6th Machine Learning and AI in Bio(Chemical) Engineering

06-07th July 2023

Conference booklet

UNIVERSITY OF CAMBRIDGE

6 July 2023

10:00 – 11:00	**Registration, welcome and refreshments**
11:00 – 12:00	**Keynote – Michael Brenner** *Scientific uses of automatic differentiation*
12:00 – 12:25	**Zsuzsanna Koczor-Benda** *High-throughput property-driven generative design of functional organic molecules*
12:25 – 12:50	**Kobi Felton** *ML-SAFT: A framework for PCP-SAFT parameter prediction*
12:50 – 14:00	**Lunch**
14:00 – 14:35	**Pietro Liò** *Generative and Graph models in chemistry and medicine*
14:35 – 15:00	**Ryan Greenhalgh** *Current methods for drug property prediction in the real world*
15:00 – 15:25	**Egon Heuson** *Enzyme activity prediction using neural networks,docking and high-throughput screening results*
15:25 – 15:35	**Break**
15:35 – 16:10	**Timur Madzhidov** *State-of-the-art on reaction prediction condition*
16:10 – 16:35	**Benoît Baillif** *Applying atomistic neural networks to bias conformer ensembles towards bioactive-like conformations*
16:35 – 17:00	**Tom Savage** *Multi-Fidelity Data-Driven Design and Analysis of Reactor and Tube Simulations (DARTS)*
17:00 – 19:00	**Networking and dinner**
19:00	**End of day 1**

7 July 2023

09:00 – 09:15	**Coffee reception**
09:15 – 10:00	**Workshop part 1** *Rapid predictive modelling without having to write code*
10:00 – 10:15	**Break**
10:15 – 11:00	**Workshop part 2** *Rapid predictive modelling without having to write code*
11:00 – 12:00	**Poster Session**
12:00 – 12:35	**Antonio Del Rio Chanona** *Building Models with Machine Learning*
12:35 – 13:00	**Jiaru Bai** *From Platform to Knowledge Graph: Distributed Self-Driving Laboratories*
13:00 – 14:00	**Lunch**
14:00 – 14:35	**Stefan Born** *Machine Learning as an integral part of an automated experimental workflow in protein engineering*
14:35 – 15:00	**Nishanthi Gangadharan** *Data-driven Dynamic Control Scheme for Antibody Producing CHO Cell Cultures in Fed Batch*
15:00 – 15:20	**Break**
15:20 – 15:45	**Emma Smith King** *Practical Machine Learning for Synthetic Chemistry*
15:45 – 16:10	**Miruta Cretu** *Standardizing chemical compounds using language models*
16:10 – 16:45	**Closing remarks**
16:45	**End of day 2**

All reported times in BST

Keynote Speakers

Dr. Michael Brenner

Michael is a faculty member in SEAS and Physics at Harvard University. He has a PhD in Physics from the University of Chicago, where he worked with Professor Leo Kadanoff. His first faculty position was at MIT in the Mathematics Department before joining the faculty of Harvard University in 2002. His research uses methods and ideas of applied mathematics to address problems in science and engineering. Current research directions range from figuring out the nature of the turbulent cascade, to understanding the rules for building materials that assemble themselves, possibly with life-like properties, to efforts to use recent advances in machine learning to facilitate scientific discovery.

Invited Speakers

Pietro Liò

Pietro Liò is a Full Professor at the department of Computer Science and Technology of the University of Cambridge and a member of the Artificial Intelligence group. He is also a member of the Cambridge Centre for AI in Medicine. His research interest focuses on developing Artificial Intelligence and Computational Biology models to understand diseases complexity and address personalised and precision medicine. Current focus is on Graph Neural Network modeling.

Timur Madzhidov

Senior Product Manager in Elsevier, responsible for the development of AI-driven tools based on Reaxys data and improvement of Reaxys data readiness for AI and ML application. Chemoinformatics specialist, researcher and educator. Topic of primary interest: reaction informatics, AI in chemistry, algorithmic chemoinformatics, chemistry-aware machine learning. Before joining Elsevier, Timur was Leading Researcher and Director of Intelligent Robochemistry lab, Kazan Federal University, Russia, group Leader in the Lab in Chemoinformatics and Molecular Modeling. Before 2022, he was a member of the Reaxys R&D collaboration network supported by Elsevier, as well as the collaboration "Machine Design of Small Molecules by Artificial Intelligence" supported by Janssen Pharmaceutics. He is a founder, lecturer (as associate professor), and former supervisor of the Master Program in Chemoinformatics and Molecular Modeling of Kazan Federal University, the first master program in chemoinformatics in Russia. Since 2013 to 2022 the program operated as the Double-Diploma with the University of Strasbourg.

Antonio Del Rio Chanona

Antonio Del Rio Chanona is the head of the Optimisation and Machine Learning for Process Systems Engineering group at the Department of Chemical Engineering, and the Centre for Process Systems Engineering, Imperial College London. His research focuses on developing and applying computer algorithms from the area of optimization, machine learning and reinforcement learning to engineering systems. The applied branch of his research looks at bioprocess control, optimization and scale-up.

Stefan Born

Technische Universität Berlin, TUB · Department of Biotechnology and Department of Mathematics, PhD.

Oral Talks

ML-SAFT: A framework for PCP-SAFT parameter prediction

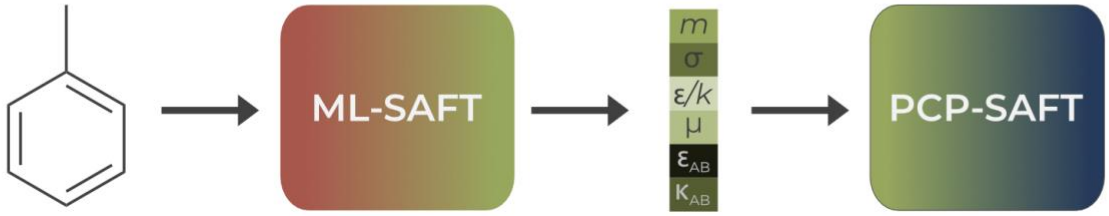
Kobi Felton[1, 2], Lukas Raßpe-Lange,[2] Jan G. Rittig, [3] Kai Leonhard, [2] Alexander Mitsos, [2,4,5]
Julian Meyer-Kirschner,[8] Carsten Knösche, [8] Alexei Lapkin[1,6,7]*
* Corresponding author

[1] Department of Chemical Engineering and Biotechnology University of Cambridge, Cambridge UK.
[2] Process Systems Engineering (AVT.SVT), RWTH Aachen University, 52074 Aachen, Germany.
[3] Institute of Technical Thermodynamics, RWTH Aachen University, 52062 Aachen, Germany
[4] Institute for Energy and Climate Research IEK-10: Energy Systems Engineering, Forschungszentrum Jülich GmbH, J¨ulich 52425, Germany
[5] JARA-ENERGY, Aachen 52056, Germany
[6]Innovation Centre in Digital Molecular Technology, Yusuf Hamied Department of
[7] Chemistry, University of Cambridge
[8] BASF SE, 67056 Ludwigshafen am Rhein, Germany

Keywords: Thermodynamics, machine learning, PC-SAFT

Abstract:

Fast and accurate prediction of fluid-phase thermodynamics is a long-standing interest of the processing engineering community. Over the last fifty years, a variety of methods have been developed ranging from group contribution methods to quantum chemical simulations to, most recently, machine learning methods. However, there is still a need for methods that can extend to a wide range of compounds without significant tuning or introspection from the end user. Group contribution methods that require careful and often manual identification of functional groups on molecules that match a database, and existing quantum mechanical (QM) methods often require significant expertise and computational cost. Machine learning methods have demonstrated promise in the prediction of thermodynamic parameters, yet many lack the thermodynamic consistency of classical thermodynamic models. Recently, it has been shown that that using a machine learning model to predict the parameters an existing classical Equation of State (EoS) can overcome the challenges of thermodynamic consistency.



In this work, we develop ML-SAFT, a framework for predicting PCP-SAFT parameters using machine learning. ML-SAFT contains several machine learning models and, most importantly, the largest database of PCP-SAFT parameters published in the literature (986 molecules). We extract data from the Dortmund Databank and develop a robust regression method to determine pure component PCP-SAFT parameters from experimental vapor pressure and liquid density data. We then train random forests, feed forward networks and message passing neural networks (MPNNs) to predict the regressed PCP-SAFT parameters.

Our results show that random forests obtain the most accurate predictions of the regressed PCP-SAFT parameters. Furthermore, the best prediction of vapor pressure on unseen molecules is obtained from the random forest. However, the best results on density predictions were obtained with parameters predicted by a MPNN. We attribute this difference to the increased representation capability of the MPNN for polar molecules, which we found to be important for density predictions.

Overall, our work demonstrates that machine learning is a powerful tool for PCP-SAFT parameter prediction. We foresee that the results shown in this work can form a baseline for future work that explores multicomponent mixture predictions using PCP-SAFT.

Applying atomistic neural networks to bias conformer ensembles towards bioactive-like conformations

Benoît Baillif[1], Jason Cole[2], Patrick McCabe[2], Andreas Bender[1]

[1] Yusuf Hamied Department of Chemistry, University of Cambridge, Lensfield Rd, CB2 1EW, Cambridge, United Kingdom

[2] Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, United Kingdom

Keywords: atomistic neural network, conformer ensemble, bioactive conformation, rigid-ligand docking, maximum common substructure

Abstract (<400 words):

The generation of energetically favourable conformations of small molecules is a common task in drug design such as during docking or pharmacophore searching. While recent conformer generators create bioactive-like conformations for most known ligands, there is currently no general method to identify them among the set of generated conformers[1], and developing methods to prioritise conformers that could represent likely target-bound poses is therefore desirable. In this work we extracted 13,460 bioactive conformations of 10,481 curated ligands in the PDBbind[2] dataset and generated up to 250 conformers for each ligand. We then trained atomistic neural networks (AtNNs) with various levels of completeness to process 3D information of generated conformers to predict the atomic root-mean-square deviation to its closest bioactive conformation (ARMSDbio). The model was compared with bioactivity-unaware ranking baselines such as a random ordering of conformers or MMFF94s energy ranking, and a bioactivity-based baseline taking the Torsion Fingerprint Deviation to the Maximum Common Substructure to the closest molecule in the training set (TFD2RefMCS). On a random ligand splitting of PDBbind, ranking conformers by the predicted $\mathrm{ARMSD}_{\mathrm{bio}}$ from the most expressive AtNNs, ComENet[3], leads to early enrichment of bioactive-like ($\mathrm{ARMSD}_{\mathrm{bio}}$ < 1 Å) conformations measured with a median BEDROC of 0.29 ± 0.02, outperforming the best bioactivity-unaware MMFF94s energy ranking baseline showing a median BEDROC of 0.18 ± 0.02, and performing on a par with the slower bioactivity-based TFD2RefMCS baseline showing a median BEDROC of 0.31 ± 0.02. Moreover, when restricting only to harder test sets to flexible molecules, the bioactivity-unaware baselines showed median BEDROCs close to 0.01, while AtNNs and TFD2RefMCS showed median BEDROCs up to 0.12. When performing rigid-ligand re-docking of PDBbind ligands with GOLD[4] using the 1% top-ranked conformers, ComENet showed a higher successful docking rate than bioactivity-unaware baselines, with a rate of 0.48 ± 0.02 compared to Generation order with a rate of 0.39 ± 0.02. Hence, the approach presented here uses AtNNs successfully to focus conformer ensembles towards bioactive-like conformations, representing an opportunity to reduce computational expense in virtual screening applications on known targets that requires input conformations.

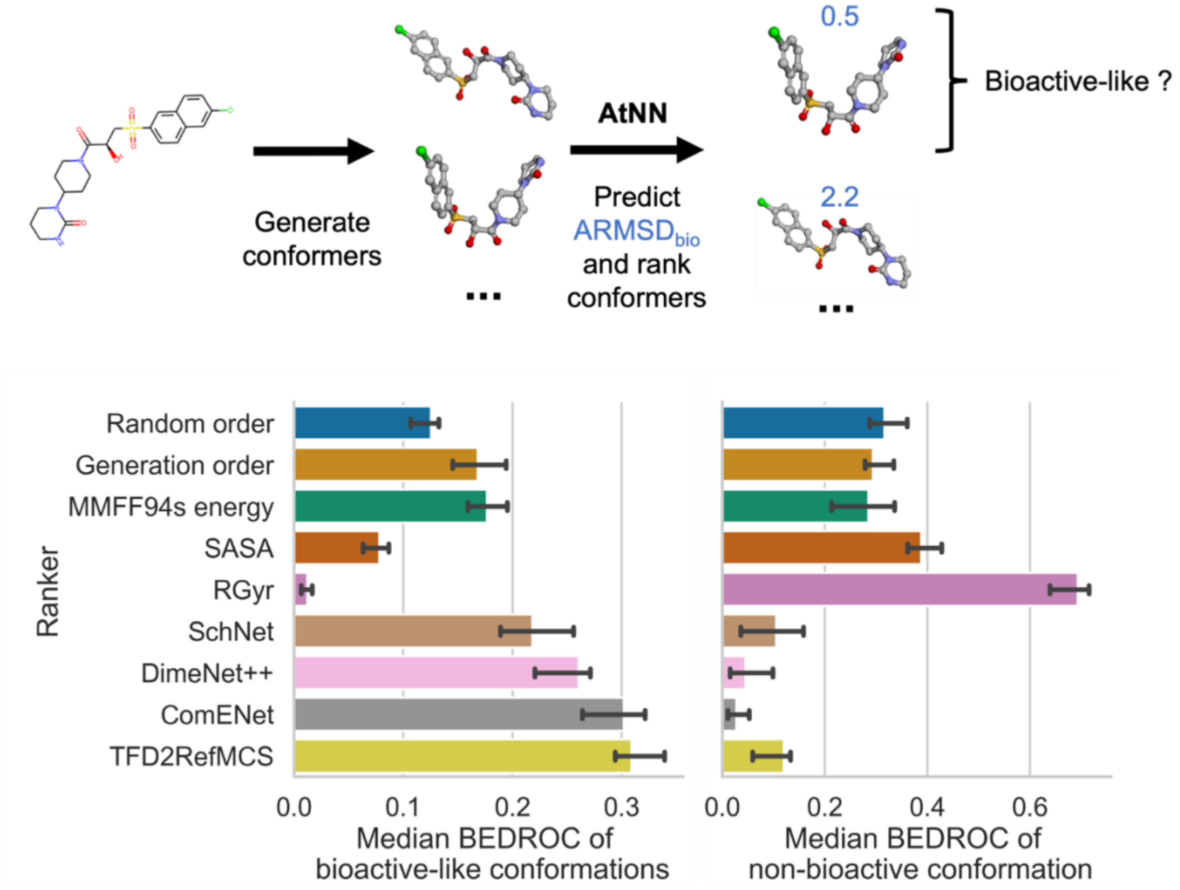
References

1.Habgood, M. Bioactive Focus in Conformational Ensembles: A Pluralistic Approach. *J. Comput. Aided Mol. Des.* **2017**, *31* (12), 1073–1083. https://doi.org/10.1007/s10822-017-0089-3.

2.Liu, Z.; Li, Y.; Han, L.; Li, J.; Liu, J.; Zhao, Z.; Nie, W.; Liu, Y.; Wang, R. PDB-Wide Collection of Binding Data: Current Status of the PDBbind Database. Bioinformatics 2015, 31 (3), 405–412. https://doi.org/10.1093/bioinformatics/btu626.

3. Wang, L.; Liu, Y.; Lin, Y.; Liu, H.; Ji, S. ComENet: Towards Complete and Efficient Message Passing for 3D Molecular Graphs. arXiv June 16, 2022. https://doi.org/10.48550/arXiv.2206.08515.

4. Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved Protein-Ligand Docking Using GOLD. Proteins Struct. Funct. Bioinforma. 2003, 52 (4), 609–623. https://doi.org/10.1002/prot.10465.



Classical and Deep Learning for Drug Property Prediction and Uncertainty Estimation: A Comparative Study

Cecilia Cabrera[1], Andrea Dimitracopoulos*[1], Jacob Green[1], Ryan Greenhalgh**[1], Maximilian Jakobs[1], Mark van der Wilk[1,2]

* Presenting author

** Corresponding author

(Authors in alphabetical order)

[1] DeepMirror

[2] Imperial College London

Keywords: Medicinal Chemistry, Drug properties prediction, Machine Learning, Deep Learning Architectures, Uncertainty Estimation

Abstract (<400 words):

The process of accurately predicting drug properties is a crucial step in drug development as it helps to de-risk laboratory testing of assets before clinical trials. Deep learning has emerged as a highly promising tool, augmenting the capabilities of machine learning models to predict absorption, distribution, metabolism, and excretion (ADME) property predictions, and thereby facilitating *in silico* screening. However, the challenge extends beyond predicting a single value for each of these properties. To enhance the confidence of chemists engaged in subsequent laboratory explorations, it is imperative for models to offer estimates of uncertainty, thereby shedding light on the dependability of the predictions. This approach enables chemists to strike a balance when examining new potential candidates for lab validation - they can appraise the trade-off between prioritising molecules exhibiting substantial potential for improvement in one or more properties of interest, and the probability of these enhancements being observed in the laboratory. In this study, we compare common deep learning architectures for featurization, such as Graph Neural Networks and Large Language Models, using semi-supervised methods along with classical fingerprinting on a variety of ADME datasets (~50). We also assess the capacity of these models to estimate uncertainty using various methods including ensembles and Bayesian inference. Our findings shed light on circumstances under which classical approaches can outperform deep learning architectures in terms of predictive power and provide insights into the trade-offs between different architectures and methods. We further discuss the implications of our findings for drug discovery and development.

Multi-Fidelity Data-Driven Design and Analysis of Reactor and Tube Simulations (DARTS)

Tom Savage[1,2], Nausheen Basha[2] Jonathan McDonough[3], Omar Matar[2], Ehecatl Antonio del Rio Chanona*[1,2]

* Corresponding author

[1] Sargent Centre for Process Systems Engineering, Imperial College London.

[2] Department of Chemical Engineering, Imperial College London.

[3] School of Engineering, Newcastle University.

Keywords: Optimisation, Simulations, Multi-fidelity, Additive Manufacturing, Data-driven

Abstract (<400 words):

The development of new manufacturing techniques such as 3D printing have enabled the creation of previously infeasible chemical reactor designs. Systematically optimizing the highly parameterized geometries involved in these new classes of reactor is vital to ensure enhanced mixing characteristics and feasible manufacturability.

Here we present a framework to rapidly solve this nonlinear, computationally expensive, and derivative-free problem, enabling the fast prototype of novel reactor parameterizations. We take advantage of Gaussian processes to adaptively learn a multi-fidelity model of reactor simulations across a number of different continuous mesh fidelities. The search space of reactor geometries is explored through an amalgam of different, potentially lower, fidelity simulations which are chosen for evaluation based on a weighted acquisition function, trading off information gain with cost of simulation. Figure 1 demonstrates the objective value as optimisation progresses, as well as time taken for each simulation.

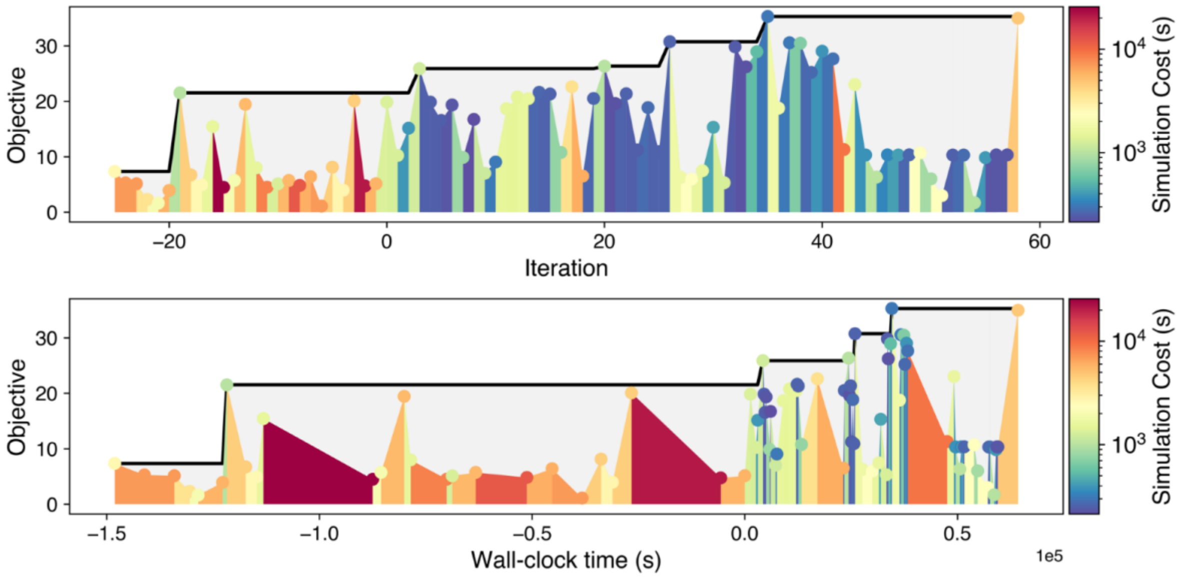


Figure 1: The number of equivalent tanks-in-series evaluated colored by the respective cost of simulation. The upper half of the figure shows these quantities against iteration and the lower half shows these quantities against wall-clock time, highlighting the importance of lower-cost simulations.

Within our framework we derive a novel criterion for monitoring the progress and dictating the termination of multi-fidelity Bayesian optimization, ensuring a high-fidelity solution is returned before experimental budget is exhausted. The class of reactor we investigate are helical-tube reactors under pulsed-flow conditions, which have demonstrated outstanding mixing characteristics, have the potential to be highly parameterized, and are easily manufactured using 3D printing. Figure 2 demonstrates the impact of the coil radius parameter on reactor mesh.

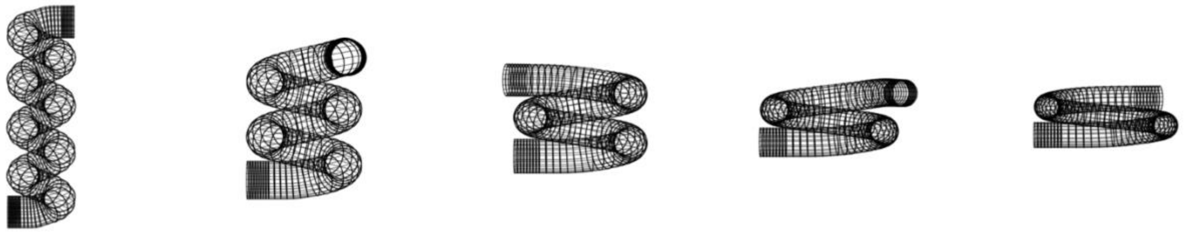


Figure 2: The effect of coil radius for a helical coil tube with a fixed length.

To validate our results, we 3D print and experimentally validate the optimal reactor geometry, confirming its mixing performance. In doing so we demonstrate our design framework to be extensible to a broad variety of expensive simulation-based optimization problems, supporting the design of the next generation of highly parameterized chemical reactors. Figure 3 demonstrates the final optimal reactor design, the 3D printed optimal reactor, and experimental validation of performance.

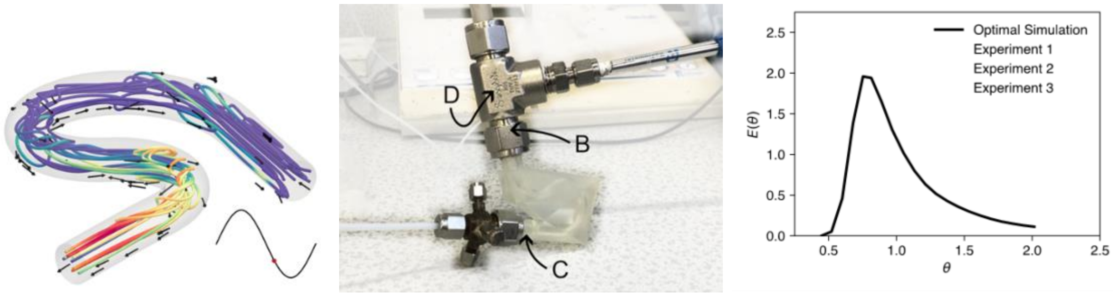


Figure 3: a) Streamlines indicating tracer concentration within the optimal reactor. b) 3D printed reactor geometry. c) The residence-time distribution predicted via CFD simulation of the solution returned from the framework, alongside 3 sets of experimental data obtained via the 3D printed reactor.

References

1.Savage T, Basha N, Matar O, EAD Chanona. Deep Gaussian Process-based Multi-fidelity Bayesian Optimization for Simulated Chemical Reactors. arXiv preprint arXiv:2210.17213. 2022 Oct 31.

2. McDonough JR, Murta S, Law R, Harvey AP. Oscillatory fluid motion unlocks plug flow operation in helical tube reactors at lower Reynolds numbers (Re≤ 10). Chemical Engineering Journal. 2019 Feb 15;358:643-57.

3.Kandasamy K, Dasarathy G, Schneider J, Póczos B. Multi-fidelity bayesian optimisation with continuous approximations. InInternational Conference on Machine Learning 2017 Jul 17 (pp. 1799-1808). PMLR

From Platform to Knowledge Graph: Distributed Self-Driving Laboratories

Jiaru Bai,[1] Sebastian Mosbach,[1,2] Connor J. Taylor,[3,4] Dogancan Karan,[2] Kok Foong Lee,[5] Simon D. Rihm,[1,2,6] Jethro Akroyd,[1,2] Alexei A. Lapkin,[1,2,4] Markus Kraft[1,2,7,8*]

* Corresponding author: mk306@cam.ac.uk (M.K.)

[1] Department of Chemical Engineering and Biotechnology, University of Cambridge, Philippa Fawcett Drive, Cambridge CB3 0AS, United Kingdom.
[2] Cambridge Centre for Advanced Research and Education in Singapore (CARES), CREATE Tower #05-05, 1 Create Way, Singapore 138602, Singapore.
[3] Astex Pharmaceuticals, 436 Cambridge Science Park Milton Road, Cambridge CB4 0QA, United Kingdom.
[4] Innovation Centre in Digital Molecular Technologies, Yusuf Hamied Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom.
[5] CMCL Innovations, Sheraton House, Cambridge CB3 0AX, United Kingdom.
[6] Department of Chemical & Biomolecular Engineering, National University of Singapore, 4 Engineering Drive 4, Singapore 117585, Singapore.
[7] School of Chemical and Biomedical Engineering, Nanyang Technological University, 62 Nanyang Drive, Singapore 637459, Singapore.
[8] The Alan Turing Institute, London NW1 2DB, United Kingdom.

Keywords: Dynamic knowledge graph, laboratory automation, digital twin, distributed laboratory, goal-driven self-optimisation

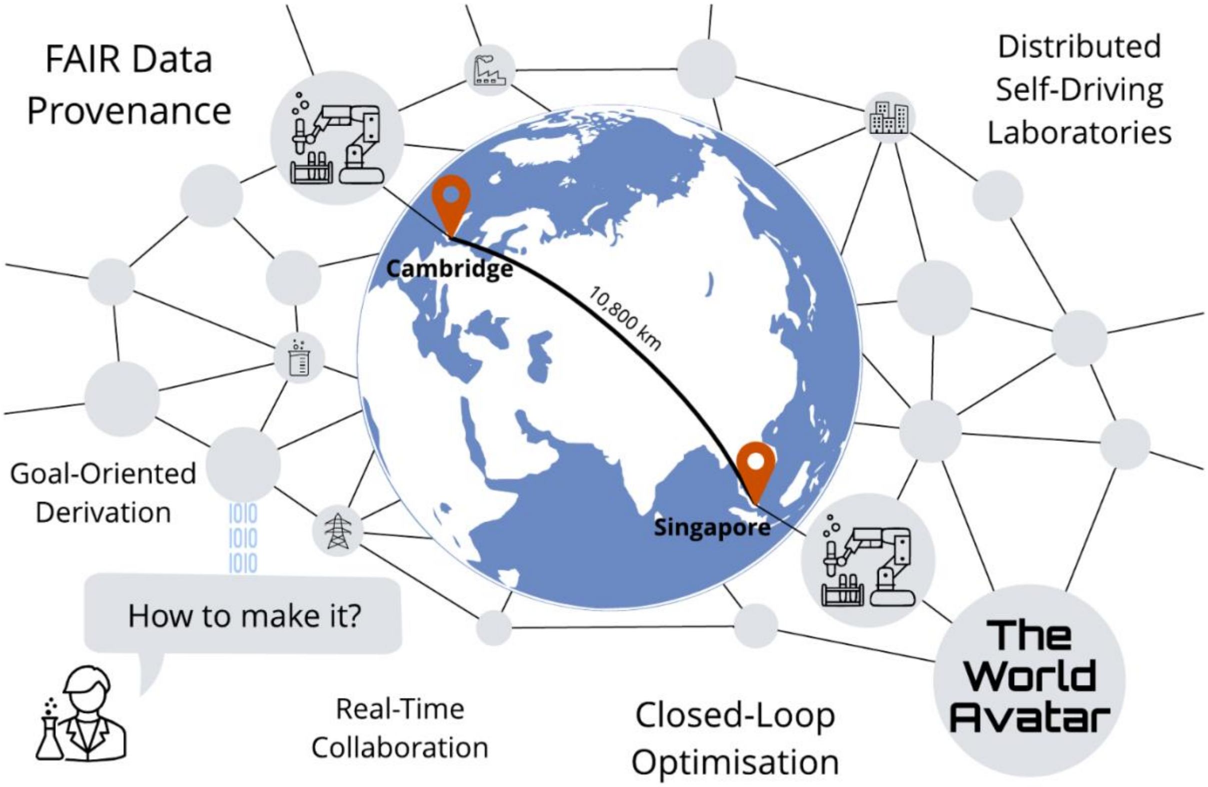
Abstract (<400 words):

The ability to integrate resources and share knowledge across organisations enables scientists to expedite the scientific discovery process, which is especially crucial in addressing emerging global challenges that require global solutions [1, 2]. In this work, we develop an architecture to enable distributed self-driving laboratories as part of The World Avatar project, an all-encompassing digital twin based on a dynamic knowledge graph. Our approach utilises ontologies to capture the data and material flows involved in design-make-test-analyse cycles, and employs autonomous agents as executable knowledge components to carry out the experimentation workflow. All data provenance is recorded following FAIR principles, ensuring its accessibility and interoperability. We demonstrate the practical application of our framework by linking two robotic setups in Cambridge and Singapore to achieve a collaborative closed-loop optimisation for a pharmaceutically-relevant aldol condensation reaction in real time. The knowledge graph evolves autonomously while progressing towards the research goals set by the scientist. The two robots effectively produced a Pareto front for the cost-yield optimisation problem over the course of two days of operation. This proof-of-concept demonstration highlights the potential of the framework to establish a globally collaborative research network and further advance scientific frontiers.

References

1.Seifrid, M. et al. Autonomous Chemical Experiments: Challenges and Perspectives on Establishing a Self-Driving Lab. Acc. Chem. Res. 2022;55(17): 2454–2466.

2.Bai, J. et al. From Platform to Knowledge Graph: Evolution of Laboratory Automation. JACS Au. 2022;2(2): 292–309.


FAIR Data
Provenance
Distributed
Self-Driving
Laboratories
Cambridge
10,800 km
Singapore
Goal-Oriented
Derivation
1010
1010
1010
How to make it?
Real-Time
Collaboration
Closed-Loop
Optimisation
The
World
Avatar

Machine Learning as an integral part of an automated experimental workflow in protein engineering

Stefan Born[1], Mark Doerr[2]

[1] Technische Universität Berlin,
[2] Universität Greifswald, Institut für Biochemie.

Keywords: FAIR data, automated experimental planning, automated machine learning, transfer learning, multitask learning

Abstract (<400 words):

The talk addresses some of the obstacles in the application of machine learning (ML) to protein engineering and relates these to requirements on software architecture and data management, which we believe to be valid beyond this domain.

Unlike classical statistical models, machine learning can only play out its full potential when large datasets can be aggregated. In the quest for new proteins or enzymes the sizes of labelled datasets for a specific task, e.g. a specific catalytic function are typically quite small. The complexity of models that can be trained on such limited data is limited as well, however models can share (some) parameters across different tasks (multitask learning, [1]) or reuse parameters trained on other tasks, possibly large scale unsupervised (transfer learning, [2,3]) or be constrained by domain knowledge [4]. Many such approaches are presently explored by scientists who manually clean and prepare the data and assemble and train the models.

After automation of experiment execution in robotic labs the automation of modelling and experimental planning is a logical next step. In order to achieve this the implicit domain knowledge of scientists has to be made explicit. Data would have to be annotated by metadata with defined semantics. Models would need the metadata to select their inputs and to determine permissible train-test splits. Metadata must include detailed information about experimental procedures, as e.g. enzymatic activity data of the same reaction type, but from different experiments or labs are often not directly comparable. A model on aggregated data from different labs would require some encoding of the conditions as an additional predictor in order to account for the differences.

Following these lines, we discuss some modelling examples with respect to the relation of FAIR data and software for automated model building, training and selection.

Ultimately predictive models would be used to take good decisions in the planning of the next experiment. We try to get a better understanding of what this means in the light of Bayesian decision theory.

At the end we give a very short overview of LARAsuite (https://gitlab.com/larasuite), which is a free and open source research data management system that addresses the problems of manual data insertion and metadata assignment by establishing radically automated processes. Data and Meta- data is mainly not entered by humans, but by machines.

References

1. Castro E, Godavarthi A, Rubinfien J, Givechian K, Bhaskar D, Krishnaswamy S. Transformer-based protein generation with regularized latent space optimization. Nat Mach Intell. 2022 Oct;4(10):840–51.
2. Rao R, Bhattacharya N, Thomas N, Duan Y, Chen P, Canny J, et al. Evaluating Protein Transfer Learning with TAPE. In: Wallach H, Larochelle H, Beygelzimer A, Alché-Buc F d', Fox E, Garnett R, editors. Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2019. Available from:

https://proceedings.neurips.cc/paper_files/paper/2019/file/37f65c068b7723cd7809ee2d31d7861c-Paper.pdf
3. Fenoy E, Edera AA, Stegmayer G. Transfer learning in proteins: evaluating novel protein learned representations for bioinformatics tasks. Briefings in Bioinformatics. 2022 Jul 18;23(4):bbac232.
4. Ao YF, Pei S, Xiang C, Menke MJ, Shen L, Sun C, u. a. Structure- and Data-Driven Protein Engineering of Transaminases for Improving Activity and Stereoselectivity. Angewandte Chemie International Edition. 2023;62(23):e202301660.

Data-driven Dynamic Control Scheme for Antibody Producing CHO Cell Cultures in Fed Batch

Nishanthi Gangadharan[1], Ayca Cankorur-Cetinkaya[2] , Matthew Cheeks[2], Alexander F Routh[1,3] , Duygu Dikicioglu[1,4*]
* Corresponding author

[1] Department of Chemical Engineering and Biotechnology, University of Cambridge, Cambridge, CB3 0AS, UK.
[2] Cell Culture and Fermentation Sciences, BioPharmaceuticals R&D, AstraZeneca, Cambridge, UK.
[3] Institute of Energy and Environmental Flows, University of Cambridge, Cambridge, CB3 0EZ, UK
[4] Department of Biochemical Engineering, University College London, London, WC1E 6BT, UK

Keywords: Fed batch bioreactor, Data-driven control, Dynamic control scheme, Machine Learning, Process Analytical Technology

Abstract:

Effective process control is a basic requirement for biopharmaceutical manufacturing to achieve high throughputs and enhanced quality control. High non-linearity and uncertainties associated with bioprocesses challenge the ability of traditional controllers to deliver satisfactory performance, thereby creating an urgent need for advanced model based control strategies for efficient bioprocess control. Process Analytical Technology (PAT) initiative has highlighted the importance of identifying critical process parameters (CPP) of a bioprocess that influence critical quality attributes (CQA), to achieve seamless integration of analytical data with real-time monitoring and control for enhanced process understanding and to overcome manufacturing challenges.

Multivariate monitoring techniques in biopharmaceuticals has resulted in the generation of large amounts of data comprising real-time measurements of critical quality and performance attributes, and if exploited efficiently can provide opportunity for developing superior control action. This study explores the different stages of development of a novel data-driven dynamic control scheme for bioprocesses in the context of antibody producing CHO cell cultures in fed-batch bioreactors. In order to harness the full potential of machine learning models for bioprocess control, we reinforced them with concepts from network theory and control theory. The models generated could predict the expected trajectory of a cell culture based on process knowledge from historic bioprocess data and propose a customized reactive control action when encountered with a deviation from the expected trajectory. The proposed closed-loop model-based multi-attribute control scheme, that combines concepts from data science, network theory and control theory, was capable of recommending sensible control actions, which ensure that the cultures remain on a pre-defined well established trajectory thereby minimising variability.

Practical Machine Learning for Synthetic Chemistry

Emma King-Smith[1,2], Alpha A. Lee*[2]
* Corresponding author

[1] Department of Chemistry, University of Cambridge.
[2] Department of Physics, University of Cambridge.

Keywords: Organic chemistry, Machine Learning, Regiochemical Prediction, High-Throughput Experimentation

Abstract (<400 words):

Synthetic chemistry has many open challenges: how reaction yields change as reactants and conditions change, [1] how molecules interact with the human body, [2] or the full underlying mechanisms of some workhorse reactions. [3] Machine learning (ML) has seen enormous strides in modeling the world's "black boxes": from image processing and recognition that rival human ability, [4] consistently beating human players in a variety of games, [5] to the amusing ruminations of the latest large language models. [6] Due to the low standardization of data, few large chemistry-focused datasets, and the mere fact that molecules are difficult systems to model, ML has historically struggled to make headway in the chemical sciences. [7] Recent developments in ML models and increased access to open-source chemistry datasets have opened the door to practical ML models, including DFT and molecular property predictions, activity predictions, and novel scaffold generation. Herein, we present two case studies utilizing recent and classic ML methods to further our predictive ability in and understanding of synthetic chemistry.

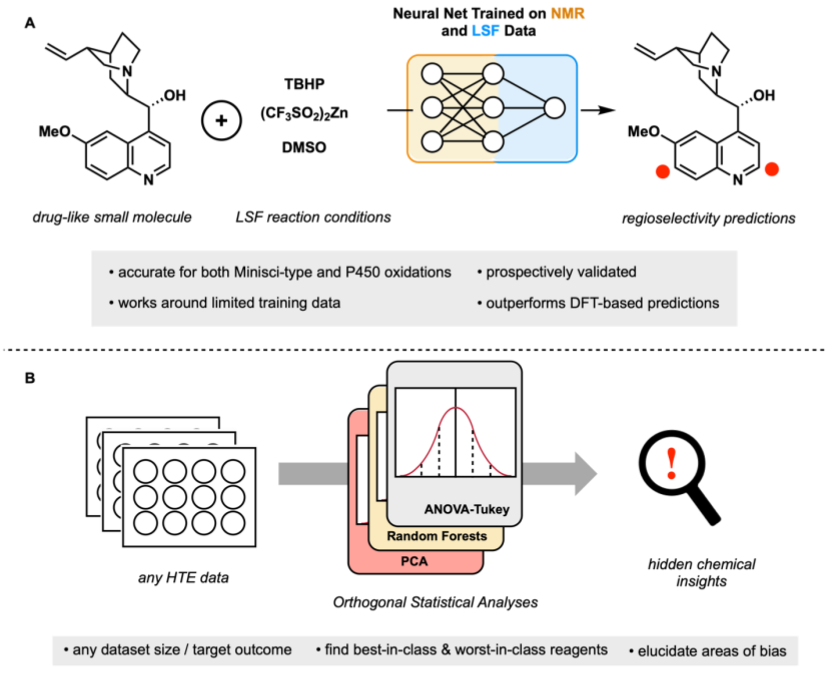

A
Neural Net Trained on NMR and LSF Data
TBHP
(CF3SO2)2Zn
DMSO
MeO
OH
N
drug-like small molecule
LSF reaction conditions
regioselectivity predictions
• accurate for both Minisci-type and P450 oxidations
• prospectively validated
• works around limited training data
• outperforms DFT-based predictions
B
any HTE data
ANOVA-Tukey
Random Forests
PCA
Orthogonal Statistical Analyses
hidden chemical insights
• any dataset size / target outcome
• find best-in-class & worst-in-class reagents
• elucidate areas of bias

Figure 1: (**A**) Graphical outline of ML predictions of Minisci and P450 LSF regiochemical outcomes. (**B**) Overview of HiTEA.

First, we investigated ML applied to chemical transformations aimed at structural diversification of drug-like molecules, late stage functionalizations (LSFs). These types of reactions are a key component of drug discovery, capable of rapidly exploring the chemical space to yield pharmacokinetically ideal compounds. [8] However, predicting the regiochemical outcomes of LSFs is still an open challenge in the field. Notably, experimental data curation is labor-intensive and time consuming. We report the development of an approach that combines a message passing neural network and ^{13}C NMR-based transfer learning to predict the atom-wise probabilities of functionalization. [9] We validated our model retrospectively and with a series of prospective experiments, showing that it accurately predicts the outcomes of Minisci-type and P450 transformations, outperforming state-of-the-art Fukui-based reactivity indices and other

graph-based ML models (Figure 1A). [10]

The second case study developed a dataset-ambivalent ML framework to analyze high-throughput experimentation (HTE) datasets. HTE has the potential to improve our understanding of organic chemistry by systematically interrogating reactivity across diverse chemical spaces. One notable bottleneck is the lack of facile analyzers which can interpret of these data's hidden chemical insights. [11] Herein we report the development of a **Hi**gh **T**hroughput **E**xperimentation **A**nalyzer (HiTEA), a robust and statistically rigorous framework which is applicable to any HTE dataset regardless of size, scope, or target reaction outcome. [12] HiTEA is validated on previously proprietary medicinal chemistry data, elucidating hidden biases and relationships between reaction components (Figure 1B).

References

1. Taylor C J, Pomberger A, Felton K C, Grainger R, Barecka M, Chamberlain T W, Bourne R A, Johnson C N, Lapkin A A. A Brief Introduction to Chemical Reaction Optimization. Chem. Rev. 2023;123(6): 3089-3126.
2. Nantasenamat C, Isarankura-Na-Ayudhya C, Naenna T, Prachayasittikul V. A practical overview of quantitative structure-activity relationship. EXCLI J. 2009;8: 74-88.
3. Sperotto E, van Klink G P M, van Koten G, de Vries J. The mechanism of the modified Ullmann reaction. Dalton Trans. 2010;39: 10338-10351.
4. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Zhiheng H, Karpathy A, Khosla A, Bernstein M, Berg A, Li F-F. ImageNet Large Scale Visual Recognition Challenge. Int. J. Comput. Vis. 2015;115: 211-252.
5. Risi S, Preuss M. From Chess and Atari to StarCraft and Beyond: How Game AI is Driving the World of AI. KI - Kunstl. 2020;34: 7-17.
6. Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, Lee P, Lee Y T, Li Y, Lundberg S, Nori H, Palangi H, Ribeiro M T, Zhang Y. Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv preprint. 2023; arXiv:2303.12712.
7. Coley C W, Green W H, Jensen K F. Machine Learning in Computer-Aided Synthesis Planning. Acc. Chem. Res. 2018;51(5): 1281-1289.
8. Moir M, Danon J J, Reekie T A, Kassiou M. An overview of late-stage functionalization in today's drug discovery. 2019;14(11): 1137-1149.
9. King-Smith E, Faber F A, Reilly U, Sinitskiy A V, Yang Q, Liu B, Hyek D, Lee A A. Predictive Minisci and P450 Late Stage Functionalization with Transfer Learning. ChemRxiv preprint. 2022; DOI: 10.26434/chemrxiv-2022-7ddw5-v2.
10. Guan Y, Coley C W, Wu H, Ranasinghe D, Heid E, Struble T J, Pattanaik L, Green W H, Jensen K F. Regio-selectivity prediction with a machine-learned reaction representation and on- the-fly quantum mechanical descriptors. Chem Sci. 2021;12: 2198-2208.
11. Mennen S M, Alhambra C, Allen C L, Barberis M, Berritt S, Brandt T A, Campbell A D, Castañón J, Cherney A H, Christensen M, Damon D B, Eugenio de Diego J, García-Cerrada S, García-Losada P, Haro R, Janey J, Leitch D C, Li L, Liu F, Lobben P C, MacMillan D W C, Magano J, McInturff E, Monfette S, Post R J, Schultz D, Sitter B J, Stevens J M, Strambeanu I I, Twilton J, Wang K, Zajac M A. The Evolution of High-Throughput Experimentation in Pharmaceutical Development and Perspectives on the Future. Org. Process Res. Dev. 2019;23(6): 1213-1242.
12. King-Smith E, Berritt S, Bernier L, Hou X, Klug-McLeod J, Mustakis J, Sach N W, Tucker J, Yang Q, Howard R, Lee A A. Probing the Chemical "Reactome" with High Throughput Experimentation Data. ChemRxiv preprints. 2022; DOI: 10.26434/chemrxiv-2022-hjnmr.

Standardizing chemical compounds using language models

Miruna Cretu*[1], Alessandra Toniato[1], Amol Thakkar[1], Amin Debabeche[1], Teodoro Laino[1], Alain C. Vaucher[1]

* Corresponding author

[1] IBM Research Zurich

Keywords: Compound standardization, natural language processing, AI

Abstract:

With the growing amount of chemical data stored digitally, it has become crucial to represent chemical compounds accurately and consistently. Harmonized representations facilitate the extraction of insightful information from datasets, and are advantageous for machine learning applications. To achieve consistent representations throughout datasets, one relies on molecule standardization, which is typically accomplished using rule-based algorithms that modify descriptions of functional groups. Here, we present the first deep-learning model for molecular standardization. We enable custom standardization schemes based solely on data, which, as additional benefit, support standardization options that are difficult to encode into rules. Our model achieves over 98% accuracy in learning two popular rule-based standardization protocols. We then follow a transfer learning approach to standardize metal-organic compounds (for which there is currently no automated standardization practice), based on a human-curated dataset of 1512 compounds. This model predicts the expected standardized molecular format with a test accuracy of 75.6%. As standardization can be considered, more broadly, a transformation from undesired to desired representations of compounds, the same data-driven architecture can be applied to other tasks. For instance, we demonstrate the application to compound canonicalization and to the determination of major tautomers in solution, based on computed and experimental data.

Enzyme activity prediction using neural networks, docking and high-throughput screening results.

Tao Jiang[1], Guillaume Darlot, Changru Ma[1], Thierry Gefflaut[2], Véronique De Berardinis[3], Sébastien Paul[4], Egon Heuson*[4]*
Corresponding author - egon.heuson@centralelille.fr

[1] E2P2L - UMI3464 CNRS, Shanghai, China.

[2] Université Clermont Auvergne, CNRS, SIGMA Clermont, ICCF, F-63000 Clermont-Ferrand, France.

[3] Génomique métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Université d'Evry, Université Paris-Saclay, 91057 Evry, France.

[4] Univ. Lille, CNRS, Centrale Lille, Univ. Artois, UMR 8181–UCCS—Unité de Catalyse et de Chimie duSolide, F-59000 Lille, France.

Keywords: transaminases, high-throughput screening, activity prediction, docking, neural network.

Abstract (<400 words):

One of the main aims of enzymatic biocatalysis is to replace conventional chemical synthesis byoffering more sustainable catalytic alternatives (solvent, temperature, etc.) for key stages in the processes. For this, it is essential to find the most efficient enzyme for the given set of conditions,and since the molecules synthesized rarely have optimized natural biosynthesis pathways, it is crucial to be able to seek new enzymes with improved activity and selectivity. Historically, there have been two opposing approaches: enzyme engineering or biodiversity exploration. Although they have proved effective to date, both are *a posteriori* method, since it is still impossible to predict an enzyme's activity from its peptide sequence alone. That said, the rapid emergence of machine learning (ML) in this field, such as the Alphafold "revolution" [1], is changing this paradigm, and several studies are beginning to move towards this goal [2–7]. The main limitationthat seems to remain is the availability of robust and curated experimental datasets describing enzyme activity for a given family, with most studies relying heavily on the often highly heterogeneous data available in international databases. That's why in the present study we wereinterested in exploiting our recent dataset around the transaminase family [8,9]. This dataset, comprising more than 25,000 activity assays performed under the same experimental conditions,on more than twenty different substrates, was generated a few years ago using a new high-throughput screening strategy to identify new transaminases suitable for synthesis. To achieve our objective, we began by attempting to correlate enzyme sequences with their activity for different substrates using neural networks. Some of the tested architectures proved effective in solving this problem once transformed into a classification problem, by grouping activities into 4 major classes. However, the high proportion of weak enzyme activities in the dataset seemed to limit the prediction accuracy for a regression-type approach. With this in mind, we decided to introduce more information at enzyme level, to establish finer correlations between their active site, substrates and activities. For this, and inspired by some recent studies using docking [2] andGNN [5,10], we started designing a new workflow which will be detailed in this talk and that is based on several ML-based available tools (Colabfold, P2Rank, Gnina, BagPype). It aims at 1) predicting the structure of our enzymes, 2) at docking the different substrates and co-factors inside the latter, and 3) at transforming the resulting 3D file into a network visualization that could be used as additional input to our neural networks.

References

1. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structureprediction with AlphaFold. Nature. août 2021;596(7873):583-9.
2. Ao YF, Pei S, Xiang C, Menke MJ, Shen L, Sun C, et al. Structure- and Data-Driven Protein Engineering of Transaminases for Improving Activity and Stereoselectivity. Angew Chem Int Ed. n/a(n/a):e202301660.
3. Boorla VS, Upadhyay V, Maranas CD. ML helps predict enzyme turnover rates. Nat Catal. 19 août 2022;5(8):655-7.
4. Vasina M, Vanacek P, Hon J, Kovar D, Faldynova H, Kunka A, et al. Functional and Mechanistic Characterization of an Enzyme Family Combining Bioinformatics and High-Throughput Microfluidics [Internet]. In Review; 2021 nov [cité 5 sept 2022]. Disponible sur: https://www.researchsquare.com/article/rs-1027271/v1
5. Gligorijević V, Renfrew PD, Kosciolek T, Leman JK, Berenberg D, Vatanen T, et al. Structure-based protein function prediction using graph convolutional networks. Nat Commun. 26 mai 2021;12(1):3168.
6. Robinson SL, Smith MD, Richman JE, Aukema KG, Wackett LP. Machine learning-based prediction of activity and substrate specificity for OleA enzymes in the thiolase superfamily. Synth Biol. 2020;5(1):ysaa004.
7. Mellor J, Grigoras I, Carbonell P, Faulon JL. Semisupervised Gaussian Process for Automated Enzyme Search.ACS Synth Biol. 17 juin 2016;5(6):518-28.
8. Gourbeyre L, Heuson E, Charmantray F, Hélaine V, Debard A, Petit JL, et al. Biocatalysed synthesis of chiral amines: continuous colorimetric assays for mining amine-transaminases. Catal Sci Technol. 2021;11:904-11.
9. Heuson E, Petit JL, Debard A, Job A, Charmantray F, de Berardinis V, et al. Continuous colorimetric screeningassays for the detection of specific L- or D-α-amino acid transaminases in enzyme libraries. Appl Microbiol Biotechnol. janv 2016;100(1):397-408.
10. Volkov M, Turk JA, Drizard N, Martin N, Hoffmann B, Gaston-Mathé Y, et al. On the Frustration to Predict Binding Affinities from Protein–Ligand Structures with Deep Neural Networks. J Med Chem. 9 juin 2022;65(11):7946-58.

Poster Session

Presenter	Poster title
Akashaditya Das	Developing a scoring system to optimise the design of CRISPR Cas12 diagnostics
Dimitrios Gerogiorgis	Dynamic Optimisation and Parameter Estimation for Biochemical Process Systems
Austin Tripp	Interpretable, uncertainty-aware machine learning with continuous Tanimoto Similarities
Ching Ching Lam	CONFPASS: fast DFT re-optimisations of structures from conformation searches
Friedrich Hastedt	A comprehensive evaluation of ML-based Retrosynthesis frameworks
Jakob Träuble	Brain Age Modelling on Viscoelastic Properties
Nicholas A. Jose	Accelerating Autonomous Experimentation with Flab
Wenyao Lyu	A Novel Framework for Automated Simultaneous Model Identification and Parameter Estimation in Kinetic Studies
Yuxuan Wang	Extrapolation is Not the Same as Interpolation
Roel J. Leenhouts	SolProp: predicting solid solubility of organic solutes for a wide range of conditions using machine learning and thermodynamics

Organisation Committee

Katie Beckwith
University of Cambridge – SynTech CDT

Henrique Magri Marçon
University of Cambridge – SynTech CDT

Francesco Ceccarelli
University of Cambridge – SynTech CDT

Ruslan Kotlyarov
University of Cambridge – SynTech CDT

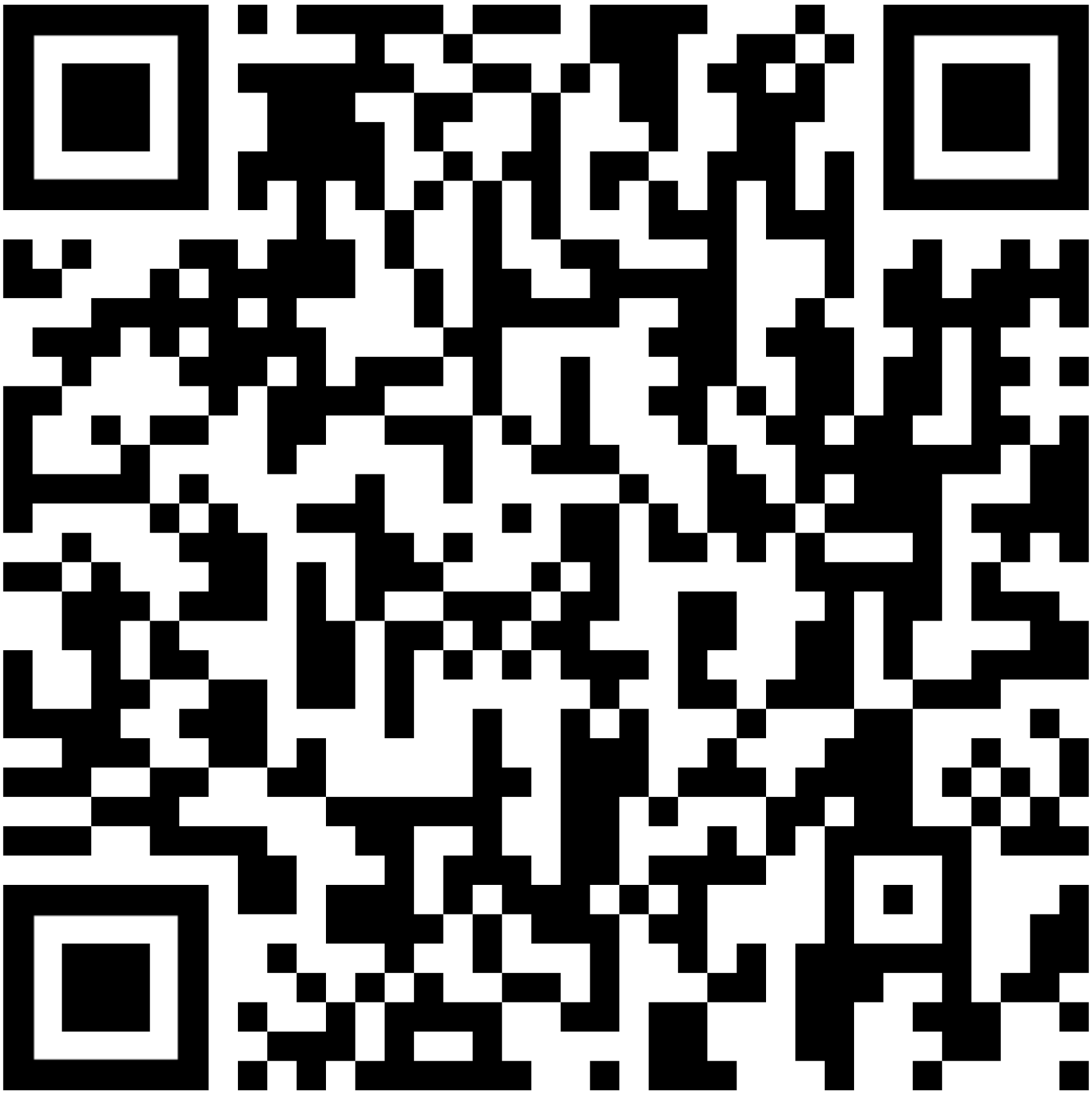
Kerstin Enright
University of Cambridge – Innovation Centre on Digital Molecular Technologies (iDMT)

Prof. Alexei Lapkin
University of Cambridge – Department of Chemical Engineering and Biotechnology

Contact information: mab@ceb.ac.uk

YOUR FEEDBACK: SCORE THE PRESENTATIONS OF DAY 1

YOUR FEEDBACK: SCORE THE PRESENTATIONS OF DAY 2



YOUR FEEDBACK: SCORE THE POSTERS

The sponsors and exhibitors

Elsevier, a global leader in information and analytics, helps researchers and healthcare professionals advance science and improve health outcomes for the benefit of society. Growing from our roots in publishing, we have supported the work of our research and health partners for more than 140 years. Elsevier offers knowledge and valuable analytics that help our users make breakthroughs and drive societal progress. Digital solutions such as ScienceDirect, Scopus, SciVal, ClinicalKey, Scibite and Sherpath support strategic research management, R&D performance, clinical decision support, and health education. Elsevier publishes over 2,800 digitized journals, including The Lancet and Cell; our 46,000+ eBook titles; and our iconic reference works, such as Gray's Anatomy. Elsevier is part of RELX, a global provider of information-based analytics and decision tools for professional and business customers.

Digital Discovery is an open access journal that publishes both theoretical and experimental research at the intersection of chemistry, materials science and biotechnology. We focus on the development and application of machine learning, AI and automation tools to unravel scientific problems, and we put data first to ensure reproducibility and faster progress for everyone.

Reaction Chemistry & Engineering

Reaction Chemistry & Engineering is an interdisciplinary journal reporting cutting-edge research focused on enhancing the understanding and efficiency of reactions. Reaction engineering leverages the interface where fundamental molecular chemistry meets chemical engineering and technology. Challenges in chemistry can be overcome by the application of new technologies, while engineers may find improved solutions for process development from the latest developments in reaction chemistry. Reaction Chemistry & Engineering is a unique forum for researchers whose interests span the broad areas of chemical engineering and chemical sciences to come together in solving problems of importance to wider society.

AstraZeneca

AstraZeneca is a global, science-led biopharmaceutical company that focuses on the discovery, development, and commercialisation of prescription medicines. Some of their primary pipelines are for the treatment of diseases in three therapy areas – Oncology, Cardiovascular, Renal & Metabolism, and Respiratory & Immunology. AstraZeneca operates in over 100 countries and its innovative medicines are used by millions of patients worldwide.

vapourtec
precision flow chemistry

***Vapourtec**, the world's leading manufacturer of flow chemistry equipment was founded in 2003 by Duncan Guthrie. Since then, Vapourtec has been at the forefront of the flow chemistry industry ever since. Headquartered in Bury St Edmunds, UK, Vapourtec design and manufacture the R-Series and E-Series flow chemistry systems that have empowered chemists throughout the world to further scientific discovery. Trusted by academics, chemists, and manufacturers around the world, the modular R-Series system has revolutionised the way many deliver the research, chemicals, and products we all rely on. With an installation base of more than 600 systems, resulting in over 970 citations in peer-reviewed publications, they continue to support their customers across the globe with the world-class products and services with which Vapourtec has become synonymous.*