Shi, Z.-C.; Nantermet, P.; Liu, Y.; Helmy, R.; Welch, C. J.; Vachal, P.; Davies, I. W.; Cernak, T.; Dreher, S. D., Nanomole-scale high-throughput chemistry for the synthesis of complex molecules. *Science* **2015,** *347* (6217), 49-53.
10. Reid, J. P.; Sigman, M. S., Holistic prediction of enantioselectivity in asymmetric catalysis. *Nature* **2019,** *571* (7765), 343-348.
11. Bergstra, J.; Yamins, D.; Cox, D. D., Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. *30th International Conference on Machine Learning (ICML 2013)* **2013**.
12. Feng, J.; Lansford, J. L.; Katsoulakis, M. A.; Vlachos, D. G., Explainable and trustworthy artificial intelligence for correctable modeling in chemical sciences. *Science Advances* **2020,** *6* (42).
13. Zakrzewski, J.; Smalley, A. P.; Kabeshov, M. A.; Gaunt, M. J.; Lapkin, A. A., Continuous-Flow Synthesis and Derivatization of Aziridines through Palladium-Catalyzed C(sp3)−H Activation. *Angewandte Chemie International Edition* **2016,** *55* (31), 8878-8883.
14. Cao, L.; Kabeshov, M.; Ley, S. V.; Lapkin, A. A., In silico rationalisation of selectivity and reactivity in Pd-catalysed C–H activation reactions. *Beilstein Journal of Organic Chemistry* **2020,** *16*, 1465-1475.

15. Nielsen, M. K.; Ahneman, D. T.; Riera, O.; Doyle, A. G., Deoxyfluorination with Sulfonyl Fluorides: Navigating Reaction Space with Machine Learning. *Journal of the American Chemical Society* **2018,** *140* (15), 5004-5008.
16. Pomberger, A.; Pedrina McCarthy, A. A.; Khan, A.; Sung, S.; Taylor, C. J.; Gaunt, M. J.; Colwell, L.; Walz, D.; Lapkin, A. A., The effect of chemical representation on active machine learning towards closed-loop optimization. *Reaction Chemistry & Engineering* **2022,** *7* (6), 1368-1379.
17. Chuang, K. V.; Keiser, M. J., Comment on "Predicting reaction performance in C–N cross-coupling using machine learning". *Science* **2018,** *362* (6416).
18. Moon, S.; Chatterjee, S.; Seeberger, P. H.; Gilmore, K., Predicting glycosylation stereoselectivity using machine learning. *Chemical Science* **2021,** *12* (8), 2931-2939.
19. Metsänen, T. T.; Lexa, K. W.; Santiago, C. B.; Chung, C. K.; Xu, Y.; Liu, Z.; Humphrey, G. R.; Ruck, R. T.; Sherer, E. C.; Sigman, M. S., Combining traditional 2D and modern physical organic-derived descriptors to predict enhanced enantioselectivity for the key aza-Michael conjugate addition in the synthesis of Prevymis™ (letermovir). *Chemical Science* **2018,** *9* (34), 6922-6927.
20. Marcou, G.; Aires de Sousa, J.; Latino, D. A. R. S.; de Luca, A.; Horvath, D.; Rietsch, V.; Varnek, A., Expert System for Predicting Reaction Conditions: The Michael Reaction Case. *Journal of Chemical Information and Modeling* **2015,** *55* (2), 239-250.
21. Fitzner, M.; Wuitschik, G.; Koller, R. J.; Adam, J.-M.; Schindler, T.; Reymond, J.-L., What can reaction databases teach us about Buchwald–Hartwig cross-couplings? *Chemical Science* **2020,** *11* (48), 13085-13093.
22. Russell, S. J.; Norvig, P.; Davis, E., *Artificial intelligence : a modern approach*. 3rd ed.; Prentice Hall: Upper Saddle River, 2010; p xviii, 1132 p.
23. Mohri, M.; Rostamizadeh, A.; Talwalkar, A., *Foundations of machine learning*. MIT Press: Cambridge, MA, 2012; p xii, 414 p.
24. McCulloch, W. S.; Pitts, W., A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics* **1943,** *5* (4), 115-133.
25. Rosenblatt, F., The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review* **1958,** *65* (6), 386-408.
26. Friedman, J. H., Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* **2001,** *29* (5).
27. Schapire, R. E., The Boosting Approach to Machine Learning: An Overview. In *Nonlinear Estimation and Classification*, 2003; pp 149-171.
28. Aschonitis, V. G.; Wu, T.; Zhang, W.; Jiao, X.; Guo, W.; Hamoud, Y. A., Comparison of five Boosting-based models for estimating daily reference evapotranspiration with limited meteorological variables. *Plos One* **2020,** *15* (6).
29. Schapire, R. E., Explaining AdaBoost. In *Empirical Inference*, 2013; pp 37-52.
30. Chen, T.; Guestrin, C., XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016; pp 785-794.
31. Dubey, S. R.; Singh, S. K.; Chaudhuri, B. B., Activation functions in deep learning: A comprehensive survey and benchmark. *Neurocomputing* **2022,** *503*, 92-108.

32. Ruiz-Castillo, P.; Buchwald, S. L., Applications of Palladium-Catalyzed C–N Cross-Coupling Reactions. *Chemical Reviews* **2016,** *116* (19), 12564-12649.
33. Ingoglia, B. T.; Wagen, C. C.; Buchwald, S. L., Biaryl monophosphine ligands in palladium-catalyzed C–N coupling: An updated User's guide. *Tetrahedron* **2019,** *75* (32), 4199-4211.
34. McCann, S. D.; Reichert, E. C.; Arrechea, P. L.; Buchwald, S. L., Development of an Aryl Amination Catalyst with Broad Scope Guided by Consideration of Catalyst Stability. *Journal of the American Chemical Society* **2020,** *142* (35), 15027-15037.
35. Tian, J.; Wang, G.; Qi, Z.-H.; Ma, J., Ligand Effects of BrettPhos and RuPhos on Rate-Limiting Steps in Buchwald–Hartwig Amination Reaction Due to the Modulation of Steric Hindrance and Electronic Structure. *ACS Omega* **2020,** *5* (34), 21385-21391.
36. Wagner, P.; Bollenbach, M.; Doebelin, C.; Bihel, F.; Bourguignon, J.-J.; Salomé, C.; Schmitt, M., t-BuXPhos: a highly efficient ligand for Buchwald–Hartwig coupling in water. *Green Chem.* **2014,** *16* (9), 4170-4178.
37. Ruiz-Castillo, P.; Blackmond, D. G.; Buchwald, S. L., Rational Ligand Design for the Arylation of Hindered Primary Amines Guided by Reaction Progress Kinetic Analysis. *Journal of the American Chemical Society* **2015,** *137* (8), 3085-3092.
38. Dennis, J. M.; White, N. A.; Liu, R. Y.; Buchwald, S. L., Breaking the Base Barrier: An Electron-Deficient Palladium Catalyst Enables the Use of a Common Soluble Base in C–N Coupling. *Journal of the American Chemical Society* **2018,** *140* (13), 4721-4725.
39. Geogheghan, K., All about that base. *Nature Chemistry* **2018,** *10* (5), 487-487.
40. Christensen, H.; Kiil, S.; Dam-Johansen, K.; Nielsen, O.; Sommer, M. B., Effect of Solvents on the Product Distribution and Reaction Rate of a Buchwald−Hartwig Amination Reaction. *Organic Process Research & Development* **2006,** *10* (4), 762-769.
41. Lei, P.; Wang, Y.; Mu, Y.; Wang, Y.; Ma, Z.; Feng, J.; Liu, X.; Szostak, M., Green-Solvent Selection for Acyl Buchwald–Hartwig Cross-Coupling of Amides (Transamidation). *ACS Sustainable Chemistry & Engineering* **2021,** *9* (44), 14937-14945.
42. Sherwood, J.; Clark, J. H.; Fairlamb, I. J. S.; Slattery, J. M., Solvent effects in palladium catalysed cross-coupling reactions. *Green Chemistry* **2019,** *21* (9), 2164-2213.
43. Gevorgyan, A.; Hopmann, K. H.; Bayer, A., Improved Buchwald–Hartwig Amination by the Use of Lipids and Lipid Impurities. *Organometallics* **2021**.
44. Baumgartner, L. M.; Dennis, J. M.; White, N. A.; Buchwald, S. L.; Jensen, K. F., Use of a Droplet Platform To Optimize Pd-Catalyzed C–N Coupling Reactions Promoted by Organic Bases. *Organic Process Research & Development* **2019,** *23* (8), 1594-1601.
45. Lau, S. H.; Yu, P.; Chen, L.; Madsen-Duggan, C. B.; Williams, M. J.; Carrow, B. P., Aryl Amination Using Soluble Weak Base Enabled by a Water-Assisted Mechanism. *Journal of the American Chemical Society* **2020,** *142* (47), 20030-20039.
46. American Chemical Society (ACS), acs.org. (accessed on September 27th, **2022**).
47. USPTO, uspto.gov. (accessed on September 27th, **2022**).
48. Jeraal, M. I.; Sung, S.; Lapkin, A. A., A Machine Learning-Enabled Autonomous Flow Chemistry Platform for Process Optimization of Multiple Reaction Metrics. *Chemistry–Methods* **2020,** *1* (1), 71-77.

49. Bradford, E.; Schweidtmann, A. M.; Lapkin, A., Efficient multiobjective optimization employing Gaussian processes, spectral sampling and a genetic algorithm. *Journal of Global Optimization* **2018,** *71* (2), 407-438.
50. Wiberg, K. B.; Rablen, P. R., Comparison of atomic charges derived via different procedures. *Journal of Computational Chemistry* **1993,** *14* (12), 1504-1518.
51. Wiberg, K. B.; Rablen, P. R., Atomic Charges. *The Journal of Organic Chemistry* **2018,** *83* (24), 15463-15469.
52. Singh, U. C.; Kollman, P. A., An approach to computing electrostatic charges for molecules. *Journal of Computational Chemistry* **1984,** *5* (2), 129-145.
53. Weinhold, F.; Landis, C. R.; Glendening, E. D., What is NBO analysis and how is it useful? *International Reviews in Physical Chemistry* **2016,** *35* (3), 399-440.
54. Knowles, R. R.; Jacobsen, E. N., Attractive noncovalent interactions in asymmetric catalysis: Links between enzymes and small molecule catalysts. *Proceedings of the National Academy of Sciences* **2010,** *107* (48), 20678-20685.
55. Winstein, S.; Holness, N. J., Neighboring Carbon and Hydrogen. XIX. t-Butylcyclohexyl Derivatives. Quantitative Conformational Analysis. *Journal of the American Chemical Society* **2002,** *77* (21), 5562-5578.
56. Charton, M., Steric effects. I. Esterification and acid-catalyzed hydrolysis of esters. *Journal of the American Chemical Society* **2002,** *97* (6), 1552-1556.
57. Taft, R. W., Linear Free Energy Relationships from Rates of Esterification and Hydrolysis of Aliphatic and Ortho-substituted Benzoate Esters. *Journal of the American Chemical Society* **2002,** *74* (11), 2729-2732.
58. Sigman, M. S.; Harper, K. C.; Bess, E. N.; Milo, A., The Development of Multidimensional Analysis Tools for Asymmetric Catalysis and Beyond. *Accounts of Chemical Research* **2016,** *49* (6), 1292-1301.
59. Harper, K. C.; Bess, E. N.; Sigman, M. S., Multidimensional steric parameters in the analysis of asymmetric catalytic reactions. *Nature Chemistry* **2012,** *4* (5), 366-374.
60. Santiago, C. B.; Guo, J.-Y.; Sigman, M. S., Predictive and mechanistic multivariate linear regression models for reaction development. *Chemical Science* **2018,** *9* (9), 2398-2412.
61. Hillier, A. C.; Sommer, W. J.; Yong, B. S.; Petersen, J. L.; Cavallo, L.; Nolan, S. P., A Combined Experimental and Theoretical Study Examining the Binding of N-Heterocyclic Carbenes (NHC) to the Cp*RuCl (Cp* = η5-C5Me5) Moiety: Insight into Stereoelectronic Differences between Unsaturated and Saturated NHC Ligands. *Organometallics* **2003,** *22* (21), 4322-4326.
62. Zhao, S.; Gensch, T.; Murray, B.; Niemeyer, Z. L.; Sigman, M. S.; Biscoe, M. R., Enantiodivergent Pd-catalyzed C–C bond formation enabled through ligand parameterization. *Science* **2018,** *362* (6415), 670-674.
63. Jackson, K.; Paton, R. *Sterimol.py; https://github.com/bobbypaton/Sterimol.*, 2017.
64. Skoraczyński, G.; Dittwald, P.; Miasojedow, B.; Szymkuć, S.; Gajewska, E. P.; Grzybowski, B. A.; Gambin, A., Predicting the outcomes of organic reactions via machine learning: are current descriptors sufficient? *Scientific Reports* **2017,** *7* (1).
65. Jiménez-Luna, J.; Grisoni, F.; Schneider, G., Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence* **2020,** *2* (10), 573-584.

Appendix

Pseudocode for adding an H atom as a point of reference for calculating steric descriptors of a target atom. To be updated.

1. Input: conformer, atom_index
2. Load the conformer and generate a mol file. Save the original structure as an RDKit object to be later used for Constrained Optimization (template = Chem.MolFromPDBFile(name.pdb), removeHs=False)
3. Generate 3D coordinates for all atoms in the original molecule using template.GetConformer().GetAtomPosition(x))
4. Initialise new positions for to-be-added H atom. Provide 8 positions as an initial location, save the file, perform constrained optimisation using *UFFConstrainedOptimize*(mol, moving_atoms, fixed_atoms, cut_off), identify energy of new conformer. Choose the lowest energy conformer amongst the 8.
5. Save the lowest energy conformer with the new H atom as a RDKit conformer.
6. Although it is a Constrained Optimisation, RDKit does not 100% fix the positions of the atoms. Thus, all atoms are returned to original positions.
7. Set the bond length of the added H atom (rdMolTransforms.SetBondLength(mol, addtoatom_index, newH_index, bond_length)
8. Return sterimol parameters, buried volume, and molecular volume (excluding the newly added H for buried volume of the index atom).

Figure 26*. Pseudocode for adding a reference H atom to a target atom (e.g., P, N) for calculating steric descriptors.*

***Table 29**. Summary of algorithms, abbreviations, and hyperparameters used in modelling the reaction data. Train/test split of 80/20 was used with randomised Monte Carlo evaluation using 10-fold cross-validation. Every model was trained with 1000 evaluations for hyperparameter optimisation.*

Algorithm	Abbreviation	Optimised hyperparameters
Dummy Regressor	DR	-
eXtreme Gradient Boosting	XGB	*eta* : uniform (0.01, 0.5) *max_depth*: quniform (3, 18, 1) *min_child_weight*: quniform (0, 10, 1) *colsample_bytree*: uniform (0.5, 1)
Natural Gradient Boosting	NGB	*learning_rate*: uniform (0.005, 0.02) *n_estimators*: quniform (10, 100, 5) *minibatch_frac*: uniform (0.01, 0.75) *col_sample*: uniform (0.5, 1.0)
Extra Trees Regressor	ETR	*max_depth*: quniform (2, 30, 2) *min_samples_split*: quniform (2, 10, 1) *criterion*: (squared_error, absolute_error) *n_estimators*: quniform (10, 200, 5)
Gradient Boosting Regressor	GBR	*learning_rate:* (0.01, 0.25, *loss*: (squared_error, absolute_error, huber) *n_estimators:* quniform (50, 500, 25) *subsample:* uniform (0.5, 1.0) *min_samples_split:* quniform (2, 10, 1) *max_depth:* quniform (3, 20, 1)
Random Forest Regressor	RFR	*n_estimator'*: quniform (100, 600, 100) *max_depth*: uniform (2, 30) *min_samples_leaf*: quniform (2, 8, 1)
Adaptive Boosting Regressor	ADA	*learning_rate:* uniform (0.005, 1.0) *loss:* (linear, exponential, square) *n_estimators:* quniform (25, 250, 10)
K-Neighbors Regressor	KNR	*n_neighbors:* quniform (5, 20, 2) *algorithm*: (auto, ball_tree, kd_tree, brute) *weights*: (uniform, distance)

		p: quniform (1, 2, 1)
Support Vector Regressor	SVR	*kernel:* (linear, poly, rbf, sigmoid) *C:* uniform (0.5, 10.0) *epsilon:* uniform (0.05, 5.0)
Ridge Regression	RR	*solver_options:* (auto, svd, cholesky, lsqr, sparse_cg, sag, saga) *alpha:* uniform (0.05, 5.0)
Artificial Neural Networks	NN	*activation_function*: (relu, linear) *dropout*: 0.2 n_hidden_layers: (2, 5) *n_neurons*: 15 *optimiser*: Adam

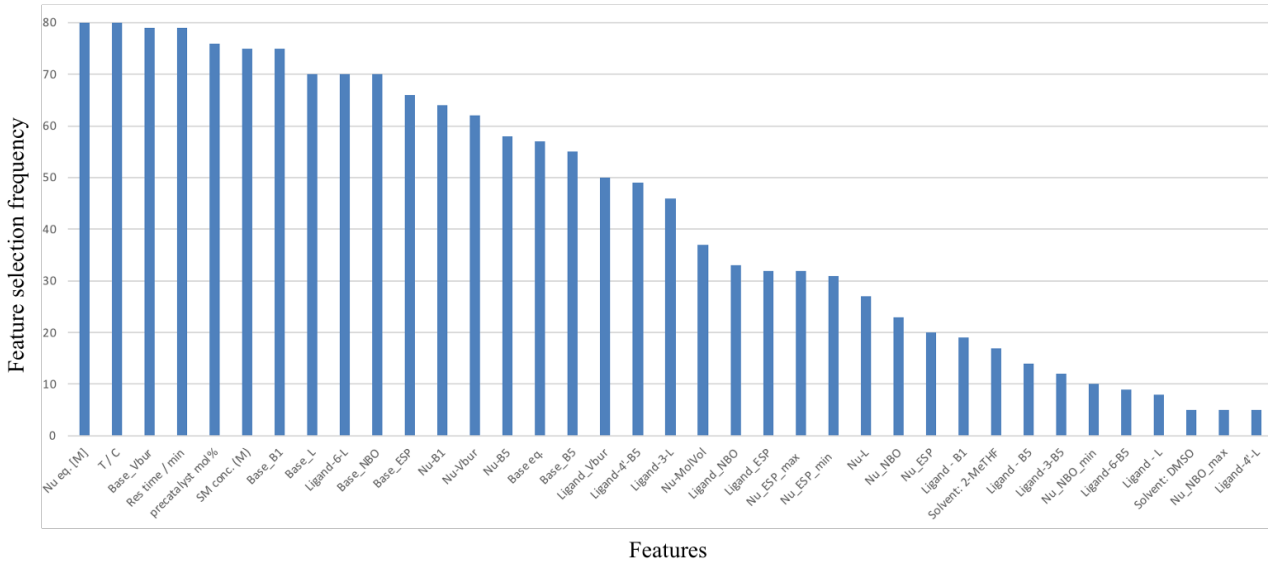


Figure 27*. An example of PFI selected rankings model features. Feature selection frequency indicates how many times a given feature was selected to be important by the selected list of models over 10 training iterations.*

***Table** 30. All the results from 10-fold CV of modelling case study 1 with three different featurisation techniques.*

All 37 features						
	373 training points			121 training points - tbu		
Model	**R^2 score**	**MAE**	**RMSE**	**R^2 score**	**MAE**	**RMSE**
DummyRegressor	-0.02	37.28	40.21	-0.03	38.32	40.57
XGBRegressor	**0.89**	**8.73**	**12.61**	**0.87**	**10.12**	**14.45**
NGBRegressor	0.79	13.84	18.09	0.73	16.60	20.67
ExtraTreesRegressor	**0.89**	**8.39**	**12.80**	**0.83**	**11.90**	**16.57**
GradientBoostingRegressor	**0.90**	**8.39**	**12.27**	**0.86**	**10.57**	**14.62**
RandomForestRegressor	0.82	11.12	16.24	0.73	15.56	20.75
AdaBoostRegressor	0.78	15.56	18.47	0.74	15.46	20.12
KNeighborsRregressor	0.68	14.22	22.12	0.64	16.75	23.70
SupportVectorRegressor	0.67	16.68	22.30	0.57	21.49	26.02
RidgeRegressor	0.67	18.98	22.61	0.69	18.45	22.29
Neural Networks	**0.87**	**9.48**	**14.05**	**0.83**	**11.11**	**15.71**

27 features						
	373 training points			121 training points		
Model	**R^2 score**	**MAE**	**RMSE**	**R^2 score**	**MAE**	**RMSE**
DummyRegressor	-0.03	37.17	40.14	-0.06	36.69	40.20
XGBRegressor	**0.91**	**8.25**	**11.77**	**0.86**	**10.86**	**14.26**
NGBRegressor	0.81	13.65	17.37	0.72	16.48	20.39
ExtraTreesRegressor	**0.90**	**8.12**	**12.30**	**0.82**	**12.12**	**16.14**
GradientBoostingRegressor	**0.91**	**8.13**	**11.78**	**0.86**	**10.811**	**14.53**
RandomForestRegressor	0.85	10.86	15.23	0.71	16.66	20.59
AdaBoostRegressor	0.78	15.86	18.54	0.70	15.92	20.74
KNeighborsRregressor	0.72	13.57	20.65	0.73	14.74	19.89
SupportVectorRegressor	0.75	14.49	19.63	0.56	21.08	25.68
RidgeRegressor	0.71	17.84	21.42	0.71	17.59	20.79
Neural Networks	**0.89**	**8.68**	**13.02**	**0.71**	**13.35**	**19.14**

One-Hot Encoding (OHE)						
	373 training points			121 training points		
Model	**R^2 score**	**MAE**	**RMSE**	**R^2 score**	**MAE**	**RMSE**
DummyRegressor	-0.02	37.62	40.64	-0.07	36.73	39.34
XGBRegressor	**0.88**	**9.69**	**13.76**	**0.75**	**13.87**	**18.87**
NGBRegressor	0.72	16.78	21.29	0.49	21.80	27.11
ExtraTreesRegressor	**0.89**	**8.77**	**13.48**	**0.74**	**14.01**	**19.28**
GradientBoostingRegressor	**0.88**	**9.26**	**13.85**	**0.71**	**14.78**	**20.49**
RandomForestRegressor	0.80	12.96	18.00	0.46	21.63	27.85
AdaBoostRegressor	0.73	17.90	21.04	0.60	18.41	23.93
KNeighborsRregressor	0.72	14.36	21.08	0.53	19.00	25.84
SupportVectorRegressor	0.71	16.64	21.62	0.41	23.73	29.21
RidgeRegressor	0.69	18.11	22.59	0.66	18.49	22.21
Neural Networks	**0.89**	**8.43**	**12.92**	**0.80**	**12.28**	**17.68**

***Table 31**. Selection frequency of features through "voting" when correlated features are removed in the pre-processing step.*

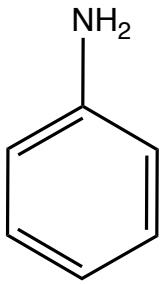
PFI selected features							
Nucleophile		**Base**		**Ligand**		**Reaction Variables**	
Nu eq.	67	Base_Vbur	57	Ligand-6-L	69	T / C	72
Nu-B1	61	Base_B1	57	Ligand-3-L	54	Res time / min	63
Nu_NBO	48	Base_L	56	Ligand_ESP	49	precatalyst mol%	60
Nu-B5	48	Base_NBO	48	Ligand - B1	46		
Nu_ESP_min	46	Base_B5	47	Ligand_Vbur	44		
Nu-Vbur	46	Base_ESP	43	Ligand_NBO	42		
Nu-L	41	Base eq.	30				
Nu_NBO_min	38						
Nu_ESP	33						
Nu_NBO_max	31						
Model selected features							
Nucleophile		**Base**		**Ligand**		**Reaction Variables**	
Nu eq.	72	Base_B1	69	Ligand_ESP	64	precatalyst mol%	72
Nu-B1	68	Base_NBO	69	Ligand-6-L	63	Res time / min	72
Nu-B5	64	Base_Vbur	67	Ligand-3-L	62	T / C	72
Nu-Vbur	57	Base_L	59	Ligand_NBO	53		
Nu_NBO	44	Base_ESP	57	Ligand_Vbur	51		
Nu_ESP_min	41	Base_B5	52	Ligand - B1	37		
Nu_ESP	21	Base eq.	51				
Nu-L	21						
Nu_NBO_max	20						
Nu_NBO_min	13						

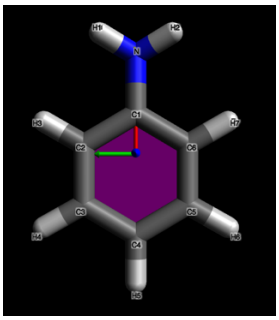
***Table 32**. Permutation feature importance selected in case study 2. Used three bases: DBU, MTBD, and BTMG.*

Feature	PFI-based selection frequency
Cat. eq.	39
Base_B1	39
Base_B5	38
Base_L	37
Base_Mol_Vol	35
Base_Vbur%	35
T / C	31
Res. time / min	23
Water ratio	15
Base_eq	15
Base_NBO	11
Base_ESP	9

DFT optimisation and descriptors calculation results

Aniline





Gaussian output

#n B3LYP/6-31G(d) Opt

Aniline optimisation

0 1

N	2.33746	0.00000	-0.07936
C	0.93868	0.00000	-0.00995
C	0.22124	1.20829	-0.00510
C	-1.17180	1.20260	0.00333
C	-1.88169	0.00000	0.00833
C	-1.17180	-1.20260	0.00333
C	0.22124	-1.20829	-0.00510
H	2.77727	0.83469	0.28939
H	2.77727	-0.83469	0.28939
H	0.76349	2.15164	-0.01311
H	-1.70566	2.14979	0.00883
H	-2.96761	0.00000	0.01631
H	-1.70566	-2.14979	0.00883
H	0.76349	-2.15164	-0.01311

MK ESP charge

Fitting point charges to electrostatic potential
Charges from ESP fit, RMS= 0.00080 RRMS= 0.07661:
ESP charges:
1

1	**N**	**-0.877117**
2	C	0.515408
3	C	-0.265234
4	C	-0.006228
5	C	-0.145986
6	C	-0.006228

7 C -0.265234
8 H 0.332019
9 H 0.332019
10 H 0.095205
11 H 0.062603
12 H 0.070967
13 H 0.062603
14 H 0.095205
Sum of ESP charges = 0.00000
ESP charges with hydrogens summed into heavy atoms:
1
1 N -0.213080
2 C 0.515408
3 C -0.170029
4 C 0.056375
5 C -0.075019
6 C 0.056375
7 C -0.170029
Charge= 0.00000 Dipole= 1.2435 0.8135 0.0000 Tot= 1.4860

NBO charge

Summary of Natural Population Analysis:

			Natural Population			
		Natural	----------			
Atom	No	Charge	Core	Valence	Rydberg	Total
N	**1**	**-0.40287**	**1.99999**	**5.40288**	**0.00000**	**7.40287**
C	2	0.16173	1.99988	3.83839	0.00000	5.83827
C	3	-0.08946	1.99994	4.08952	0.00000	6.08946
C	4	-0.01449	1.99994	4.01455	0.00000	6.01449
C	5	-0.06723	1.99994	4.06729	0.00000	6.06723
C	6	-0.01449	1.99994	4.01455	0.00000	6.01449
C	7	-0.08946	1.99994	4.08952	0.00000	6.08946
H	8	0.17737	0.00000	0.82263	0.00000	0.82263
H	9	0.17737	0.00000	0.82263	0.00000	0.82263
H	10	0.03032	0.00000	0.96968	0.00000	0.96968
H	11	0.03396	0.00000	0.96604	0.00000	0.96604
H	12	0.03296	0.00000	0.96704	0.00000	0.96704
H	13	0.03396	0.00000	0.96604	0.00000	0.96604
H	14	0.03032	0.00000	0.96968	0.00000	0.96968

```
================================================================================
======
  * Total *  -0.00000     13.99958     36.00042     0.00000     50.00000
```

***Table 33**. Summary of descriptors and values used for parameterisation of nucleophiles in the case study 1.*

Molecule	L	B1	B5	%V_{bur}	ESP	NBO	Molecular volume
Anilline	6.32	2.33	*3.66*	29.42	-0.88	-0.4	89.48
Benzamide	6.34	2.1	6.13	39.23	-0.89	-0.45	108.55
Morpholine	5.25	1.74	4.9	38.46	-0.69	-0.31	81.75
Phenylethylamine	7.91	2.04	*6.01*	26.62	-0.87	-0.39	120.36

***Table 34**. Summary of descriptors and values used for parameterisation of bases.*

Base	L	B1	B5	%V_{bur}	ESP	NBO	Molecular Volume
BMTG	6.2	3.06	5.11	65.83	-0.64	-0.36	176.9
BTTP	5.22	3.23	8.06	54	-0.63	-0.5	291.35
DBU	5.27	1.74	7.18	37.8	-0.65	-0.33	146.58
MTBD	5.27	1.76	6.85	38.52	-0.62	-0.38	142.48
P2Et	6.34	2.26	6.45	44.94	-0.96	-0.96	313.7
TEA	4.06	2.51	4.64	44.5	-0.38	-0.27	113.32
TMG	6.05	2.07	4.24	51.45	-0.69	-0.46	114.43

***Table 35**. Summary of selected atoms and axes, and respective descriptors calculated for the ligands.*

Ligand	Target atom or position	Direction	Index	L	B1	B5	Vbur%
Alphos	P	Pd→P	29	6.79	1.7	15.77	79.17
	3	C22→O	24→28	5.55	3.03	11	66.73
	6	C25→H20	27→78	9.15	2.96	11.31	61.19
	4'	C12→C50	14→54	6.8	3.73	8.89	71.55
Ephos	P	Pd→P	22	9.38	4.66	7.77	76.60
	3	C15→O	15→19	5.2	2.91	9.18	67.47
	6	C18→H19	18→57	6.15	2.81	8.96	56.1
	4'	C4→C2	4→2	4.25	3.7	8.58	58
t-BuBrettPhos	P	Pd→P	10	8.16	4.63	7.31	81.48
	3	C2→O1	3→2	4.55	2.91	9.58	69.12
	6	C5→O2	7→6	8.6	3.02	9.32	57.99
	4'	C20→C26	23→29	4.25	4.56	7.39	59.04
t-BuXPhos	P	Pd→P	21	7.94	4.79	7.51	78.30
	3	C20→H27	20→57	4.6	2.76	8.99	59.59
	6	C17→H24	17→54	6	2.7	9.1	55.95
	4'	C4→C2	4→2	4.25	3.55	7.18	58.14

***Table 36**. Dataset generated in-house for case study 2.*

Exp. No:	T / ° C	Res. time / min	Water ratio / %	Base choice	Base eq.	Cat. eq.	Yield / %
1	40	36.2	30	BTMG	1.50	0.1	35
2	137	45	30	BTMG	1.50	0.05	30
3	59	37.8	6	DBU	1.50	0.02	1
4	138	28.1	30	MTBD	1.50	0.05	22
5	128	10.2	30	BTMG	1.50	0.1	27
6	87	30	15	BTMG	1.50	0.05	19
7	122	7.6	15	BTMG	1.50	0.1	33
8	112	23.3	15	MTBD	1.50	0.05	18
9	97	19.8	6	MTBD	1.50	0.1	26
10	146	12.5	30	MTBD	1.50	0.05	20
11	82	21.8	15	BTMG	1.50	0.02	2
12	107	42.5	6	MTBD	1.50	0.02	5
13	74	26.8	15	BTMG	1.50	0.1	53
14	51	6.3	6	BTMG	1.50	0.05	7
15	115	34.1	6	DBU	1.50	0.1	12
16	71	46.4	15	MTBD	1.50	0.05	12
17	34	17.2	30	DBU	1.50	0.05	2
18	62	40.1	30	DBU	1.50	0.1	18
19	90	16	30	BTMG	1.50	0.02	2
20	47	49.8	15	DBU	1.50	0.05	3
21	30	23.3	30	MTBD	1.50	0.05	5
22	95	30.5	15	DBU	1.50	0.02	3
23	48	40.5	15	DBU	1.50	0.1	9
24	84	39.3	30	DBU	1.50	0.05	13
25	71	6.3	30	DBU	1.50	0.1	15
26	126	42.9	30	DBU	1.50	0.02	6
27	79	36.4	6	BTMG	1.50	0.02	2
28	138	42.5	6	DBU	1.50	1	9
29	81	12.1	30	DBU	1.50	0.05	7
30	39	24.5	30	DBU	1.50	0.1	5
31	112	18.2	30	MTBD	1.50	0.02	11
32	150	44.9	6	DBU	1.50	0.02	4
33	97	15.4	15	BTMG	1.50	0.05	34
34	112	32.0	30	DBU	1.50	0.02	5
35	41	14.8	15	DBU	1.50	0.02	1
36	101	16.7	6	DBU	1.50	0.02	3
37	137	18.1	6	DBU	1.50	0.02	3
38	127	11.7	6	DBU	1.50	0.05	7
39	30	49.3	15	DBU	1.50	0.02	0
40	30	5.2	6	DBU	1.50	0.02	0
41	30	41.5	15	DBU	1.50	0.02	0
42	95	48.1	6	DBU	1.50	0.02	3
43	138	37.1	15	DBU	1.50	0.02	5

44	122	5.0	6	DBU	1.50	0.02	4
45	150	9.3	30	DBU	1.50	0.1	32
46	87	26.2	15	DBU	1.50	0.02	3
47	129	49.3	15	BTMG	1.50	0.05	37
48	80	15.1	15	DBU	1.50	0.05	9
49	125	5.0	30	DBU	1.50	0.02	7
50	149	37.7	15	DBU	1.50	30	7
51	59	37.7	30	DBU	1.50	0.02	1
52	150	5.2	6	BTMG	1.50	0.05	17
53	30	5.0	30	DBU	1.50	3	8
54	118	49.8	15	DBU	1.50	3	17
55	89	39.9	30	DBU	1.50	0.1	27
56	85	5.0	6	MTBD	1.50	0.02	2
57	133	49.9	15	DBU	1.50	0.02	6
58	30	6.4	30	DBU	1.50	0.02	0
59	33	10.8	15	DBU	1.50	6	0
60	145	42.2	30	DBU	1.50	0.02	6
61	83	5.0	30	DBU	1.50	0.02	1
62	31	8.7	30	DBU	1.50	0.02	0
63	61	47.8	6	DBU	1.50	0.02	1
64	31	50.0	6	DBU	1.50	3	0
65	150	5.2	15	DBU	1.50	6	3
66	59	36.5	15	BTMG	1.50	0.1	48
67	150	50.0	6	DBU	1.50	0.02	3
68	90	22.7	15	BTMG	1.50	0.02	5
69	149	50.0	6	MTBD	1.50	0.02	7
70	150	15.5	6	MTBD	1.50	0.02	5
71	31	47.4	15	DBU	1.50	0.02	1
72	73	27.2	15	BTMG	1.50	0.1	65
73	73	27.3	15	BTMG	1.50	0.1	60
74	67	33.2	15	BTMG	1.50	0.1	67
75	72	36.1	15	BTMG	1.50	0.1	64
76	63	41.1	15	BTMG	1.50	0.1	63
77	101	31.5	15	BTMG	1.50	0.1	53
78	72	30.6	6	BTMG	1.50	0.1	57
79	73	18.6	15	BTMG	1.50	0.1	56
80	30	34.2	15	BTMG	1.50	0.1	29
81	89	39.2	15	BTMG	1.50	0.1	58
82	66	36.4	6	BTMG	1.50	0.1	62
83	70	36.3	15	BTMG	1.50	0.1	67
84	65	31.8	15	BTMG	1.50	0.1	60
85	68	34.9	15	BTMG	1.50	0.1	59
86	71	50.0	6	BTMG	1.50	0.1	54
87	67	29.9	15	BTMG	1.50	0.1	59
88	70	41.0	15	BTMG	1.50	0.1	64
89	74	38.9	15	BTMG	1.50	0.1	57

90	64	41.1	15	BTMG	1.50	0.1	63
91	79	38.7	15	BTMG	1.50	0.1	59
92	70	36.2	15	BTMG	1.50	0.1	59
93	82	28.4	15	BTMG	1.50	0.1	51
94	147	24.3	15	BTMG	1.50	0.1	31

Chapter 5

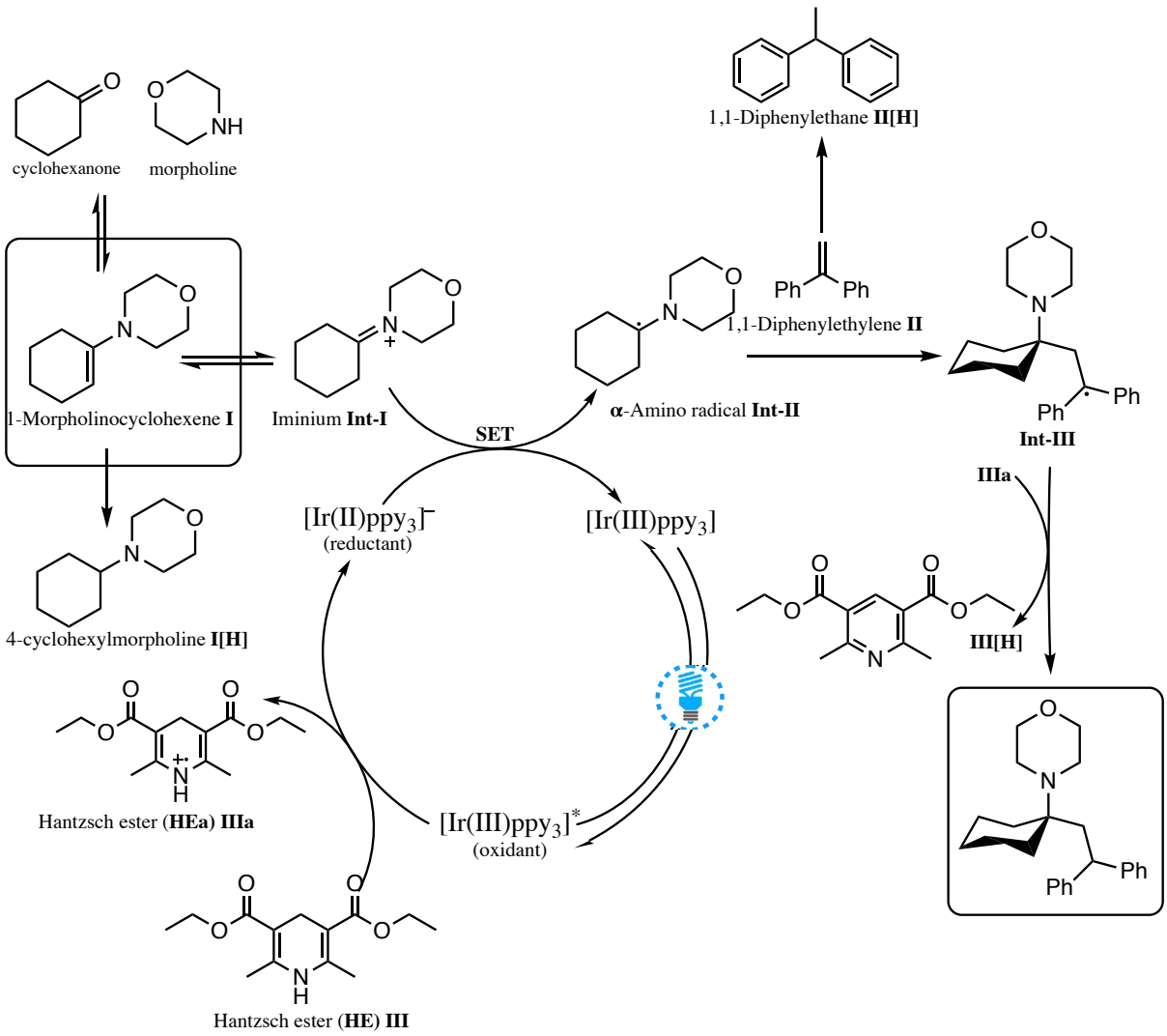
Machine learning-driven optimisation of photoredox amine synthesis in flow

Introduction

Photoredox catalysis

Photoredox catalysis plays a significant role in synthesis of pharmaceutically relevant compounds such as $C(sp^3)$-$C(sp^3)$-rich tertiary amines. Recently, visible-light photocatalysis has seen an increased adoption in academia and industry, owing it to accessible synthesis of pharmaceutically relevant compounds under milder conditions for transformations such as $C(sp^3)$-$C(sp^3)$ bond formation that would normally require harsh conditions, long reactions times, and multi-step synthesis.[1] Visible-light photocatalysis is a process of single-electron transfer (SET) or energy transfer (ET) from an excited state of a photocatalyst, as a result of absorbing low-energy visible light, to organic substrates. For instance, the proposed mechanism for photocatalytic tertiary amine synthesis, starting from an enamine and alkene in the presence of $Ir(ppy)_3$ photocatalyst, shown in Scheme 6 highlights these steps.[2]

Starting with 1-morpholinocyclohexene (**I**), the first step, possibly during the reaction mixture preparation, is acid-catalysed iminium ion intermediate (**Int-I**) formation. Once visible-light reaches catalyst species, photocatalyst $Ir(ppy)_3$ is excited, leading to the long-lived photoexcited $[Ir(III)ppy_3]^*$ oxidant (1.9 μs lifetime),[3] which is quenched by Hantzsch ester (**III**) resulting in $[Ir(II)ppy_3]^-$ reductant and the associated Hantzsch ester radical cation **IIIa**. Combined with iminium ion intermediate (**Int-I**), the species $[Ir(II)ppy_3]^-$ is sufficiently reducing to undergo SET to regenerate the ground state photocatalyst and α-amino radical intermediate (**Int-II**). Added to 1,1-diphenylethylene (**II**), the α-amino radical intermediate leads to the formation C-C bond in the form a benzylic radical **Int-III**. Followed by hydrogen-atom transfer (HAT) with Hantzsch ester radical cation **IIIa**, **Int-III** leads to the formation of final tertiary amine product 4-(1-(2,2-diphenylethyl)cyclohexyl)morpholine.[2]



Scheme 6. *Proposed mechanism for Ir-catalysed photoredox amine synthesis in this project adopted from ref.*[2]

The advancements in the field have mainly been triggered by deeper understanding of the underlying mechanisms as highlighted in Scheme 6, a wider range of photocatalysts, and improved light-emitting diode (LED) technology in terms of energy efficiency, cost, and accurate temperature control. Since the first reported application of photoredox catalysis in organic chemistry more than 40 years ago,[4] the field has been significantly expanded with the developments by academic groups of Jensen,[5, 6] Noel,[7, 8] Gaunt,[2, 9, 10] MacMillan,[11, 12] Stephenson,[13, 14] and others. Despite this progress, with main research focused on discovery of novel molecules and transformations, which are often performed in batch, there is a significant gap between discovery and developing robust and scalable processes. Batch experiments, though convenient and easier to set up, usually require long reaction times (10-50 h), often are not reproducible, and difficult to automate and scale. Specifically, given the sensitivity of

photoredox reactions to water, air, and light, and the identified competing side reactions (Scheme 6), there exists a need to develop a workflow to efficiently generate *a priori* knowledge (e.g., optimal reactor and lamp choices, solubility predictions, featurisation of discrete variables with relevant molecular descriptors) to develop robust and scalable processes.

Continuous flow reactors

Highlighted by Noel and colleagues as nine good reasons for continuous flow photocatalysis, transferring results in batch to continuous flow systems allow for reduced reaction time (due to improved irradiation of the reaction solution), easier scalability, precise control of time and temperature, reproducibility, and faster mixing.[8, 15] This ability to control and accurately tune the physical parameters helps avoid side reactions, allow for hazardous material use, and makes it easier to conduct multi-step and multiphase reactions. Seeberger and colleagues demonstrated dehalogenation of α-chlorophenacylacetates in a continuous-flow microreactor (FEP capillary, 750 μm, 4.7 mL, 100 psi BPR, 17W LED lamps) using $Ru(bpy)_3Cl_2$ photocatalyst.[16] Compared to the results in batch (<50% yield in 24 h), authors achieved 82% yield under 30 mins in flow. Similarly, Noel and co-workers developed a continuous-flow microreactor (PFA capillary tubing, 500 μm ID, 1/16“ OD, 200 μL V_{max}, blue LED) for trifluoromethylation and perfluoroalkylation of five-membered heterocycles. Using gaseous CF_3I for trifluoromethylation reaction, all substrates were converted to respective products in good yield (55-95%) in 8-16 mins (*vs* 12-72 h in batch).[7] A cloud-inspired continuous-flow microreactor development was reported by Khan group.[17] Authors used glass beads (diameter ~75 μm, refractive index n_D=1.52) packed inside PFA reactor tubes to promote light scattering. Under three separate reactor tubing ID (1 mm, 5 mm, and 10 mm) for four different transformations, the glass bead reactor resulted in a superior result compared to the single-phase version under the same reaction conditions.

Self-optimisation algorithms

While use of continuous-flow reactors allows for faster reaction times and more efficient process control, holistic and robust process development of novel transformations is still a laborious and complex task. This is mostly due to the difficulties in identifying the underlying chemical and physical parameters that affect the process objective(s), quantifying the nonlinear interactions between them, and lack of data. In order to address these challenges and accelerate

the optimisation process, recent years has seen an increased implementation of "self-optimisation" systems (i.e., automated experimental systems to perform optimisation of a reaction, and, potentially, also separation, without human intervention).[18] When optimising for a single objective over continuous variables only, several optimisation algorithms such as Nelder-Mead-Simplex,[19] SNOBFIT,[20] and steepest descent[21] have been implemented for different reactions (e.g., nanoparticle synthesis,[22] heterogenous catalytic reactions[23]). With no prior model of the chemical reactions, these "black-box" approaches focus on relationships between input and output variables or objective(s). However, these algorithms are either local search algorithms with limited scalability to larger systems (e.g., Nelder-Mead-Simplex),[24] require expensive derivative estimations and are inaccurate for noisy systems (e.g., steepest descent), or have slow convergence despite successful implementations as a global search algorithm (SNOBFIT).[25, 26] Moreover, most of the published work on self-optimisation are limited to optimising up to five continuous variables for a single objective.[27]

Bayesian optimisation approaches, such as MOAL,[28] TS-EMO,[29] MVMOO,[30] and Google Vizier[31] construct non-parametric statistical models, Gaussian processes, using a sequential active learning approach (re-training a model when new experimental observations become available). Bayesian optimisation approaches utilise all available data to build statistical models and are thus data efficient. The models provide quantification of uncertainty that is used to solve the inherent exploration–exploitation trade-off within the derivative-free optimisation. Extensions of TS-EMO algorithms have been successfully implemented for multi-objective self-optimisation[32-34] and solvent selection.[35, 36] For a holistic and scalable process development, an ideal algorithm should optimise for multiple competing objectives, account for the inherent exploration-exploitation trade-off by iteratively minimising prediction uncertainty, and optimise for continuous and discrete variables simultaneously.

Algorithm development

Nomadic Evolutionary Multiobjective Optimisation (NEMO) is a Bayesian optimisation (BO) algorithm, developed internally in our group, designed to efficiently optimise complex black-box systems for multiple objectives over continuous and discrete variables simultaneously (Figure 28).[37] First, unlike other BO algorithms that employ a single surrogate model (often Gaussian processes), NEMO fits several black-box models – Gaussian processes, neural networks (concrete, Bayesian, etc.), XGBoost, and NGBoost on the training data for each

black-box objective. Then, Latin Hypercube Sampling,[38, 39] a statistical method to sample from a given space subject to a condition of sampling a single point in a given row and column. Once the surrogate models are trained on the training dataset, the model with the best prediction accuracy from step 1 (e.g., XGBoost) is used to predict the outcomes of LHS sampled points. In our implementation, five conditions were selected per iteration to be verified experimentally in the laboratory to accelerate the optimisation process. Compared to some other BO algorithms, which would select the top five conditions based on the highest Expected Hypervolume Improvement (EHVI) of each individual point, which could result in all five points being close to each other (i.e., five suggested reaction conditions being similar to each other), NEMO optimises for the highest EHVI for the combination of points. Moreover, once combination of multiple conditions is selected, the SciPy minimize function utilising the L-BFGS-B algorithm is called to refine the identified points to further increase EHVI. The sampled conditions with the predicted outcome and the associated uncertainty accounts for the exploration and exploitation trade-off for development of a robust and predictive model, instead of potentially getting stuck at, and exploiting, local optima. Finally, suggested conditions are verified experimentally and fed back to the algorithm until optimisation criteria is achieved (Results and Discussion). It is important to highlight that training, hyperparameter optimisation, model selection, and sampling of best conditions are fully automated within NEMO. An overview of NEMO's architecture is given in Figure 28.

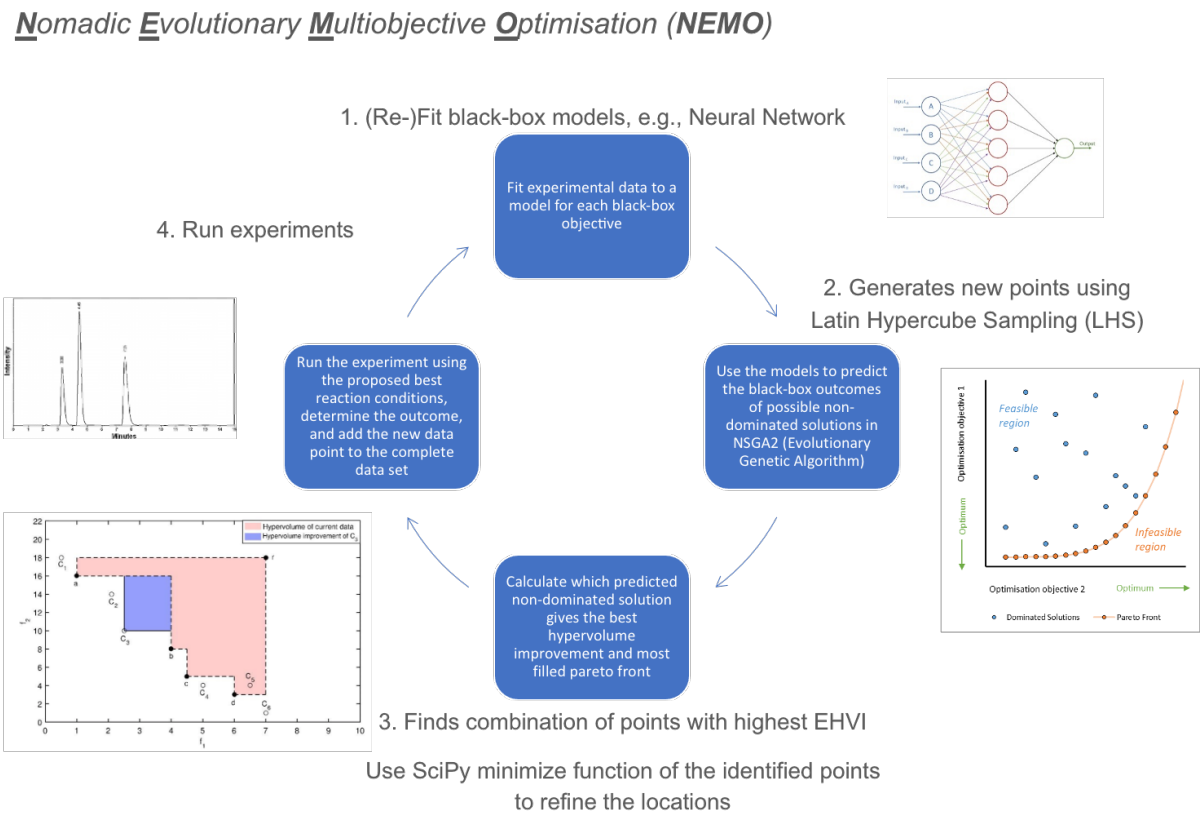


Figure 28. *An overview of the NEMO algorithm.*

Molecular descriptors for solvent selection

When expanded into discrete variables including solvents, reagents, or ligands, most of the self-optimisation algorithms mentioned above struggle with the curse of dimensionality and inefficiencies (in predictive accuracy) of black-box optimisation algorithms. A common approach to overcome this problem has been demonstrated using molecular descriptors to map the discrete variables onto a continuous space. Parameterisation of discrete variables with relevant descriptors could increase the predictive accuracy of surrogate models and unlock new mechanistic insights. Earlier works in describing partition behaviour of solutes in solvents was demonstrated by Zissimos *et al*. on comparing experimental Abraham descriptors with Klampt's COSMOments, often used for quantitative structure-property relationship (QSPR) and linear free energy relationships (LFER), to explain solvation phenomena for data set of 470 compounds.[40] The authors found a large overlap of chemical information content between the two sets and computationally calculated Klampt descriptors could be used to replace experimentally calculated Abraham parameters. In comparison, Sevcik *et al*. used 17 descriptors as an input and was only able to predict experimental Abraham parameter S using multiple linear regression modelling.[41]

Use of Abraham parameters for solvent selection in reactions, instead of QSPR, was demonstrated by Adjiman and colleagues in a computer-aided molecular design (CAMD) study. Starting with six solvents in the training dataset and a linear regression model, the authors found nitromethane to be the best solvent for maximising reaction rate constant in different solvents for Menschutkin reaction after five iterations.[42] Amar *et al*. expanded the descriptor space to 17 to optimise for high conversion (>90%) and high diastereoselectivity (>60%) in asymmetric hydrogenation reaction to synthesize Brivaracetam, a new anti-epileptic drug produced by UCB.[35] Out of 34 solvents explored based on human intuition and experience, only the outcome of a single solvent met both objectives whilst 7 out of 18 solvents suggested by the TS-EMO algorithm[29] matched both objectives, proving the capability of the algorithm to handle discrete variables and its use for solvent selection. While this approach worked well for identifying ideal solvent candidates, all the continuous variables such as temperature and concentrations of substrates were kept constant during the solvent screening. Similarly, Chonghuan *et al*. used 15 descriptors for solvents in a Mitsunobu transesterification reaction to produce isopropyl benzoate.[36] Using Autoencoder for dimensionality reduction and artificial neural networks (ANNs) as a surrogate model to design the experiments, the final surrogate model identified 1-chloropentane (93% yield) as a promising solvent for the single objective of maximising for reaction yield.

Even though the use of larger number of descriptors could provide more information about solvents (for better differentiation), the concept of "plethora of descriptors"[40] comes at the cost of sampling from high dimensional space and potentially introducing (reaction) irrelevant descriptors. Additionally, dimensionality reduction techniques such as Principal Component Analysis (PCA) or Autoencoder could reduce the descriptor space with the risk of losing solvability and reaction relevant descriptors. Moreover, despite the successful demonstration of algorithm-guided solvent selection for certain transformations, to the best of the authors knowledge, reaction optimisation for discrete, specifically solvents, and continuous variables simultaneously has not been demonstrated for photoredox amine synthesis.

The theory behind the selected descriptors (i.e., sigma moments) and the methodology to calculate them are provided in detail in Chapter 3.

Introduction to the case study

In this work, we demonstrated machine learning-driven optimisation of photoredox tertiary amine synthesis (Scheme 6) for six continuous variables (e.g., concentration, temperature, residence time) and 20 solvents, parameterised using five sima moments, in semi-automated continuous-flow setup. Starting with 120 solvents, the workflow (Figure 29) included multiple steps of *a priori* knowledge generation (e.g., solubility predictions and measurements, UV-Vis and actinometry studies) to narrow down the discrete space. A novel Bayesian optimisation algorithm, Nomadic Evolutionary Multiobjective Optimisation (NEMO), was used to identify and populate the Pareto front for reaction objectives - yield and cost.

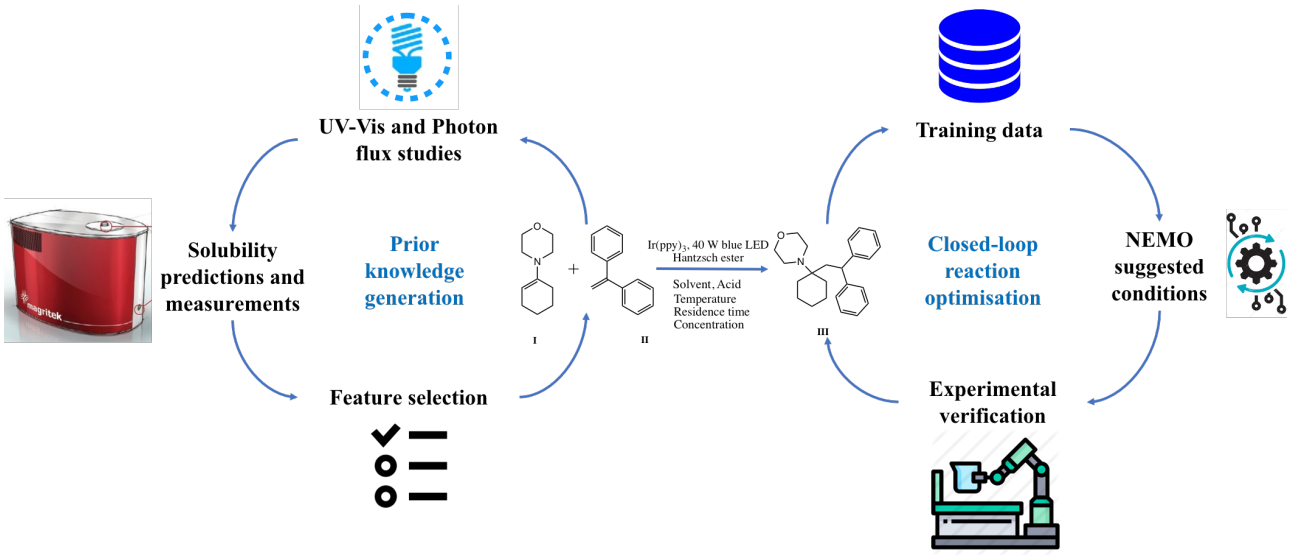


Figure 29. *An overview of the workflow used in generating a priori knowledge and optimising the reaction in this project.*

Contributions

Contributions to the chapter is as follows. Dr. Simon Sung and Dr. Mohammed Jeraal developed the Bayesian optimisation algorithm NEMO used in this project. Dr. Danilo Russo advised on the conceptual development of the project. All the optimisation and solubility measurement experiments, method development, analysis, prior knowledge generation studies (e.g., UV-Vis, solubility predictions, actinometry studies), model training, and feature analysis were carried out by me.

Methods and Materials

Materials

Reagents 1-morpholino-1-cyclohexene, >97%; 1,1-diphenylethylene, 98%; diethyl 1,4-dihydro-2,6-dimethyl-3,5-pyridinedicarboxylate, >98%; tris(2-phenylpyridinato)iridium(III), >98% were purchased from TCI Chemicals and used as received. 1,2-Dichloroethane, anhydrous, 99.8%; 1,3-dimethyl-2-imidazolidinone, absolute, ≥99.5%; 1,3-dimethyl-3,4,5,6-tetrahydro-2(1*H*)-pyrimidinone, absolute, ≥99.0%, furfuryl alcohol, 98%; tetrahydrofurfuryl alcohol, 99%; 2-methyltetrahydrofuran, anhydrous, ≥99.0%; 2-propanol, anhydrous, 99.5%; acetonitrile, anhydrous, 99.8%; benzyl alcohol, anhydrous, 99.8%; N,N-dimethylformamide, anhydrous, 99.8%; dimethyl sulfoxide, anhydrous, ≥99.9%; cyclohexanone, *ReagentPlus*, 99.8%; ethanol, anhydrous, ≥99.5%; ethyl acetate, anhydrous, 99.8%; N,N-dimethylacetamide, anhydrous, 99.8%; 1-methyl-2-pyrrolidinone, anhydrous, 99.5%; acetone, HPLC plus, ≥99.9%; tetrahydrofuran, anhydrous, ≥99.9%; dichloromethane, anhydrous, ≥99.8%; propionic acid, ACS reagent, ≥99.5%; naphthalene, 99% ; mesitylene, 98% were purchased from Sigma Aldrich and used as received. Non-anhydrous solvents were bubbled with nitrogen for 45 minutes and stored over activated molecular sieves in a Schlenk tube. For a fair comparison, cost of solvents was calculated based on the 2 L solvent cost (unless not available by a supplier) in the cost objective function.

Procedures

Experiments were carried out in flow using Vapourtec R2 Series in 10 mL UV-150 reactor. Given the solubility and mixing issues highlighted in Results and Discussion, individual reaction solutions were prepared inside glovebox before each reaction and loaded directly from the vial using Gilson 271 Liquid Handler. The solids – Hantszch ester (diethyl 1,4-dihydro-2,6-dimethyl-3,5-pyridinedicarboxylate), catalyst (*fac*-Ir(ppy)$_3$), and internal standard (naphthalene, 0.02 mmol) were stored in glovebox and the relative amounts were added to a 10 mL (crimp) vial. 0.2 mmol (34 μL) of 1-morpholino-1-cyclohexene **I** was added with relative amounts of 1,1-diphenylethylene **II** (0.2 – 0.4 mmol), propionic acid (0.02 – 0.2 mmol) and the respective anhydrous solvent to prepare 5 mL final solution. Vial was closed, wrapped with parafilm, and transferred to a sonicator bath to promote faster dissolving of the solids (< 5 mins). Before pumping the solution to the reactor, inert nitrogen line was connected to the vial to replace the pumped solution and avoid air. Reactions were conducted in UV-150 reactor

with 470 nm LED lamp and 6 bar BPR for the given residence time and temperature. External chiller was used to achieve lower temperatures. Reaction crude was collected and 25 μL of crude reaction mixture was diluted to 1.0 mL in acetonitrile and analysed using HPLC. Experimental setup is given in Figure 32 in the Appendix. Full data generated in the optimisation is given in Table 44 in the Appendix.

Analytical Methods

Dissolved Hantszch ester quantity in 38 solvents was measured using Magritek Spinsolve 60 ULTRA Benchtop NMR with mesitylene as an internal standard. The analysis condition was conducted using 1D EXTENDED+ at 2 scans, 6.4 s acquisition time, 2 min repetition time, and 90 pulse angle.

Reaction composition was analysed using HPLC (Shimadzu LC-20A, D2 lamp with PDA detector; Eclipse Plus C18, 95Å, 3.0 x 100 mm, 3.5 μm column). Injection volume of 5 μL, oven temperature of 30 °C, and total flowrate of 1.0 mL min^{-1} with acetonitrile/water (45%/55%) was found to be optimal. Acetonitrile concentration was increased to 80% in 17 mins, then to 98% in in a minute, held at 98% for another minute before reducing it to 50% in a minute and holding it at 50% for another minute. The product *vs* internal standard calibration plot is provided in Figure 33 in the Appendix.

Results and Discussion

Prior knowledge generation

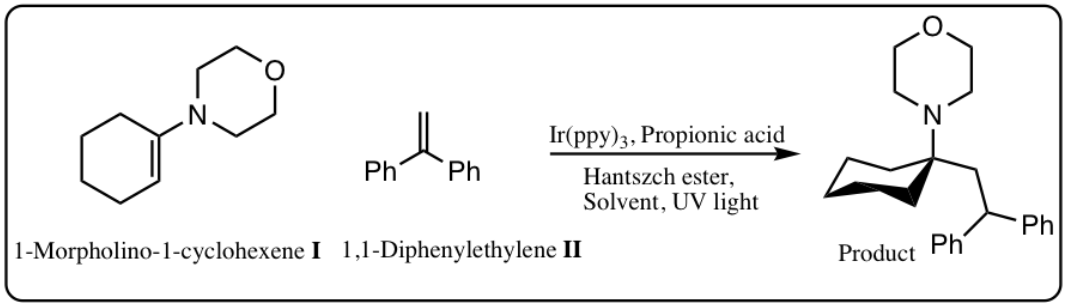
Identification and parameterisation of reaction space

Lower and upper bounds for the six continuous reaction variables are provided in Table 37. 1-morpholine-1-cyclohexene **I** (40 mM) was used as the liming reactant and equivalence of 1.0 – 2.0 was used for 1,1-diphenylethylene **II** and Hantzsch ester (HEH) based on the proposed reaction mechanism (Scheme 6)[2]. If Hantzsch ester solubility in a particular solvent was less than 80 mM (i.e., 2.0 eq. for Hantzsch ester), then the solubility value was used as the upper limit for Hantzsch ester to maintain a clear solution for higher light penetration. For instance, 1.51 eq. was used as the upper limit for Hantzsch in 2-propanol (IPA) given solubility for HEH in IPA was measured to be 60.47 mM (Table 40, Appendix). Clear reaction solution is important for achieving high and uniform light penetration to promote faster reactions and to

avoid light scattering, which introduces an unknown parameter in the model given the variations in the particle size and quantity in different solvents, and for the algorithm to learn the relevant descriptors (i.e., intrinsic solvent properties). Lower and upper bounds of 0.5 - 5.0 mol% was selected for the catalyst $Ir(ppy)_3$ based on the UV-vis absorption spectra measurements (Figure 35, Appendix), solubility in various solvents,[43] and the range reported in literature[7,13]. Due to its catalytic role for reaction initiation and iminium ion (**Int-I**) intermediate formation, propionic acid range was set as 0.1 – 1.0 equivalence. The ranges for time and temperature were set based on the preliminary data and the reactor capability, respectively.

Starting with an initial list of 115 solvents, the discrete variable space is reduced based on several criteria (e.g., removal of inherently basic solvents such as pyridine given the reaction is acid catalysed) as given in Table 37. Hantzsch ester and $Ir(ppy)_3$ catalyst solubilities were the most important criteria in selecting candidate solvents. Even though suspensions can be pumped using the Vapourtec peristaltic pump, presence of solids blocks and scatters light. Therefore, solvents that can achieve clear solution in 40 mM reaction scale was selected based on the measured solubility values with the full list provided in Table 45, Appendix. As explained in Introduction, five sigma moments were used for parameterisation of solvents. Since each descriptor value and range are different, the descriptors were first standardised before implementing PCA. The solvent descriptors were reduced to two and three principal components retaining 78.6% and 93.5% information, respectively (Figure 34, Appendix). Moreover, contribution of each of the original five descriptors to new components are given in Figure 34 in Appendix. For example, first component is highly dominated by electrostatic polarity, asymmetry of sigma profile, and hydrogen bond acceptor strength. The second and the third components are mainly defined by hydrogen bond donor strength and surface area, respectively, showing that all the five components are included in the 3-component space.

Table 37. *Reaction scheme, the lower and upper bounds for reaction parameters.*



Variable	Lower bound	Upper bound
1,1-Diphenylethylene (**II**) / eq.	1.0	2.0
Hantzsch ester / eq.	1.0	2.0
fac-Ir(ppy)$_3$ / mol%	0.5	5.0
Propionic acid / eq.	0.1	1.0
Temperature / C	10	50
Residence time / min	4	60
Solvent descriptor	**Meaning**	
Sig0	Surface area	
Sig2	Electrostatic polarity	
Sig3	Asymmetry of sigma profile	
Hb_acc	Hydrogen bond acceptor strength	
Hb_don	Hydrogen bond donor strength	
Criteria	**Example**	**Number of solvents**
Liquid at 10 °C	Sulfalone	111
Basic	Pyridine	108
Water sensitive	Water	107
(Strongly) acidic	Triflic acid	101
UV cut off limit	None	101
Solubility criteria	Methylformate	20

UV-Vis and photon flux studies

Beer-Lambert law (*eq. 1*) states that light penetration decreases exponentially over a distance (l) for a given molar extinction coefficient (ε) at a certain concentration (c), suggesting that reactions might only take place at the surface (of the reaction medium), leaving a significant area of "dark" zones inside the reaction medium that could lead to quenching of excited states or side reactions, especially if the reaction medium is limited by photon flux. Quantifying the influence of photon flux was demonstrated by Corcoran et al. when transferring optimal conditions to larger scale reactors.[44] Comparing 10, 60, and 150 mL reactors, the authors found that reaction yield versus absorbed photon equivalents followed the same line for all three reactors, suggesting that reaction is driven by absorbed photon equivalence, not residence time alone. Similar study was reported by Lévesque et al. over the lamp choices.[45] For lamp intensities of 30, 60, and 82.5 W, product yield versus equivalents of emitted photons followed

the same trajectory for three different lamps. Both of these studies highlight the importance of operating under conditions that is not limited by photon flux.

$$A = \log_{10} T = \log_{10} \frac{I_0}{I} = \varepsilon c l \qquad (1)$$

In this study, three LED lamps from Vapourtec (365 nm, 420 nm, and 470 nm) were tested. Based on a preliminary data, the highest yield was achieved with 470 nm lamp under fixed conditions. Based on the UV-Vis absorption spectra (Figure 35, Appendix) measured from 200 nm to 800 nm, the *fac*-Ir(ppy)$_3$ catalyst absorbs under all three lamps with measured excitation coefficient of 1413.5 M^{-1} cm^{-1} at 470 nm. Although both of the reactants absorb around 310-320 nm, and hence does not affect light penetration, Hantzsch ester absorbs at 425 nm, affecting light penetration under the LEDs with the wavelength of 365 nm and 420 nm. Moreover, using lower energy lamps could avoid potential side reactions as it is less likely for other reaction species to absorb under longer wavelength. Combining preliminary data with the UV-Vis absorption spectra, 470 nm LED lamp was used for optimisation. Under 470 nm lamp, 1mM photocatalyst concentration in 1 mm tubing ID results in absorbance of 0.14. This allows for a theoretical scalability to 10 mm ID (i.e., 100x volume increase from 10 mL reactor to 1 L scale) or increase of catalyst concentration to 14 mM (Abs = 2) whilst maintaining a clear solution without sacrificing on the uniform light absorption.

A model to calculate the amount of photon radiated to the reaction medium is important to choose the appropriate reactor type and light source in accordance with the objectives (e.g., space time yield maximization) the reaction is evaluated at.[46-48] Photon flux received in Vapourtec UV-150 photochemical reactor was studied using potassium ferrioxalate actinometer (Figure 36, Appendix). A model was developed to calculate absorption depending on the reactor depth, light source, wavelength and intensity, and irradiation (residence) time based on the actinometer (potassium ferrioxalate) conversion to ferrous ion (Fe^{+2}). Following the ferrous ion calibration (Figure 37, Appendix) and the model equation (equation 7 and 8, Appendix), the standard 1 mm ID Vapurtec tubing in flow resulted in 30.9x more photon flux (Einstein s^{-1}) or 6.18x actinometric intensity of absorbed photon (Einstein s^{-1} L^{-1}) over Kessil Blue lamp in batch (4 mL vial) as summarised in Table 41 in the Appendix. Experimental procedure, measurement techniques, and equations are reported in detail in the Appendix.

Solubility predictions using COSMOtherm and measurements using benchtop NMR

Previous sections highlight the importance of working in a transparent region to optimise for photon flux in the reaction medium. Running reactions as a suspension, though manageable in terms of hardware, significantly reduces light penetration and makes it difficult to quantify light scattering, thus increasing the risk of irreproducibility. For a reductive deiodination reaction, Nguyen et al. reported 750% scale up of the reaction whilst simultaneously decreasing the *fac*-Ir(ppy)$_3$ catalyst loading by 2,000% and Hantzsch ester equivalence from 2.0 to 1.1 eq. without a significant loss in reaction efficiency. Although attributed to serendipity, this is likely due to the presence of catalyst and Hantzsch ester in solid forms in the reaction. At the smaller scale, the catalyst concentration in the reaction corresponded to 2.5 mM whilst it was only 0.38 mM in the larger scale. Considering the measured solubility of 0.41 mM for Ir(ppy)$_3$ in acetonitrile,[43] the authors conducted the larger scale reaction under fully transparent conditions, promoting higher light penetration. Therefore, in our study, experimental solubility results reported by Jesperson *et al*. was used to decide the upper limit for *fac*-Ir(ppy)$_3$ (Table 43, Appendix) in case algorithm suggested values were above the solubility limit.[43]

Solubility values for Hantzsch ester were predicted using COSMO*therm* for all 107 solvents from the original list.[49] The software requires a reference list for solvents based on experimental measurements. Although measuring *via* solvent drying method is simpler, it mostly works well with volatile solvents. For this reason, benchtop NMR with mesitylene as an internal standard was used for solubility measurements. Both approaches led to similar results, confirming the use case of either approach to generate solubility data (Table 45, Appendix). COSMO*therm* solubility prediction accuracies were compared using 9, 18, and 35 references. Although prediction accuracy improved with more experimentally measured references, predicted values using COSMO*therm* were used as a qualitative guide to select the solvents to be verified experimentally. Using 40 mM reaction concentration as a criteria and based on the measured solubility values, 20 solvents were selected to be used in the final solvent candidates list for optimisation. For certain solvents, solubility limit was used as the upper bound in running the reaction to achieve a clear solution. Procedure for both measurement approaches (Figure 38), accuracy comparisons (Table 42), and full solubility data (Table 45), both predicted and measured, are provided in the Appendix.

Training dataset collection and algorithm-guided optimisation

Latin Hypercube sampling (LHS) was implemented separately in the 3-component space to obtain eight solvents to generate the training dataset for the NEMO algorithm. Objective of the training dataset was to provide as much information about the reaction to the algorithm as possible. This is why several iterations of sampling were repeated to then decide the best set of solvents (Table 38). Similarly, LHS was implemented on six continuous variables with lower and upper bounds (Table 37) as a range to be sampled from. Six conditions per solvent were selected and conducted as a training dataset generation. List of solvents used in training dataset with respective conditions for the highest and the lowest yield in each solvent is given in Table 38. NEMO was trained on the dataset to suggest five conditions per iteration based on the highest EHVI for the combination of five conditions. In terms of describing solvents, both five sigma moments and three principal components generated by applying PCA on five sigma moments were benchmarked. XGBoost was selected as the best model for yield using three principal components with test RMSE of 3.66% and test R^2 of 0.77. When using five sigma moments, NGBoost was selected as the best model for yield with model performance of test RMSE = 3.03% and test R^2 of 0.84. Therefore, optimisation was conducted using six continuous variables and five sigma moments to describe solvents.

***Table 38**. Part of the training dataset with highest and lowest values for yield in each solvent. DMF: N,N-dimethylformamide, DMSO: dimethyl sulfoxide, NMP: 1-methyl-2-pyrrolidinone, DCM: dichloromethane, THFA: tetrahydrofurfuryl alcohol, EA: ethyl acetate.*

Entry	Alkene eq.	Ir(ppy)3 mol%	HE eq.	Acid eq.	T / C	Time / min	Solvent	Cost / £	Yield / %
1	1.81	1.14	1.56	0.27	28	68	DMF	1.40	26
2	1.19	2.33	1.81	0.72	43	5.5	DMF	2.05	5
3	1.81	1.14	1.56	0.27	28	68	DMSO	1.96	8
4	1.19	2.33	1.81	0.72	43	5.5	DMSO	2.62	2
5	1.81	1.14	1.24	0.27	28	68	Acetone	1.25	3
6	1.31	1.50	1.24	0.94	38	38	Acetone	1.42	0
7	1.69	2.92	1.44	0.49	35	113	NMP	2.38	19
8	1.19	2.33	1.60	0.72	43	5.5	NMP	2.09	7
9	1.70	3.05	1.44	0.50	33	25	DCM	2.19	71
10	1.06	1.45	1.07	0.60	48	53	DCM	1.25	19
11	1.69	2.92	1.35	0.49	35	113	Cyclohexanone	2.08	10
12	1.19	2.33	1.35	0.72	43	5.5	Cyclohexanone	1.75	0
13	1.56	3.05	1.19	0.38	18	98	THFA	2.19	0
14	1.44	0.55	1.28	0.83	23	23	THFA	0.89	0
15	1.94	0.85	1.94	0.16	13	83	EA	1.17	0
16	1.56	0.85	1.19	0.38	18	98	EA	1.01	0

Conditions suggested by NEMO (orange points, Figure 30a) during the optimisation had combination of new solvents (e.g., acetonitrile), low yield conditions, existing solvents (e.g., DCM), and high yield conditions, suggesting a balance between exploration and exploitation. For purely exploitatory algorithms, a selected model could have suggested conditions near the best result (e.g., 71% yield) provided in the training data (Table 38, row=9). Over a few iterations, algorithm suggested conditions focused on and around the experimentally identified Pareto front. After 5 iterations (25 data points), NEMO suggested conditions had 5 points on the experimentally identified Pareto versus 3 points found in the training dataset of 48 points, highlighting the efficiency of NEMO to identify and populate the Pareto front.

In terms of stopping criteria, three conditions were used – Pareto front population, hypervolume improvement, and model predictive accuracy. When the optimisation was initiated with 48 data points, initial learning curve was the steepest at the start, suggesting new conditions were sampled based on the highest EHVI (Figure 39a, Appendix). Over 5 iterations, hypervolume improvement converged, suggesting the model has minimised uncertainties in unknown areas and achieved a high prediction accuracy. This was also validated with test RMSE value when 80/20 split was implemented with 5-fold cross-validation (CV). RMSE for

predicted values for yield was 3.7%, a low number considering 1.80% experimental error from HPLC analysis (Figure 33, Appendix), and the R^2 score of 0.96. Moreover, visual analysis of the Pareto front confirms that most of the NEMO suggested conditions were on or around the Pareto front, compared to low yield and high-cost conditions in the training dataset. All three conditions implemented as a stopping criteria confirm that optimal conditions were found with a highly predictive model after balancing for exploration and exploitation, and the Pareto front to find trade-offs between the objectives reaction yield and cost. Full data generated during the optimisation is given in Table 44 in the Appendix.

Benchmarking NEMO using pool-based sampling

In order to benchmark the learning efficiency of NEMO, the algorithm performance was evaluated using different dataset sizes. During the sampling, NEMO was forced to sample one condition per iteration from the pool of experimentally validated dataset based on highest EHVI. In the first benchmark, NEMO was trained on 16 data points, two conditions per each of eight solvents (Figure 30c). The training dataset included a point on the Pareto front (Table 38, row 9). During the optimisation, most of the points sampled by NEMO (orange points) were on or around the Pareto front, highlighting the learning efficiency of NEMO. This was reflected on the hypervolume learning over experiment number (Figure 39b, Appendix). After 20+ sampled points, hypervolume improvement has converged, suggesting that NEMO maximised the hypervolume learning. In the second benchmark, NEMO was trained using 14 data points, two points per each of seven solvents, excluding DCM to avoid presence of any points on the Pareto front (Figure 30d). Regardless, most of the sampled points were on or around the Pareto front and the hypervolume improvement was maximised in 20+ sampled points. Using small dataset size in both cases, with or without Pareto front points in the training dataset, NEMO maximised the hypervolume learning and populated the Pareto in total of ~40 experiments.

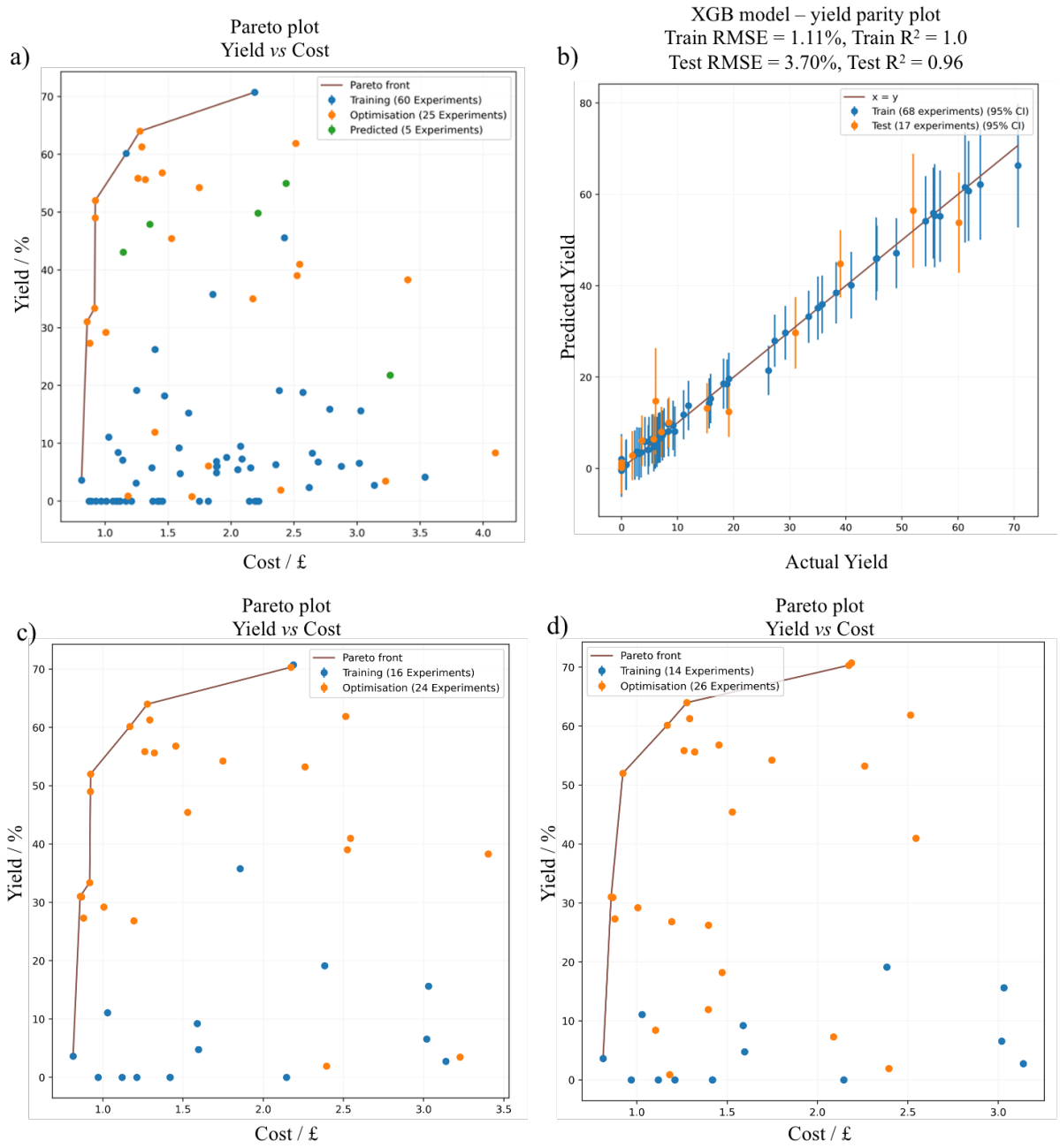


***Figure 30**. a) Optimisation plot (yield vs cost) and Pareto front population using NEMO. b) Parity plot for actual vs predicted yield using the best model. Pool-based sampling and benchmarking NEMO with (c) 16 and (d) 14 training data points.*

Explainable AI

Opening black-box models to extract insights and learn or validate physical knowledge about the reaction has gathered attention in the concept of explainable AI to bridge the gap between physical models and black-box models.[50, 51] We implemented two approaches – partial dependence plot (PDP) and permutation feature importance (PFI) to analyse the sensitivity of

objectives, yield and cost, to continuous variables and solvent descriptors. In PFI, importance of a feature is calculated based on the model score when that feature is randomly shuffled.[52] PDP, on the other hand, shows whether a relationship between a feature or two features with the target is linear, monotonic or more complex.[53] Both approaches identified sig3, asymmetry of σ-profile, as the most important variable for yield (Figure 31). Lower values of sig3 corresponded to higher yield, identifying dichloromethane (DCM) and 1,2-dichloroethane (DCE) as the best solvents amongst 20 candidates included in optimisation. Solvent molecular area was selected as the second most important descriptor for solvents based on permutation feature importance. In terms of continuous variables, Hantzsch ester equivalence and 1,1-diphenylethylene **II** equivalence had the highest feature importance. This was followed by residence time and *fac*-Ir(ppy)$_3$ catalyst mol%. It is important to note that correlated features could have a low PFI or PDP values since the influence of one parameter could be masked by the other. For instance, higher yield can be achieved at low catalyst mol% with a longer residence time since the reaction is driven by absorbed photon equivalence, which is the product of residence time and dissolved catalyst quantity. Acid equivalence and temperature had small influences on yield, with higher values correlated with a relatively higher yield. Cost is an analytical function and was influenced by catalyst quantity the most, followed by the solvent cost. Partial dependence plots for all individual variables are provided in Figure 40-41 in the Appendix.

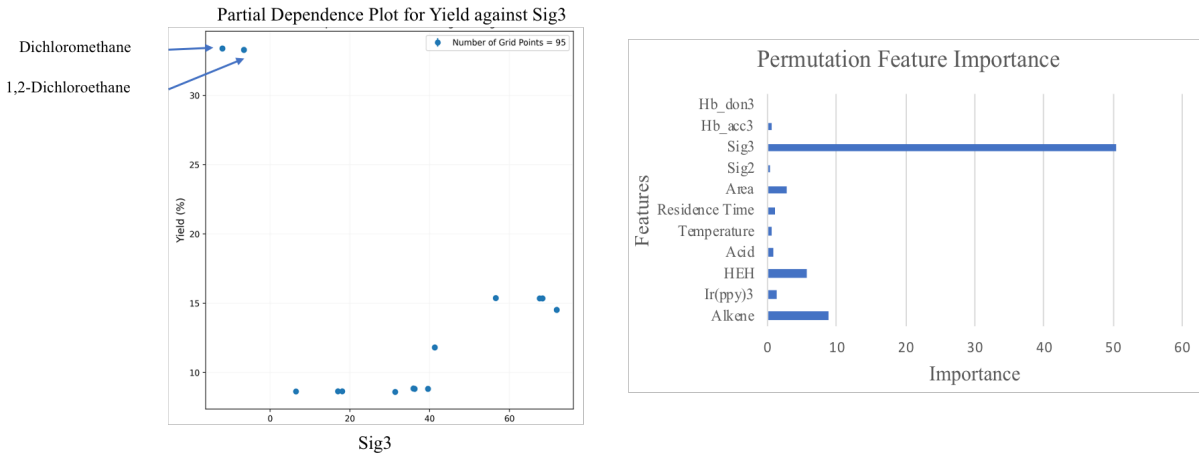


***Figure 31**. Partial dependence plot for Sig3, asymmetry of σ-profile, against yield (left) and permutation feature importance of all variables for yield (right).*

Conclusions

A workflow to generate *a priori* knowledge and developing a robust process in a semi-automated manner using Bayesian optimisation algorithm has been demonstrated for photoredox amine synthesis. Starting with 115 solvent candidates for discrete variables, the solvent candidates were shortlisted to 20 based on multiple screening criteria. UV-Vis and photon flux studies provided a guideline for optimal reactor, lamp, and tubing selection. Solubility predictions were analysed using COSMO*therm* software. Although quantitative values of the predictions were not used in optimisation, qualitative ranking of solvents based on the predicted solubility values served as a guideline to experimentally validate solubility values for Hantzsch ester using benchtop NMR. A recently developed algorithm NEMO was used to build a robust model, using five black-box surrogate models and sampling based on highest EHVI for combination of suggested conditions. A Pareto front was identified for the objectives, reaction yield and cost, and was populated based on recommended conditions by NEMO. Although the reaction was not directly optimised for productivity, the highest productivity value achieved equals to ~12g/day scale up using Vapourtec and was ~25x higher than the optimal batch result achieved in-house, with the theoretical scale up of reactor from 10 mL to 1 L without sacrificing on light penetration.

References

1. Li, P.; Terrett, J. A.; Zbieg, J. R., Visible-Light Photocatalysis as an Enabling Technology for Drug Discovery: A Paradigm Shift for Chemical Reactivity. *ACS Medicinal Chemistry Letters* **2020,** *11* (11), 2120-2130.
2. Trowbridge, A.; Reich, D.; Gaunt, M. J., Multicomponent synthesis of tertiary alkylamines by photocatalytic olefin-hydroaminoalkylation. *Nature* **2018,** *561* (7724), 522-527.
3. Flamigni, L.; Barbieri, A.; Sabatini, C.; Ventura, B.; Barigelletti, F., Photochemistry and Photophysics of Coordination Compounds: Iridium. In *Photochemistry and Photophysics of Coordination Compounds II*, 2007; pp 143-203.
4. Van Bergen, T. J.; Hedstrand, D. M.; Kruizinga, W. H.; Kellogg, R. M., Chemistry of dihydropyridines. 9. Hydride transfer from 1,4-dihydropyridines to sp3-hybridized carbon in sulfonium salts and activated halides. Studies with NAD(P)H models. *The Journal of Organic Chemistry* **2002,** *44* (26), 4953-4962.
5. Hsieh, H.-W.; Coley, C. W.; Baumgartner, L. M.; Jensen, K. F.; Robinson, R. I., Photoredox Iridium–Nickel Dual-Catalyzed Decarboxylative Arylation Cross-Coupling: From Batch to Continuous Flow via Self-Optimizing Segmented Flow Reactor. *Organic Process Research & Development* **2018,** *22* (4), 542-550.
6. Pomberger, A.; Mo, Y.; Nandiwale, K. Y.; Schultz, V. L.; Duvadie, R.; Robinson, R. I.; Altinoglu, E. I.; Jensen, K. F., A Continuous Stirred-Tank Reactor (CSTR) Cascade for Handling Solid-Containing Photochemical Reactions. *Organic Process Research & Development* **2019,** *23* (12), 2699-2706.
7. Straathof, N. J. W.; Gemoets, H. P. L.; Wang, X.; Schouten, J. C.; Hessel, V.; Noël, T., Rapid Trifluoromethylation and Perfluoroalkylation of Five-Membered Heterocycles by Photoredox Catalysis in Continuous Flow. *ChemSusChem* **2014,** *7* (6), 1612-1617.
8. Cambié, D.; Bottecchia, C.; Straathof, N. J. W.; Hessel, V.; Noël, T., Applications of Continuous-Flow Photochemistry in Organic Synthesis, Material Science, and Water Treatment. *Chemical Reviews* **2016,** *116* (17), 10276-10341.
9. Henry Blackwell, J.; Harris, G. R.; Smith, M. A.; Gaunt, M. J., Modular Photocatalytic Synthesis of α-Trialkyl-α-Tertiary Amines. *Journal of the American Chemical Society* **2021,** *143* (39), 15946-15959.
10. Deneny, P. J.; Kumar, R.; Gaunt, M. J., Visible light-mediated radical fluoromethylation via halogen atom transfer activation of fluoroiodomethane. *Chemical Science* **2021,** *12* (38), 12812-12818.
11. Prier, C. K.; Rankic, D. A.; MacMillan, D. W. C., Visible Light Photoredox Catalysis with Transition Metal Complexes: Applications in Organic Synthesis. *Chemical Reviews* **2013,** *113* (7), 5322-5363.
12. Shaw, M. H.; Twilton, J.; MacMillan, D. W. C., Photoredox Catalysis in Organic Chemistry. *The Journal of Organic Chemistry* **2016,** *81* (16), 6898-6926.
13. Nguyen, J. D.; D'Amato, E. M.; Narayanam, J. M. R.; Stephenson, C. R. J., Engaging unactivated alkyl, alkenyl and aryl iodides in visible-light-mediated free radical reactions. *Nature Chemistry* **2012,** *4* (10), 854-859.

14. Narayanam, J. M. R.; Stephenson, C. R. J., Visible light photoredox catalysis: applications in organic synthesis. *Chem. Soc. Rev.* **2011,** *40* (1), 102-113.
15. Plutschack, M. B.; Pieber, B.; Gilmore, K.; Seeberger, P. H., The Hitchhiker's Guide to Flow Chemistry. *Chemical Reviews* **2017,** *117* (18), 11796-11893.
16. Bou-Hamdan, F. R.; Seeberger, P. H., Visible-light-mediated photochemistry: accelerating Ru(bpy)32+-catalyzed reactions in continuous flow. *Chemical Science* **2012,** *3* (5).
17. Zheng, L.; Xue, H.; Wong, W. K.; Cao, H.; Wu, J.; Khan, S. A., Cloud-inspired multiple scattering for light intensified photochemical flow reactors. *Reaction Chemistry & Engineering* **2020,** *5* (6), 1058-1063.
18. Fabry, D. C.; Sugiono, E.; Rueping, M., Online monitoring and analysis for autonomous continuous flow self-optimizing reactor systems. *Reaction Chemistry & Engineering* **2016,** *1* (2), 129-133.
19. McMullen, J. P.; Jensen, K. F., Integrated Microreactors for Reaction Automation: New Approaches to Reaction Development. *Annual Review of Analytical Chemistry* **2010,** *3* (1), 19-42.
20. Jeraal, M. I.; Holmes, N.; Akien, G. R.; Bourne, R. A., Enhanced process development using automated continuous reactors by self-optimisation algorithms and statistical empirical modelling. *Tetrahedron* **2018,** *74* (25), 3158-3164.
21. Moore, J. S.; Jensen, K. F., Automated Multitrajectory Method for Reaction Optimization in a Microfluidic System using Online IR Analysis. *Organic Process Research & Development* **2012,** *16* (8), 1409-1415.
22. Krishnadasan, S.; Brown, R. J. C.; deMello, A. J.; deMello, J. C., Intelligent routes to the controlled synthesis of nanoparticles. *Lab on a Chip* **2007,** *7* (11).
23. Ley, S. V.; Fitzpatrick, D. E.; Ingham, R. J.; Myers, R. M., Organic Synthesis: March of the Machines. *Angewandte Chemie International Edition* **2015,** *54* (11), 3449-3464.
24. Lagarias, J. C.; Reeds, J. A.; Wright, M. H.; Wright, P. E., Convergence Properties of the Nelder--Mead Simplex Method in Low Dimensions. *SIAM Journal on Optimization* **1998,** *9* (1), 112-147.
25. McMullen, J. P.; Jensen, K. F., An Automated Microfluidic System for Online Optimization in Chemical Synthesis. *Organic Process Research & Development* **2010,** *14* (5), 1169-1176.
26. Holmes, N.; Akien, G. R.; Savage, R. J. D.; Stanetty, C.; Baxendale, I. R.; Blacker, A. J.; Taylor, B. A.; Woodward, R. L.; Meadows, R. E.; Bourne, R. A., Online quantitative mass spectrometry for the rapid adaptive optimisation of automated flow reactors. *Reaction Chemistry & Engineering* **2016,** *1* (1), 96-100.
27. Mateos, C.; Nieves-Remacha, M. J.; Rincón, J. A., Automated platforms for reaction self-optimization in flow. *Reaction Chemistry & Engineering* **2019,** *4* (9), 1536-1544.
28. Echtermeyer, A.; Amar, Y.; Zakrzewski, J.; Lapkin, A., Self-optimisation and model-based design of experiments for developing a C–H activation flow process. *Beilstein Journal of Organic Chemistry* **2017,** *13*, 150-163.
29. Bradford, E.; Schweidtmann, A. M.; Lapkin, A., Efficient multiobjective optimization employing Gaussian processes, spectral sampling and a genetic algorithm. *Journal of Global Optimization* **2018,** *71* (2), 407-438.

30. Manson, J. A.; Chamberlain, T. W.; Bourne, R. A., MVMOO: Mixed variable multi-objective optimisation. *Journal of Global Optimization* **2021,** *80* (4), 865-886.
31. Golovin, D.; Solnik, B.; Moitra, S.; Kochanski, G.; Karro, J.; Sculley, D., Google Vizier. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '17*, 2017; pp 1487-1495.
32. Jorayev, P.; Russo, D.; Tibbetts, J. D.; Schweidtmann, A. M.; Deutsch, P.; Bull, S. D.; Lapkin, A. A., Multi-objective Bayesian optimisation of a two-step synthesis of p-cymene from crude sulphate turpentine. *Chemical Engineering Science* **2022,** *247*.
33. Schweidtmann, A. M.; Clayton, A. D.; Holmes, N.; Bradford, E.; Bourne, R. A.; Lapkin, A. A., Machine learning meets continuous flow chemistry: Automated optimization towards the Pareto front of multiple objectives. *Chemical Engineering Journal* **2018,** *352*, 277-282.
34. Jeraal, M. I.; Sung, S.; Lapkin, A. A., A Machine Learning-Enabled Autonomous Flow Chemistry Platform for Process Optimization of Multiple Reaction Metrics. *Chemistry–Methods* **2020,** *1* (1), 71-77.
35. Amar, Y.; Schweidtmann, Artur M.; Deutsch, P.; Cao, L.; Lapkin, A., Machine learning and molecular descriptors enable rational solvent selection in asymmetric catalysis. *Chemical Science* **2019,** *10* (27), 6697-6706.
36. Zhang, C.; Amar, Y.; Cao, L.; Lapkin, A. A., Solvent Selection for Mitsunobu Reaction Driven by an Active Learning Surrogate Model. *Organic Process Research & Development* **2020,** *24* (12), 2864-2873.
37. Sung, S.; Jeraal, M. I.; Lapkin, A. A., Nomadic Evolutionary Multiobjective Optimisation (NEMO) algorithm, https://github.com/simonsung06/NEMO **2022**.
38. McKay, M. D.; Beckman, R. J.; Conover, W. J., A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. *Technometrics* **1979,** *21* (2).
39. Tang, B., Orthogonal Array-Based Latin Hypercubes. *Journal of the American Statistical Association* **1993,** *88* (424).
40. Zissimos, A. M.; Abraham, M. H.; Klamt, A.; Eckert, F.; Wood, J., A Comparison between the Two General Sets of Linear Free Energy Descriptors of Abraham and Klamt. *Journal of Chemical Information and Computer Sciences* **2002,** *42* (6), 1320-1331.
41. Svozil, D.; Ševčík, J. G. K.; Kvasnička, V., Neural Network Prediction of the Solvatochromic Polarity/Polarizability Parameter. *Journal of Chemical Information and Computer Sciences* **1997,** *37* (2), 338-342.
42. Struebing, H.; Ganase, Z.; Karamertzanis, P. G.; Siougkrou, E.; Haycock, P.; Piccione, P. M.; Armstrong, A.; Galindo, A.; Adjiman, C. S., Computer-aided molecular design of solvents for accelerated reaction kinetics. *Nature Chemistry* **2013,** *5* (11), 952-957.
43. Jespersen, D.; Keen, B.; Day, J. I.; Singh, A.; Briles, J.; Mullins, D.; Weaver, J. D., Solubility of Iridium and Ruthenium Organometallic Photoredox Catalysts. *Organic Process Research & Development* **2019,** *23* (5), 1087-1095.
44. Corcoran, E. B.; McMullen, J. P.; Lévesque, F.; Wismer, M. K.; Naber, J. R., Photon Equivalents as a Parameter for Scaling Photoredox Reactions in Flow: Translation of Photocatalytic C−N Cross-Coupling from Lab Scale to Multikilogram Scale. *Angewandte Chemie International Edition* **2020,** *59* (29), 11964-11968.

45. Lévesque, F.; Di Maso, M. J.; Narsimhan, K.; Wismer, M. K.; Naber, J. R., Design of a Kilogram Scale, Plug Flow Photoreactor Enabled by High Power LEDs. *Organic Process Research & Development* **2020,** *24* (12), 2935-2940.
46. Aillet, T.; Loubiere, K.; Dechy-Cabaret, O.; Prat, L., Accurate Measurement of the Photon Flux Received Inside Two Continuous Flow Microphotoreactors by Actinometry. *International Journal of Chemical Reactor Engineering* **2014,** *12* (1), 257-269.
47. Hatchard, C.; Parker, C. A., A new sensitive chemical actinometer-II. Potassium ferrioxalate as a standard chemical actinometer. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* **1956,** *235* (1203), 518-536.
48. Lehóczki, T.; Józsa, É.; Ősz, K., Ferrioxalate actinometry with online spectrophotometric detection. *Journal of Photochemistry and Photobiology A: Chemistry* **2013,** *251*, 63-68.
49. COSMO*therm*, Version C3.0, Release 17.01; COSMOlogic GmbH & Co. KG.
50. Feng, J.; Lansford, J. L.; Katsoulakis, M. A.; Vlachos, D. G., Explainable and trustworthy artificial intelligence for correctable modeling in chemical sciences. *Science Advances* **2020,** *6* (42).
51. Jiménez-Luna, J.; Grisoni, F.; Schneider, G., Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence* **2020,** *2* (10), 573-584.
52. Breiman, L., Random forests. *Machine learning* **2001,** *45* (1), 5-32.
53. Greenwell, B. M.; Boehmke, B. C.; McCarthy, A. J., A simple and effective model-based variable importance measure. *arXiv preprint arXiv:1805.04755* **2018**.
54. Loponov, K. N.; Lopes, J.; Barlog, M.; Astrova, E. V.; Malkov, A. V.; Lapkin, A. A., Optimization of a Scalable Photochemical Reactor for Reactions with Singlet Oxygen. *Organic Process Research & Development* **2014,** *18* (11), 1443-1454.

Appendix

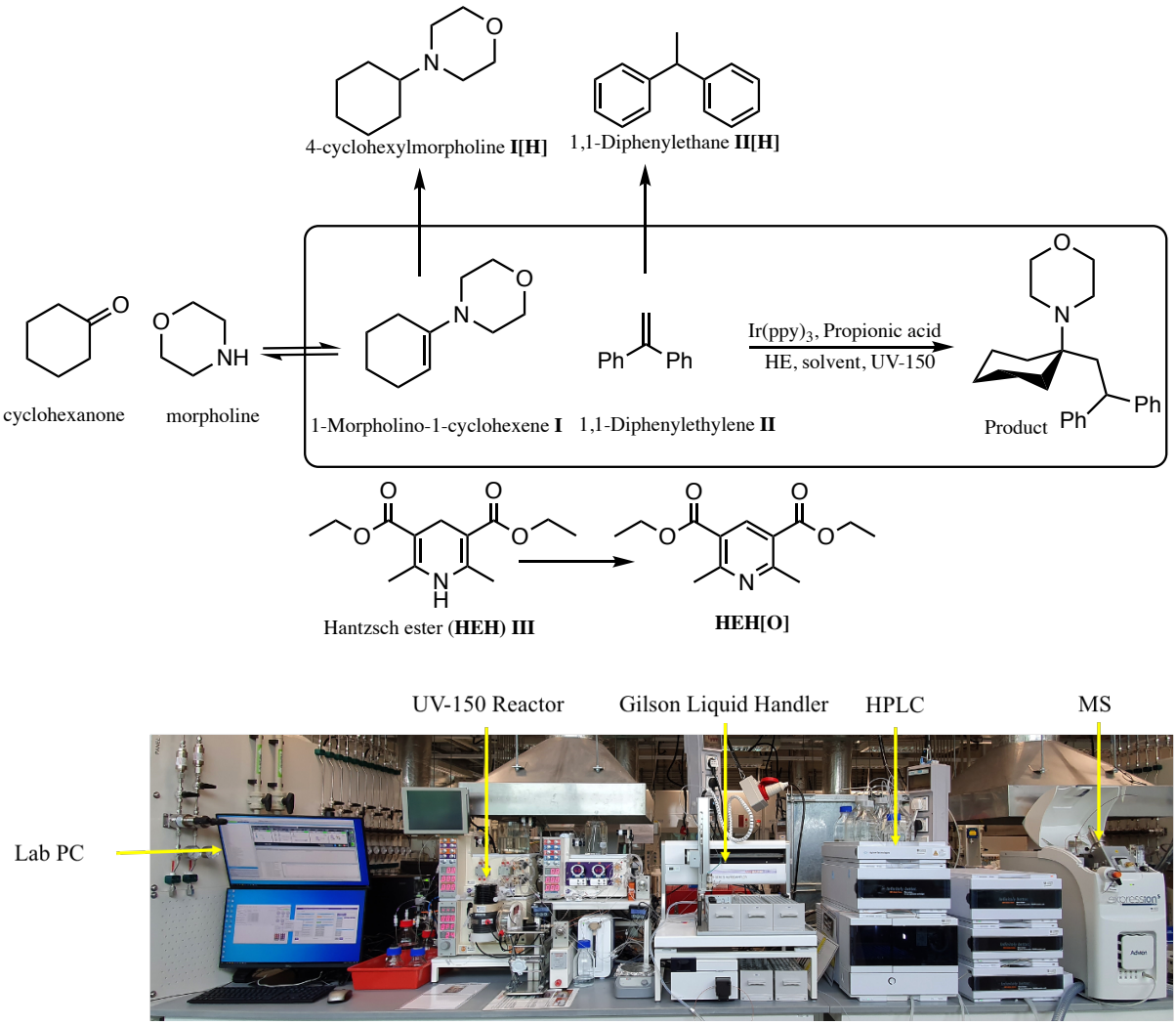


***Figure 32**. Reaction chemistry and experimental setup for photoredox amine synthesis with identified competing side reactions.*

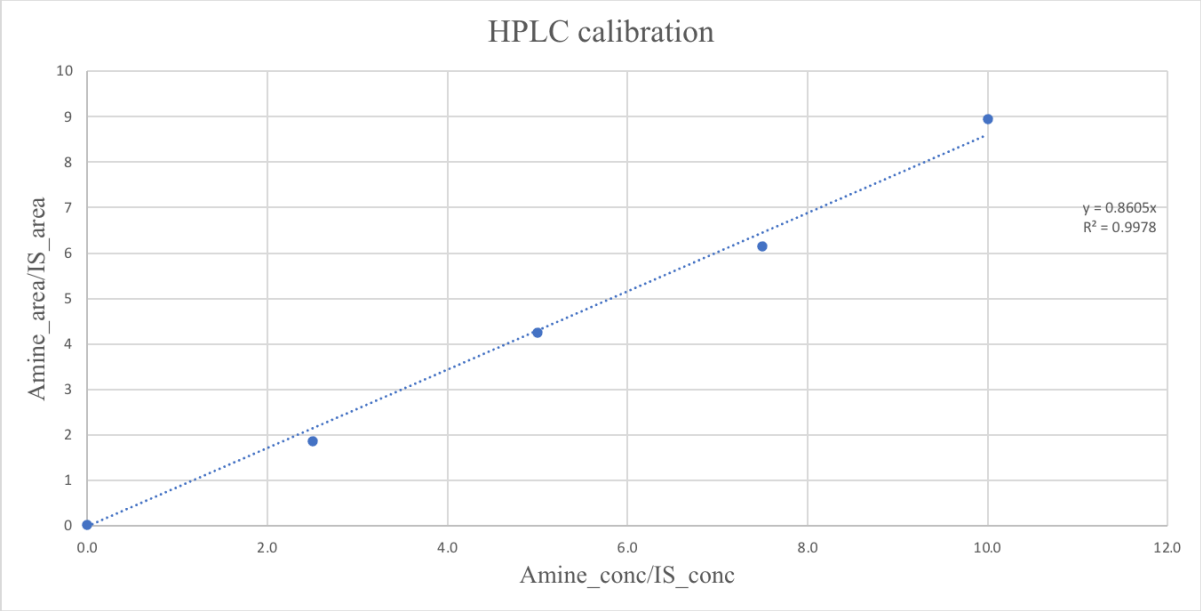


***Figure 33**. HPLC calibration for amine product (**III**) vs mesitylene internal standard. Standard deviation for yield based on analysing the same vial three times was 1.80%.*

Component	Information / %	Dimensions	Information / %
PC1	47.7	2D	78.6
PC2	30.9	3D	93.5
PC3	14.9		

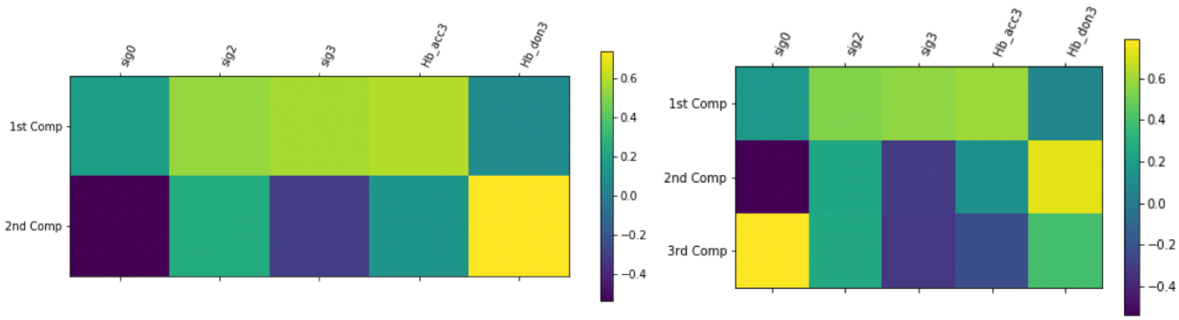


***Figure 34**. Information percentage retained after implementing PCA on five sigma moments (top). Contribution of each of the initial descriptors to the principal components (bottom).*

***Table 39**. List of solvent candidates used in optimisation and respective sigma moments calculated using COSMOtherm.*

Solvent No	Molecule	Abbv	Area	sig2	sig3	Hb_acc3	Hb_don3
1	1,2-dichloroethane	DCE	118.7231	38.945	-6.58605	0	0
2	1,3-dimethyl-2-imidazolidinone	DMI	154.8425	64.9879	63.6522	3.6463	0
3	1,3-dimethyltetrahydropyrimidin-2(1h)-one	DMPU	168.3184	62.7928	67.5844	4.4743	0
4	2-furanmethanol	Furufyrl alcohol	134.4065	66.98897	6.457333	2.112433	1.598867
5	2-methyltetrahydrofuran	2-MeTHF	132.8992	31.94015	30.65855	2.60355	0
6	2-propanol	IPA	108.1837	46.9444	21.06115	2.86455	1.1897
7	acetonitrile	MeCN	83.3227	48.4098	16.9976	0.901	0
8	benzylalcohol	BnOH	151.4705	60.32833	6.469025	2.006225	1.572775
9	dimethylformamide	DMF	117.0929	60.0196	56.594	3.8646	0
10	dimethylsulfoxide	DMSO	112.8855	78.6819	71.8756	4.7426	0
11	cyclohexanone		140.5918	46.7592	41.2752	2.8451	0
12	ethanol	EtOH	89.36075	46.6555	18.08005	2.7002	1.3905
13	ethylacetate	EA	134.4695	54.34033	36.22283	2.136233	0
14	n,n-dimethylacetamide	DMA	134.0875	61.5682	63.605	4.3094	0
15	n-methyl-2-pyrrolidinone	NMP	140.6449	63.1739	68.2801	4.3578	0
16	propanone		103.3383	47.4029	35.9321	2.6007	0
17	tetrahydrofurfurylalcohol	THFA	141.433	65.53496	39.59406	4.34038	1.14161
18	THF	THF	113.6072	31.6911	31.345	2.6497	0
19	CH2Cl2	DCM	99.3069	28.9636	-11.9892	0	0.1088
20	dimethylisosorbide		201.5833	86.60345	65.88202	5.075417	0

***Table 40**. List of final solvents used during the optimisation and the associated cost.*

Solvent No	Molecule	Abbv.	Hantzsch ester pred. solublity / mM	Hantzsch measured solubility / mM	Price £ / L	Price £ / 5 mL (rxn vol)	Purchase vol / L
1	1,2-dichloroethane	DCE	27.91	38.06	69.5	0.3475	2
2	1,3-dimethyl-2-imidazolidinone	DMI	91.88	309.25	648	3.24	0.5
3	1,3-dimethyltetrahydropyrimidin-2(1h)-one	DMPU	156.43	357.02	370	1.85	1
4	2-furanmethanol	Furufyrl alcohol	36.82	86.45	99.9935	0.499968	1
5	2-methyltetrahydrofuran	2-MeTHF	39.57	122.21	141	0.705	2
6	2-propanol	IPA	16.98	60.47	44.4	0.222	2
7	acetonitrile	MeCN	11.25	19	99	0.495	2
8	benzylalcohol	BnOH	42.04	91.2	203	1.015	2
9	dimethylformamide	DMF	99.22	178.33	84.5	0.4225	2
10	dimethylsulfoxide	DMSO	30.98	159.51	198	0.99	2
11	cyclohexanone		41.51	-	41.4	0.207	1
12	ethanol	EtOH	30.28	45.02	80	0.4	2
13	ethylacetate	EA	40.79	80.56	53.5	0.2675	2
14	n,n-dimethylacetamide	DMA	158.03	330.22	75.5	0.3775	2
15	n-methyl-2-pyrrolidinone	NMP	172	390.82	99.5	0.4975	2
16	propanone		31.65	49.71	66.8	0.334	1
17	tetrahydrofurfurylalcohol	THFA	36.74	-	58.0754	0.290377	1
18	THF	THF	84.82	93.96	73.5	0.3675	2
19	CH2Cl2	DCM	118.66	64.11	46.25	0.23125	2
20	dimethylisosorbide		15.52	-	114.4	0.572	2.5

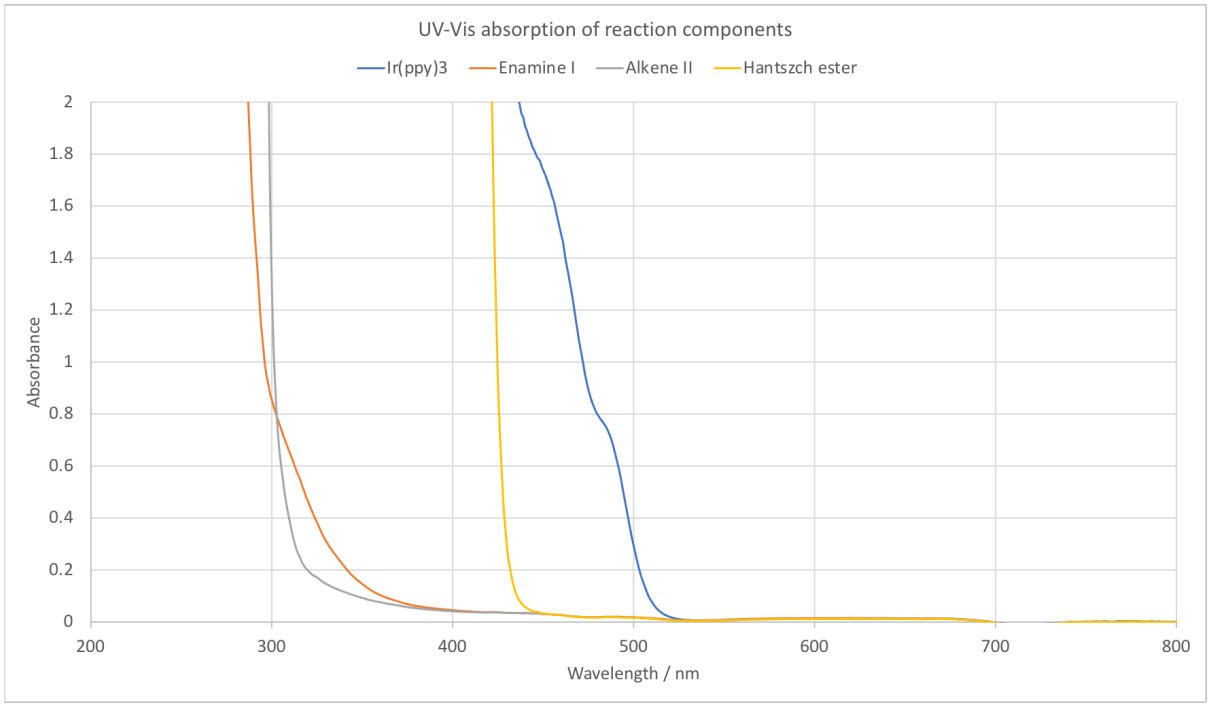


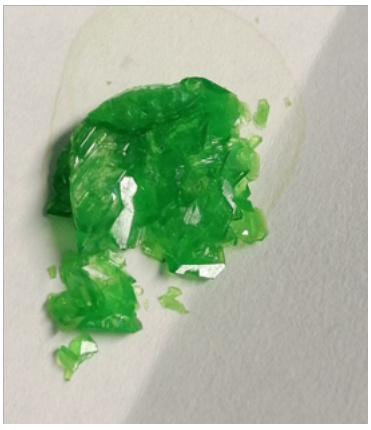
***Figure 35**. UV-Vis absorption of reaction components.*

Photon flux study

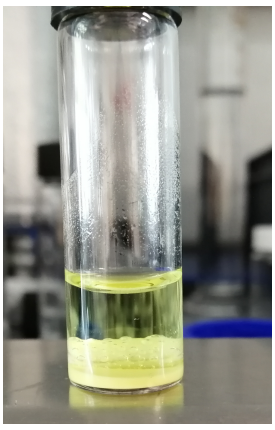
Experimental

$$2FeCl_3 + 6K_2C_2O_4 \rightarrow 2\ K_3[Fe(C_2O_4)_3] + 6KCl$$

Potassium ferrioxalate was synthesised by mixing 1.5 mol L^{-1} $6K_2C_2O_4\ H_2O$ and 1.5 mol L^{-1} $FeCl_3$ at 3:1 volume ratio. The mixture was rigorously stirred to give potassium ferrioxalate precipitate, which was filtered, recrystallised three times with water, and dried overnight in oven at 40 °C. For actinometry studies in batch (4.0 mL glass vial) using Kessil Blue lamp and in flow using Vapourtec UV-150 reactor (10.0 mL), appropriate amounts of potassium ferrioxalate solutions (C_{A0}), specifically 0.006 mol L^{-1} (2.947 g) and 0.15 mol L^{-1} (73.68 g), were prepared in 0.05 mol L^{-1} H_2SO_4. For analysis, based on the amount of ferrous ion produced, 650 µL of irradiated solution was mixed with 26 µL of o-phenanthroline solution (0.05 mol L^{-1}), 1150 µL of CH_3COONa (1.0 mol L^{-1}), 750 µL of H_2SO_4 (0.5 mol L^{-1}), and analysed using UV-Vis. O-phenanthroline concentration was at least three times higher than Fe^{+2} concentration to ensure complex formation. Absorption coefficient of $Fe^{+2}(phen)_3$ complex at 512 nm was measured to be 9,918 L mol^{-1} cm^{-1}, compared to 10,910 L mol^{-1} cm^{-1} reported by Lehoczki *et al*. and 10,980 L mol^{-1} cm^{-1} (at 510 nm) reported by Aillet *et al*.



$K_3[Fe(C_2O_4)_3]$x3H_2O



$K_3[Fe(C_2O_4)_3]\ precipite$

Figure 36. *(left) potassium ferrioxalate crystal and (right) precipitate.*

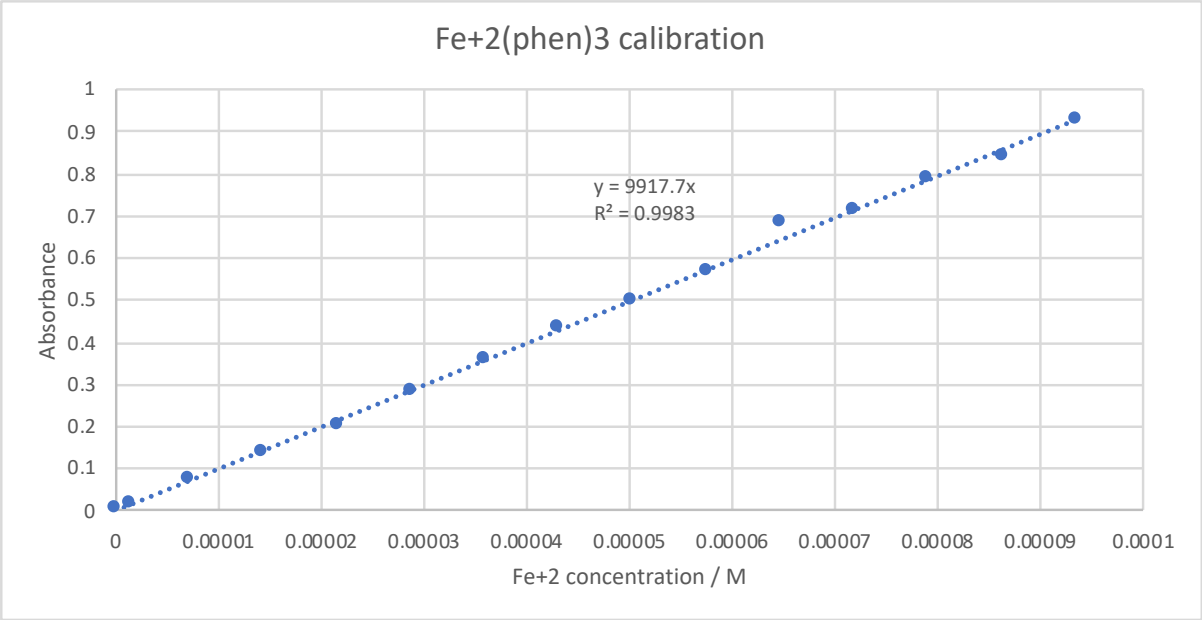


Figure 37*. Calibration line for Fe^{+2} complex formation with 1,10-phenanthroline at 512 nm.*

Modelling

Photodecomposition of potassium ferrioxalate in water is given as following.

$$2[Fe(C_2O_4)_3]^{3-} \rightarrow 2Fe(C_2O_4) + 3[C_2O_4]^{2-} + 2CO_2$$

Assuming a perfect mixing in the continuous flow reactor in Vapourtec, potassium ferrioxalate decomposition can be described as the following and the consumption rate of potassium ferrioxalate (A), and conversion to ferrous ion (B), is linearly related to the quantum yield (φ_λ / mol einstein^{-1}) of the reaction and the mean absorbed photon flux density ($< L^a_{p,\lambda} >$ / einstein m^{-3} s^{-1}) at wavelength λ nm (equation 3).

$$A \xrightarrow{h\nu} B \tag{2}$$

$$-\frac{dC_A}{dt} = \varphi_\gamma < L^A_{p,\lambda} > \tag{3}$$

$$\langle L^A_{p,\lambda} \rangle = \frac{q_{p,\gamma}}{V_r} f_\lambda \tag{4}$$

Where f_λ is the fraction of the absorbed light and $q_{p,\gamma}$ (einstein s^{-1}) is the photon flux received over the entire volume in a reactor V_r (m^3). The fraction of the absorbed light could be calculated based on the molar absorption coefficient κ_λ (m^2 mol^{-1}) of A in a given path length l (m) as following:

$$f_\lambda = 1 - e^{-A_e} = 1 - e^{-\kappa_\lambda C_A l} \quad (5)$$

The received photon flux over the entire reactor volume depends on the reactor material transmittance (T_λ) and the incoming photon flux ($q_{p0,\lambda}$) at the surface, and can be expressed as

$$q_{p,\lambda} = q_{p0,\lambda}\, T_\lambda \quad (6)$$

For fluoropolymer (FEP) reactor tubing of Vapourtec UV-150 reactor, material transmittance at 512 nm was taken as 1.0.

For monochromatic light source, reactor material transmittance (T_λ) is equal to 1.0 and is constant, which means $q_{p,\lambda} = q_{p0,\lambda}$. Moreover, both $q_{p,\gamma}$ and κ_λ are constant. Integrating equation 3 from C_{A0} to C_A, and $t = 0$ to $t = \tau$, based on equations 3-6 leads to integrated equation 7 to quantify the photon flux $q_{p,\gamma}$ (einstein s^{-1}) received over the absorbing volume.

$$\left(\varphi_\gamma \frac{q_{p,\gamma}}{V_r}\right)\tau = C_{A0}X + \frac{1}{\kappa_\lambda l}\ln\left[\frac{1 - e^{-\kappa_\lambda C_{A0} l}}{1 - e^{-\kappa_\lambda C_{A0}(1-X) l}}\right] \quad (7)$$

Where $C_A = C_{A0}(1 - X)$ for residence time $\tau = \frac{V_r}{Q}$.

For a polychromatic light source, the light wavelength λ is not constant, which means both $q_{p,\gamma}$ and κ_λ are not constant. The received photon flux needs to be integrated over discrete wavelengths intervals $\Delta\lambda_i$, together with the lamp density function g_λ at various wavelengths. As demonstrated by Aillet et al.[46] received photon flux for a polychromatic light source can be quantified as

$$\frac{dX}{dt} = \frac{1}{C_{A0}}\left(\frac{q_{p,0}}{V_r}\right)\sum\nolimits_{\Delta\lambda_i}[T_{\lambda_i}\varphi_{\lambda_i} g_{\lambda_i}(1 - e^{-\kappa_{\lambda_i} C_{A0}(1-X) l})] \quad (8)$$

Where $q_{p,0}$ is the total incoming photon flux. For each wavelength λi and the power $M_{p,\lambda}$ emitted at that wavelength, the density function g_λ of the lamp can be express as following:

$$g_\lambda = \frac{M_{p,\lambda}}{\sum_{\lambda_i} M_{p,\lambda_i}} \quad (9)$$

***Table 41**. Comparison of photon flux received in batch under Kessil blue lamp and in flow under 470 nm LED. The results are compared with different lamps and reactor setups reported by Aillet et al.[46] and Loponov et al.[54] (i.e. CARES → this project)*

Reactor type	**Light type**	**Wavelength / nm**	**Lamp**	**Reactor volume /mL**	**Photon flux (x10^6) / Einstein s^{-1}**	**Actinometric intensity of absorbed photons (x10^4) / Einstein L^{-1} s^{-1}**
Aillet et. al. microphotoreactor - flow	Polychromatic	-	High pressure Hg	0.81	26.2	323.46
Aillet et. al. microphotoreactor - flow	Monochromatic	365	UV-LED	0.54	0.382	7.07
Aillet et. al. immersion well	Polychromatic	-	High pressure Hg	225	47.6	2.16
CARES batch	Polychromatic	-	Kessil Blue lamp	2	0.2	1.00
CARES VT	Monochromatic	470	UV-150	10	6.18	6.18
Loponov et. al.	Monochromatic	420	FL			5.7
	Monochromatic	524	LED			4.6
	Polychromatic		Xe arc			2.3
	Polychromatic		Hg MP (IW)			1.2

Solubility predictions and measurements

Experimentally reported solubility values for $Ir(ppy)_3$ catalyst for common solvents was used.[43] Solubility of Hantzsch ester was measured using direct solvent evaporation technique. First, a suspension of Hantzsch ester in a given solvent with pre-weighed vial was prepared and stirred for 1 h. The suspension was filtered to accommodate a clear solution (1 mL). The solvent was dried by blowing air over the solution. Final weight of the vial was measured to calculate the total dissolved amount of Hantzsch ester. Similarly, 1 mL clear solution was prepared to be analysed via benchtop NMR with mesitylene as an internal standard.

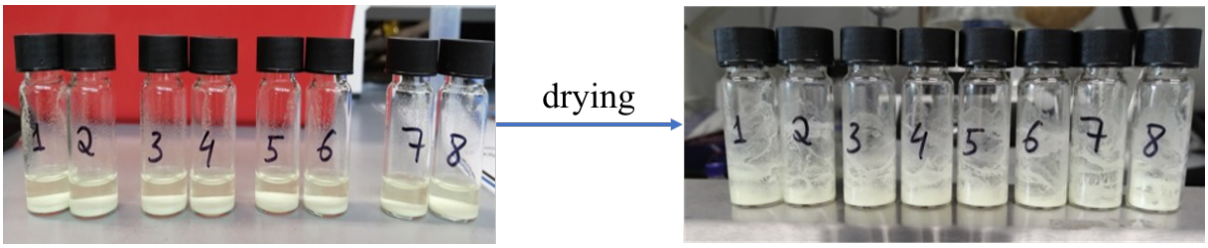


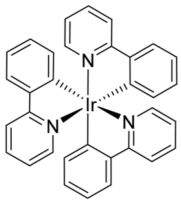
***Figure 38**. Preparation of Hantzsch ester suspensions and drying process.*

***Table 42**. Comparison of solubility results via air drying vs benchtop NMR analysis.*

Solvent	Solubility (mg /mL) - measured using solvent drying	Solubility (mg /mL) - measured using benchtop NMR
Methanol	5.60	5.60
Diethyl ether	1.00	1.00
Chloroform	36.00	38.54
Acetone	10.40	12.59
Hexane	0.01	0.01
Acetonitrile	4.40	4.81
Diethyl carbonate	2.80	4.50
Dichloromethane	15.00	16.24
N,N-dimethylformamide	52.80	45.17
1-butanol	6.80	5.66

Table 43. *Maximum solubility of the catalyst fac-Ir(ppy)3 in commonly used solvents. Reproduced from ref*[43]

***fac*-Ir(ppy)$_3$**		
CAS # 94928-86-6		
Solvents	**Molar Concentration**	**ppm**
Acetone[19]	$6.0x10^{-4}$	$5.0x10^{2}$
Acetonitrile[19]	$4.1x10^{-4}$	$3.5x10^{2}$
Dichloromethane[19]	$7.3x10^{-3}$	$3.6x10^{3}$
N,N-Dimethylformamide[19]	$1.5x10^{-3}$	$1.0x10^{3}$
Dimethylsulfoxide[19]	$3.7x10^{-3}$	$2.2x10^{3}$
Ethyl Acetate[20]	$3.4x10^{-4}$	$2.5x10^{2}$
Methanol[20]	$1.1x10^{-5}$	$9.4x10^{0}$
Methyl-t-butyl ether[20]	$6.2x10^{-5}$	$5.5x10^{1}$
N-Methyl 2-pyrrolidinone[19]	$5.2x10^{-2}$	$3.3x10^{4}$
Tetrahydrofuran[19]	$2.1x10^{-3}$	$1.6x10^{4}$
Toluene[20]	$5.9x10^{-4}$	$4.4x10^{2}$
Water[19]	-	<1
4:1 Acetonitrile:Water[21]	-	10-100
1:1 Acetonitrile:Water[21]	-	10-100



COSMOtherm for solubility predictions

Available solvents in the COSMO*therm* database of 1401 solvents were loaded automatically in the software with BP_TZVPD_FINE parameterisation. If a certain solvent was not available, conformer search and structure optimisation were performed using COSMO*conf* program. Depending on the run, experimental references values were for Hantzsch ester solubility in 9, 18, or 35 solvents were provided. Choices of solvents in the reference list were chosen to maximise the diversity of solvent classes.

Benchmarking results using pool-based sampling

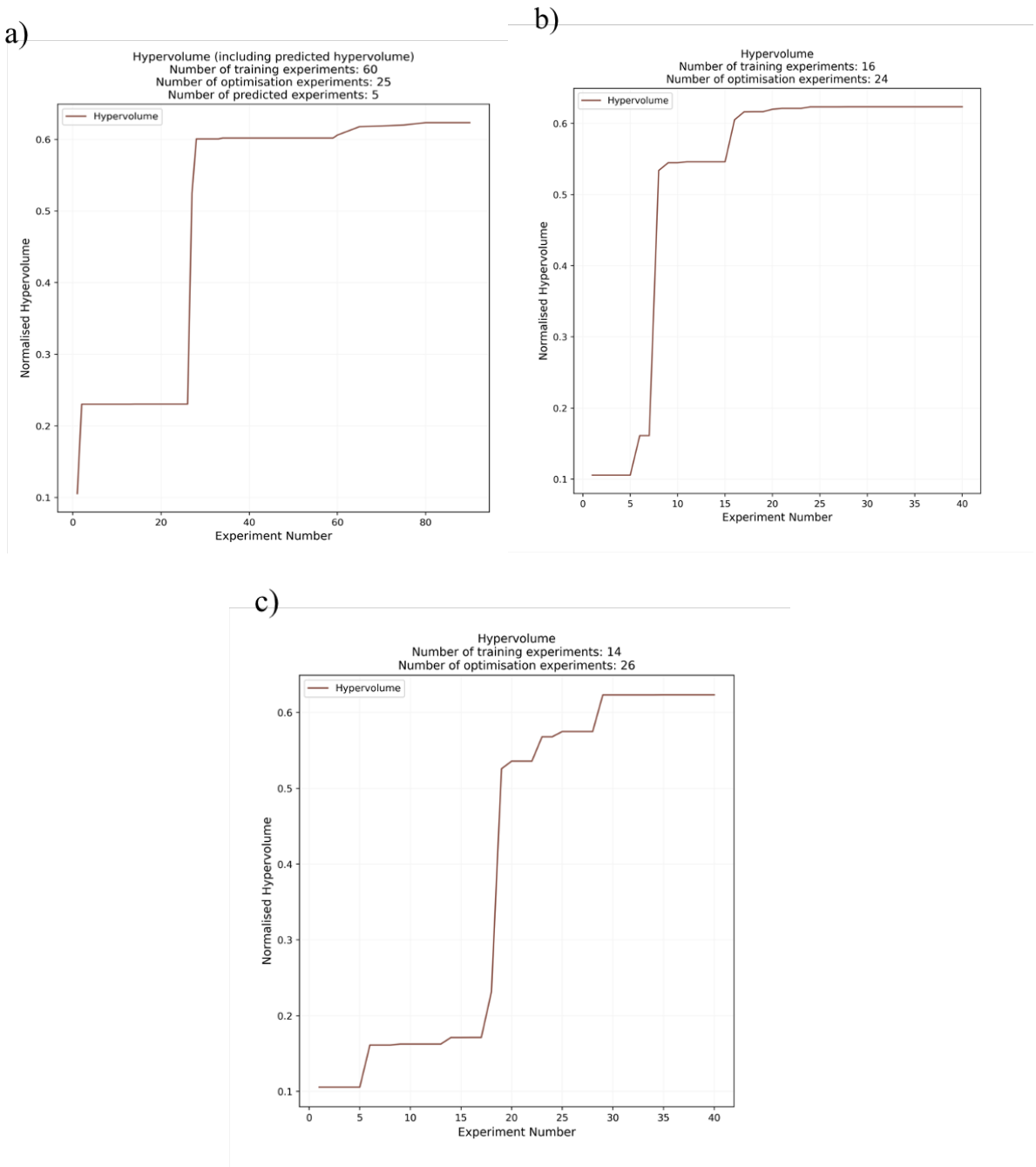


***Figure 39**. Hypervolume improvement over experiment number for a) full optimisation, b) benchmarking using 16 training points, including a point on the Pareto front, and c) 14 training points with no Pareto points included.*

Explainable AI

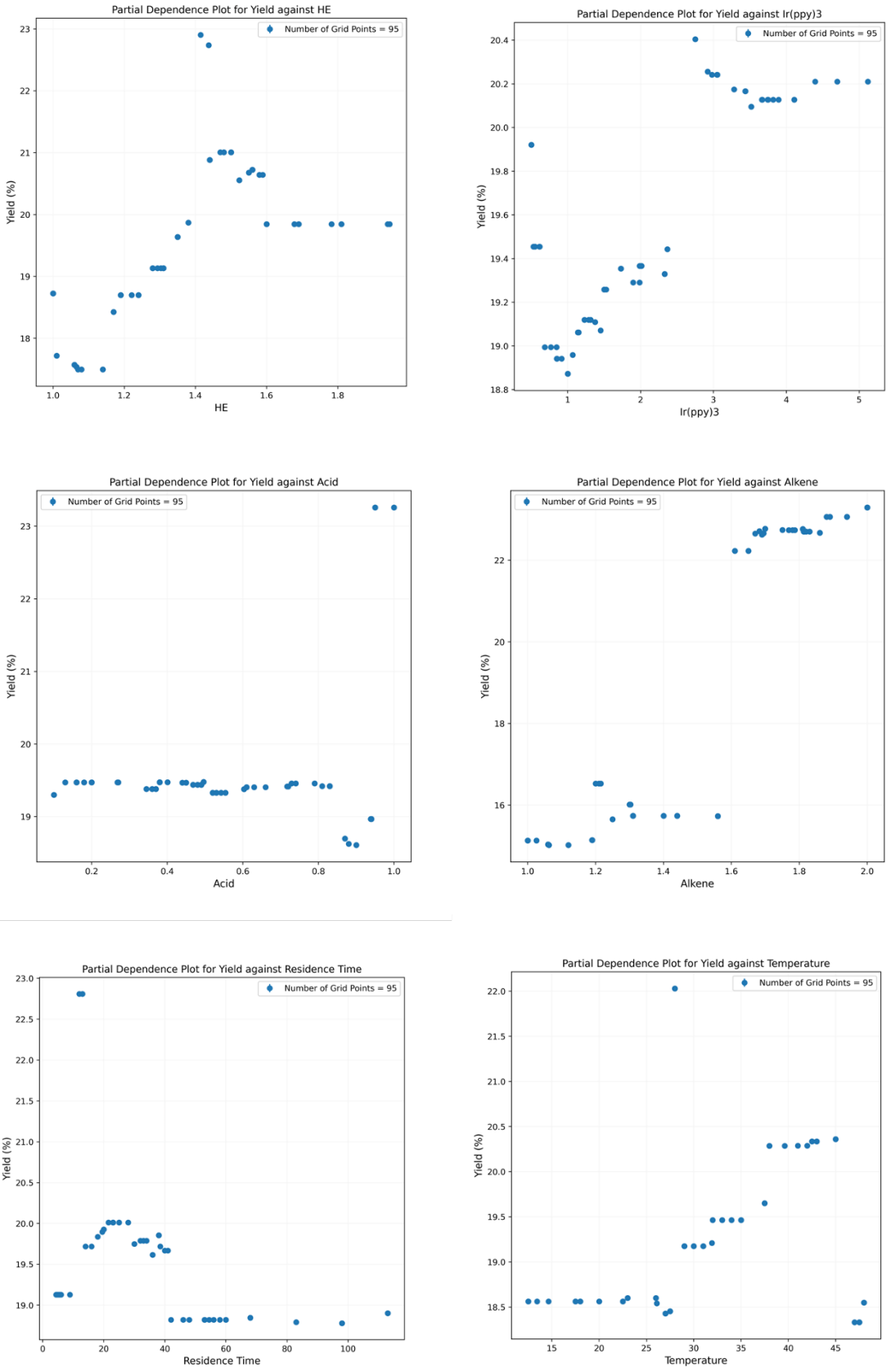


Figure 40. *Partial dependence plot for continuous variables against yield.*

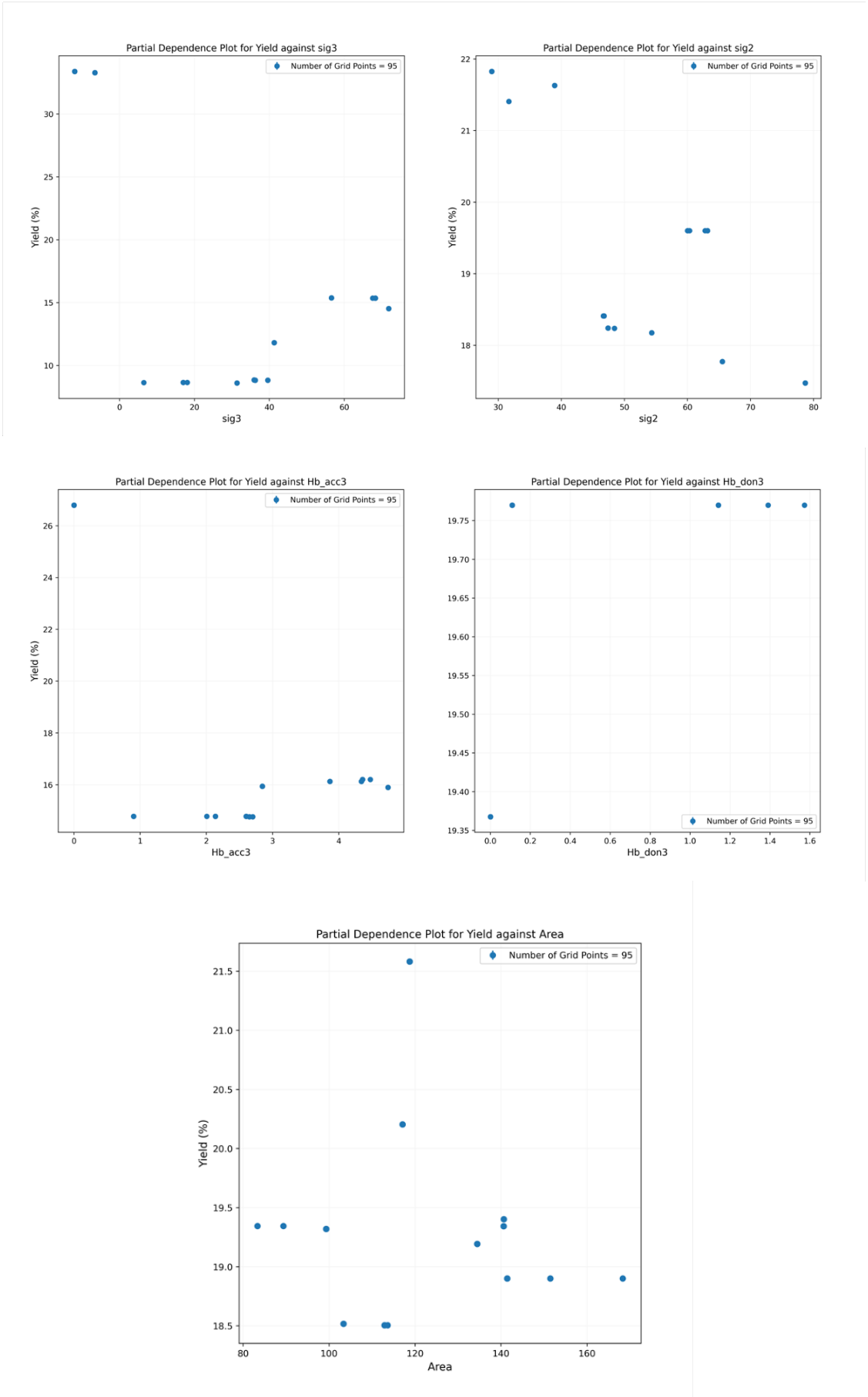


***Figure 41**. Partial dependence plot for all solvent descriptors against yield.*

***Table 44**. Data generated during the training and optimisation*

Rxn No	Rxn type	Alkene eq.	$Ir(ppy)_3$ mol%	HEH eq.	Acid eq.	T / C	Time / min	Solvent	Cost / £	Yield / %
1	training	1.06	1.73	1.06	0.61	18	98	DMF	1.588	9.20
2	training	1.44	0.55	1.31	0.83	23	23	DMF	1.029	11.09
3	training	1.81	1.14	1.56	0.27	28	68	DMF	1.397	26.23
4	training	1.31	3.75	1.69	0.94	38	38	DMF	2.783	15.90
5	training	1.19	2.33	1.81	0.72	43	5.5	DMF	2.053	5.47
6	training	1.56	3.52	1.19	0.38	48	53	DMF	2.571	18.81
7	training	1.56	3.52	1.19	0.38	18	98	DMSO	3.138	2.76
8	training	1.44	0.55	1.31	0.83	23	23	DMSO	1.596	4.77
9	training	1.81	1.14	1.56	0.27	28	68	DMSO	1.964	7.58
10	training	1.69	2.92	1.44	0.49	35	113	DMSO	2.875	6.03
11	training	1.31	4.11	1.69	0.94	38	38	DMSO	3.540	4.18
12	training	1.19	2.33	1.81	0.72	43	5.5	DMSO	2.621	2.36
13	training	1.06	1.73	1.06	0.61	48	53	DMSO	2.155	5.79
14	training	1.56	1.50	1.19	0.38	18	98	Acetone	1.419	0.00
15	training	1.44	0.55	1.24	0.83	23	23	Acetone	0.927	0.00
16	training	1.81	1.14	1.24	0.27	28	68	Acetone	1.247	3.14
17	training	1.69	1.50	1.24	0.49	35	113	Acetone	1.433	0.00
18	training	1.31	1.50	1.24	0.94	38	38	Acetone	1.423	0.00
19	training	1.19	1.50	1.24	0.72	43	5.5	Acetone	1.418	0.00
20	training	1.06	1.50	1.06	0.61	48	53	Acetone	1.378	0.00
21	training	1.56	3.52	1.19	0.38	18	98	NMP	2.646	8.30
22	training	1.44	0.55	1.31	0.83	23	23	NMP	1.104	8.43
23	training	1.81	1.14	1.56	0.27	28	68	NMP	1.472	18.20
24	training	1.69	2.92	1.44	0.49	35	113	NMP	2.383	19.13
25	training	1.31	4.11	1.60	0.94	38	38	NMP	3.030	15.62
26	training	1.19	2.33	1.60	0.72	43	5.5	NMP	2.088	7.29
27	training	1.06	1.73	1.06	0.61	48	53	NMP	1.663	15.26
28	training	1.81	1.07	1.56	0.27	28	68	DCM	1.168	60.14
29	training	1.70	3.05	1.44	0.50	33	25	DCM	2.187	70.71
30	training	1.30	3.44	1.68	0.94	38	38	DCM	2.424	45.55
31	training	1.19	2.37	1.68	0.72	43	5.0	DCM	1.856	35.78
32	training	1.06	1.45	1.07	0.60	48	53	DCM	1.251	19.16
33	training	1.94	4.70	1.35	0.16	13	83	Cyclohexanone	3.019	6.57
34	training	1.56	3.52	1.19	0.38	18	98	Cyclohexanone	2.355	6.31
35	training	1.44	0.55	1.31	0.83	23	23	Cyclohexanone	0.813	3.65
36	training	1.81	1.14	1.35	0.27	28	68	Cyclohexanone	1.141	7.10
37	training	1.69	2.92	1.35	0.49	35	113	Cyclohexanone	2.075	9.51

38	training	1.31	4.11	1.35	0.94	38	38	Cyclohexa none	2.692	6.80
39	training	1.19	2.33	1.35	0.72	43	5.5	Cyclohexa none	1.749	0.00
40	training	1.06	1.73	1.06	0.61	48	53	Cyclohexa none	1.372	5.77
41	training	1.94	3.05	1.28	0.16	13	83	THFA	2.220	0.00
42	training	1.56	3.05	1.19	0.38	18	98	THFA	2.191	0.00
43	training	1.44	0.55	1.28	0.83	23	23	THFA	0.891	0.00
44	training	1.81	1.14	1.28	0.27	28	68	THFA	1.211	0.00
45	training	1.69	2.92	1.28	0.49	35	113	THFA	2.145	0.00
46	training	1.31	3.05	1.28	0.94	38	38	THFA	2.203	0.00
47	training	1.19	2.33	1.28	0.72	43	5.5	THFA	1.819	0.00
48	training	1.06	1.73	1.06	0.61	48	53	THFA	1.456	0.00
49	training	1.94	0.85	1.94	0.16	13	83	EA	1.166	0.00
50	training	1.56	0.85	1.19	0.38	18	98	EA	1.010	0.00
51	training	1.44	0.55	1.31	0.83	23	23	EA	0.874	0.00
52	training	1.81	0.85	1.56	0.27	28	68	EA	1.089	0.00
53	training	1.69	0.85	1.44	0.49	35	113	EA	1.063	0.00
54	training	1.31	0.85	1.69	0.94	38	38	EA	1.101	0.00
55	training	1.19	0.85	1.81	0.72	43	5.5	EA	1.119	0.00
56	training	1.06	0.85	1.06	0.61	48	53	EA	0.970	0.00
57	training	2.00	1.00	1.50	0.20	28	20	DMSO	1.885	6.89
58	training	2.00	1.00	1.50	0.20	30	40	DMSO	1.885	4.91
59	training	2.00	1.00	1.50	0.20	30	60	DMSO	1.885	5.97
60	training	2.00	1.00	1.50	1.00	30	60	DMSO	1.890	6.05
61	Optimisation	1.25	1.00	1.00	0.37	33	14	DCM	1.006	29.21
62	Optimisation	1.21	0.62	1.00	0.95	45	34	DCE	0.922	49.01
63	Optimisation	1.00	0.77	1.00	0.90	29	56	DCM	0.879	27.31
64	Optimisation	1.86	1.23	1.58	0.81	20	4.65	DCM	1.261	55.83
65	Optimisation	1.00	0.85	1.00	0.48	26	36	DCM	0.917	33.38
66	Optimisation	1.30	0.68	1.07	0.87	3	28	DCM	0.858	31.03
67	Optimisation	1.56	5.12	1.52	0.53	26	18	DCE	3.403	38.28
68	Optimisation	1.67	1.99	1.47	0.63	32	48	DCE	1.747	54.22
69	Optimisation	1.65	3.44	1.48	0.79	27	32	DCE	2.514	61.89
70	Optimisation	1.12	3.74	1.17	0.36	30	38	DMPU	4.098	8.38
71	Optimisation	1.61	0.61	1.58	0.16	34	33	DCM	0.923	52.00
72	Optimisation	1.75	3.67	1.60	0.74	48	5	DCM	2.544	40.97
73	Optimisation	1.65	1.99	1.08	0.45	32	32	THF	1.690	0.79
74	Optimisation	1.86	1.53	1.55	0.88	27	58	DCE	1.528	45.42
75	Optimisation	1.89	3.28	1.14	0.52	47	6	THF	2.394	1.92
76	Optimisation	2.00	1.31	1.50	0.20	29	20	DCM	1.292	61.26
77	Optimisation	1.77	3.89	1.78	0.47	45	30	BnOH	3.482	3.48
78	Optimisation	2.00	2.00	1.00	0.20	29	40	MeCN	1.820	6.10
79	Optimisation	1.82	1.38	1.59	0.73	42	41	DCE	1.454	56.8

80	Optimisation	1.03	0.92	1.30	0.55	45	16	EtOH	1.182	0.87
81	Optimisation	2.00	1.29	1.50	0.20	28	20	DCM	1.277	63.98
82	Optimisation	1.20	2.98	1.06	0.52	45	46	DCE	2.173	35.02
83	Optimisation	2.00	1.15	1.50	0.20	31	12	DCE	1.320	55.62
84	Optimisation	1.40	3.67	1.58	0.18	31	25	DCM	2.524	39.03
85	Optimisation	2.00	1.15	1.50	0.20	29	60	DMF	1.395	11.93
86	Optimisation	1.83	2.98	1.38	0.95	30	4.35	DCE	2.260	53.24
87	Optimisation	1.88	2.75	1.56	0.44	41	13	DCE	2.173	70.31
88	Optimisation	1.78	3.82	1.01	0.45	34	36	MeCN	2.774	0.43
89	Optimisation	1.06	1.15	1.60	0.66	27	42	DCM	1.193	26.82
90	Optimisation	1.00	0.53	1.00	0.10	29	33	DCE	0.867	30.93

***Table 45**. Predicted and measured Hantzsch ester solubility in various solvents.*

No	Solvent	COSMO*therm* predicted solubility / mM	Experimentally measured solubility / mM
1	formicacid	532.86	
2	hexamethylphosphoramide	174.29	
3	n-methyl-2-pyrrolidinone	172.00	390.82
4	n,n-dimethylacetamide	158.03	330.22
5	1,3-dimethyltetrahydropyrimidin-2(1h)-one	156.43	357.02
6	ch2cl2	118.66	64.11
7	dimethylformamide	99.22	178.33
8	1,3-dimethyl-2-imidazolidinone	91.88	309.25
9	thf	84.82	93.96
10	dioxane	82.09	51.66
11	propionicacid	76.79	25.64
12	chcl3	53.81	152.17
13	cyclopentanone	52.86	
14	aceticacid	51.05	41.75
15	1-methyl-pyrrolidine	50.29	
16	methanol	42.30	22.11
17	benzylalcohol	42.04	
18	cyclohexanone	41.51	
19	ethylacetate	40.79	80.56
20	methylacetate	40.25	
21	2-methyltetrahydrofuran	39.57	122.21
22	2-furanmethanol	36.82	86.45
23	tetrahydrofurfurylalcohol	36.74	
24	ethylformate	33.79	

25	propanone	31.65	49.71
26	dimethylsulfoxide	30.98	159.51
27	ethanol	30.28	45.02
28	propanol	29.95	
29	n-propylacetate	29.11	
30	isobutanol	29.07	
31	methylpropionate	28.92	30.08
32	1,2-dichloroethane	27.91	
33	1-methoxy2-propanol	27.76	
34	butanone	26.92	52.45
35	1-butanol	26.25	22.35
36	isopropylacetate	25.34	28.94
37	2-methoxyethanol	25.29	
38	isopentanol	25.27	
39	2-hydroxypropanoicacidethylester	24.62	
40	dimethoxymethane	24.25	52.80
41	diglyme	24.11	
42	methylformate	23.72	
43	1,2-dimethoxyethane	23.32	
44	1-pentanol	22.63	
45	dihydro-5-methyl-2(3h)-furanone	21.55	
46	dimethylcarbonate	21.13	
47	4-methyl-2-pentanone	20.38	
48	diethylcarbonate	19.63	17.76
49	benzene	19.43	
50	n-butylacetate	18.77	
51	aceticacid-2-methylpropylester	17.96	
52	tert-butylacetate	17.94	
53	2-butanol	17.44	
54	2-propanol	16.98	60.47
55	2-hexanone	16.70	
56	Dibutyl_Isosorbide_Ether	15.52	
57	1-heptanol	15.07	
58	2-methyl-2-butanol	14.86	
59	isoamylacetate	14.21	
60	lacticacid	14.08	
61	n-pentylacetate	13.69	
62	1-octanol	13.22	
63	ethylsuccinate	12.37	
64	dimethyl_adipate	11.85	
65	1,2-ethanedioldiacetate	11.61	
66	acetonitrile	11.25	19.00
67	methyl-t-butylether	10.79	

68	diethylether	10.25	3.95
69	cyrene	9.47	
70	anisole	9.01	
71	butylenecarbonate	8.49	
72	cyclopentyl-methyl-ether	7.98	
73	aceticanhydride	7.82	
74	chlorobenzene	7.16	
75	1,2-dichlorobenzene	6.90	10.10
76	toluene	6.80	4.74
77	glycerol-triacetate	6.63	
78	nitromethane	6.58	
79	Dibutyl_Succinate	6.40	
80	aceticacid-2-ethylhexylester	5.95	
81	propyleneglycol	5.94	
82	1-chlorobutane	5.93	
83	methyl-tert-amylether	5.85	
84	tetrahydro-2,2,5,5-tetramethylfuran	5.22	
85	propylenecarbonate	4.43	
86	1,3-propanediol	4.41	
87	diisopropylether	4.18	
88	1,2-dimethylbenzene	3.80	
89	1,4-dimethylbenzene	3.73	
90	2-ethoxy-2-methyl-propane	3.67	
91	1,3-dimethylbenzene	3.64	
92	trifluoromethylbenzene	2.96	
93	isopropylbenzene	2.73	
94	1-methyl-4-isopropylbenzene	1.79	
95	glycol	1.73	
96	cs2	1.58	
97	methyloleate	1.44	4.36
98	triethylamine	1.42	1.58
99	dipentene	1.20	
100	glycerol	0.55	0.00
101	cyclohexane	0.23	0.79
102	pentane	0.22	
103	methylcyclohexane	0.21	
104	hexane	0.18	0.04
105	2,2,4-trimethylpentane	0.18	
106	n-heptane	0.15	
107	h2o	0.04	0.79

Chapter 6

Concluding remarks

This thesis aimed at developing workflows for robust process development of novel and complex chemical transformations, assisted with machine learning algorithms, cheminformatics software, and laboratory automation tools. We demonstrated the applicability of these methodologies in black-box optimisation of bio-waste to functional molecules synthesis (Chapter 2), selection of alternative solvents, using molecular informatics and similarity metrics, for new processes balancing process performance and SHE impacts (Chapter 3), holistic modelling of literature data using machine learning and DFT-based descriptors of Buchwald-Hartwig amine synthesis (Chapter 4), and Bayesian optimisation of discrete and continuous variables for a robust process development of sensitive photoredox amine synthesis in flow (Chapter 5). This section provides a summary of research achievements as well as limitations, and suggestions for future work.

Summary of key contributions and new knowledge created

(1) Multi-objective Bayesian optimisation of eight continuous variables for a complex reaction networks in the form of a bio-waste

The study described in Chapter 2 demonstrated the black-box optimisation of acid-catalysed ring opening reaction in batch and radical dehydrogenation reaction in flow without any prior physico-chemical mechanistic information.[1] Using an extended version of Bayesian optimisation algorithm TS-EMO,[2] which allowed for a secondary independent sampling for variable reaction time for more efficient data generation, the algorithm accounted for the exploration-exploitation trade-off as observed in the algorithm-guided experimental points during the optimisation. Results showed a cluster of optimal conditions around the experimentally identified Pareto points to identify the trade-offs between the objectives yield and conversion. Moreover, the model hyperparameters revealed further information about the sensitivity of reaction objectives to the input variables, allowing for extraction of physical

insights about the reaction. Considering the previously reported literature on optimisation of reactions for up to five continuous variables,[3] this work demonstrated development of robust processes for functional molecules synthesis from bio-waste routes, which often have high chemical complexity and reaction variables, as an alternative solution to traditional processes (e.g., prior purification) in de-carbonisation of the chemical supply chain.

(2) Selection of solvent alternatives based on molecular informatics, similarity metrics, and SHE guides

Whilst most of the reported solvent selection guides focus on safety, health, and environmental (SHE) impacts, or screen solvents based on a manually selected list of commonly used solvents for a given reaction, we highlighted the need for balancing process performances and SHE impacts in solvent selection and the need for a solvent selection guide that could also suggest "non-obvious" solvents. Using only a handful of computationally generated descriptors for the library of 120 commonly used solvents and the library of 457 solvents, results showed a significant overlap with the solvent selection guides reported by AstraZeneca[4] and Syngenta.[5] The interactive tools developed by these companies required extensive list of experimental and computational solvent parameters, and did not account for the choice of "similarity" metrics, which could significantly affect the choice of an alternative (i.e., "similar") solvent suggested for a given study.[6-8] We found that, for various transformations (e.g., Menschutkin reaction,[9, 10] solvolysis of t-butyl chloride[11]), rankings of solvents based on experimental data matched with solvent suggestions in our workflow based on combination of similarity metrics (i.e., similarity fusion) and parameterisation techniques. In other words, solvents that led to good reaction performances were recommended to be alternatives to each other, often identifying solvents that might not be selected based on domain knowledge and conventional way of describing solvent "classes". This demonstrated the applicability of alternative solvent selection guide on experimental data, which could assist in designing initial list of solvents for efficient process development, saving time and cost, compared to traditional way of selecting solvents.

(3) Holistic modelling of reaction data using DFT-based descriptors and machine learning

In Chapter 4, we developed a workflow to utilise publicly available data and physically meaningful descriptors for holistic modelling of Buchwald-Hartwig amine synthesis.[12] The

pipeline included a dozen machine learning algorithms (e.g., Neural Networks, XGBoost, Random Forest) to predict reaction yield, and explainable AI tools (e.g., permutation feature importance) to extract the most important and physically meaningful descriptors. This allowed for generation and validation of *a priori* knowledge before optimisation of a similar transformation in the laboratory. Indeed, transferability of the model selected results (i.e., important descriptors for parameterisation of reaction components) proved useful for a different Buchwald-Hartwig amination reaction in the lab, optimised using a fully automated flow setup. Moreover, the workflow included a novel way to calculate the steric descriptors of ligands, bases, and nucleophiles. Despite the increased number of features compared to a simple One-Hot Encoding (OHE) approach, DFT-based descriptors allowed for physical interpretability of the results, required less data to achieve high predictive accuracy, and demonstrated model generalisability. Compared to a single algorithm-based modelling, training ten ML algorithms allowed flexibilities around the most accurate algorithm selection depending on the reaction dataset size, distribution, and type. The results demonstrated that optimisation of new reactions in the laboratory could be accelerated using publicly available datasets and our ML-driven workflow.

(4) Bayesian optimisation of discrete and continuous variables for multiple objectives for a photoredox amine synthesis reaction in flow

In Chapter 5, we developed a novel way to simultaneously optimise discrete variables (e.g, solvents) and continuous variables (e.g., residence time, concentration, temperature) for objectives reaction yield and cost. Demonstrated on a highly sensitive photoredox amine synthesis reaction,[13] combination of *a priori* knowledge, continuous flow setup, and a recently developed Bayesian optimisation algorithm NEMO[14] allowed for development of robust reaction with reproducible results. Generation of a priori knowledge included solubility predictions and measurements of Hantzsch ester solid in 115 commonly used solvent candidates, UV-Vis absorbance of reaction components for optimal reaction concentration selection, and photon flux studies to quantify the influence of light for optimal reactor, lamp, and tubing selection. Using five black-box surrogate models and fully automated training and sampling in NEMO, algorithm suggested conditions led to identification and increase of Pareto points. Even though the reaction was not optimised for productivity, the highest productivity value achieved equals to ~12g/day production capacity of the final product using 10 mL

Vapourtec reactor, and the theoretical scalability of the reactor from 10 mL to 1 L without being limited by photon flux.

Research limitations

Some of the limitations of the developed workflows were highlighted in respective chapters. For instance, as stated in Chapter 3, list of suggested alternative solvents for a given reaction solvent is unlikely to generalise to all transformations. Solvents could play different roles depending on the reaction,[15] hence, reaction specific descriptors for solvents would be different. Even though the solvent candidates list was diversified using different similarity metrics (i.e., similarity fusion) and parameterisation techniques (i.e., data fusion) to capture combination of various solvent influences, it is unlikely to capture the complexity and interaction of reaction species. Therefore, it serves as a tool to assist process chemists with the decision space and is ideal for use with domain knowledge. Moreover, it might not always be feasible to balance process performance and SHE impact of solvents, as shown in Chapter 5, where the best identified solvents for high reaction yield were halogenated solvents dichloromethane and 1,2-dichloroethane.

Most of the machine learning algorithms require large dataset for high prediction accuracy. Therefore, holistic modelling of reactions, which is required for developing a robust model as opposed to quantifying the influence of a single reaction component, require large reactions data, which may not always be available for novel chemical transformations. This challenge was party addressed using dozens of ML algorithms, which are different in their learning efficiencies and data requirements. This poses a need for either more accurate data mining and extraction tools from large databases and publicly available data, and/or efficient generation of experimental data in-house. Moreover, it is computationally expensive, and potentially introduces variations, to calculate DFT-based descriptors for complex molecules such as the Pd-L complex in Chapter 4. This required development of a new workflow to add a reference H atom and optimise the structure with a new point of reference. Even though the distance to the point of reference (e.g., Pd → P was mimicked using H → P with Pd-P bond distance) was accurate, using H atom instead of Pd may lead to certain variations in the descriptor accuracy.

For sensitive reactions, identifying and finalising the reaction space requires significant amount of *a priori* knowledge as demonstrated in Chapter 5. For instance, solubility of Hantzsch ester solid, which affects the light attenuation and hardware configuration if present as a suspension

in the reaction, was predicted using COSMO*therm*[16] software in 115 solvents. Even though the results served as a qualitative suggestion to narrow down the solvent space for experimental measurements, the computational results were not accurate to be used quantitatively in reaction optimisation.

Outlook and suggestions for further work

In this section, future work regarding individual steps of the core research workflow developed in this project is described.

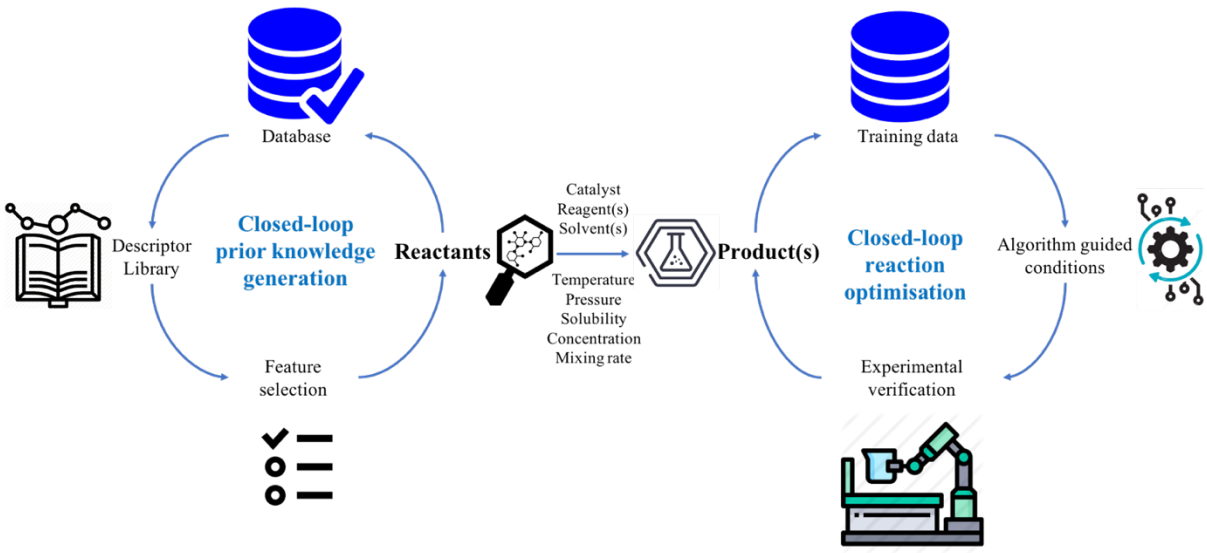


Figure 42. Schematic illustration of process development workflow developed in this project.

(1) Automated extraction of relevant reactions data and use of negative data

Collecting reaction data for a given chemical transformation has been demonstrated before in the form of hand-crafted literature data and based on automated classification of reactions. Although it was partly developed internally and was not included in the thesis, relevant data could be extracted from databases such as USPTO[17] and Reaxys[18] as reported before.[19, 20] This can be done using substructure (e.g., neighbouring atoms of the reaction centre) based search for matching electrophiles, nucleophiles, and final products with the target reaction. However, it is important to note that reactions data from various sources are often performed at different scales and setups (e.g., batch versus flow), and this, alongside continuous variables such as the reaction temperature, are often not reported and should be considered in reaction modelling.

Moreover, often, only successful (e.g., high yield) reactions are reported. This significantly narrows down the prediction range for a machine learning algorithm. For instance, if majority of the reported reactions for a given reaction type have above 70% yield, then it is unlikely for a model to learn the underlying chemical interactions in the reaction, and a naive guess of 85% yield for every condition would result in a high model prediction accuracy. To avoid this bias, we reported the average yield of ~40% for the holistic modelling of Buchwald-Hartwig amination reaction in Chapter 4. Therefore, it is important to include, either through existing electronic laboratory notebooks (ELNs), thesis, or new databases such as the Open Reaction Database (ORD),[21] negative data in reaction modelling. This, however, is not an issue when the data is generated in-house and the optimisation algorithm accounts for the extrapolation-exploitation trade-offs as demonstrated in Chapter 2 and Chapter 5.

(2) Efficient generation of molecular descriptors

Most of the previously reported modelling approaches utilise computationally demanding DFT-based descriptors, often based on Guassian. The increase in open-source cheminformatics packages such as Mordred,[22] Psi4,[23] and PySCF,[24] widely increased the list of potential descriptors and computational accessibility. Internally, preliminary results of modelling substrate screening reaction using Mordred-based descriptors, combined with quantum chemistry simulations using Psi4 demonstrated promising, and identified certain descriptors (e.g., HOMO, LUMO, reaction centre hybridisation state) that could be relevant to the reaction being studied. Alternatively, increased number of databases for molecular informatics, such as sigma profile database[25] or the Cambridge Crystallographic Data Centre (CCDC),[26] could provide a more efficient way of gathering molecular descriptors for a reaction.

(3) Adaptive feature engineering tools

Identifying reaction specific descriptors for all the reaction components play a significant role in model robustness, as stated several times in the thesis. Currently, most of the feature selection and engineering methods depend on domain knowledge, combined with time consuming hand-tuning of important descriptors. However, this could be accelerated using automated (combinatorial) trial of available list of descriptors for improved model performance, and/or more efficient feature selection tools such as the Boruta algorithm[27] and recursive feature elimination techniques, and statistical tools such as mutual information-based feature selection (MIFS).[28] This is especially important when a long list of descriptors is

available, and the model accuracy suffers from the "curse of dimensionality". Moreover, an interactive tool that allows for execution of these mentioned techniques with no coding requirement would significantly speed up the modelling.

(4) Clustering workflow for better sampling of discrete variables in DoE

When generating the initial training dataset, Latin Hypercube Sampling (LHS) was commonly used to sample from the discrete variables space mapped onto the continuous space using descriptors.[29] When the list of discrete variables and their parameterisation descriptors are small, a simple 3D visual using principal component analysis (PCA)[30] could reveal potential clustering of discrete variables and allow for intuition-driven selection of initial candidates to maximise for model learning. However, for large datasets, such as 457 solvents library, using LHS may not be the most efficient technique. Alternatively, various unsupervised ML algorithms (e.g., K-Means, GaussianMixtureModels, AffinityPropagation, DBSCAN) could be used to group discrete variables based on their "similarity" to each other.[31, 32] Sampling based on clustering of discrete variables could lead to more efficient sampling and representation of the discrete variables space.

(5) Laboratory automation and online analysis tools

Given the continuous advancements in laboratory automation in academia[33-40] and industry (e.g., Chemspeed, DeepMatter, Synthace, and Automata), more and more repetitive tasks in the lab could be automated. Whilst designing (i.e., identification and selection of) the reaction space, deciding the hardware setup, and initial validation of reaction execution still depend heavily on domain knowledge and expertise of process chemists, automation tools could help accelerate certain tasks such as substrate and condition screening.

(6) Multistep optimisation of reactions, including consideration of downstream processing

As elaborated in literature review in Chapter 2 and 5, most of the previous research was based on optimisation of single step processes. However, for a complete end to end process, process requirements of the next steps and downstream processing should be in consideration. For instance, a solvent used in step 1 of a reaction for high yield may not be the ideal solvent for step 2 of a reaction or for purification of the final product. Recently, several reports have demonstrated expansion to multistep process optimisation, and is more likely to be the main focus on research moving forward.[41-43]

Closing remarks

This thesis opened highlighting the underlying complexity of process development in the pharmaceuticals and fine chemicals industries, and highlighted the need for more time and cost efficient methods for process development. The main objective of this project was to integrate recent advancements in machine learning algorithms, cheminformatics packages, and laboratory automation tools, combined with ever increasing scientific data, into chemical research for transforming novel reactions to robust processes. The demonstrated workflows on several case studies reinforced the use of such technologies in research and development to help identify complex interactions in chemistry, assist with decision making, and free up more time by automating repetitive tasks in the laboratory.

References

1. Jorayev, P.; Russo, D.; Tibbetts, J. D.; Schweidtmann, A. M.; Deutsch, P.; Bull, S. D.; Lapkin, A. A., Multi-objective Bayesian optimisation of a two-step synthesis of p-cymene from crude sulphate turpentine. *Chemical Engineering Science* **2022,** *247*.
2. Bradford, E.; Schweidtmann, A. M.; Lapkin, A., Efficient multiobjective optimization employing Gaussian processes, spectral sampling and a genetic algorithm. *Journal of Global Optimization* **2018,** *71* (2), 407-438.
3. Mateos, C.; Nieves-Remacha, M. J.; Rincón, J. A., Automated platforms for reaction self-optimization in flow. *Reaction Chemistry & Engineering* **2019,** *4* (9), 1536-1544.
4. Diorazio, L. J.; Hose, D. R. J.; Adlington, N. K., Toward a More Holistic Framework for Solvent Selection. *Organic Process Research & Development* **2016,** *20* (4), 760-773.
5. Piccione, P. M.; Baumeister, J.; Salvesen, T.; Grosjean, C.; Flores, Y.; Groelly, E.; Murudi, V.; Shyadligeri, A.; Lobanova, O.; Lothschütz, C., Solvent Selection Methods and Tool. *Organic Process Research & Development* **2019,** *23* (5), 998-1016.
6. Bajusz, D.; Rácz, A.; Héberger, K., Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics* **2015,** *7* (1).
7. Bender, A.; Jenkins, J. L.; Scheiber, J.; Sukuru, S. C. K.; Glick, M.; Davies, J. W., How Similar Are Similarity Searching Methods? A Principal Component Analysis of Molecular Descriptor Space. *Journal of Chemical Information and Modeling* **2009,** *49* (1), 108-119.
8. Holliday, J. D.; Hu, C.-Y.; Willet, W., Grouping of Coefficients for the Calculation of Inter-Molecular Similarity and Dissimilarity using 2D Fragment Bit-Strings. *Combinatorial Chemistry & High Throughput Screening* **2002,** *5* (2).
9. Abraham, M. H.; Grellier, P. L., Substitution at saturated carbon. Part XX. The effect of 39 solvents on the free energy of Et3N, EtI, and the Et3N–EtI transition state. Comparison

with solvent effects on the equilibria Et3N + EtI ⇌ Et4N+I–and Et3N + EtI ⇌ Et4N++ I–. *J. Chem. Soc., Perkin Trans. 2* **1976,** (14), 1735-1741.
10. Ganase, Z. An experimental study on the effects of solvents on the rate and selectivity of organic reactions. Imperial College London, 2015.
11. Zhou, T.; Qi, Z.; Sundmacher, K., Model-based method for the screening of solvents for chemical reactions. *Chemical Engineering Science* **2014,** *115*, 177-185.
12. Baumgartner, L. M.; Dennis, J. M.; White, N. A.; Buchwald, S. L.; Jensen, K. F., Use of a Droplet Platform To Optimize Pd-Catalyzed C–N Coupling Reactions Promoted by Organic Bases. *Organic Process Research & Development* **2019,** *23* (8), 1594-1601.
13. Trowbridge, A.; Reich, D.; Gaunt, M. J., Multicomponent synthesis of tertiary alkylamines by photocatalytic olefin-hydroaminoalkylation. *Nature* **2018,** *561* (7724), 522-527.
14. Sung, S.; Jeraal, M. I.; Lapkin, A. A., Nomadic Evolutionary Multiobjective Optimisation (NEMO) algorithm, https://github.com/simonsung06/NEMO **2022**.
15. Buncel, E.; Stairs, R. A.; Wilson, H., *The role of the solvent in chemical reactions*. Oxford University Press: Oxford ; New York, 2003; p ix, 159 p.
16. COSMO*therm*, Version C3.0, Release 17.01; COSMOlogic GmbH & Co. KG.
17. USPTO, uspto.gov. **(accessed on September 27th, 2022)**.
18. Reaxys, reaxys.com **(accessed on September 27th, 2022)**.
19. Fortunato, M. E.; Coley, C. W.; Barnes, B. C.; Jensen, K. F., Machine learned prediction of reaction template applicability for data-driven retrosynthetic predictions of energetic materials. In *SHOCK COMPRESSION OF CONDENSED MATTER - 2019: Proceedings of the Conference of the American Physical Society Topical Group on Shock Compression of Condensed Matter*, 2020.
20. Schneider, N.; Lowe, D. M.; Sayle, R. A.; Landrum, G. A., Development of a Novel Fingerprint for Chemical Reactions and Its Application to Large-Scale Reaction Classification and Similarity. *Journal of Chemical Information and Modeling* **2015,** *55* (1), 39-53.
21. Kearnes, S. M.; Maser, M. R.; Wleklinski, M.; Kast, A.; Doyle, A. G.; Dreher, S. D.; Hawkins, J. M.; Jensen, K. F.; Coley, C. W., The Open Reaction Database. *Journal of the American Chemical Society* **2021,** *143* (45), 18820-18826.
22. Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T., Mordred: a molecular descriptor calculator. *Journal of Cheminformatics* **2018,** *10* (1).
23. Turney, J. M.; Simmonett, A. C.; Parrish, R. M.; Hohenstein, E. G.; Evangelista, F. A.; Fermann, J. T.; Mintz, B. J.; Burns, L. A.; Wilke, J. J.; Abrams, M. L.; Russ, N. J.; Leininger, M. L.; Janssen, C. L.; Seidl, E. T.; Allen, W. D.; Schaefer, H. F.; King, R. A.; Valeev, E. F.; Sherrill, C. D.; Crawford, T. D., Psi4: an open-source ab initio electronic structure program. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2012,** *2* (4), 556-565.
24. Sun, Q.; Berkelbach, T. C.; Blunt, N. S.; Booth, G. H.; Guo, S.; Li, Z.; Liu, J.; McClain, J. D.; Sayfutyarova, E. R.; Sharma, S.; Wouters, S.; Chan, G. K. L., PySCF: the Python-based simulations of chemistry framework. *WIREs Computational Molecular Science* **2017,** *8* (1).

25. Mullins, E.; Oldland, R.; Liu, Y. A.; Wang, S.; Sandler, S. I.; Chen, C.-C.; Zwolak, M.; Seavey, K. C., Sigma-Profile Database for Using COSMO-Based Thermodynamic Methods. *Industrial & Engineering Chemistry Research* **2006,** *45* (12), 4389-4415.
26. Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C., The Cambridge Structural Database. *Acta Crystallographica Section B Structural Science, Crystal Engineering and Materials* **2016,** *72* (2), 171-179.
27. Kursa, M. B.; Rudnicki, W. R., Feature Selection with the Boruta Package. *Journal of Statistical Software* **2010,** *36* (11).
28. Hanchuan, P.; Fuhui, L.; Ding, C., Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2005,** *27* (8), 1226-1238.
29. Tang, B., Orthogonal Array-Based Latin Hypercubes. *Journal of the American Statistical Association* **1993,** *88* (424).
30. Pearson, K., LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **1901,** *2* (11), 559-572.
31. Hadipour, H.; Liu, C.; Davis, R.; Cardona, S. T.; Hu, P., Deep clustering of small molecules at large-scale via variational autoencoder embedding and K-means. *BMC Bioinformatics* **2022,** *23* (S4).
32. Cassar, J., A big data approach for clustering large chemical datasets. **2019**.
33. Jeraal, M. I.; Sung, S.; Lapkin, A. A., A Machine Learning-Enabled Autonomous Flow Chemistry Platform for Process Optimization of Multiple Reaction Metrics. *Chemistry–Methods* **2020,** *1* (1), 71-77.
34. Clayton, A. D.; Manson, J. A.; Taylor, C. J.; Chamberlain, T. W.; Taylor, B. A.; Clemens, G.; Bourne, R. A., Algorithms for the self-optimisation of chemical reactions. *Reaction Chemistry & Engineering* **2019,** *4* (9), 1545-1554.
35. Holmes, N.; Akien, G. R.; Savage, R. J. D.; Stanetty, C.; Baxendale, I. R.; Blacker, A. J.; Taylor, B. A.; Woodward, R. L.; Meadows, R. E.; Bourne, R. A., Online quantitative mass spectrometry for the rapid adaptive optimisation of automated flow reactors. *Reaction Chemistry & Engineering* **2016,** *1* (1), 96-100.
36. Taylor, C. J.; Baker, A.; Chapman, M. R.; Reynolds, W. R.; Jolley, K. E.; Clemens, G.; Smith, G. E.; Blacker, A. J.; Chamberlain, T. W.; Christie, S. D. R.; Taylor, B. A.; Bourne, R. A., Flow chemistry for process optimisation using design of experiments. *Journal of Flow Chemistry* **2021,** *11* (1), 75-86.
37. Coley, C. W.; Thomas, D. A.; Lummiss, J. A. M.; Jaworski, J. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H.; Hicklin, R. W.; Plehiers, P. P.; Byington, J.; Piotti, J. S.; Green, W. H.; Hart, A. J.; Jamison, T. F.; Jensen, K. F., A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **2019,** *365* (6453).
38. MacLeod, B. P.; Parlane, F. G. L.; Brown, A. K.; Hein, J. E.; Berlinguette, C. P., Flexible automation accelerates materials discovery. *Nature Materials* **2021,** *21* (7), 722-726.
39. Shi, Y.; Prieto, P. L.; Zepel, T.; Grunert, S.; Hein, J. E., Automated Experimentation Powers Data Science in Chemistry. *Accounts of Chemical Research* **2021,** *54* (3), 546-555.

40. Ley, S. V.; Fitzpatrick, D. E.; Ingham, R. J.; Myers, R. M., Organic Synthesis: March of the Machines. *Angewandte Chemie International Edition* **2015,** *54* (11), 3449-3464.
41. Schotten, C.; Manson, J.; Chamberlain, T. W.; Bourne, R. A.; Nguyen, B. N.; Kapur, N.; Willans, C. E., Development of a multistep, electrochemical flow platform for automated catalyst screening. *Catalysis Science & Technology* **2022,** *12* (13), 4266-4272.
42. Clayton, A. D.; Schweidtmann, A. M.; Clemens, G.; Manson, J. A.; Taylor, C. J.; Niño, C. G.; Chamberlain, T. W.; Kapur, N.; Blacker, A. J.; Lapkin, A. A.; Bourne, R. A., Automated self-optimisation of multi-step reaction and separation processes using machine learning. *Chemical Engineering Journal* **2020,** *384*.
43. Sung, S.; Jeraal, M. I.; Lapkin, A. A., *(Manuscript in process for submission)*. **2022**.