Molecular Graphics and Modelling Society (MGMS)

AI for Pharma: Data, Infrastructure and Algorithms

Date: 5th September 2025

Venue: KW002 King William Court, University of Greenwich, Old Royal Naval College, Park Row, London SE10 9LS

Program and Abstracts

9:20-9:50	Registration and Coffee
9:50-10:00	Introduction: Jiayun Pang
Session 1	Chair: Jiayun Pang
10:00-10:30	Data to Performance: AI workflow activities at the Pistoia Alliance Christian Baber, Chief Portfolio Officer, Pistoia Alliance
10:30-11:00	ChEMBL's approach to improve data preparedness for AI applications Barbara Zdrazil, ChEMBL Team Coordinator, EMBL-EBI
11:00-11:25	Poster presenters, lightning talks
11:25-11:30	RSC law group information session Andy Nicoll, Partner, Elkington+Fife
11:30 - 12:00	Coffee Break
Session 2	Chair: Christian Baber
12:00-12:30	Integrating Artificial Intelligence into the drug discovery pipeline Gian Marco Ghiandoni, Associate Principal AI Engineer, AstraZeneca Prakash Rathi, Director of Engineering, Data Analytics and AI, AstraZeneca
12:30-13:00	Finetuning large language models for organic reaction prediction – A multi-task modular approach Jiayun Pang, Associate Professor in Computational Chemistry, University of Greenwich
13:00-13:05	AIChemy information session Chris Mellor, AIchemy Hub Manager, Imperial College
13:05-14:30	Lunch and Posters
Session 3	Chair: Gian Marco Ghiandoni
14:30-15:00	Assessing the chemical intelligence of large language models Nicholas Runcie, DPhil Student, University of Oxford
15:00-15:30	From unstructured data to actionable knowledge Thierry Hanser, Founder and Consultant AI and Cheminformatics, Ixelis
15:30-15:35	MGMS information session Steve Maginn, Chair of MGMS, Chemical Computing Group
15:35	Networking Drinks

With support from

UKRI Engineering and Physical Sciences Research Council
Funding Support by EPSRC
Grant Ref: EP/Y004167/1

>_Alchemy AI FOR CHEMISTRY

PISTOIA ALLIANCE

Talk 1

J. Christian Baber, Chief Portfolio Officer, Pistoia Alliance

Christian is a computational chemist with a PhD in AI-driven synthetic accessibility and de novo compound design. He continued this work with a postdoctoral fellowship at Osaka University before transitioning to industry, where he has led informatics, predictive modelling, and automation teams across startups and large pharma. Christian previously led Scientific Computing & Informatics at Shire/Takeda and the global Scientific & Pharmaceutical Data, Informatics and Systems function in Janssen R&D. Currently as Chief Portfolio Officer of the Pistoia Alliance, he is responsible for all the projects, communities and training run under the umbrella of the Alliance from the initial idea, through execution, to delivery of artifacts and sustainability of products.

Data to Performance: AI workflow activities at the Pistoia Alliance

J. Christian Baber, Pistoia Alliance

There has been a lot of focus on LLMs and Generative AI in recent years but to be practical you need AI ready data to use and a way to measure performance in order to show value from such models. The Pistoia Alliance is where leading companies work together to solve pre-competitive problems. This talk will discuss a number of the completed and ongoing activities of the Alliance covering the length of the AI workflow for Pharmaceutical R&D

Talk 2

Barbara Zdrazil, ChEMBL Team Coordinator, EMBL-EBI

Barbara is an accomplished expert in cheminformatics and computational drug discovery with nearly 20 years of experience. She earned her PhD in Pharmaceutical Chemistry at the University of Vienna in 2006 and completed her postdoctoral research at the University of Dusseldorf. As a group leader at the University of Vienna she achieved her Habilitation in Pharmacoinformatics in 2019. Since 2021, Barbara has been part of the European Bioinformatics Institute (EMBL-EBI), first as Safety Data Scientist and Consultant for Open Targets and, in 2022, became ChEMBL Team Coordinator within the Chemical Biology Services Team. From 2024-2025, she acted as the Interim Team Leader for Chemical Biology Resources at EMBL-EBI. She also serves as Editor-in-Chief of the Journal of Cheminformatics, contributing her expertise to advancing the field of Cheminformatics.

ChEMBL's approach to improve data preparedness for AI applications

Barbara Zdrazil, EMBL-EBI

ChEMBL is an open-access bioactivity database widely used in drug discovery and AI-driven research. With the growing diversity of deposited datasets and the increasing demands from machine learning applications, the ChEMBL team has implemented new measures to strengthen data quality, provenance, and reusability. Guided by FAIR and TRUST principles, recent updates include improved metadata and ontological annotations, refined data curation pipelines to guarantee high data quality and improved data provenance. Collaborative efforts such as EUbOPEN have provided large-scale, high-quality probe and chemogenomics datasets, helping to establish improved data deposition processes and meta data annotations for new and existing data types. Current work focuses on making ChEMBL more AI-ready by expanding assay annotation and improving data quality controls. Integration efforts such as the BioChemGraph project further extend ChEMBL's utility by linking structural and bioactivity data. These developments illustrate how a mature data resource can evolve to meet the needs of data-driven drug discovery while ensuring openness and reliability.

Talk 3

Gian Marco Ghiandoni, Associate Principal AI Engineer, AstraZeneca

Gian Marco is a cheminformatics and machine learning engineer at AstraZeneca whose role is to bridge science and technology to accelerate drug discovery programs. He is involved in leading the development of chemistry services and molecular property and retrosynthesis prediction capabilities. Beyond his core role, he is an active contributor across different areas of research, an industrial supervisor in graduate programmes, and a committee member of the Molecular Graphics and Modelling Society.

Integrating Artificial Intelligence into the Drug Discovery Pipeline

Gian Marco Ghiandoni[†], Prakash Chandra Rathi[†]

[†]Augmented DMTA Platform, Data Analytics and AI, R&D IT, AstraZeneca, The Discovery Centre (DISC), Francis Crick Avenue, Cambridge CB2 0AA, United Kingdom

The past decade has seen a sharp rise in research on Artificial Intelligence (AI) for drug discovery. (Hasselgren & Oprea, 2024) New methods are published daily, reinforcing the idea that AI will play a pivotal role in delivering tomorrow's pharmaceuticals. (Catacutan et al., 2024) However, bringing AI into real-world discovery pipelines faces practical challenges in terms of data foundations, infrastructure, and integration. (Ghiandoni et al., 2024) These include securing robust training data, re-engineered pipelines for continuous deployment, and optimised hardware and inference capabilities. In this talk, we describe the journey of our platform, Augmented DMTA (Design-Make-Test-Analyse), in delivering AI into preclinical drug discovery at AstraZeneca.

Talk 4

Jiayun Pang, Associate Professor in Computational Chemistry, University of Greenwich

Jiayun is an Associate Professor at the University of Greenwich, London. She received her PhD in Computational Chemistry from the University of Birmingham and completed her PDRA training at the University of Manchester. Jiayun's current research explores large language models and Natural Language Processing (NLP) algorithms for chemical reaction prediction. Her ongoing research focuses on addressing key issues to enable the effective use of pretrained language models in chemistry, including tokenisation, finetuning efficiency, decoding algorithms and multi-task learning.

Finetuning Large Language Models for Organic Reaction Prediction – A Multi-Task Modular Approach

Jiayun Pang[1], Ahmed Zaitoun[1], Xacobe Couso Cambeiro[1], and Ivan Vulić[2]

1. School of Science, Faculty of Engineering and Science, University of Greenwich, Medway Campus, Central Avenue, Chatham Maritime, ME4 3RL, UK. E-mail: j.pang@gre.ac.uk
2. Language Technology Lab, University of Cambridge, 9 West Road, Cambridge CB3 9DA, UK.

Large language models (LLMs) are typically used for downstream tasks through transfer learning. In this scenario, LLMs, pre-trained on vast amounts of raw data, are finetuned for new tasks using a small number of labelled data. We have developed a flow of general finetuning, task-specific finetuning and inference using two pre-trained LLMs, namely FlanT5 and ByT5 to address a range of organic reaction prediction tasks, including reagents prediction, retrosynthesis, and forward reaction prediction. We have focused on several critical aspects of this framework: i) Robust and adequate tokenisation. ii) Training data efficiency for generalisation.[1] iii) Modularity and parameter efficient finetuning (PEFT), aiming to update extremely small and specialised components of the general finetuned model for specific tasks while maintaining the overall structure of the underlying model. [unpublished results]

We tested our models on the general organic reactions as represented in the USPTO datasets and the more novel and challenging C-H functionalisation reaction, achieving state-of-the-art GPU efficiency and model versatility and adaptability. While there is some variation that stems from the choice of input preprocessing, tokenization, and model size, the training data size is instrumental to the final performance of the general fine-tuned models.[1] Our unpublished results also highlight the potential of the PEFT approach: Despite finetuning using under 1% of the full finetuning parameters, PEFT models achieve top 1 accuracy slightly above the full parameter finetuning in retrosynthesis and reagent prediction tasks for the USPTO-TPL dataset and lead to significant improvement in predicting the products of C-H functionalization reactions. Our work operates at the intersection of several broad areas of AI and chemistry research including multi-task learning, transfer learning and computer-assisted synthesis planning. Ultimately, we are interested in developing a multi-scale AI-model to aid chemical reaction planning and optimization.

References:

[1] Pang and Vulić. Specialising and analysing instruction-tuned and byte-level language models for organic reaction prediction. (2025) Faraday Discuss., 256, 413-433.

Talk 5

Nicholas Runcie, DPhil Student, University of Oxford

Nicholas completed his Master's degree in medicinal and biological chemistry at the University of Edinburgh. During his final year he worked at AstraZeneca in the computational chemistry team, applying generative modelling tools to drug projects. He is now a DPhil student in the Oxford Protein Informatics Group supervised by Fergus Imrie and Charlotte M. Deane. His research focuses on using large language models to perform chemical reasoning.

Assessing the Chemical Intelligence of Large Language Models

Nicholas Runcie, University of Oxford

Reasoning models are large language models (LLMs) that have been fine-tuned to articulate step-by-step reasoning before providing an answer, significantly enhancing their problem-solving capabilities. Such models have already demonstrated substantial improvements over non-reasoning LLMs in domains like mathematics and software engineering. To evaluate their performance in chemistry, we introduce ChemIQ, a benchmark designed to assess the ability of LLMs to interpret and reason about molecular structures. We evaluate multiple state-of-the-art reasoning models and compare them against baseline non-reasoning LLMs. Our results show that reasoning models consistently outperform the baseline models and have acquired multiple new capabilities, such as the ability to write IUPAC names of molecules and perform structure elucidation from NMR data. Furthermore, we found evidence that the reasoning process resembles that of human chemists. Our results demonstrate that the latest reasoning models can, in some cases, perform advanced chemical reasoning.

Talk 6

Thierry Hanser, Founder and Consultant AI and Cheminformatics, Ixelis

Thierry completed his PhD at the University of Strasbourg, where he applied Machine Learning techniques to automate the extraction of organic reaction knowledge from chemical databases. He continued his research during a postdoctoral fellowship in retrosynthetic expert systems at Harvard University, working with Nobel Laureate Prof. E.J. Corey, followed by a second postdoc in de novo drug design at the University of Leeds with Prof. Peter Johnson. Thierry later returned to the University of Strasbourg to teach cheminformatics and founded Ixelis, a software company dedicated to molecular information systems. In 2006, he joined Lhasa Limited, where he now leads the *Molecular Informatics and AI* group—an interdisciplinary team of nine passionate scientists. His team focuses on developing innovative methods to predict the adverse effects of chemicals using cheminformatics and AI. They also pioneered a novel approach to Federated Learning in drug discovery and are currently integrating Generative AI into next-generation solutions for drug safety assessment.

From unstructured data to actionable knowledge

Thierry Hanser

In an era where 90% of available data is unstructured, AI is revolutionising access to strategic insight through advanced knowledge extraction, understanding, and agentic large language model (LLM) frameworks. At the intersection of machine learning, natural language processing, and autonomous reasoning, this approach is rapidly becoming pivotal in the global AI landscape. This presentation explores how agentic LLM systems, retrieval-augmented generation (RAG), and specialised AI tools transform scientific literature and technical documents into actionable knowledge. Real-world applications span drug discovery, safety monitoring, and regulatory acceleration, with further use-cases in cosmetics, agrochemical, and chemical industries. Innovation will greatly benefit from the collaboration of human and AI reasoning. As hybrid and augmented intelligence become essential for scientific progress, AI-driven automated knowledge extraction stands out as a cornerstone in this transformative era.

Posters

1	**Software toolkits for *in silico* screening of polymer excipients used in small molecule formulation and drug delivery** Hannah Turney, KCL
2	**Holo-like conformation selection using a computer vision-based deep-learning model** Yu-Yuan Yang, Queen Mary University of London
3	**Computational Insights into Ion-Mobility Mass Spectrometry for RNA Therapeutics** Vladimir Kozyrev, Loughborough University
4	**Generative AI as a tool for metabolite identification** Alex Porter, Syngenta
5	**Generative Machine Learning for Automating Structure Elucidation in Synthesis** Zheqi Jin, University of Bristol
6	**Finetuning Large Language Models for Prediction of C-H Functionalisation Selectivity** Ahmed M. Zaitoun, University of Greenwich

Poster 1

Software toolkits for *in silico* screening of polymer excipients used in small molecule formulation and drug delivery

Hannah Turney, *Micaela Matta*

Department of Chemistry, King's College London, Strand Building (East Wing), Surrey Street, Westminster, WC2R 2LS

The use of polymers as excipients in small molecule pharmaceutical formulations is an established approach for the controlled delivery of drugs. However, designing safe and effective formulations is resource-intensive and delays product delivery to the clinic, primarily due to the sensitivity of polymer substructure to delivery properties.

Molecular dynamics (MD) simulations can reveal information about the conformational changes, binding interactions, and dynamical properties of molecules. However, limitations remain in the routine application of MD simulations to polymer excipients, such as the difficulty in parameterizing larger polymers and ensuring the transferability of force fields across different polymer chemistries.

The emergence of neural network potentials (NNPs) for the estimation of force field parameters marks a transformative shift in how molecular systems are parameterized, offering an opportunity to completely shift away from traditional atom-typing approaches. In this project, we create a robust and scalable building[1] and parameterization workflow for polymer excipients using NNP force fields to overcome the limitations of semi-empirical methods. This research incorporates existing open-source molecular dynamics tools from organizations such as the Open Force Field Consortium[2] to promote FAIR data practices.

With our established workflow, we perform systematic molecular dynamics simulations of different polymer:drug systems to yield kinetic and mechanical parameters predicting excipient suitability for drug delivery. This enables fast, accurate, and reproducible profiling of polymers in the context of drug formulation design. We collaborate with experimental formulation development teams at Johnson&Johnson Innovative Medicine to validate our models and drive the design of our polymer:drug candidate systems.

References:
[1] H. N. Turney, M. Matta, *JOSS,* (2025), **10(110)**, 8053.
[2] L. Wang, P. Kumar Behara, M.W. Thompson, T. Gokey, Y. Wang, J.R. Wagner, D.J. Cole, M.K. Wilson, M.R. Shirts, D.L. Mobley, *J. Phys Chem,* (2024) **128**, 7043-7067.

Poster 2

Holo-like conformation selection using a computer vision-based deep-learning model

Yu-Yuan Yang[1,2], Richard W. Pickersgill[2], Arianna Fornili[3,4]

1. Digital Environment Research Institute, Queen Mary University of London, E1 1HH

2. School of Biological and Behavioural Sciences, Queen Mary University of London, E1 4NS

3. School of Physical and Chemical Sciences, Queen Mary University of London, E1 4NS

4. The Thomas Young Centre for Theory and Simulation of Materials, WC1E 6BN

Email: yu-yuan.yang@qmul.ac.uk

In structure-based drug design, multiple protein conformations from molecular dynamics (MD) simulations capture binding site dynamics, and a selection of the ligand-binding (holo)-like conformations elevated the success rate of drug discovery. However, identifying holo-like conformations in a protein ensemble without prior knowledge of the holo state remains challenging. This research aims to develop a deep learning tool to analyze protein binding sites and detect the regions suited for chemical fragment binding through semantic segmentation (FragBEST: **Frag**ment-**B**ased protein **E**nsemble semantic **S**egmentation **T**ool). The predictions from this tool can be utilized to identify holo-like conformations within MD trajectories. Here, we showcased FragBEST on MD simulations of the complex between cardiac myosin (CM) and the first-in-class cardiac myotrope, omecamtiv mecarbil (OM). FragBEST-Myo (FragBEST for Myosin) can segment the binding site with an accuracy of ~95% and a mean Intersection over Union (mIoU) > 0.75 on testing sets of the different states of CM. A combined score from the four FragBEST-Myo-derived descriptors successfully ranked the conformations, detecting the holo-like conformations on both unbiased and steered ligand-free (apo) MD trajectories. Our method outperformed random sampling by achieving a higher recovery rate of the holo conformation in OM ensemble docking. The predicted labels from our optimal FragBEST-Myo serve as a valuable guide for users to validate potential chemical fragment binding. With larger and more diverse datasets, our approach can be scaled and modified to transform additional systems for drug design.

Poster 3

Computational Insights into Ion-Mobility Mass Spectrometry for RNA Therapeutics

Vladimir Kozyrev[1]; Anna Croft[1]; Christof Jäger[2]; Martina Pannuzzo[2]
[1]*Loughborough University, Loughborough, Leicestershire LE11 3TU*
[2] *Data Science and Modelling, Pharmaceutical Sciences R&D AstraZeneca, Gothenburg, Pepparedsleden 1, SE-431 83 Mölndal, Sweden*

Corresponding author: v.kozyrev@lboro.ac.uk

Oligonucleotide therapeutics are frequently modified with phosphorothioate (PS) linkers to improve stability and pharmacokinetics. [1] PS substitution at each backbone linkage introduces a chiral center and yields diastereomeric mixtures (Rp/Sp) that complicate synthetic control and downstream characterization. Ion mobility mass spectrometry (IM-MS) offers the possibility of separation of diastereoisomeric fragment ions, yet the assignment of peaks to specific stereochemical configurations remains laborious. To accelerate the accurate profiling of PS stereochemistry in RNA therapeutics, complementary computational strategies are required to support experimental ion mobility assignments.

In this work, we perform gas-phase molecular dynamics (MD) simulations on oligonucleotide fragments with linker PS modifications. From the MD snapshots, theoretical CCS values are calculated using the trajectory method. By comparing computed CCS distributions with experimental IM-MS data, we aim at establishing a methodology for reliable differentiation between Rp and Sp configurations. This computational workflow assists and speed up peak assignment accuracy and paves the way for automated integration of MD-derived CCS predictions into high-throughput IM-MS characterization of therapeutic oligonucleotides.

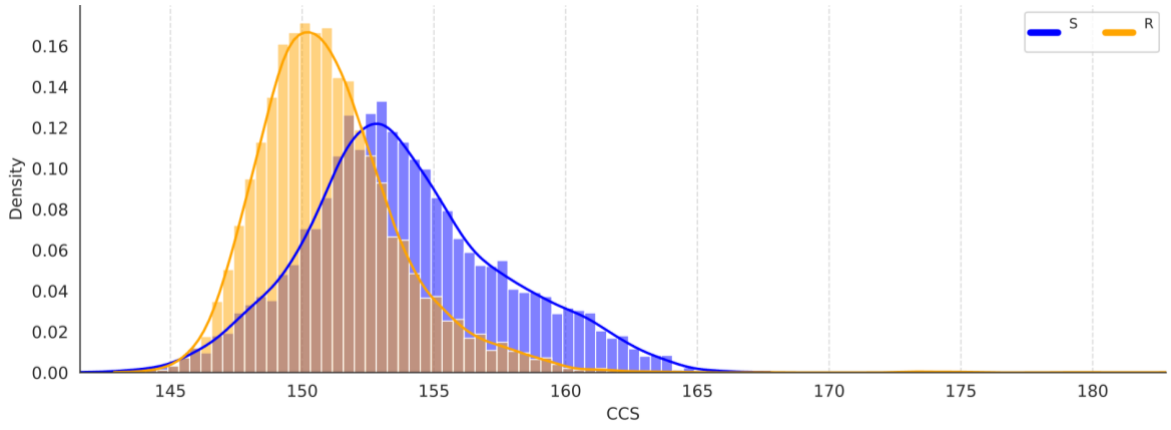

Figure 1: *Distribution of theoretical CCS values for S and R isomers.*

References

[1] Egli, M., & Manoharan, M. (2023). *Nucleic Acids Research*, *51*(6), 2529–2573.

Poster 4

Generative AI as a tool for metabolite identification

Alex Porter

Digital Chemical Design, Jealott's Hill International Research Centre, Syngenta

Understanding metabolism plays a crucial role in the development and registration of agrochemicals and pharmaceuticals.[1,2] Identifying the structures of metabolites generated in complex real-world studies remains a significant challenge, particularly when those metabolites are the product of multiple successive transformations from the starting material. In this study, we propose a novel approach for metabolite identification by combining generative chemistry, reinforcement learning (RL) and cheminformatics pipelines to tackle this problem.

Our method employs a generative model trained on known metabolite structures and enumerated chemical space around a specific unknown metabolite. The enumerated structures are generated using the opensource molecular graph generator Surge[3] and then filtered using NextMove's Arthor[4]. The generative agent then iteratively constructs metabolite candidates, and the generated structures are evaluated using a scoring function. This scoring function consists of: the presence of known substructures; the total mass of the species, and other relevant structural information from spectroscopic studies of the unknown metabolite. Through reinforcement learning the agent learns to generate high-scoring metabolite structures through trial-and-error exploration guided by the scoring function.

By leveraging the power of generative chemistry and reinforcement learning, our approach demonstrates the potential of this technology to provide predictions of metabolite structures.

References:

1. Chaleckis, R. *et al.* (2019) 'Challenges, progress and promises of metabolite annotation for LC–MS-based metabolomics', *Current Opinion in Biotechnology*, 55, pp. 44–50. doi:10.1016/j.copbio.2018.07.010.
2. Gowda, G.A. and Djukovic, D. (2014) 'Overview of mass spectrometry-based metabolomics: Opportunities and challenges', *Methods in Molecular Biology*, pp. 3–12. doi:10.1007/978-1-4939-1258-2_1.
3. McKay, B.D., Yirik, M.A. and Steinbeck, C. (2022) 'Surge: A fast open-source chemical graph generator', *Journal of Cheminformatics*, 14(1). doi:10.1186/s13321-022-00604-9.
4. *Arthor* (no date) *NextMove Software | Arthor*. Available at: https://www.nextmovesoftware.com/arthor.html (Accessed: 28 August 2024).

Poster 5

Generative Machine Learning for Automating Structure Elucidation in Synthesis

Zheqi Jin[a], *Mohammad Golbabaee*[b], *Craig Butts*[a]

a. School of Chemistry, University of Bristol, Cantock's Cl, Bristol, BS8 1TS
b. School of Engineering Mathematics and Technology, University of Bristol, Ada Lovelace Building, Tankard's Cl, Bristol, BS8 1TW

Accurately elucidating molecular structures from Nuclear magnetic resonance (NMR) spectroscopy is of pivotal importance in chemical synthesis and drug discovery. Existing approaches, such as Computer-Aided Structure Elucidation (CASE)[1], are based on library searches and therefore struggle to interpret unseen molecules. With the growing availability of large-scale datasets from both experiments and simulations, machine learning offers a powerful framework for learning from the relationship between NMR spectroscopy data and molecular structures.

In this work, we inverted our in-house Graph Transformer Network (GTN) model, IMPRESSION[2], which was originally designed to predict NMR parameters (chemical shifts and scalar couplings) for small molecules, and demonstrated its capability to reconstruct molecular bonding patterns directly from NMR data.
Our initial architecture, IMPRESSION-OneShot, performed bond prediction in a single step but achieved only 37% molecular accuracy, primarily due to the absence of information from NMR-inactive nuclei (e.g., ^{16}O) in solution state. To address this limitation, we developed a second framework, IMPRESSION-Stepwise, which locates NMR-inactive nuclei in a sequential manner and substantially improved molecular accuracy to over 80%. To further enhance molecular accuracy and robustness, new IMPRESSION variants capable of predicting bond order are under development.
In conclusion, this work highlights the potential of machine learning, and specifically the IMPRESSION framework, to advance automated molecular structure elucidation from NMR data beyond traditional approaches.

1. M. Elyashberg and D. Argyropoulos, *Magn. Reson. Chem.*, 2021, **59**(7), 669-690.
2. C. Yiu, B. Honoré, W. Gerrard, J. Napolitano-Farina, D. Russell, I. M. Trist, R. Dooley and C. P. Butts, *Chem. Sci.*, 2025, **16**, 8377-8382.

Poster 6

Finetuning Large Language Models for Prediction of C-H Functionalisation Selectivity

Ahmed M. Zaitoun, Xacobe C. Cambeiro, Jiayun Pang

School of Science, Faculty of Engineering and Science, University of Greenwich, Af0193o@gre.ac.uk

C–H functionalisation selectivity plays a critical role in enhancing the activity of existing drugs (known as Late-Stage Functionalisation) and selectively designing new drug candidates. However, accurately predicting this selectivity remains a significant challenge due to the subtle electronic energy differences of C–H bonds within a molecule.

Transformer-based sequence-to-sequence (seq2seq) language models have shown promises in organic chemistry prediction for a variety of tasks, such as forward reaction products, reaction yields, reagents, retrosynthesis, and reaction classifications [5]. In this study, we investigate the ability of two pre-trained seq2seq language models, ByT5 and FlanT5, to predict the products of C–H functionalization reactions. These models, derived from Google's ByT5 [2] and FlanT5 [3] architectures, have been pre-trained and fine-tuned with different strategies to enhance their compatibility with chemical data [4]. We curated a dataset of 390 C–H functionalisation reactions from literature. These reactions all involve ligand-to-metal charge transfer (LMCT) with a hydrogen atom transfer (HAT) agent. We used 90% of the dataset for training and 10% for testing. After fine-tuning using our C-H dataset, FlanT5 achieved a top-1 accuracy of 41.03%, while ByT5 reached 15.38% in C-H functionalisation selectivity. To further enhance predictive accuracy, we implemented a Parameter-Efficient Fine-Tuning (PEFT) approach [4], which aims to mitigate Catastrophic Forgetting by preserving the general knowledge learned from pretraining. Interestingly, ByT5 showed a more substantial improvement, reaching 43.59% in top-1 accuracy, while FlanT5 achieved a marginal increase, also reaching 43.59% in the top-1 prediction. For comparison, a previous study using a similar approach achieved top-1 accuracy of 60.81% using a dataset of 1,041 C–H borylation reactions, [1] indicating that an increased dataset size may enhance performance.

Further analysis shows that our language models are able to predict valid structures (i.e. some incorrect predictions still exhibit chemical common sense) and can reach 60% in top-5 accuracy. We have used SHAP (Shapley Additive Explanations) analysis, an explainable AI method, to visualize how different functional groups in the reactants contribute to the predicted products. It indicates that the models have the ability to identify key reagents involved in the reactions. Our findings demonstrate that both FlanT5 and ByT5 can be effectively fine-tuned with a relatively small number of reactions to improve their predictive power for novel reactions.

References

[1] Kotlyarov, R. 2024, Journal of Chemical Information and Modeling, 64,10, 4286–4297

[2] Xue, L. 2022, arXiv:2105.13626

[3] Chung, H. W, 2022, arXiv:2210.11416

[4] Vulic,I, 2024, arXiv, 2405.10625

[5] Lu, J. 2022, Journal of Chemical Information and Modeling, 62,6, 1376–1387