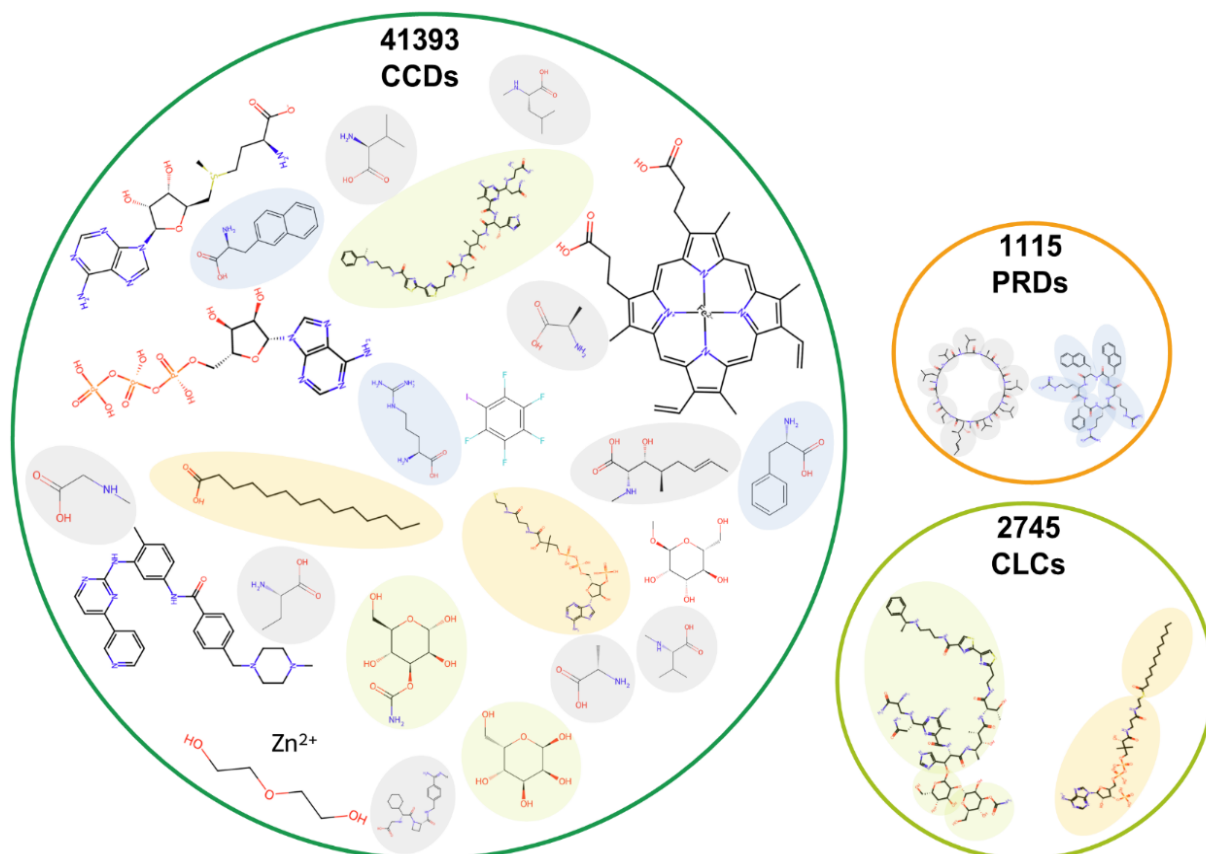


NEWSLETTER

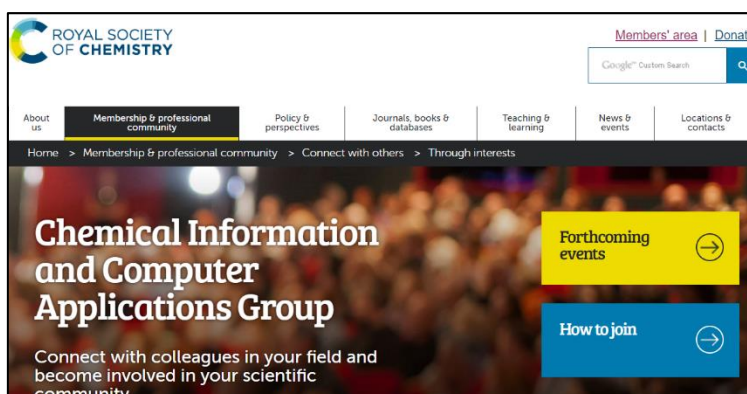
Winter 2023-24

CICAG aims to keep its members abreast of the latest activities, services and developments in all aspects of chemical information, from generation through to archiving, and in the computer applications used in this rapidly changing area, through meetings, newsletters and professional networking.

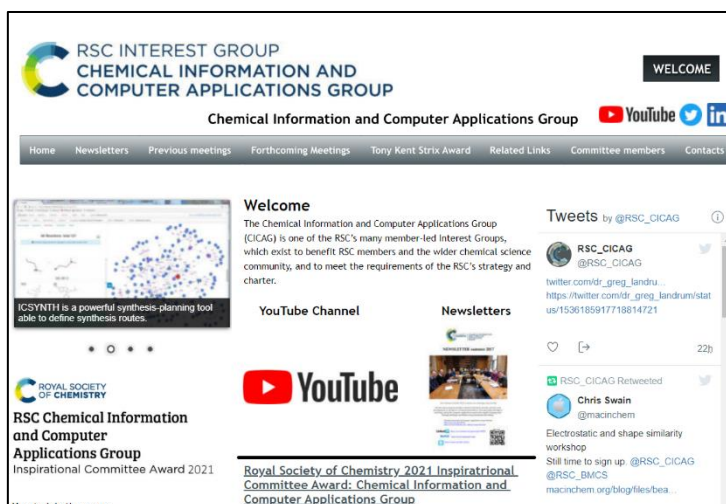


Total number of unique small molecules in the Protein Data Bank (PDB) as of 5 Dec 2023. See the full article on p. 11-16.

CICAG Websites and Social Media



<http://www.rsc.org/CICAG>



<http://www.rscicag.org>



<https://www.youtube.com/c/RSCCICAG>



<https://www.linkedin.com/groups/1989945/>



[@rsccicag@sciencemastodon.com](mailto:rsccicag@sciencemastodon.com)



[@RSC_CICAG](https://twitter.com/RSC_CICAG)
https://twitter.com/RSC_CICAG

Contents

| | |
|--|----|
| Chemical Information and Computer Applications Group Chair's Report | 4 |
| Eight Good Reasons to Join CICAG | 5 |
| CICAG Planned and Proposed Future Meetings | 6 |
| Meeting Report: The Ninth Joint Sheffield Conference on Chemoinformatics | 6 |
| UKeIG: Martin White wins Jason Farradane and Tony Kent Strix 2023 Memorial Awards | 7 |
| Meeting Reports: UK QSAR Autumn 2023 Meeting..... | 9 |
| Accessing Enhanced Small Molecule Information in the Protein Data Bank in Europe..... | 11 |
| Meeting Report: Solutions in Science (SinS-1) Conference Report and Invitation to SinS-2 | 16 |
| Catalyst Science and Discovery Centre News | 17 |
| Meeting Report: RSC-CICAG and RSC-BMCS 6th Artificial Intelligence in Chemistry Symposium, 4-5 September 2023 | 20 |
| The Summer of Science Festival at Nantwich Museum | 43 |
| Cheminformatics: A Digital History – Part 4. Ladies First..... | 46 |
| William Jorgensen Wins 2024 Arthur C. Cope Award for Organic Chemistry | 50 |
| RSC Databases Update: ChemSpider..... | 52 |
| Book Review: Architects of Memory..... | 53 |
| Cheminformatics and Chemical Information Books..... | 54 |
| Cambridge Structural Database (CSD) Updates | 56 |
| News from the American Chemical Society Chemical Information Division (CINF)..... | 56 |
| Call for Nominations for the 2025 Skolnik Award | 57 |
| Physical Sciences Data Infrastructure (PSDI) News..... | 58 |
| News from CAS | 61 |
| The Development of the Chemist's Notebook – Meeting Announcement | 62 |
| AI in Drug Discovery 2023 – A Highly Opinionated Literature Review (Part I) | 63 |
| Other Chemical Information News | 73 |

Contributions to the CICAG Newsletter are welcome from all sources – please send to the Newsletter Editor
Dr Helen Cooke FRSC: email helen.cooke100@gmail.com

Chemical Information and Computer Applications Group Chair's Report

Contribution from RSC-CICAG Chair Dr Chris Swain, email: swain@mac.com

The second part of this year has been very busy with CICAG involved in several meetings since the last Newsletter, and reports on these meetings are in this edition.

- Ninth Chemoinformatics Conference, Sheffield, 19-21 June 2023
- Solutions in Science, Cardiff, 4-6 July 2023
- 6th RSC-BMCS / RSC-CICAG Artificial Intelligence in Chemistry, Cambridge 4-5 September 2023
- UK QSAR Autumn meeting, Liverpool, 14 September 2023

We are also in early planning for several other events: 7th RSC-BMCS/RSC-CICAG Artificial Intelligence in Chemistry (16-18 September 2024), Markush Centenary meeting, Molecular Simulation and Free Energies, (Spring 2024) and a Python for Chemists course.

We discussed postponing the 7th AI in Chemistry meeting, or having a smaller one-day meeting, but due to overwhelming feedback from delegates the 2024 event will take place in September at Churchill College, Cambridge. It will be a 2.5-day event with the first half day set aside as an introduction to artificial intelligence for newcomers to this exciting area.

It is perhaps worth mentioning that CICAG provides [bursaries](#) to attend these meetings. CICAG and BMCS provided 11 bursaries for the 6th AI in Chemistry meeting, the most we have given for any meeting to date.

I attended the 2023 RSC Interest Groups Meeting (18 October, Nottingham), the first face-to-face meeting of interest groups for several years and the first since the reorganisation of the RSC interest groups and subject communities. It was a great opportunity to meet other interest groups and a common talking point was the challenges in making such contacts. CICAG have co-organised conferences with several other groups and we are always open to new opportunities to interact.

As RSC subscription forms are hitting the doorstep, I thought I'd mention that CICAG came into existence in 2007 and currently has 725 members from 43 different countries. The community is growing by 60-70 members per year. Whilst RSC members can join up to five interest groups free of charge, in practice many members do not take up this opportunity. You can request to join a group via [email](#), telephone (01223 432141) or the RSC [website](#).

Recently, Pat Walters wrote an excellent [blog post](#) highlighting some of the issues faced by computational and cheminformaticians as they move from academia to industry. In particular, the lack of mentors for those working in smaller companies or start-ups where they might be the only computational scientist. In larger companies it is possible to interact with more experienced computational chemists as well as medicinal chemists and other scientists. One way to build a network outside the company is to join a group like CICAG, go to the conferences and meetings, offer a talk or poster, or even offer to help with organisation of meetings. As we move towards more small organisations and virtual companies, building a wider network is ever more important.

Social media became an increasingly important way for communicating with CICAG members (and non-members) during lockdown and the trend continues. [Twitter](#) (X) now has 1625 followers, and the CICAG account is now curated by Samantha Pearman-Kanza who has greatly enhanced the content and advertising of

CICAG's and other relevant meetings. [LinkedIn](#) now has 638 members, we also have a [Mastodon](#) account (74 followers) which may become more important in the future. The CICAG [website](#) is often updated and we would be very interested to hear suggestions for additional content for all channels.

CICAG's [YouTube](#) channel now has 1106 subscribers and contains the 13 video presentations from AI4Proteins meetings in addition to all 20 of the [Open-Source Tools for chemistry](#) workshops. These videos have proved to be very popular and have been watched 38,500 times. The most popular video, a [workshop on DataWarrior](#), has been watched nearly 7000 times. If anyone has suggestions for additional workshops feel free to get in touch.

For CICAG, the 6th Artificial Intelligence in Chemistry Symposium (4-5 September 2023) was the highlight of the year, and the meeting report in this Newsletter is a comprehensive review of the fabulous work presented. This meeting grows in size every year and underlines the key role of AI/ML in the chemical sciences. This Newsletter also includes meeting reports from the Ninth Joint Sheffield Conference on Chemoinformatics and the UK QSAR autumn meeting, all meetings are high spots in the cheminformatics calendar and CICAG is delighted to support attendance with bursaries.

We also have a new contribution to the "Cheminformatics: A digital history" series, from Wendy Warr. This is another fascinating insight into the history of cheminformatics and Wendy has plenty of anecdotes to tell of the early days. I can't help but feel that at some point we should combine all of these stories into a book.

We also have contributions from CCDC and ChemSpider, and are looking forward to further updates from them in future issues.

Once again, I'd like to invite contributions to the CICAG Newsletter that would be of interest to the CICAG community. Please contact the Newsletter Editor, [Helen Cooke](#), or me to discuss your ideas.

Eight Good Reasons to Join CICAG

Contribution from Dr Chris Swain, RSC-CICAG Chair, email: swain@mac.com and Dr Helen Cooke, RSC-CICAG Newsletter Editor.



The Chemical Information and Computer Applications Group (CICAG) is one of the RSC's many member-led Interest Groups, which exist to benefit RSC members and the wider chemical science community, and to meet the requirements of the RSC's strategy and charter. CICAG aims to support users and producers of chemical information, data, and computer applications to advance excellence in the field.

Why join CICAG when you renew your RSC membership?

1. It's an excellent vehicle for learning and information/knowledge exchange during a period of rapid change in our field, e.g. keeping up with developments in AI and machine learning.

- You'll be part of a group committed to gathering and sharing insights with other members and the wider chemical information and computer applications community.
- You'll get early alerts about CICAG conferences and events.
- You can access and contribute to resources, such as the [CICAG Newsletter](#), our YouTube channel, and social media, e.g. [LinkedIn](#).
- There are opportunities to apply for bursaries to attend meetings and events.
- Via CICAG's committee, you'll have the chance to influence RSC chemical information strategy.
- It's a great way to network with people with whom you share interests.
- There are often opportunities to join the CICAG Committee and help further CICAG's aims, and organise events and activities.

From 2024, some membership categories can now join up to five interest groups free of charge. In addition to joining interest groups when renewing their membership, RSC members can join or change interest groups between the renewal periods. To do this, please visit the RSC's [Membership and professional community](#) or contact the RSC at membership@rsc.org.uk.

CICAG Planned and Proposed Future Meetings

The table below provides a summary of CICAG's planned and proposed scientific and educational meetings. For more information, please contact CICAG's Chair, Dr Chris Swain, swain@mac.com.

| Meeting | Date | Location | Further Information |
|---|-----------------|------------------|---|
| Molecular Simulation and Free Energies | 14 June 2024 | Burlington House | https://www.rsc.org/events/detail/78099/molecular-simulations-for-chemistry |
| 7th Artificial Intelligence in Chemistry Meeting | 16-18 Sept 2024 | Cambridge | https://www.rscbmcs.org/events/aichem7/ |
| Centenary of Markush Structures | TBD | Burlington House | Details to follow |
| CICAG AGM | TBD | Burlington House | Date to coincide with a meeting or event |
| Python for Chemists | TBD | TBD | Details to follow |

Meeting Report: The Ninth Joint Sheffield Conference on Chemoinformatics

Contribution from Justin Lübbbers, Department for Algorithmic Molecular Design Center for Bioinformatics, University of Hamburg, email: justin.luebbbers@uni-hamburg.de

My name is Jutin Lübbbers, I am a doctoral student at the University of Hamburg. Under the guidance of Professor Matthias Rarey, I work with the Algorithmic Molecular Design group. My research concentrates on developing algorithms to better understand and improve chemical fragment spaces. Existing examples of these libraries, such as the REALSpace by Enamine, encode between 10^8 and 10^{26} compounds, which makes them non-numerable and, therefore, incompatible with traditional search and analysis tools. My recent project, SpaceProp2, builds upon the existing SpaceProp algorithm that enables the computation of property distributions for chemical fragment spaces. SpaceProp2 extends on the capabilities of SpaceProp by computing

structure-based distributions and by providing example compounds to boost the explainability of the results. These distributions are used to compare the composition of different chemical fragment spaces and uncover possibilities for their optimisation with the goal to provide better virtual search spaces for the early stages of drug discovery projects.



Justin Lübbbers receives the certificate for the Peter Willett Award for Outstanding Poster Presentation from Professor Willett at the Sheffield Conference (photo courtesy of Dr Wendy A. Warr).

From 19-21 June 2023, I attended the Ninth Chemoinformatics Conference in Sheffield – the first scientific conference in my academic journey. As an attendee, I had the opportunity to present a poster about my SpaceProp2 project. The conference was an insightful event. It hosted a variety of presentations and posters, each reflecting the innovative work being done in chemoinformatics across academia and industry. It was enlightening to observe current research trends, discover potential overlaps between projects, and learn about common challenges faced by different groups. For me, it was particularly encouraging to learn how many fellow researchers are working with chemical fragment spaces. What I learned about their approaches to the topic, and especially their demand for specific software solutions in this area, will certainly guide my research in the years to come. In addition to the knowledge and experience I gained from the presentations, it was the chance to connect with other researchers and the organisation of the whole event, including social gatherings such as the conference

dinner and the trip to the local museum, that made the conference a great experience which I will remember for a long time.

After two days of exciting academic presentations and interesting discussions with fellow researchers, I had the honour to receive the Peter Willett award for outstanding poster presentation, funded by the RSC-CIGAG. I want to express my deep gratitude to the judging panel and the RSC-CIGAG for recognising my work with this prestigious award. In addition, I want to thank Professor Mathias Rarey and our industrial partner Dr Uta Lessel from Boehringer Ingelheim for supporting me in my studies. Finally, I want to thank Professor Val Gillet and the organising committee for such a memorable event. I already look forward to the next one.

UKeiG: Martin White wins Jason Farradane and Tony Kent Strix 2023 Memorial Awards

Contribution from Gary Horrocks, UKeiG, CILIP, email: info.uk eig@cilip.org.uk



The UK electronic information Group (UKeiG) is pleased to announce that the winner of the prestigious international Jason Farradane and Tony Kent Strix Memorial Awards for 2023 is Martin White, a Fellow of the Royal Society of Chemistry, a Fellow of the British Computer Society and an Honorary Fellow of CILIP – the library and information association. He recently retired from his prominent information consultancy role.

The Jason Farradane Award is presented in recognition of an outstanding contribution to the library and information science profession and the Tony Kent Strix Memorial Award is given in recognition of an outstanding contribution to the field of search and information retrieval.

Starting out as an Information Officer in the metallurgical industry in 1970, Martin White's subsequent career has involved electronic publishing, high-technology market research, and information and knowledge management consulting before setting up Intranet Focus Ltd in 1999. He is a pioneer of the business-critical importance of effective enterprise solutions for information and knowledge search and discoverability.

The judging panels for both awards would like to congratulate Martin on his prolific and significant leadership and contribution to the profession on multiple fronts.

- International information management, intranet and enterprise search consultant for over fifty organisations with complex corporate challenges, including the International Money Fund, World Bank, NATO, United Nations, European Commission and a number of major pharmaceutical companies
- Presentations and workshops at conferences in 15 countries
- Author of ten books on intranets, enterprise search and information management
- A Visiting Professor at the iSchool, University of Sheffield since 2002 and close links with City, University of London since 1977
- Dedication to the development and growth of the UK's information profession through his energetic contribution to the Institute of Information Scientists – a predecessor to CILIP.

Martin was shocked but delighted to receive the news:

“After a career of over fifty years in information science, receiving two awards reflecting the work of Tony Kent and Jason Farradane in the same year is a great honour. Kent and Farradane both played a crucial role in the development of my career. I am immensely proud.”

Udo Kruschwitz, Professor of Information Science, University of Regensburg writes:

“Martin has demonstrated major, sustained and influential achievements in the information retrieval and information science community, bridging the gap between industry and academia with a continuous, longstanding effort in forming and shaping a community of practitioners and academics in the field of search.”

Dr Sandra Ward BSc PhD Cert Ed Hon FCLIP concurs:

“Martin has dedicated his career to information science, information management and promoting the necessity for organisations to use these skills to deliver organisational benefits through well-constructed Intranets and Internets completely aligned with business objectives. He is the only person I know to use Information Scientist as the profession on his passport.”

Professor Charles Oppenheim BSc, PhD, PG Diploma in Information Science, Cert. Ed., DSc, Hon FCLIP, AUMIST, FRPSL applauds the news:

“Martin has an international reputation. He has authored many notable books on information consultancy and related topics and is co-author of a highly regarded history of the Institute of Information Scientists. As a result of his numerous professional activities, talks, conference

presentations and writings he has become one of the best known and most relied upon senior members of the library and information science profession.”

Martin celebrated both of his awards in a special Zoom lecture on 5 December 2023, which will soon be available online.

The UKeiG awards judging panels would like to thank colleagues who submitted nominations, and we look forward to your submissions in 2024. The excellence and quality of the entries is proof positive that the information retrieval community is thriving.

More about the awards

The Tony Kent Strix Award was inaugurated in 1998 by the Institute of Information Scientists. It is now presented by UKeiG in partnership with the International Society for Knowledge Organisation UK (ISKO UK), the Royal Society of Chemistry Chemical Information and Computer Applications Group (RSC-CICAG) and the British Computer Society Information Retrieval Specialist Group (BCS IRSG).

The Jason Farradane Award is presented in recognition of an outstanding, creative and enterprising contribution to the wider library and information profession. It honours Jason Farradane, who first made an impact on the library and information science community with a paper on the “scientific approach to documentation” presented at a Royal Society Scientific Information Conference in 1948. He was instrumental in establishing the Institute of Information Scientists in 1958, alongside the first academic information science courses in 1963 at the precursor to City University, London, where he became Director of the Centre for Information Science in 1966.

Links

[SearchResearch Online](#)

[Evolution and impact: a history of the Institute of Information Scientists 1958-2002](#)

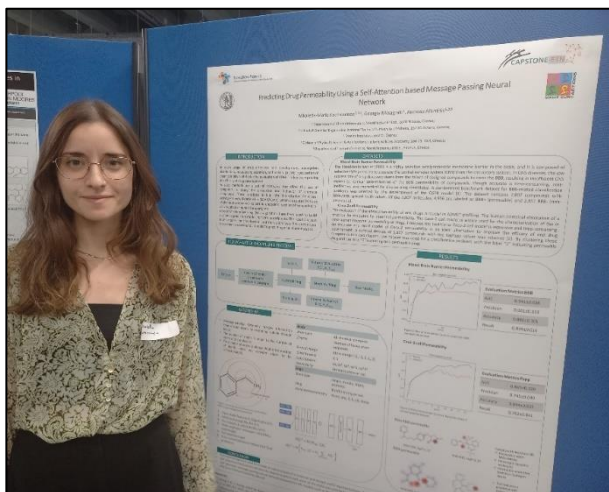
[Martin White – LinkedIn profile](#)

[The Search Network](#)

Meeting Reports: UK QSAR Autumn 2023 Meeting

Contribution from Nikoletta-Maria Koutroumpa, National Technical University of Athens, Greece
email: nkoutroumpa@mail.ntua.gr

The University of Liverpool hosted the UK QSAR Autumn 2023 Meeting on 14 September 2023. I am grateful for being awarded a travel bursary for the meeting, funded by the Royal Society of Chemistry Chemical Information and Computer Applications Group, giving me the opportunity to attend insightful talks from distinguished and young scientists in the field. I am a second-year PhD student at the National Technical University of Athens, Greece in the Department of Chemical Engineering and my research project focuses on discovering new compounds with the use of machine learning and artificial intelligence.



Nikoletta-Maria Koutroumpa with her poster.

At this meeting I had a poster presentation featuring part of my PhD work on predicting molecular properties with the use of deep learning. In more detail, the results of predicting drug permeability using a self-attention-based message-passing neural network were presented. The algorithm outperforms other machine learning models examined. More specifically, Random Forest and Multi-layer Perceptron were tested for their predictive accuracy, and they achieved an area under the curve (AUC) of 0.83, 0.81 for blood-brain barrier permeability, and 0.79, 0.75 for Caco-2 cell permeability, respectively. The self-attention message passing neural network outperformed these machine learning algorithms, resulting in an AUC of 0.94 and 0.87 for each permeability dataset, respectively. Furthermore, since the self-attention layer

assigns a degree of importance to substructures of the molecules, it provides interpretable results that can help researchers design novel, advanced molecules. In the case of permeability, substructures that increase permeability are highlighted, such as lipophilic substituents. In this respect, we can easily visualise the results and identify substituents of the molecules that affect their ability to cross cell membranes.

This one-day meeting was held in person and the agenda included seven highly interesting presentations and poster sessions. One presentation that I distinguished was by Rachael Pirie from NextMove Software, entitled "Can you hear the shape of a drug?" which covered the interesting topic of retrieving molecular surface shape descriptors using the theory of Riemannian geometry. The day closed by awarding the best poster to Madeleine Taylor from the University of Strathclyde, presenting reference interaction site model descriptors in HPLC screening. Attending the UK QSAR meeting, I got a broad spectrum of the *in silico* methods used in different areas of science. Since it was the first meeting at which I presented a part of my PhD project, it was a great opportunity for me to get to know and interact with senior scientists and fellow PhD students in this field.

Contribution from Miguel Garcia-Ortegon, Yusuf Hamied Department of Chemistry, University of Cambridge

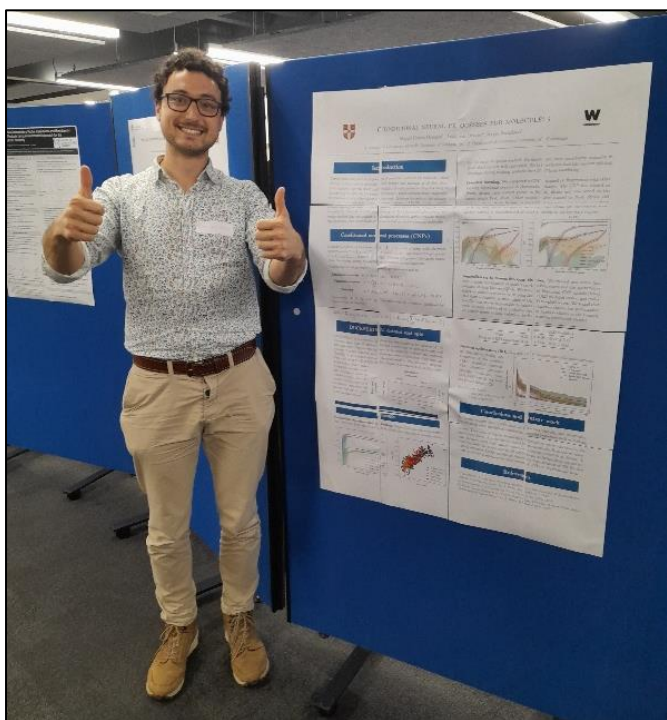
email: miguel.garcia.ortegon@gmail.com

This September I was lucky to attend the UK QSAR Autumn 2023 Meeting in Liverpool. It was a great experience. Not only did I learn a great deal about cheminformatics and drug discovery, but I also met incredible researchers and inspiring people who made me feel fortunate to work in this field.

My name is Miguel Garcia-Ortegon and I am a final-year PhD student at the University of Cambridge. I work on machine learning for drug discovery under the joint supervision of Sergio Bacallado, Andreas Bender and Carl Rasmussen. My project and supervisors are very interdisciplinary, which is refreshing. I am always interested in meeting like-minded people who enjoy working at the frontiers of machine learning, chemistry and biology.

I learnt about the UK QSAR meeting from my supervisor Andreas Bender, who is an active member of the UK cheminformatics community. We submitted our work on conditional neural processes for docking score prediction. Neural processes are models for meta-learning, which means they can be applied to multi-task datasets where each task has very few labelled datapoints. This setting could be highly applicable to drug discovery datasets, which encompass many protein targets and molecules but are highly sparse. In this project,

we benchmarked neural processes on the synthetic dataset of docking scores DOCKSTRING, with a view to applying them to real bioactivities in the future.



*Miguel Garcia-Ortegon with his poster.
Photo: courtesy of Benoit Baillif.*

Excitingly, my submission was accepted for a poster presentation (yay!). In addition, I was awarded the travel bursary by the Royal Society of Chemistry CICAG interest group, which really facilitated my attendance. I am very grateful to CICAG for their generous sponsorship.

My trip to Liverpool started amusingly due to a succession of train strikes, delays and cancellations. I had to travel from Cambridge to Liverpool via London and transfer four times! Sadly, in one of the transfers, the delays got the better of me and I forgot my poster at the station. However, the UK QSAR meeting organiser, Professor Neil Berry, was incredibly kind and printed my poster at the very last minute, using A4 sheets, which we later pasted together. The result was almost as good as the original, as you can see in the picture. A big shout-out and thanks to Neil for coming to the rescue!

The actual conference was really enjoyable and instructive. I highly benefited from the posters and talks. The talks that I found most interesting were by Professor Adam Nelson from the University of Leeds, who presented a chemistry-first strategy for high-throughput phenotypic screening, and by Dr Elena de Orbe from AstraZeneca, who presented methods to identify amino acids where antibodies can be modified selectively. Finally, what I enjoyed the most was meeting other PhD students in the field, who could potentially become collaborators in the future. After the conference, a group of students visited the Liverpool Metropolitan Cathedral, which had a very interesting architecture inspired by Latin American culture.

In summary, attending the UK QSAR meeting was a very enjoyable and productive experience. I highly recommend it to any student in cheminformatics!

Accessing Enhanced Small Molecule Information in the Protein Data Bank in Europe

Contribution from Dr Preeti Choudhary, Senior Bioinformatics Scientist, Protein Data Bank in Europe team, EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK, email: cypreeti@ebi.ac.uk

Protein Data Bank in Europe ([PDBe](#)), a core partner of the [worldwide Protein Data Bank \(wwPDB\)](#) consortium, collects, curates, and provides access to the global repository of macromolecular structure models in the [Protein Data Bank \(PDB\)](#). Managed by the wwPDB consortium, PDB houses over 200,000 PDB entries, with approximately 75% of these structures having at least one bound small molecule. Some of these small molecules

are present due to experimental necessities, while others serve as ligands playing diverse biological roles, such as cofactors, metabolites, substrates, and inhibitors.

Unique small molecules in PDB

To ensure standardised and accurate data on these small molecules, the wwPDB maintains a Chemical Component Dictionary (CCD). This comprehensive reference resource contains data for all unique chemical components, including individual amino acids, nucleotides, and ligands found in PDB entries. The CCD provides descriptions of chemical properties such as stereochemical assignments, chemical descriptors (chemical formula, chemical weight, SMILES and InChI), systematic chemical names, chemical connectivities, and idealised 3D coordinates (generated using Molecular Networks' Corina, and if there are issues, OpenEye's OMEGA). Unique CCD identifiers are used to identify all instances of a specific small molecule in the PDB. For example, ATP, the CCD identifier for adenosine triphosphate, may appear in various conformations when complexed with different macromolecules, in different conditions (e.g. pH, temperature), and/or determined using different experimental techniques.

During refinement and annotation, complex ligands are often fragmented into individual chemical components (CCDs), posing challenges in identification and mapping to other databases. In 2013, the wwPDB introduced the [Biologically Interesting Molecule Reference Dictionary \(BIRD\)](#) to address this issue. This dictionary encompasses entries from the Peptide-like Molecules Reference Dictionary (PRD), each assigned a unique identifier and accompanied by comprehensive descriptions for every chemically distinct peptide-like inhibitor or antibiotic molecule. These entries provide details on the composition, connectivity, chemical structure description, and functions of the respective molecules. For instance, vancomycin, a glycopeptide antibiotic, is accurately identified using the PRD identifier PRD_000204. In 2020, the wwPDB [standardised carbohydrate representation](#) addressing the fragmentation of carbohydrate polymers into individual components (monosaccharides). While BIRD and carbohydrate remediation partially resolves issues for certain peptide-like inhibitors/antibiotic ligands and carbohydrates, numerous fragmented multi-component ligands remain unresolved.

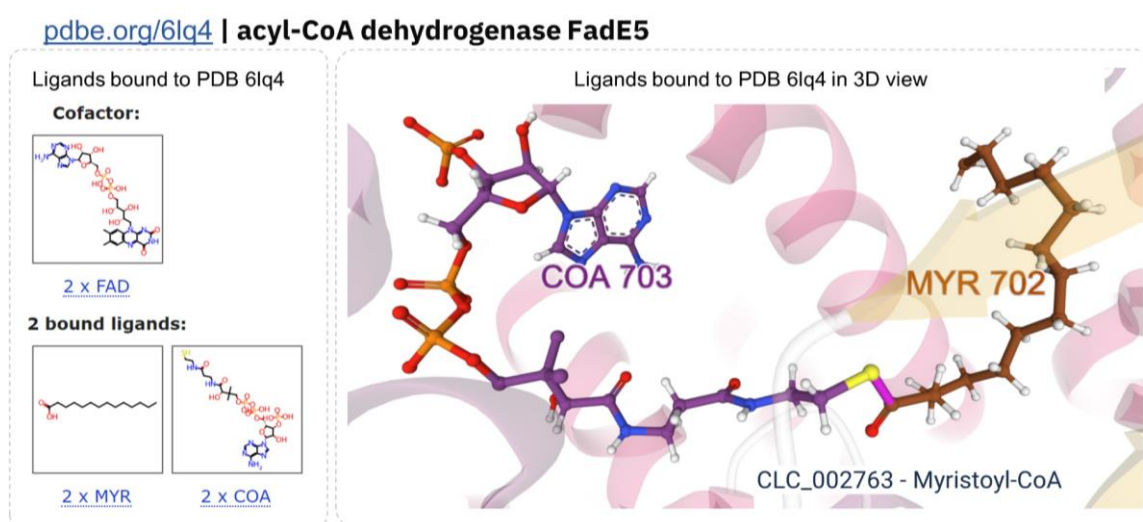


Figure 1. Illustration of a Covalently Linked Component (CLC) in PDB entry 6lq4. In the archival representation (left box), Coenzyme A (COA) and Myristic acid (MYR) are individually annotated as separate ligands bound to the protein. However, the 3D view (right box) reveals the complete molecule, Myristoyl-Coenzyme A, featuring a covalent link between the two components – CoA (carbons coloured purple) and MYR (carbons coloured brown). This complete structure is the substrate for N-myristoyltransferase and is identified using the CLC identifier CLC_002763.

To address this fragmentation challenge holistically, PDBe recently introduced Covalently Linked Components (CLC), a novel class of reference small molecules. These facilitate the identification of ligands composed of multiple covalently linked Chemical Components (CCDs) throughout the entire PDB archive. CLCs offer a more accurate and comprehensive representation of these multi-component ligands, bridging the gaps left by fragmented CCDs not represented in the BIRD dictionary. PDBe has streamlined the identification and analysis of CLCs by assigning them a unique identifier based on InChIKey. For instance, Myristoyl-CoA, composed of covalently linked CCDs COA and MYR, can now be recognised as a single molecule using the CLC identifier CLC_002763.

The identification of CLCs representing chemically complete ligands facilitates mapping to other chemical databases such as PubChem, ChEMBL and KEGG. These additional mappings were not possible before because of the fragmented representation using multiple CCDs. The CLCs along with the wwPDB CCD and BIRD (PRD entries) reference dictionaries provide a comprehensive set of unique small molecule ligands found in the PDB as shown in Figure 2.

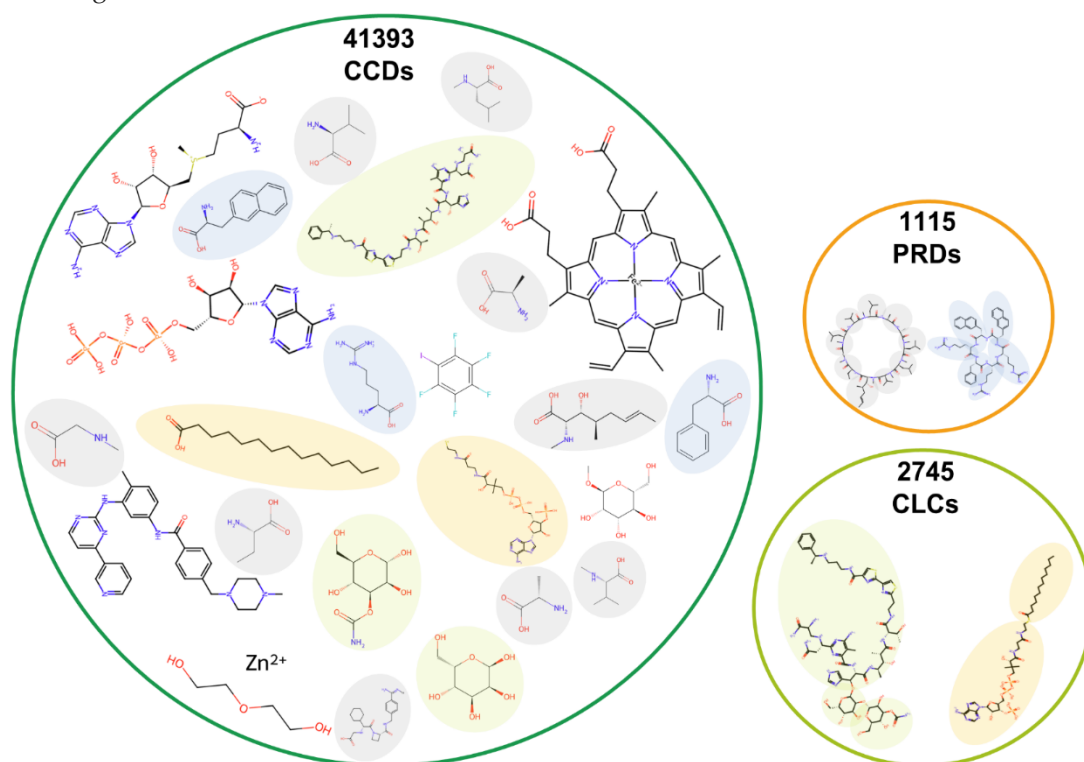


Figure 2: Total number of unique small molecules in the Protein Data Bank (PDB) as of 5 Dec 2023, categorised into Chemical Components Dictionary (CCD), Peptide-like molecules Reference Dictionary (PRD), and Covalently Linked Components (CLC). Notably, CLC and PRD comprise multi-component ligands, each composed of two or more Chemical Components (CCDs).

Enriched small molecule information

The PDBeChem pipeline enriches the small-molecule (CCD, PRD and CLC) files with additional metadata and cross-references to other external resources. For each CCD, PRD and CLC molecule, PDBe provides high-quality 2D depiction, and idealised 3D model coordinates in .sdf, .cml, and .pdb data formats. The reference PDBx/mmCIF files for individual CLC and PRD molecules are also enriched with the following additional metadata:

- Regenerated idealised conformers using RDKit
- Physicochemical properties

- Murcko scaffolds
- Substructures/fragments
- Synonyms
- Cross-references to other external small molecule databases like ChEMBL, PubChem, KEGG, and DrugBank through UniChem
- Additional information on descriptions, synonyms, taxonomy, and known targets for the small molecules mapped to DrugBank

All the above data is available through the [PDBe FTP area for small molecules](#) with specific file information in its [README](#) file.

The PDBeChem pipeline uses [PDBe CCDUtils](#), an RDKit-based toolkit, for handling and analysing small molecules in the PDB. Users can leverage the [tutorial](#) to compute the above data or conduct customised analyses. For more information on PDBe CCDUtils, please refer to the [following publication](#).

Functional annotation of small-molecules

PDBe annotates ligands based on their functional role as cofactor-like, reactant-like or drug-like molecules wherever possible. For instance, coenzyme A is annotated as a cofactor-like molecule in *Clostridium acetobutylicum* thiolase (PDB id: 4x14), Glutamine as reactant-like in Glutamine-dependent NAD(+) synthetase (PDB id: 6ofc), and Pexidartinib as a drug in FMS Kinase domain (PDB id: 4r7h). [Cofactor-like annotations](#) rely on structural similarity to template molecules representing 27 cofactor classes and their association with curated cofactor-binding EC numbers in PDB entries. Reactant-like annotations are based on structural similarity to reactants (substrates or products) in the [Rhea database](#), a curated resource of biochemical reactions. Ligands are annotated as drug-like molecules through structural similarity to drugs in the [DrugBank database](#).

To explore all cofactor-like molecules and their associated PDB entries for a specific cofactor class, users can utilise [PDBe's advanced search](#) or [API endpoints](#). These functional annotations are visually highlighted in green within the ligand gallery on the [PDBe-KB](#) aggregated view pages for a given protein. For example, COA is highlighted as a cofactor-like for the enzyme Acetyl-CoA acetyltransferase, UniProt accession: [A0A0R4I970](#).

Related ligands

Using the [PARITY method](#), all CCDs with 60% or more similarity to a given CCD are identified and made available via the [API endpoint](#). The API endpoint also includes information about stereoisomers and ligands with the same scaffold.

Ligand interactions with macromolecules

PDBe calculates ligand-macromolecule interactions using [PDBe Arpeggio](#) based on the [Arpeggio](#) tool originally developed in the Blundell group at Cambridge University. There are four main categories of interactions determined by Arpeggio: (I) atom–atom; (II) atom–plane; (III) plane–plane; and (IV) plane–group. These interactions can be visualised using 2D and 3D interactive components on the ligand [environment details](#) section of the PDBe entry page (Figure 3), offering a comprehensive view of macromolecule residue-level interactions for each ligand atom.

Accessing data via PDBe API end-points

All the ligand-related information showcased on PDBe and PDBe-KB webpages is powered by API endpoints, offering programmatic access. To know how to retrieve information, including ligand binding residues,

functional roles, and similar ligands, using various API endpoints in Python programming language, please refer to our [API tutorial](#).

Environment details NEW

GDP 401 bound to chain A

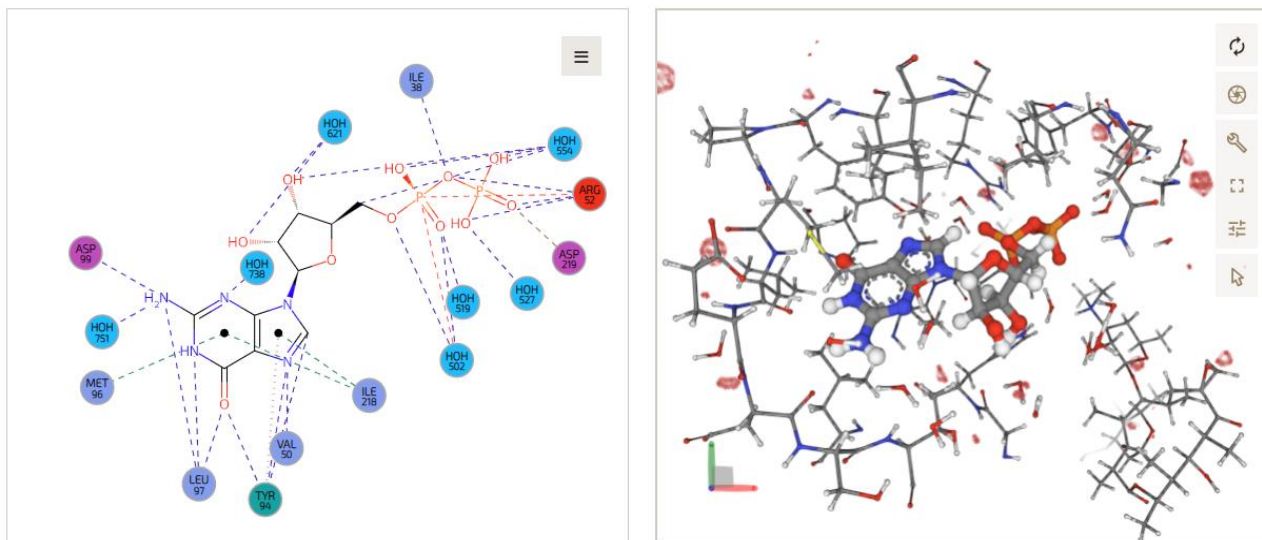


Figure 3: LigEnv (left) and Mol* (right) components, as displayed for a [ligand GDP in PDB entry 5igr](#). The viewer displays atomic-level interactions between ligands and macromolecular binding sites, based on calculated interactions from Arpeggio software. The information displayed in the LigEnv viewer also interacts with an adjacent Mol* viewer, allowing you to easily highlight key binding residues in the 3D structure. These components are interactive, with the highlighting of specific ligand or protein residue in the LigEnv viewer also highlighted in the Mol*.

Upcoming improvements

The long-established [PDBChem service](#) provides a generic browsing interface for ligands in the wwPDB CCD, but it lacks support for PRD and CLC and misses the enriched data from the PDBChem pipeline. PDBe is developing PDBe-KB Aggregated views of ligands, to deliver a more comprehensive view of ligands and their interactions by aggregating data across the PDB for a given ligand. Additionally, PDBe is working on new ligand-related API endpoints and tutorials for PDB-wide protein-ligand interaction data analysis. Additional functional annotation of 'metabolite-like' ligands in PDB is being integrated based on their mapping to [Metabolomics Workbench](#).

PDBe is working with the [ChEMBL](#) and [CCDC](#) teams to enhance data integration across PDB, ChEMBL, and CSD resources. CCDC has developed a pipeline that intersects data from CSD, PDB, and ChEMBL, and this data is being incorporated into [UniChem](#). PDBe and CCDC are also working on annotating binding sites using [fragment hotspot maps](#) generated from CSD data for all the proteins in PDB. Additionally, a dedicated pipeline is being developed to integrate target information and bioactivity data from ChEMBL for all protein-ligand complexes in the PDB. The resulting data will soon be accessible via FTP and subsequently on our web pages as well.

Acknowledgements

We acknowledge the following individuals who have made significant contributions to the work described in the article: Ibrahim Roshan Kunnakkattu, Lukas Pravda, Nurul Nadzirin, Oliver S. Smart, Qi Yuan, Stephen Anyango, Sreenath Nair, Mihaly Varadi and Sameer Velankar from PDBe team. Additionally, we extend our

gratitude to David Lowe and Ian Bruno from CCDC team as well as Melissa F. Adasme and Andrew Leach from ChEMBL team, for their collaborative work on the BBSRC funded BioChemGraph project (BB/T01959X/1).

Meeting Report: Solutions in Science (SinS-1) Conference Report and Invitation to SinS-2

Contribution from Dr Sam Whitmarsh, Director Analytical Science and Digital Transformation, CatSci Ltd, Cardiff, email: sam.whitmarsh@catsci.com; Professor John Langley, Director of Southampton Chemistry Analytical Solutions and Head of the Mass Spectrometry Facility, University of Southampton, email: gjl@soton.ac.uk

The Solutions in Science (SinS) Conference held in July 2023 in Cardiff emerged as a unifying platform for the Separation Science Group meetings, fostering collaboration across academia and industry. With over 150 delegates, SINS-1 showcased cutting-edge analytical work in molecular characterisation, emphasising sustainable analytical science. The conference brought together diverse scientific communities through shared plenary lectures and collaboration with various RSC interest groups and external bodies.

SinS-1, organised by the RSC Separation Science Group and supported by the RSC-CICAG, Analytical Chemistry Trust fund, BMSS, CAMS, ChromSoc, IFST, The Institute of Physics and the British Society for NanoMedicine, featured a comprehensive exhibition with leading instrument suppliers and solution providers. The focus on sustainability extended to meeting construction, with an emphasis on promoting early career scientists through various levels of participation.



Plenary speakers included Dr Francisco Pena-Pereira, Dr Pernilla Sorme and Dr Paul Ferguson, who addressed real-time analysis; minimising the environmental impact of work in labs; and reducing the environmental impact of analytical methods. The conference covered sessions on food, environment, one health, clinical and forensics, molecular characterisation, hyphenated techniques, green techniques, and emerging contaminants, featuring a diverse range of presenters.

The event showcased engaging flash oral presentations, parallel sessions for open discussions, and vendor presentations offering practical tips. Prize winners, recognising the next generation of analytical science

solutions, included poster awardees Etienne Kant, Molly Wilson and Helena Rapp Wright, along with oral presentation winners Rebecca Baker, Elena Bandini and Caitlin Chapman.

The synergy between analytical science and the digital sciences continues to develop as domain expertise transfer and collaboration between the two groups expands. SinS-1 highlighted this further, AI and new digital approaches to data handling featuring prominently in the meeting. Professor Chiara Cordero discussed the application of computer vision algorithms to interrogate 2D-Gas Chromatography (GC) data. The output of the models was fed through a chemometric data processing pipeline to characterise and categorise the quality of hazelnuts and this approach showed potential for a sensomics-based expert system (SEBES). Paul O'Nion also highlighted the application of AI approaches to complex 2D-GC datasets from a commercial perspective, to simplify and accelerate workflows in flavour profiling and taint analysis. Viktoriia Turkina described the challenges in non-targeted high resolution mass spectrometry workflows common in the life sciences and environmental sciences. Conventional methods rely on matching spectral libraries but these approaches have limitations. Turkina's work described the inclusion of a molecular feature prediction model based on a fragmentation tree prediction model. This matched SMILES strings and MS fragmentation patterns from commercially available databases. The model then allowed prediction of MS fragmentation patterns from unknowns and the assignment of a probability score for different structures. The new approach compared favourably with commercially available tools and was found to be able to rank candidates previously indistinguishable including the potential to differentiate between isomers. The conference closed with two plenary lectures: Dr Saer Samanipour highlighted advances in data-driven approaches for predictive analytical support and Professor Steve Conlan talked about real-time genomic analysis using handheld devices.

SinS-1 was notably well rated by exhibitors and attendees alike with a 100% rating of excellent or good, and we are pleased to announce that SinS-2 will be held in Cardiff on 8-10 July 2025. The organisers strongly welcome further contributions that show the potential of new AI and ML data processing methods and encourage you to get in touch. If you have a topic of interest that you would like presented at SinS-2 or you feel that your Interest Group would want to be part of this event, and perhaps join the Scientific Committee please contact our conference Chair [Professor John Langley](#). Equally, if you are from industry and would like to find out more about exhibiting and sponsoring the event can see the [current floor plan here](#) and a selection of [sponsorship opportunities here](#). Please contact info@ilmexhibitions.com for any event enquiries.

Catalyst Science and Discovery Centre News

Contribution from Dr Diana Leitch, Trustee, Catalyst Science and Discovery Centre and CICAG Committee Member, email: diana.leitch@googlemail.com



As 2023 reaches its end there is much to report on the varied and sometimes challenging year which staff and trustees at the Catalyst Science Discovery Centre and Museum in Widnes have experienced.

Sadly I regret to report that three very long-term and elderly supporters died during the year and we deeply regret their loss and also that of one of our younger trustees. Mr Alex Cowan and Dr Ralph Hodge CBE had been involved with Catalyst since 1987 when we were founded as the Halton Museum of the Chemical Industry. Alex worked for Halton Borough Council and Ralph was CEO of ICI Chemical and Polymers Division and one of our original Patrons. Mrs Rona Collins was the widow of another of our founders in 1987, Alderman John Collins, who worked tirelessly with Alex Cowan on our behalf. Finally our fellow trustee, Rachel Wilkinson,

who had worked for BASF and her own legal company had just left our Board of Trustees and died in June at an early age. I attended all their funerals to represent Catalyst.

On a much more positive note the Education staff were set KPI targets for the calendar year (January-December 2023) by the then CEO Dr Lee Juby, and as I write have exceeded them all and are still, as of 7 December, taking last minute bookings so they will increase above these figures below.

- Target for school visits – 200 and already have had 203
- Target to deliver workshops (to schools and the public visitors at weekends and in holidays) – 400 (440 have already been delivered)
- Target for sleepovers – 50 and we have already had 61

At one time we wondered if schools and youth groups would return after covid and the cost of living effects on organisations, but they have and there is growing demand for our services in promoting STEM to young people. Some of our industrial supporters are helping schools to visit us as the cost of coach transport has escalated in the last year and are providing funding for schools in their operational area to come for the day.

Demand for educational visits became so great during the year that we needed an extra member in the educational team to satisfy demand. Thanks to special funding from two well wishers and long-term Catalyst supporters we were able to appoint an extra young Educational Assistant. Emily started a couple of months ago to join Shanna, our existing Educational Assistant and Lucinda Lewis, the Education Manager. All our education work is supported by a band of dedicated volunteers and Weekend Presenters.

In August we were delighted to welcome two groups from mosques in Rochdale who came as part of the Rochdale Science Initiative which is supported by the RSC, and people of all ages thoroughly enjoyed themselves using Catalyst's educational facilities.

We are still welcoming groups and visitors and will be doing so right up to Christmas when we close for a couple of days, so were delighted that our Community Engagement Officer, Clare Lightfoot, after much work was able to persuade John Lewis in Liverpool to donate a large Christmas tree and related decorations to light up one of our public areas. Here in the photo you will see Lucinda in the middle, Clare at the right and one of our education volunteers, Rosemary, on the left.



We have two more Christmas trees, one in our café and shop area and another in our INEOS Inovyn Theatre where on 14 December 2023 we are welcoming visitors to watch and participate in the streaming from the Royal Institution in London of the RI Christmas Lecture on AI which is being given by expert Mike Woolridge. The attendees won't just see what we may all see on our TV screens over Christmas but how the lecture is filmed and an interview with the Director of the Royal Institution.

On Sunday 3 December we welcomed 50 visitors to our fourth floor Observatory to watch the demolition of the first set of cooling towers of the Fiddlers Ferry Power Station near Warrington, Cheshire, on the banks of the River Mersey. Fiddlers Ferry, one of the last coal-fired power stations, was built in 1971. It was decommissioned nearly three years ago and the land has been acquired by Peel Holdings for housing and light industrial development. However, despite being there on a bitterly cold morning from 7.45am for the 'blow down' we

were beaten by the weather as thick fog covered the whole area and our birds-eye panoramic view couldn't happen. We heard the boom and the bang at 9.40am as the towers collapsed and saw the clouds of dust through the fog. I had been asked to give an interview on Radio Merseyside on Monday morning about the event which I did and was able to explain that despite the disappointment of not seeing the collapse the camaraderie among the visitors and staff and the wonderful 'bacon butties' provided by our Café Manager, Denise, meant everyone enjoyed themselves. The visitors left asking when they could buy tickets to see the remaining four cooling towers come down in January 2024. We had also had a sleepover on Saturday night/Sunday morning so the Beavers who were still there having breakfast and a workshop were able to be involved.



So our educational provision is vibrant and greatly helped by the new facilities which were funded by UKRI/Wellcome and installed during the pandemic in 2020-21. Our Interactive Gallery where the Specialist Periodic Table, funded by the RSC and RSC-CICAG, and other hands-on exhibits are located, are always popular.

When we have sponsors and donors we add their names to the Periodic Table and during a recent visit by Mottie and Dr Maggie Kessler, joint Managing Directors of 2M Chemicals, they had chosen molybdenum (Mo) and magnesium (Mg)

as their sponsored elements. Here you will see them in the photo pointing at their elements which appeared on the screen.

Meanwhile our Heritage and Museum side, as we are a joint facility, is going from strength to strength. Our Collections Archivist, Judith Wilde, regularly receives new items to add to our archive and artefacts collections of the chemical industry. I wrote in the last CICAG Newsletter in Summer 2023 about the paintings which originated from the 'Portraits of an Industry' project commissioned by ICI during World War II. We had three of the portraits but we now have four. Alan Utteridge from Wiltshire brought the painting of his grandfather, James Cunningham, and has given it to us. In the photo you will see him presenting it to Judith. James worked at Paints Division at Slough in Buckinghamshire and was involved in the development of varnishes and paints. The hunt goes on for more portraits and I am thinking of writing an article for the magazine 'Who do You Think You Are' about the rest before they disappear for ever.



Another recent acquisition is a very rare document sent to us by Professor Neil Alford MBE, FREng, who is professor of Physical Electronics and Thin Film Materials at Imperial College. It has the signatures of two very powerful women at the time and is dated 1 April 1987. It is The Queen's Award for Technological Achievement

awarded to ICI Advanced Materials Group of Imperial Chemical Industries PLC. Who are the two women – well Queen Elizabeth II and Margaret Thatcher.

The Gossage and the Brunner families continue to be great supporters of Catalyst and I am delighted to report that for the first time ever I met members of the Mond family when they came to the North West for the 150th Anniversary Weekend of the founding of Brunner Mond in 1873 at Winnington near Northwich. My talk to launch the weekend on 20 October 2023 was a great success as were visitors seeing the Catalyst archives which Judith Wilde had taken to Northwich for the public to see there. Members of both the Brunner and Mond families came to Catalyst on 21 October, the Mond family for the first time, since their great grandfather Ludwig Mond, left in 1873 from the building where he had worked with John Brunner and William Gossage. I had managed to organise a special visit for them to the house in Widnes where Ludwig and his wife, Frida Mond, lived and where their two sons, Robert Mond, the chemist and archaeologist and Alfred Mond, the politician and later the 1st Lord Melchett were born. They couldn't believe that the current owner of the house allowed them to see the bedroom.

Project Synergy to secure nearly £1 million from NLHF to improve and update our heritage areas and change how we present the chemical heritage of the area to new audiences continues apace.

Sadly our CEO, Dr Lee Juby FIET, left us in August 2023 to return to the hydrogen industry where he had been working before. He had done a wonderful job in the year he was with us. In the meantime our Education Manager, Lucinda, has been Acting CEO until our newly appointed CEO arrives to take up her post at the end of January 2024. Our new Chair of Trustees, Hugh Dowding, and all the Board of Trustees and Catalyst staff look forward to her arrival.

Finally – breaking news. A student at Manchester Metropolitan University has had her application for a Collaborative Doctoral Award (CDA) approved. Her proposal which was accepted is entitled “Designing sensory environments and engagement in science museums to provide inclusive science learning experiences for audiences with ADHD and ASD”. She is partnering with Catalyst for this work and will concentrate on the age group 7-11 year olds. This group forms a substantial part of our younger visitors. We are really pleased to be involved in this project over the next year.

May I wish all members of CICAG and other readers of this Newsletter warm Seasonal Greetings wherever you may be in the UK or the world and keep up the great work in the chemical information and computer applications roles we all have. If you heard Martin White speaking two days ago on receiving the Jason Farradine and Tony Kent Awards he summed up for me what we are all there to do.

Do come and visit Catalyst if you can.

Meeting Report: RSC-CICAG and RSC-BMCS 6th Artificial Intelligence in Chemistry Symposium, 4-5 September 2023

Contribution from Morgan C. Thomas, Yusuf Hamied Department of Chemistry, University of Cambridge; from December 2023: Computational Sciences, Universitat Pompeu Fabra, email: morgancole.thomas@upf.edu

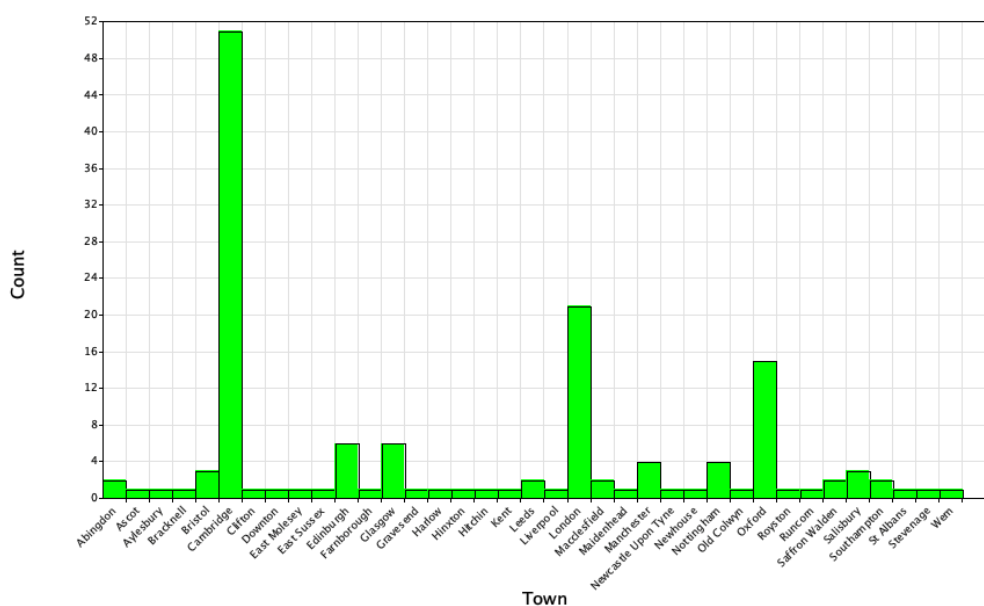
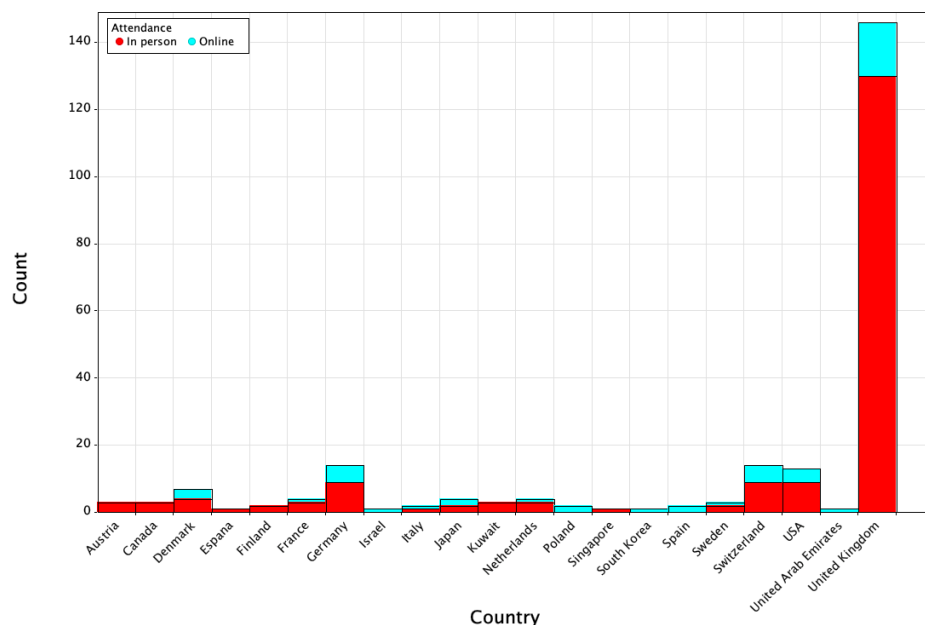
The 6th Artificial Intelligence in Chemistry Symposium, held at Churchill College (Cambridge University, UK) was organised by the Royal Society of Chemistry's Chemical Information and Computer Applications Group (RSC-CICAG) and the Royal Society of Chemistry's Biological and Medicinal Chemistry Sector (RSC-BMCS).

The organising committee was comprised of Samantha Hughes (AstraZeneca, Co-Chair), Garrett M. Morris (University of Oxford, Co-Chair), Chris Swain (Cambridge MedChem Consulting), Nathan Brown (Healx) and Kim Jelfs (Imperial College London). AstraZeneca and OpenBioSim kindly sponsored the conference for which 225 individuals registered with 180 in-person participants. In addition to the organising committee, Hannah Bruce-McDonald (Charm Therapeutics) and Kathryn Giblin (AstraZeneca) also contributed as chairs.

Meeting attendance

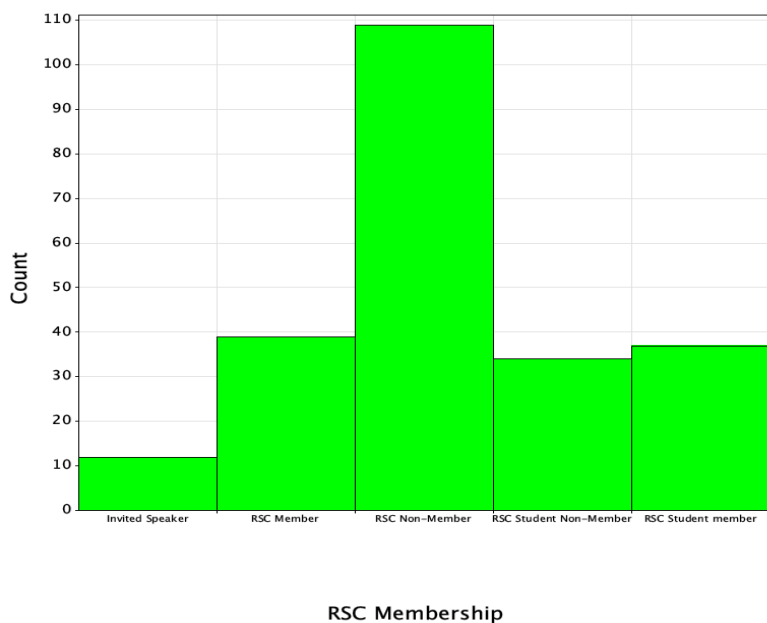
(compiled by Chris Swain)

This year's event co-organised by Royal Society of Chemistry Chemical Information and Computer Applications group (CICAG) and Biological and Medicinal Chemistry Sector (BMCS) was the biggest yet with 231 attendees, with 185 in person and 46 online. Whilst virtual attendees do miss out on some of the interactions that in-person attendees enjoy it does mean the geographical reach of the meeting is much wider and those with difficulty travelling can attend. In total attendees came from 21 different countries and whilst the UK accounted for the majority there were significant numbers from the USA, Germany and Switzerland.



Within the 146 UK attendees 51 were from Cambridge, with substantial numbers from London (21), Oxford (15), Edinburgh (6) and Glasgow (6), but it was encouraging to see attendees from all across the UK.

Whilst the majority of attendees were not RSC members it was very encouraging to see that just over half of the 71 students who attended



were RSC members. These were supported by a record number of 11 bursaries given by CICAG and BMCS. Bursaries are a key element in supporting attendance and are available for all CICAG and BMCS meetings.

Encouraging diversity is one of the important elements of the meeting and the gender ratio among session chairs and speakers was 2:1 M:F, which was similar to the ratio among the attendees (2.6:1), but interestingly the ratio among the students was roughly 1:1.

Of the 231 attendees 88 were from academic labs the others from a wide variety of different industries, non-profit and Government labs.

Scientific content

Language is the future of chemistry (Keynote 1)

Andrew White, University of Rochester and Future House, USA

Andrew White commenced the 6th Artificial Intelligence in Chemistry Symposium with his presentation on the role and use of language for chemistry in the future. He began by sharing his book on the key principles of deep learning for molecules and materials¹ and a recent review of his that largely covers the topic of this talk.² Then Andrew introduced the concept of inductive biases and how they relate to molecules based on the deep learning architecture chosen³ and how indeed graphs are a good natural representation of molecules that can capture equivariances such as permutation equivariance (a swap in atom input should result in an equivalent swap in model output, for example, atomic charge), translation and rotation invariances (a shift in input should not result in any resulting change in model output); necessity and complexity of equivariance/invariance also depends on the choice of 2D or 3D representations. However, challenges remain such as the Weisfeiler-Lehman isomorphism test where two different molecules (for example, decaline and bicyclopentyl) have indistinguishable node environments (with respect to the graph) despite having different atomic environments (with respect to chemistry). Such problems can be overcome by injecting stereoinformation into the node features, but more complicated isomorphism challenges exist like different helicene rotamers in 3D space. Andrew's main takeaway was that "molecules are not simple graphs" and deep graph architectures must be applied appropriately.⁴ Andrew also shared a high-profile example, namely AlphaFold,⁵ where removing model invariance only resulted in a marginal decrease in performance of 0.2, and so he also raised the question "does it matter?". Moreover, molecules exist in dynamic states and not only a point cloud which further complicates the modelling problem.

Andrew then steered his presentation towards theoretically simpler problems such as the relationship between chemical synthesis, product, and related properties and how large language models (LLMs) can be used with fine-tuning⁶ to address these problems in chemistry,⁷ especially in low-data scenarios. Andrew explained that in their lab they integrate Bayesian optimisation with LLMs to select or acquire data optimally.⁸ Andrew

demonstrated how his team extract uncertainty estimation from the token probabilities during model inference, how LLMs can undergo in-context learning at inference only (without the need to update model parameters) by providing training data within prompts, for example, achieving a Pearson correlation of 0.921 on a solubility prediction task from IUPAC name, and how uncertainty and in-context learning can be combined for Bayesian optimisation of reaction catalysis. In a from-scratch experimental trial, he showed that a catalyst performing at effectively thermodynamic maximum was found within six iterations using these techniques. Importantly, Andrew mentioned that the identified catalyst exists in the literature and so could be present in the LLM training data and that, moving forward, more careful experiments are needed to investigate the novelty of proposed materials.

Andrew then asked, “So why do LLMs work?” followed by stating that language has been honed for thousands of years in the real-world and that if we are able to model it, we should utilise the power of language. Following this logic, that we should teach other ML models to emit language such that the output can be used for training, fine-tuning or in-context learning of LLMs. As a more specific example, how can we extract any structure-activity relationships of molecules and provide natural language explanations. Before proceeding Andrew clarifies “What is an explanation”,⁹

1. Justification: reasoning for using a prediction, like test error.
2. Interpretability: “the degree to which an observer can understand the cause of a decision”.
3. Explanation: a presentation of information intended for humans that give the context and cause for a prediction.

With respect to chemistry, Andrew described how molecular counterfactuals (molecules containing a small structural change to a reference but a large difference in prediction) can be used to inform what change may be required;¹⁰ this can be approximated by chemical space enumeration.¹¹ Alternatively, how an interpretable surrogate model can be used where the features are representable by language (for example, different functional groups or substructures).¹² T-statistics can then be used to look at the significance of features to the prediction and the corresponding language explanation can be summarised by an LLM and used to aid prompts.

In the last section of Andrew’s talk, he introduced neuro-symbolic computing¹³ (aka Agents) that provide access to tools available for LLMs to use, which could be seen as an alternative to multi-modal models. One example Andrew shared was providing key scientific papers from PubMed as context before asking a literature question,¹⁴ which prevents hallucination of references. In the context of chemistry, this could be tools such as web search, PubChem search, Python, RDKit, synthesis planner etc. and Andrew went on to show an example of how these tools could be used by an LLM to design a compound with a similar profile to Dasatinib and purchase it. Andrew then described his work on ChemCrow¹⁵ which is an Agent integrated with a variety of chemistry tools that demonstrated the ability to conduct end-to-end virtual screening to identify a chromophore with a specified wavelength. This process included data cleaning, preprocessing, featurisation, Random Forest model training, inference, and finally synthetic route planning. ChemCrow showed improved performance overall on such chemistry-related tasks compared to GPT-4. Andrew concluded by introducing his start-up Future House which is a non-profit trying to improve the performance of such Agents and to conduct wet-lab experimentation. Finally, Andrew stated that “we are on the cusp of a revolution in science – we can connect our data, research papers, the internet, and models within one framework”.

References

- (1) White, A. D. Deep Learning for Molecules and Materials. *LiveCoMS*. **2022**, 3(1), 1499. DOI: 10.33011/livecoms.3.1.1499
- (2) White, A. D. The future of chemistry is language. *Nat. Rev. Chem.* **2023**, 7, 457–458. DOI: 10.1038/s41570-023-00502-0
- (3) Battaglia, P. W. et al. Relational inductive biases, deep learning, and graph networks. *arXiv*. **2018**. DOI: 10.48550/arXiv.1806.01261

- (4) Joshi, C. K. et al. On the Expressive Power of Geometric Graph Neural Networks. *arXiv*. **2023**. DOI: 10.48550/arXiv.2301.09308
- (5) Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, 596, 583–589. DOI: 10.1038/s41586-021-03819-2
- (6) Dinh, T. et al. LIFT: Language-Interfaced Fine-Tuning for Non-Language Machine Learning Tasks. *arXiv*. **2022**. DOI: 10.48550/arXiv.2206.06565
- (7) Jablonka, K. M. et al. Is GPT-3 all you need for low-data discovery in chemistry? *ChemRxiv*. **2023**. DOI:10.26434/chemrxiv-2023-fw8n4
- (8) Ramos, M. C. et al. Bayesian Optimization of Catalysts With In-context Learning. *arXiv*. **2023**. DOI: 10.48550/arXiv.2304.05341
- (9) Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *AIJ*. **2019**, 267, 1-38. DOI: 10.1016/j.artint.2018.07.007
- (10) Wellawatte, G. P. et al. A Perspective on Explanations of Molecular Prediction Models. *J. Chem. Theory Comput.* **2023**, 19(8), 2149-2160. DOI: 10.1021/acs.jctc.2c01235
- (11) Nigam, A. et al. Beyond generative models: superfast traversal, optimization, novelty, exploration and discovery (STONED) algorithm for molecules using SELFIES. *Chem. Sci.* **2021**, 12, 7079-7090. DOI: 10.1039/D1SC00231G
- (12) Gandhi, H. A. et al. Explaining structure-activity relationships using locally faithful surrogate models. *ChemRxiv*. **2022**. DOI: 10.26434/chemrxiv-2022-v5p6m-v2
- (13) Karpas, E. et al. MRKL Systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning. *arXiv*. **2022**. DOI: 10.48550/arXiv.2205.00445
- (14) Jin, Q. et al. PubMedQA: A Dataset for Biomedical Research Question Answering. *arXiv*. **2019**. DOI: 10.48550/arXiv.1909.06146
- (15) Bran, A. M. et al. ChemCrow: Augmenting large-language models with chemistry tools. *arXiv*. **2023**. DOI: 10.48550/arXiv.2304.05376

Geometric ML: from Euclid to drug design (Keynote 2)

Michael Bronstein, University of Oxford, UK

In the next keynote presentation, Michael Bronstein commenced by reiterating a thought-provoking quote: “Symmetry is an idea by which humanity, throughout the ages, has endeavoured to comprehend and instil order, beauty, and perfection”. Subsequently, he delved into the historical origins of symmetry, tracing its roots back to approximately 370 BC as a geometric concept attributed to the works of Plato and Euclid. Modern geometry underwent consolidation through the efforts of F. Klein, who conceptualised it as a space defined by transformation groups. For instance, classic Euclidean geometry can be transformed through rigid transformations. Within the realm of machine learning, an assortment of architectural approaches exists, each tailored to distinct tasks and data types. Geometric deep learning, the focus of Michael’s work, seeks to elucidate the foundational principles underlying these various architectures, with the aim of unifying them and spawning new avenues for research.

Michael proceeded to touch upon the curse of dimensionality, briefly noting that high-dimensional inputs necessitate an exorbitant number of samples for effective learning. Neural networks (NNs) derive their efficacy from possessing strong inductive biases. Michael illustrated how Convolutional Neural Networks (CNNs) leverage shift equivariance, a property stemming from associated symmetries. Graph Neural Networks (GNNs) confront two distinct symmetries: the internal symmetry of the domain and the external symmetry of the data, encompassing permutation symmetry and rotation/translation symmetry. Transitioning more specifically to the domain of GNNs, Michael emphasised their utility in modelling interactions within systems such as social networks or molecular structures. Such systems can be represented computationally through matrices, including feature matrices and adjacency matrices. However, this approach imposes an arbitrary node order which must be accounted for. GNNs conduct permutation-invariant learning of node features by employing node pooling with an equivariant concatenation strategy. These node features can be acquired through convolutional, attentional, and message-passing techniques – he asked “what problems can we model with this these characteristics?”

Michael then highlighted the Weisfeiler-Lehman test, a method that iteratively examines node types based on their neighbours until no further changes occur. However, he noted that this test is inadequate for certain scenarios, exemplified by instances like decalin and bicyclopentyl, where chemically different atoms are indistinguishable nodes using this procedure. As a solution, GNNs can enhance their expressiveness by incorporating more intricate tests. For example, the Weisfeiler-Lehman test is unable to assess cycles, so cycle-related information, such as triangles, can be embedded to enhance a GNN's predictive capabilities for molecular properties, reducing predictive error.

Proceeding to the topic of Topological Message Passing, Michael highlighted the inclusion of different dimensions.¹ This approach encompasses 0-dimensional nodes, 1-dimensional nodes with edges, and 2-dimensional nodes with edges and rings (simplices). Other techniques for bolstering expressiveness were also mentioned, including the addition of positional encoding, equivariant GNNs, subgraph GNNs, or cellular GNNs. Additionally, the question of whether graphs are essential was raised, with alternatives like DeepSet/PointNet, suitable for scenarios involving point clouds or atomic arrangements, discussed.

Michael then delved into the field of manifolds and geometric graphs, emphasising their significance due to their capacity to represent the surfaces of three-dimensional objects while disregarding their internal structures. In chemistry, this abstraction simplifies problems like protein interactions by eliminating irrelevant internal details. Manifolds feature both local symmetries (local gauge transformations, such as coordinates on the manifold) and global symmetries (deformations). Manifolds are only Euclidean locally, necessitating distance and angle metrics in tangent space, which are intrinsic quantities, meanwhile any deformation on the manifold is isometric. To integrate manifolds with NNs, Geodesic CNNs, permit the local representation of signals in tangent space, rendering them invariant to deformations. However, the challenge arises in establishing a local reference frame, given the ambiguity in manifold definitions. Hence, gauge-equivariant filters are designed to maintain invariance to surface deformations. Manifolds result in an intermediate graph representation, with the initial neighbour being arbitrary but following this they are more standardised, offering more structure than molecular graphs.

Michael then raised some practical examples, such as protein folding. However, he stressed the need to model the sequence-to-structure and structure-to-function relationships, noting that similar sequences do not necessarily correlate with similar folds. Moreover, concepts like lock-and-key still need to be considered for predicting protein-ligand. He also mentioned how GNNs have been very popular in the discovery of new molecular structures such as antibiotics,² however, moving to other problems like protein-protein interactions (PPIs) is a different challenge. For PPIs, geometric deep learning on meshes can be used.³ Furthermore, he discussed their use on *de novo* protein design using MaSIF, which was experimentally tested with PD-L1 binders and for SARS-COV-2 spike binder design.⁴

Anticipating the future of deep learning, Michael alluded to the rise of generative models, specifically the use of diffusion models.⁵ These models learn a denoising process from input data that undergoes noising to a Gaussian distribution. These can be conditioned on protein pocket 3D structure and equivariant GNNs⁶ can be used to 'wiggle' atoms into the pocket to generate new *de novo* linkers as in DiffLinker.⁷

In conclusion, Michael emphasised the power of geometric deep learning in molecular modelling, underlining several successful case studies. However, experimental validation remains a crucial. He concluded with the quote: "The knowledge of certain principles easily compensates the lack of knowledge of certain facts" –Claude Adrien Helvetius.

References

- (1) Bodnar, C. et al. Weisfeiler and Lehman Go Topological: Message Passing Simplicial Networks. *arXiv*. **2021**. DOI: 10.48550/arXiv.2103.03212
- (2) Stokes, J. M. et al. Deep Learning Approach to Antibiotic Discovery. *Cell*. **2020**, 180(4), 688-702. DOI: 10.1016/j.cell.2020.01.021
- (3) Gainza, P. et al. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat. Methods*. **2020**, 17, 184–192. DOI: 10.1038/s41592-019-0666-6
- (4) Gainza, P. et al. De novo design of protein interactions with learned surface fingerprints. *Nature*. **2023**, 617, 176–184. DOI: 10.1038/s41586-023-05993-x
- (5) Schneuing, A. et al. Structure-based Drug Design with Equivariant Diffusion Models. *arXiv*. **2022**. DOI: 10.48550/arXiv.2210.13695
- (6) Satorras, V. C. et al. E(n) Equivariant Graph Neural Networks. *arXiv*. **2021**. DOI: 10.48550/arXiv.2102.09844
- (7) Igashov, I. et al. Equivariant 3D-Conditional Diffusion Models for Molecular Linker Design. *arXiv*. **2022**. DOI: 10.48550/arXiv.2210.05274

Explainable prediction of catalysing enzymes from reactions using multilayer perceptrons

Daniel Probst, École Polytechnique Fédérale de Lausanne, Switzerland

Daniel started his presentation by introducing the need to obtain explanations for predictions and the corresponding field of Explainable AI (XAI).¹ In brief, he points out that XAI involves 1) post-hoc explainable, 2) model specific, 3) multi-layer neural networks (NNs) and 4) feature relevance, explanation, and visualisation. He then mentioned that the primary focus of his presentation is on post-hoc explainable model-specific approaches with feedforward neural networks (FFNNs), including feature relevance, explanation, and visualisation. These are necessary to ensure symmetry between dry-lab and wet-lab scientists.

The presentation then delved into the application of XAI for reaction yield prediction, specifically employing the Differential Reaction FingerPrint (DRFP)² in combination with XGBoost. This outperformed large language models like BERT resulting in state-of-the-art performance. Daniel then posed the question “How can we advance this further?”. The proposed solution involved using Shapley Additive Explanations (SHAP) to approximate the contribution of individual features to the model’s predictions. He found this particularly beneficial in identifying critical features for yield prediction but also acknowledged that SHAP tends to perform well with Random Forest and tree-based models, but its use with NNs is somewhat more of an approximation.

Daniel focused on the application of these techniques to biocatalysis, an area with decent yield data. He used a 1-layer Multi-Layer Perceptron (MLP) DRFP mapping with atom-wise contribution aggregation to explain his predictions. Daniel emphasised that in this case study, model interpretability held more significance than model performance. The explanations did shed some light on misclassifications by examining negative contributions to the prediction. This was helpful to identify issues with both the data and model architecture. For example, the training data for sulfohydrolase reactions lacked the atomic environments found in the test set. Daniel highlighted that such insights were valuable for evaluating, enriching, or augmenting training data. The framework and tools are available as a Python package under the name ‘theia-pypi’ with additional information and resources accessible on Daniel’s [GitHub page](#). Additionally, a collaborative notebook is available (t.ly/UkAkS) to facilitate further exploration and experimentation.

Daniel concluded by referencing work by Saebi et al.³ that also used DRFP resulting in SOTA on real-world datasets for yield prediction reinforcing the success of this approach with respect to both performance and explanation.

References

- (1) Arrieta, A. B. et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82-115. DOI: 10.1016/j.inffus.2019.12.012
- (2) Probst, D. et al. Reaction classification and yield prediction using the differential reaction fingerprint DRFP. *Digital Discovery* **2022**, *1*, 91-97. DOI: 10.1039/D1DD00006C
- (3) Saebi, M. et al. On the use of real-world datasets for reaction yield prediction. *Chem. Sci.* **2023**, *14*, 4997-5005. DOI: 10.1039/D2SC06041H

Transformer models for synthesis prediction in the pharmaceutical domain

Samuel Genheden, AstraZeneca, UK

Samuel started by introducing his team's previous work on AiZynthFinder¹ which uses reaction templates and a neural network to search and rank retrosynthetic routes using Monte Carlo Tree Search to identify optimal routes, based on work by Segler et al.² However, alternative, template-free (TF), approaches exist that are made possible by using large language models and framing it as a translation problem³ i.e., reactants as input and products as outputs or vice versa. The advantages of TF synthesis compared to template-based (TB) include the lack of necessity for atom-mapping and template extraction steps, as well as better scaling to large chemical space and possibility to extrapolate outside known chemical space. He then mentioned that within AstraZeneca (AZ), AiZynthFinder is a well-established TB baseline that chemists use daily, so value must be added to supersede this approach, especially considering that this TB approach provides the ability to search internal or literature data for reaction data containing a particular and specific template.

To investigate the potential value added by large language models, a Transformer model (Chemformer⁴) was trained with 30M Reaxys reactions, 10M NextMove reactions and 2M AZ electronic lab notebook reactions resulting in a final corpus of 18M unique reactions.

First Samuel posed the question, "How much do one-step retrosynthesis models impact route prediction?". He then showed that they found a disconnect between the performance of single-step synthesis prediction and successful multi-step route finding, highlighting that single-step performance is not necessarily predictive of overall performance when integrated into full retrosynthesis planning.^{5,6} Overall, USPTO-50K was insufficient as a benchmark and a single-step model could result in an over 25% change in the success rates of retrosynthesis planning, also leading to the selection of different chemistry along the route. Then Samuel asked, "Is Chemformer a chemically viable option to replace TB retrosynthesis?". Upon further investigation it was confirmed that the Chemformer performance varied based on reaction class, for example, more complex rearrangements don't perform as well in exact matching and round trip accuracy. Compared to TB, the Chemformer TF approach performs better on single-step prediction by exact matching and round trip accuracy. For multi-step performance, 23.4% of 5k compounds designed by AZ chemists were only identified by the TF approach compared to 0.7% only identified by the TB approach. However, the routes suggested by the TF approach are slightly less feasible than TB according to the forward reaction prediction model. The Chemformer was able to utilise 4,656 different and more diverse templates not present in the TB approach, however, these were also deemed less feasible and in-fact, 25% of these 'new' templates are actually multi-step reactions with two distinct reaction sites. So not all novel templates are exploiting novel chemistry.

Overall, Samuel shared what he thinks is the next step, stating that he thinks that the Chemformer is a viable option and will be used to replace or augment the TB approach adding value to retrosynthesis planning within AstraZeneca. With regards to implementation, it is more expensive to run the Chemformer in production and it comes with the caveat of being unable to link back to experimental and literature cases.

References

- (1) Genheden, S. et al. AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. *J. Cheminform.* **2020**, *12*, 70. DOI: 10.1186/s13321-020-00472-1
- (2) Segler, M. et al. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature.* **2018**, *555*, 604–610. DOI: 10.1038/nature25978
- (3) Schwaller, P. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *CS Cent. Sci.* **2019**, *5*(9), 1572–1583. DOI: 10.1021/acscentsci.9b00576
- (4) Irwin, R. et al. Chemformer: a pre-trained transformer for computational chemistry. *Mach. Learn. Sci. Technol.* **2022**, *3*, 015022. DOI: 10.1088/2632-2153/ac3ffb
- (5) Hassen, A. K. et al. Mind the Retrosynthesis Gap: Bridging the divide between Single-step and Multi-step Retrosynthesis Prediction. *arXiv.* **2022**. DOI: 10.48550/arXiv.2212.11809
- (6) Torren-Peraire, P. et al. Models Matter: The Impact of Single-Step Retrosynthesis on Synthesis Planning. *arXiv.* **2023**. DOI: 10.48550/arXiv.2308.05522

Transforming chemistry with Transformers

Kevin Maik Jablonka, University of Jena, Germany

Kevin commenced his presentation stating that the fundamental requirement for machine learning is data, most of which is in a language format, especially in the first instance of communication. This holds true also for chemistry, for example synthesis instructions. Nowadays, we can simply ask large language models (LLMs) like ChatGPT – so he poses the questions “what role do LLMs play in chemistry?” and “why is text so powerful?”. He then splits the roles that LLMs can play in chemistry into four different analogies: an all-knowing professor, a director, a curator, and a teacher. Additionally, investigating the latent vectors of similar words it is found they exist in similar latent space; therefore, it is possible to rank materials based on proximity in the latent space to discover novel materials with similar or even favourable properties.¹

He then likened LLMs to an all-knowing professor due to their ability to conduct text completion in a predictive sense and even given tabular data in a textual representation.² This can outperform state-of-the-art Random Forest models with as few as 50 data points and applies to a variety of tasks (including classification, regression, and inverse design) and representations (including SMILES, SELFIES, IUPAC, etc.). He stated that the IUPAC name is the most performant representation but that fine-tuning with multiple representations increases accuracy and provides a measure of confidence based on the variance of prediction across representations. Finally, he mentions that out-of-distribution molecules can be generated and properties can be optimised with careful iterative training. Practically, these models are incredibly easy to use with a scikit-learn-like API in Python.

Next, he showed how they can also be likened to directors with the use of Agents (also mentioned previously in Andrew White’s presentation), i.e., LLMs with access to other tools (for example, ReACT,³ Toolformer,⁴ MRKL system⁵). He also mentioned that a large advantage of LLMs is actually mapping predictions to procedures to be acted upon, which cannot be done by many ML models. Finally, he demonstrated how LLMs could be combined with the rigid structure of electronic lab notebooks (ELNs) as a virtual assistant to query ELNs aiding flexibility, for example, a semantic search, to dynamically create interfaces etc.

Kevin moved on to how LLMs could also be used as curators by parsing unstructured text into structured text like JSON to describe knowledge graphs, or process synthesis instructions into robotic friendly automated code. Moreover, how they can also be used as teachers for students to provide personalised feedback, for example to summarise lectures notes or compose test questions and model answers. Tools like Whisper⁶ can be used to classify a lecture video.

Overall Jablonka concluded by acknowledging the power and multiple roles LLMs have in chemistry, most of which are covered in a recent publication.⁷ He added that perhaps new opportunities can be found in foundation models for chemistry from different data-sources. This is also being spearheaded by the [ChemNLP](#) project in collaboration with Michael Pieler and Andrew White.

References

- (1) Tshitoyan, V. et al. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **2019**, *571*, 95–98. DOI: 10.1038/s41586-019-1335-8
- (2) Jablonka, K. M. Is GPT all you need for low-data discovery in chemistry? *ChemRxiv*. **2023**. DOI: 10.26434/chemrxiv-2023-fw8n4-v2
- (3) Yao, S. et al. ReAct: Synergizing Reasoning and Acting in Language Models. *arXiv*. **2023**. DOI: 10.48550/arXiv.2210.03629
- (4) Shick, T. et al. Toolformer: Language Models Can Teach Themselves to Use Tools. *arXiv*. **2023**. DOI: 10.48550/arXiv.2302.04761
- (5) Karpas, E. et al. MRKL Systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning. *arXiv*. **2022**. DOI: 10.48550/arXiv.2205.00445
- (6) Radford, A. et al. Robust Speech Recognition via Large-Scale Weak Supervision. *arXiv*. **2022**. DOI: 10.48550/arXiv.2212.04356
- (7) Jablonka, K. M. et al. 14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon. *Digital Discovery*. **2023**, Advance Article. DOI: 10.1039/D3DD00113J

Discovery of synthesisable organic materials

Steven Bennett, Imperial College London, UK

Steven commenced his conference presentation by framing it as an exploration of the interplay between theoretical predictions and experimental validations in the field of Porous Organic Cages (POCs). POCs are organic molecules that form stable cages with an internal cavity that can be used in applications such as gas storage, molecular separations, and catalysis. Their ease of synthesis (for example, by imine condensation of suitable pre-cursors) and ease of post functionalisation provides an advantage compared to zeolite alternatives. However, many topologies and geometries exist¹ that are dependent on precursor and reaction conditions which need to be incrementally optimised. He then stated that high-throughput methods have accelerated discovery since 2018, but machine learning can also be used to predict the properties of organic cages for further acceleration of discovery.

Steven explained how precursors can be used to predict final cage properties.² For example, how different structural features can be extracted to identify cage property relationships or how Variational Autoencoders (VAEs) can be used to identify precursors in similar latent space that are likely to yield analogous POC geometries.³ However, real-world challenges in POC research remain, such as component availability, synthesis complexity, and the formulation of the final material. Reactions can yield various products, including polymers, which may differ from the predicted outcomes. Furthermore, POCs properties may not correlate with prediction and in particular, the stability of these materials in the solid state is a crucial concern.

The presentation then delved into precursor selection, a difficult challenge due to the size of chemical space and synthetic requirements and cost. Steven then showed how common techniques like SAscore⁴ and SCscore⁵ failed to replicate chemical intuition for 12,553 binary hand-labelled precursors. Therefore, they directly modelled these labels with a Random Forest (RF) model resulting in 84% precision.⁶ Steven then used this model to filter out the least predicted synthesisable 99% of a large database of possible precursors down to 28,185 precursors whose cage configuration was then predicted using stk⁷ and pyWindow.⁸ This *in silico* approach resulted in 100s of possible cages with predicted shape-persistence. Notably, the RF model filtering was

predicted to result in larger cages compared to using SAScore or SCScore, suggesting new discoveries using this approach.

Steven's presentation then moved on to the fact that there are still unknowns: "Will the precursors react to form the desired product, or side-products? Will it converge? Will the cage collapse?". To tackle this, Steven presented an automated pipeline to systematically combine precursors in a high-throughput manner, analysed by NMR and LCMS, followed by an automated Python pipeline to assess cage topology. This crucially assesses if the reactions have completed conversion and which topology has been formed. In total, approximately 140 reactions were categorised into 10 different success classifications. During the screen, two completely novel cages were discovered. Post-hoc analysis highlighted that different precursors are more likely to generate stable cages, for example, more flexible diamines form many cages, but they tend to collapse, whereas rigid diamines formed fewer cages but when they did, they were more shape-persistent. Steven then shared a slide showing a diverse range of cage topologies that were identified from 4-15Å and stating that some literature cages were rediscovered, although transferring these out of solution and into formulation were still a concern.

References

- (1) Santolini, V. et al. Topological landscapes of porous organic cages. *Nanoscale*. **2017**, *9*, 5280-5298. DOI: 10.1039/C7NR00703E
- (2) Szczypiński, F. T. et al. Can we predict materials that can be synthesised? *Chem. Sci.* **2021**, *12*, 830-840. DOI: 10.1039/D0SC04321D
- (3) Zhou J. et al. Deep Generative Design of Porous Organic Cages via a Variational Autoencoder. *ChemRxiv*. **2023**. DOI: 10.26434/chemrxiv-2023-ggnz0
- (4) Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform.* **2009**, *1*(8). DOI: 10.1186/1758-2946-1-8
- (5) Coley, C. W. et al. SCScore: Synthetic Complexity Learned from a Reaction Corpus. *J. Chem. Inf. Model.* **2018**, *58*(2), 252–261. DOI: 10.1021/acs.jcim.7b00622
- (6) Bennett, S. et al. Materials Precursor Score: Modeling Chemists' Intuition for the Synthetic Accessibility of Porous Organic Cage Precursors. *J. Chem. Inf. Model.* **2021**, *61*(9), 4342–4356. DOI: 10.1021/acs.jcim.1c00375
- (7) Turcani, L. et al. stk: An extendable Python framework for automated molecular and supramolecular structure assembly and discovery. *J. Chem. Phys.* **2021**, *154*, 214102. DOI: 10.1063/5.0049708
- (8) Miklitz, M.; Jelfs, K. E. pywindow: Automated Structural Analysis of Molecular Pores. *J. Chem. Inf. Model.* **2018**, *58*(12), 2387–2391. DOI: 10.1021/acs.jcim.8b00490

Data-driven materials discovery (Keynote 3)

Jacqui Cole, University of Cambridge, UK

Jacqui commenced the third keynote talk by introducing historical and current approaches to material discovery: serendipity, trial-and-error, and systematic materials design. She emphasised that in order to move towards systematic materials design we need data, which is pivotal to conducting data-driven discovery. She then proposed the notion of having comprehensive data encompassing all possible molecules and their associated properties to establish structure-function relationships, and that a search engine could then be used to probe all of chemical property space. In reality, Jacqui acknowledged we do not have this information available but perhaps we do in a highly fragmented form, i.e., the scientific literature. Therefore, there is a need to consolidate this for effective utilisation in materials discovery. Jacqui highlighted the work she has conducted towards this end by the development of tools to facilitate data extraction from scientific literature in the form of image data (ImageDataExtractor¹), chemical data (ChemDataExtractor^{2,3}), parsing chemical schematics (ChemSchematicResolver⁴), and reaction data (ReactionDataExtractor^{5,6}).

Her presentation then delved into the specifics of ChemDataExtractor,² a software tool designed to process documents and extract chemical names and related properties into a chemical database. Jacqui explained the

methodology as natural language processing which involves sentence splitting, word tokenisation, grammatical tagging, parsing into a hierarchical tree, and interdependency resolution to associate synonyms and numerical data with molecular structures. The new version (version 2.0³) also allows customisation of ontologies. As a demonstration of its use, Jacqui focused on the energy sector and photovoltaics. She posed the question of whether the software could aid in identifying new light-harvesting materials, which in practice is usually a combination of two molecules to optimise absorption over the full spectrum of sunlight wavelength emission. The software successfully extracted 9431 dye candidates, which underwent a filtering process resulting in the selection of five dyes with favourable properties. These dyes, initially not designed for photovoltaics, were experimentally validated, and achieved near-industry-standard efficiency (92% of the organometallic standard).

Additionally, Jacqui highlighted several side applications of ChemDataExtractor. For example, the creation of computational databases from experimental data,⁷ enabling property prediction and accuracy validation. Alternatively, it is also applicable to device data and not just molecules.⁸

Transitioning to a different application, Jacqui discussed the use of ChemDataExtractor in mining data related to batteries, which yielded a dataset of 292k entries. However, the software encountered difficulties in identifying whether materials were anodes, cathodes, or electrolytes. To address this issue, Jacqui described the development of an interactive module using extracted data and a Transformer model to classify the components correctly. This module was applied to the original dataset, facilitating comprehensive data mining.⁹ This advancement led to the creation of BatteryDataExtractor,¹⁰ which combined ChemDataExtractor with BERT and incorporated confidence scores to enhance material and property identification. The process involved iterative question-and-answer prompts to extract relationships.

Jacqui then raised the question of “What’s next?” and mentioned photocatalysis for water splitting as a more complex endeavour due to the multitude of conditions and potential products involved. This challenge was addressed by categorising data into photocatalyst, cocatalyst, additives, and light sources. However, a key challenge here is that reactions are often depicted in scientific literature as figures. So, Jacqui developed ReactionDataExtractor.^{5,6} This tool identifies arrows, molecule names, conditions, and molecular graphs using DECIMER.¹¹ Jacqui concluded by briefly mentioning the functionality of other tools such as ImageDataExtractor for electron microscopy image analysis.

References

- (1) Mukaddem, K. T. et al. ImageDataExtractor: A Tool to Extract and Quantify Data from Microscopy Images. *J. Chem. Inf. Model.* **2020**, *60*(5), 2492–2509. DOI: 10.1021/acs.jcim.9b00734
- (2) Swain, M. C.; Cole, J. M. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. *J. Chem. Inf. Model.* **2016**, *56*(10), 1894–1904. DOI: 10.1021/acs.jcim.6b00207
- (3) Mavračić, J. et al. ChemDataExtractor 2.0: Autopopulated Ontologies for Materials Science. *J. Chem. Inf. Model.* **2021**, *61*(9), 4280–4289. DOI: 10.1021/acs.jcim.1c00446
- (4) Beard, E. J.; Cole, J. M. ChemSchematicResolver: A Toolkit to Decode 2D Chemical Diagrams with Labels and R-Groups into Annotated Chemical Named Entities. *J. Chem. Inf. Model.* **2020**, *60*(4), 2059–2072. DOI: 10.1021/acs.jcim.0c00042
- (5) Wilary, D. M.; Cole, J. M. ReactionDataExtractor: A Tool for Automated Extraction of Information from Chemical Reaction Schemes. *J. Chem. Inf. Model.* **2021**, *61*(10), 4962–4974. DOI: 10.1021/acs.jcim.1c01017
- (6) Wilary, D. M.; Cole, J. M. ReactionDataExtractor 2.0: A Deep Learning Approach for Data Extraction from Chemical Reaction Schemes. *J. Chem. Inf. Model.* **2023** (in press). DOI: 10.1021/acs.jcim.3c00422
- (7) Beard, E.J. et al. Comparative dataset of experimental and computational attributes of UV/vis absorption spectra. *Sci. Data.* **2019**, *6*, 307. DOI: 10.1038/s41597-019-0306-0

- (8) Beard, E.J.; Cole, J. M. Perovskite- and Dye-Sensitized Solar-Cell Device Databases Auto-generated Using ChemDataExtractor. *Sci. Data.* **2022**, *9*, 329. DOI: 10.1038/s41597-022-01355-w
- (9) Huang, S.; Cole, J. M. BatteryBERT: A Pretrained Language Model for Battery Database Enhancement. *J. Chem. Inf. Model.* **2022**, *62*(24), 6365–6377. DOI: 10.1021/acs.jcim.2c00035
- (10) Huang, S.; Cole, J. M. BatteryDataExtractor: battery-aware text-mining software embedded with BERT models. *Chem. Sci.* **2022**, *13*, 11487-11495. DOI: 10.1039/D2SC04322J
- (11) Rajan, K. et al. DECIMER: towards deep learning for chemical image recognition. *J. Cheminform.* **2020**, *12*, 65. DOI: 10.1186/s13321-020-00469-w

Protein-ligand co-folding: predicting the structure of the protein-ligand complex from the protein sequence and SMILES string of the ligand

Laksh Aithani, Charm Therapeutics, UK

The second day of the conference commenced with Laksh Aithani introducing recent advancements in the field of protein-folding prediction, in-particular highlighting AlphaFold¹ and RosettaFold² as examples. He underscored a significant challenge not yet addressed by these methods, which is the ability to account for the presence of ligands and their impact on protein conformation. Addressing this challenge is crucial for advancing the field of protein folding. This prompted the development of DragonFold by Charm Therapeutics, a novel protein-ligand co-folding model. DragonFold considers both the protein sequence and the 2D ligand topology (via SMILES), folding them together in a coordinated end-to-end manner. It leverages both publicly available data and proprietary in-house data sources. DragonFold shows promise in achieving higher accuracy than traditional docking methods, while also being capable of accommodating induced fit to both protein backbone and side chains.

He then shared how a compute cluster of 128 NVIDIA A100 GPUs are utilised to train DragonFold, which makes use of data from the Protein Data Bank (PDB), with more than 25% of structures containing a co-crystal ligand. Laksh then transitioned to the evaluation of DragonFold's performance. He noted that while time-split evaluation is common in the community, it may not be sufficiently rigorous. Key benchmarks for performance comparison included Glide,³ DiffDock,⁴ and DiffDock-ESMFold.⁵ Notably, DragonFold sets itself apart by only requiring the protein sequence representation, as opposed to DiffDock, which necessitates the entire 3D protein structure. The test set in this evaluation comprised 250 structures extracted from the EquiBind⁶ test split, a common benchmark dataset, and the metrics employed were ligand root mean squared deviation (RMSD) from ground truth. The results Laksh showcased highlighted DragonFold's performance, with 24% of predicted ligand poses achieving an <1Å RMSD compared to Glide (20%), DiffDock (19%), and DiffDock-ESMFold (8%). Laksh reiterated that this is particularly noteworthy considering that DragonFold relies only on protein sequence as input. When the threshold was relaxed to <2Å RMSD, DiffDock with a holo protein structure outperformed DragonFold achieving 48% of predictions within 2Å RMSD compared to DragonFold's 41%. Note that this approach to measuring model performance is not without limitation as later discussed by Martin Buttenschoen.

However, Laksh emphasised that there is a contribution of protein folding accuracy in this evaluation (only DragonFold and DiffDock-ESM start from protein sequence), revealing that DragonFold predicted 191/250 protein folds with high Global Distance Test (GDT) scores. For example, when only considering protein structures where AlphaFold performs well (GDT>0.9), DragonFold's ligand RMSD performance improved (28% with ligands <1Å RMSD). He further demonstrated similar results with DragonFold's performance: performance on protein subsets with higher GDT scores (0.8, 0.85, and 0.9) incrementally increased, eventually reaching 38% <1Å with a GDT >0.9. Laksh also discussed the use of confidence scores, which correlate with ligand RMSD accuracy (i.e., low confidence indicated an incorrect structure, while high confidence correlated with high accuracy), showcasing an advantage over methods like Glide.

Transitioning to case studies, Laksh presented examples involving TEAD inhibitors⁷ and WDR91.⁸ The former exhibited successful redocking, while the latter struggled to accurately recreate the binding pose – therefore, performance is still very case dependent. DragonFold was then highlighted for its additional utility in conducting kinome scans (i.e., identifying potential off-targets) by observing a significant overlap of folding confidence with known off-target proteins. Moreover, DragonFold was used for virtual screening, identifying micromolar hits through the co-folding of hundreds of compounds from a commercial library. It additionally identified selective inhibitors by screening a commercial library against three kinases and selecting compounds that exhibited confident folding only on one kinase, 170 compounds were screened identifying hits that selectively inhibited only the correct kinase.

References

- (1) Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, 596, 583–589. DOI: 10.1038/s41586-021-03819-2
- (2) Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, 373, 871-876. DOI: 10.1126/science.abj8754
- (3) Friesner, R. A. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, 47(7), 1739–1749. DOI: 10.1021/jm0306430
- (4) Corso, G. et al. DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking. *arXiv*. **2022**. DOI: 10.48550/arXiv.2210.01776
- (5) Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*. **2023**, 379, 1123-1130. DOI: 10.1126/science.ade2574
- (6) Stärk, H. et al. EquiBind: Geometric Deep Learning for Drug Binding Structure Prediction. *arXiv*. **2022**. DOI: 10.48550/arXiv.2202.05146
- (7) Hagenbeek, T.J. et al. An allosteric pan-TEAD inhibitor blocks oncogenic YAP/TAZ signaling and overcomes KRAS G12C inhibitor resistance. *Nat. Cancer*. **2023**, 4, 812–828. DOI: 10.1038/s43018-023-00577-0
- (8) Ahmad, S. et al. Discovery of a first-in-class small molecule ligand for WDR91 using DNA-encoded chemical library selection followed by machine learning. *bioRxiv*. **2023**. DOI: 10.1101/2023.08.21.552681

Recent advancements in DECIMER.ai and automated mining from scientific literature in COCONUT 2

Kohulan Rajan, Friedrich Schiller University, Germany

Kohulan discussed the presence of data issues in the COCONUT¹ database, noting that many of the underlying databases are no longer available online. This inspired the DECIMER² project to undertake the task of capturing 2D chemical structures and converting them into a machine-learning-readable format in order to automatically mine data and check for completeness. While alternative optical recognition models often rely on rule-based methods, DECIMER (Deep Learning for Chemical Image Recognition) stands out as it focuses on an image classification task, converting images of chemical structures into SMILES strings.

The DECIMER workflow encompasses several components:

1. Segmentation tool:³ This tool was trained using 1820 manually annotated data points to identify segments in chemical structures. However, Kohulan acknowledged that incorrect segments could sometimes be identified. The goal was to separate different structures into distinct classes.
2. Image Classifier: The image classifier determines whether a segmented image represents a chemical structure or not.
3. Image Transformer:⁴ This component encodes an image of a 2D chemical structure into a SMILES string using an EfficientNet-V2 Encoder CNN with a Transformer Decoder DNN. Training this model posed a challenge, particularly with regards to obtaining sufficient training data.

4. RanDepict:⁵ RanDepict employs various tools to generate different representations of chemical structures, including Markush structures. This approach facilitated the curation of training data for the image Transformer.

Kohulan highlighted an intriguing discovery made during the project: the Transformer could effectively interpret hand-drawn chemical structures without explicit training, achieving an average Tanimoto similarity coefficient of 0.71. To further enhance this capability, synthetic hand-drawn images were embedded into RanDepict to expand the curated training dataset.

Training of the DECIMER model occurred on Google TPUs, with each training epoch taking less than 10 hours. Performance evaluation revealed that DECIMER was state-of-the-art, with approximately 70% identical predictions. Extensive benchmarking was conducted against external datasets. Moreover, testing on augmented data did not impact the image Transformer's performance, an advantage over rule-based methods for which this test negatively impacted performance.

Another challenge posed was how to extract the exact orientation of a chemical structure. One potential approach explored was using CXSMILES, which showed some promise. However, there was also a possibility to use a compressed form⁶ which Kohulan is currently investigating.

Moreover, the DECIMER approach can be used to provide a confidence score for each particular SMILES token to interpret its overall prediction.

Kohulan concluded by shifting the focus back to COCONUT and introduced COCONUT v2, an ongoing project aimed at revamping COCONUT with a new user interface. The project also involved the creation of depictions using the Cheminformatics Python Microservice pipeline⁷ with extended citation capabilities. Additionally, user submission interfaces and APIs were in development as part of this initiative. Lastly, [DECIMER.ai](#)⁸ is accessible to test with a user uploaded PDF.

References

- (1) Sorokina, M. et al. COCONUT online: Collection of Open Natural Products database. *J. Cheminform.* **2021**, *13*, 2. DOI: 10.1186/s13321-020-00478-9
- (2) Rajan, K. et al. DECIMER: towards deep learning for chemical image recognition. *J. Cheminform.* **2020**, *12*, 65. DOI: 10.1186/s13321-020-00469-w
- (3) Rajan, K. et al. DECIMER-Segmentation: Automated extraction of chemical structure depictions from scientific literature. *J. Cheminform.* **2021**, *13*, 20. DOI: 10.1186/s13321-021-00496-1
- (4) Rajan, K. et al. DECIMER 1.0: deep learning for chemical image recognition using transformers. *J. Cheminform.* **2021**, *13*, 61. DOI: 10.1186/s13321-021-00538-8
- (5) Brinkhaus, H.O. et al. RanDepict: Random chemical structure depiction generator. *J. Cheminform.* **2022**, *14*, 31. DOI: doi.org/10.1186/s13321-022-00609-4
- (6) Mayfield, J. Data Compression of INCHIKEYS and 2D Coordinates. *NIH Virtual Workshop on InChI*. **2021**. URL: https://www.nextmovesoftware.com/talks/Mayfield_DataCompressionOfInChIKeysAnd2dCoordinates_NIHINCHI_202103.pdf
- (7) Chandrasekhar, V. et al. Cheminformatics Microservice: unifying access to open cheminformatics toolkits. (8) *ChemRxiv*. **2023**. DOI: 10.26434/chemrxiv-2023-hk8zn-v2
- (8) Rajan, K., et al. DECIMER.ai: an open platform for automated optical chemical structure identification, segmentation and recognition in scientific publications. *Nat. Commun.* **2023**, *14*, 5045. DOI: 10.1038/s41467-023-40782-0

PoseBusters: An evaluation of the physical plausibility of the prediction of deep learning-based docking methods

Martin Buttenschoen, University of Oxford, UK

Martin initiated his presentation by sharing the primary motivation behind his work, which predominantly revolved around the use of Root Mean Square Deviation (RMSD) as an evaluation metric for *de novo* docking deep learning models. However, he highlighted a crucial limitation of RMSD is that it fails to capture unusual geometry or constrained ligands effectively. To address this limitation, Martin introduced PoseBusters,¹ an open-source Python package that utilises RDKit to enhance the evaluation of 3D generative models for pose prediction. PoseBusters assesses various aspects of chemical validity and consistency, including sanitisation, molecular formula, bonds, intramolecular validity (e.g., bond lengths, bond angles, internal clashes, and energy), and intermolecular validity (e.g., protein-ligand clashes).

To delve deeper into PoseBusters' capabilities, Martin described how they conducted re-docking experiments using the Astex Diverse set.² They benchmarked several docking tools, including Vina,³ GOLD,⁴ DeepDock,⁵ Uni-Mol,⁶ TANKBind,⁷ DiffDock,⁸ and Equibind.⁹ Standard docking protocols were applied, with some using a predefined search space, while others, particularly deep learning based (DL-based) methods, took a nearby surrounding area or the whole structure (i.e., blind docking).

When considering RMSD alone, DiffDock outperformed all other methods with 72% of predictions achieving $<2\text{\AA}$ RMSD to the ground truth pose. However, when assessing chemical validity, this performance dropped to just 47% and Gold emerged as the top performer with 64% achieving $<2\text{\AA}$ RMSD. Notably, TANKBind's performance dropped significantly from 59% to 5.9%. This revealed that good RMSD values alone were not predictive of plausibility, and that DL-based methods had not yet surpassed traditional docking methods. Furthermore, Martin highlighted that DL-based methods were generally trained on similar datasets. PoseBusters therefore introduced a challenging new test set of 328 structures published after 2021; results on this test set were significantly worse for all DL-based methods. Conversely, classical methods were only marginally affected. For instance, DiffDock (most performant DL-based method) dropped from 47% to 14%, meanwhile Gold only dropped from 64% to 48% (most performant classical method).

Martin elucidated that common failure modes in the DL-based methods were related to issues with bond lengths and protein-ligand clashes. Some DL-based methods predicted excessively long bonds, and all DL-based methods exhibited steric clashes between the protein and ligand. In contrast, classical methods showed fewer instances of these issues, reaffirming that DL-based methods had not yet surpassed classical approaches.

Overfitting was another concern with regards to DL-based models. Martin and his team measured sequence identity and divided the dataset into three buckets: 0-30, 30-95, and 95-100 sequence identity. This had minimal impact on classical methods but significantly affected the performance of DL-based methods, which performed poorly when dealing with more dissimilar sequences than present in the training data. For example, DiffDock resulted in $<2\text{\AA}$ RMSD values of 0.78%, 12%, and 24% compared to GOLD with values of 46%, 56%, and 47% at different sequence identities respectively.

Finally, Martin concluded that DL-based methods clearly do not learn some physics-based principles. However, by minimising the ligands using forcefields like AMBER, while keeping the protein fixed, some performance is rescued.

References

(1) Buttenschoen, M. et al. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *arXiv*. 2023. DOI: 10.48550/arXiv.2308.05777

- (2) Hartshorn, M. J. et al. Diverse, High-Quality Test Set for the Validation of Protein–Ligand Docking Performance. *J. Med. Chem.* **2007**, *50*(4), 726–741. DOI: 10.1021/jm061277y
- (3) Trott, O. et al. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **2010**, *31*(2), 455–461. DOI: 10.1002/jcc.21334
- (4) Verdonk, M. L. et al. Improved protein–ligand docking using GOLD. *J. Comput. Chem.* **2003**, *52*(4), 609–623. DOI: 10.1002/prot.10465
- (5) Isert, C. et al. Structure-based drug design with geometric deep learning. *Curr. Opin. Struct. Biol.* **2023**, *79*, 102548. DOI: 10.1016/j.sbi.2023.102548
- (6) Zhou, G. et al. Uni-Mol: A Universal 3D Molecular Representation Learning Framework. *ICLR*. **2023**.
- (7) Lu, W. et al. TANKBind: Trigonometry-Aware Neural Networks for Drug-Protein Binding Structure Prediction. *bioRxiv*. **2022**. DOI: 10.1101/2022.06.06.495043
- (8) Corse, G. et al. DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking. *arXiv*. **2022**. DOI: 10.48550/arXiv.2210.01776
- (9) Stärk, H. EquiBind: Geometric Deep Learning for Drug Binding Structure Prediction. *arXiv*. **2022**. DOI: 10.48550/arXiv.2202.05146

IMPRESSION – Graph Transformer Network for the prediction of NMR parameters

Chun Yee Calvin Yu, University of Bristol, UK

Calvin's presentation commenced with an introduction to IMPRESSION (Intelligent Machine PRedicting Shift Scalar Information Of Nuclei), a machine learning model designed to address the forward prediction challenge of chemical structure to atomic NMR chemical shifts. This prediction is valuable for validating the structure and conformation of final products in chemical reactions. Traditional methods involve running computationally expensive Density Functional Theory (DFT) calculations and matching the results to NMR shifts, however, employing machine learning could significantly accelerate this process.

Calvin started by explaining that some initial work had been carried out by Will Gerrard¹ which involved a kernel ridge regression model to predict chemical shifts. The model achieved relatively high accuracy with prediction deviation within 0.23 ppm for ¹H and 2.45 ppm for ¹³C. However, a separate model was required for each parameter (i.e., one for ¹H and another for ¹³C). Furthermore, it was constrained by a limited applicability domain and wasn't very scalable.

In an attempt to overcome these limitations, Calvin's team adopted a graph transformer network to predict multiple NMR parameters simultaneously for a much broader range of atoms, as well as for improved scaling. In the graph representation, a fully connected graph was employed, where attention was applied. Nodes represented atom types and edges interatomic distance, bond distance and coupling type. Importantly, this representation is invariant to rotation and translation. Notably, different attention layers were used for different atom types and parameters. Diverse data (shown by chemical space projection complementarity) were sampled from various online databases such as CSD² and ChEMBL.³ Then RDKit conformers were generated and DFT (mPW9PW91 for geometry optimisation and wb97X-D for tensor calculation which is scaled to chemical shift)⁴ calculations were performed to obtain the training labels, before splitting the data into train and test sets.

Calvin moved on to results showing that the most recent model achieved remarkable accuracy, with a deviation of 0.09 ppm for ¹H and 1.01 ppm for ¹³C, equivalent to a 1% error range. This level of accuracy was theoretically sufficient to discern conformational differences. To measure confidence, 5-fold cross-validation was utilised to map variance, which correlated with the absolute error. This analysis highlighted that allenes were frequently mispredicted, which Calvin suggested could be due to bond order, which was not incorporated as an input feature into the model.

Calvin also mentioned the transferability of the model to predict different properties, for example, molecular charge. This also achieved high performance, demonstrating a low Mean Absolute Error (MAE) of 0.003 for H and 0.011 for C, along with a strong correlation to true values.

An exciting aspect discussed was the inverse problem: given an NMR spectrum, can the correct chemical structure be generated? This is the current focus of the research group, and preliminary results using a Generative Variational Autoencoder (GVAE) showed that it currently works only 2% of the time. In conclusion, Calvin emphasised that the machine learning model they developed could reasonably predict chemical shifts with confidence as well as be used to predict other atomic properties.

References

- (1) Gerard, W. et al. IMPRESSION – prediction of NMR parameters for 3-dimensional chemical structures using machine learning with near quantum chemical accuracy. *Chem. Sci.* **2020**, *11*, 508-515. DOI: 10.1039/C9SC03854J
- (2) Gaulton, A. et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*(D1), D1100-D1107. DOI: 10.1093/nar/gkr777
- (3) Groom, C.R. et al. The Cambridge Structural Database. *Acta Crystallogr. B.* **2016**, *72*, 171-179. DOI: 10.1107/S2052520616003954
- (4) Guan, Y. et al. Real-time prediction of ¹H and ¹³C chemical shifts with DFT accuracy using a 3D graph neural network. *Chem. Sci.* **2021**, *12*, 12012-12026. DOI: 10.1039/D1SC03343C

Industrial applications of few-shot learning in early drug discovery

Lukas Friedrich, Merck Healthcare KgaA, Germany

Lukas initiated his presentation by revisiting the drug discovery pipeline and emphasising the importance of Quantitative Structure-Activity Relationship (QSAR) modelling in identifying lead compounds. He highlighted that as the drug discovery process progresses, less data is available due to the increased cost of experimental assays. Therefore, from an *in silico* perspective, it's crucial to predict the bioactivity against a target as accurately and especially as early as possible. Lukas raised the question of which methods could be applied at different stages of drug discovery. Examples included physics-based methods like docking that could be employed when the protein structure is available, similarity searches that can be used to find analogues for hit identification, or machine learning and deep learning techniques that can be utilised, particularly in lead optimisation when a substantial amount of data is available. However, Lukas posed whether there exists a method that relies on only a few data points, allowing its use even earlier in hit identification?

He proceeded to describe the concept of 'few-shot learning', explaining that it involves query samples and a support set (which comprises a small dataset of additional examples – none of which are contained in the training dataset). The approach involves the model learning a similarity function and then applying this function to the support set to determine which elements in the support set are similar to the query. Transferring this approach to drug discovery, a training set of molecular structures is used with different tasks, such as binary classification. A query sample is introduced, representing a task that was not part of the training dataset, alongside a support set of which the learned similarity function is applied, ultimately providing a prediction label for the unknown task.

Lukas then introduced MHNfs¹ (molecule embedding for few-shot learning), the model used in the work presented. This model consists of several components, including an encoder to embed molecular fingerprints, a context module (consisting of a Hopfield layer) to enrich information and relationships between the support and query, an attention model for information exchange, and a similarity module that calculates a weighted mean over similarity values of the support set to the query.

The MHNfs model was benchmarked using the FSMol² dataset. The results demonstrated that it achieved state-of-the-art performance in some tasks compared to many other methods. An ablation study emphasised the importance of the context module and cross-attention module on performance, meanwhile a dataset domain shift to the Tox21 dataset³ did not impact performance as much as other methods (although Lukas noted that the performance was still not sufficient for production-level use).

In the context of industry, traditional approaches such as similarity search, property prediction, and docking are more routinely used and are therefore more appropriate baselines. When given only eight active and eight inactive molecules, few-shot learning outperformed a Random Forest model (trained on the same 16 molecules) and similarity search methods. These results provided confidence for Lukas and his team to use few-shot learning to enrich screening in projects that aren't structurally enabled. Moreover, Lukas and his team are currently implementing this approach in the open-source community Hugging Face.⁴

References

- (1) Schimunek, J. et al. Context-enriched molecule representations improve few-shot drug discovery. *arXiv*. 2023. DOI: 10.48550/arXiv.2305.09481
- (2) Stanley, M. et al. FS-Mol: A Few-Shot Learning Dataset of Molecules. *NIPS*. 2021.
- (3) Richard, A. M. et al. The Tox21 10K Compound Library: Collaborative Chemistry Advancing Toxicology. *Chem. Res. Toxicol.* 2021, 34(2), 189–216. DOI: 10.1021/acs.chemrestox.0c00264
- (4) Hugging Face home page. <https://huggingface.co/> (accessed 18/09/23)

Artificial Intelligence applied to drug discovery

Noor Shaker, Glamorous AI and X-Chem, UK



Noor Shaker presenting at the Symposium.

Noor started by introducing X-Chem, a company specialising in DNA-encoded library screening (with over 100 DNA-encoded libraries and 12 clinical candidates) that acquired GlamorousAI to integrate artificial intelligence to enhance the DNA-encoded library (DEL) platform via ArtemisAI. ArtemisAI serves as an interface to process, analyse, predict, and generate chemical data. A key emphasis of ArtemisAI was the

elimination of human bias from decision-making. Molecular representations and models are meticulously investigated, and the best-performing models are selected based on objective criteria rather than following 'fashionable' trends. This approach results in a collection of 5-10 models that form an ensemble to provide the most accurate answers with confidence estimates.

Noor then described how DEL can be utilised to screen billions (7.5B) of compounds in a single test tube, yielding numerous hits (up to 100M). Here machine learning (ML) comes into play, using the binding data to train the ensemble of models. These models can subsequently be employed for screening other public or proprietary databases.

Moving on to some case studies, Noor described that three years ago, X-Chem collaborated with Google on a DEL screen. This work was replicated using AI, and 59 predicted compounds were purchased, with approximately 50% turning out to be active (Noor presented the diversity and novelty of chemical space achieved). To conduct a similar evaluation, X-Chem searched the literature to identify hits against a target which were then randomised into Enamine (decoys) to measure retrieval of these hits in a retrospective manner. The use of ML models for this task resulted in a 63.5% enrichment rate. Similar results were found for other protein targets, with enrichment values ranging from 0 to 811-fold, including HPK1, which showed an enrichment of 501-fold. However, Noor stated that the success of this approach can heavily depend on the chemical space overlap between the training data and its relevance to the task.

Noor transitioned to the broader application of ML in predicting other chemical properties, mentioning the Therapeutics Data Commons (TDC)¹ challenges. The results showed near state-of-the-art performance, excelling in some properties while lagging slightly behind in others. It was noted that literature models tend to struggle with generalisation in an industrial setting. However, the automated processes of ArtemisAI allows data to select the most appropriate models.

The presentation moved on to generative models for designing KEAP1 modulators. In this case study, a distinct binding pocket was treated as a zero-shot problem, with only one fragment to expand within the pocket. Generative models proposed billions of compounds within 2-3 weeks, which were then filtered to a more manageable number. Thousands of these compounds were screened using physics-based models like docking, NN scores, and FEP calculations, leading to the selection of tens of final compounds. Three compounds were ultimately selected, and their closest analogues found in ZINC² were tested, resulting in the identification of real binders.

Noor concluded by providing an example of the ArtemisAI user interface, demonstrating how a chemical structure can be uploaded and various properties of interest can be predicted or used to guide *de novo* design for proposing new molecules.

References

- (1) Huang, K. et al. Artificial intelligence foundation for therapeutic science. *Nat. Chem. Biol.* **2022**, *18*, 1033–1036. DOI: 10.1038/s41589-022-01131-2
- (2) Irwin, J. J. et al. ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *J. Chem. Inf. Model.* **2020**, *60*(12), 6065–6073. DOI: 10.1021/acs.jcim.0c00675

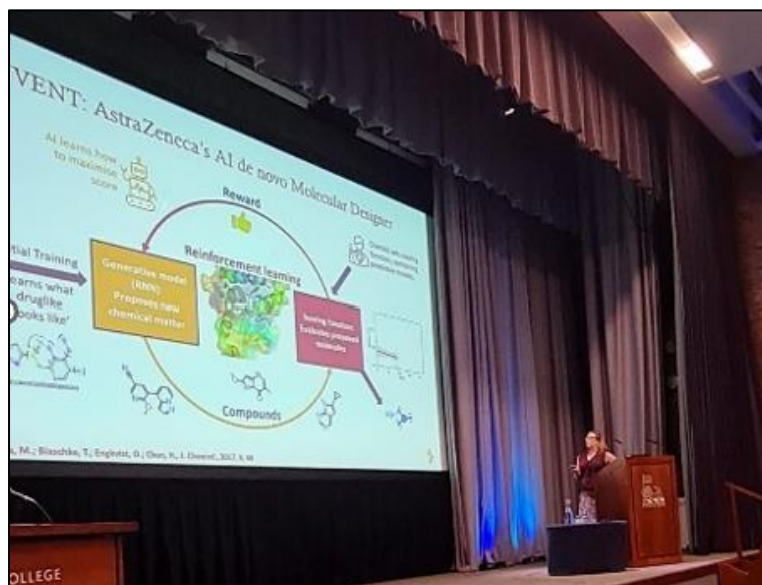
De novo design using generative AI methods: application to the discovery of new compounds for oncology projects

Kathryn Giblin, AstraZeneca, UK

Kathryn began by outlining that this presentation would discuss REINVENT^{1,2} (a language-based generative model), scaffold hopping, LibINVENT³ (a library-focussed modification to REINVENT), and other developments in the field. REINVENT was described as a recurrent neural network pre-trained on a dataset of compounds in SMILES format, including public domain and AstraZeneca (AZ) data. This pre-trained model is used to propose *de novo* chemical structures that undergo iterative optimisation via reinforcement learning, as *de novo* molecules are evaluated by scoring functions. This can be described as a pseudo actor-critic method.

Kathryn then shared how AZ has been using REINVENT internally for approximately four years for both exploration and exploitation of chemical space. Exploration being analogous to scaffold hopping often utilising 3D structure-based scoring functions meanwhile exploitation is more analogous to lead optimisation, emphasising small conservative changes to chemical structure. In this time, REINVENT has been applied, in

close collaboration with medicinal chemistry, to over 75% of projects at AstraZeneca and has contributed 11 novel scaffolds to various projects. Moreover, REINVENT has evolved into a flexible tool-based platform, including LibINVENT,³ LinkINVENT,⁴ and a Transformer⁵ for lead optimisation. For example, LibINVENT can be used to expand specific vectors from a scaffold, LinkINVENT for core hopping and designing linkers between fragments (ideal for PROTAC design), and Transformers for small property-based structural changes in lead optimisation.



Kathryn Giblin presenting at the Symposium.

Kathryn then delved into scaffold hopping and its implementation at AZ using REINVENT. Traditional methods involve virtual screening (VS) and exhaustive screening of a library, while generative models implicitly explore theoretical chemical space. The challenge lies in optimising the search in this implicit chemical space, which can be non-trivial. Transfer learning accelerates optimisation, reducing the process from requiring around 900 steps to 200 steps, albeit at the cost of limiting chemical space. Kathryn illustrated these concepts by presenting two case studies. In the first study from 2019, REINVENT was used to identify a backup series for a kinase. Transfer learning was

used to train a kinase-specific model and kinase-general model (that was more diverse). Reinforcement learning with QSAR scoring functions were employed, leading to the discovery of several scaffolds: 1) a known active scaffold that could be optimised to a pIC₅₀ value of 8.3, 2) a scaffold with an unprecedented kinase hinge binding motif, however, synthesis proved difficult, and 3) a new active scaffold that was further optimised but did not translate to the desired phenotype (the latter was not modelled during reinforcement learning).

In the second case study, LibINVENT was used to modify a predefined scaffold, significantly speeding up the optimisation process to requiring fewer than 100 steps. LibINVENT is an encoder-decoder model that takes an input scaffold and proposes a modified final product, leveraging knowledge from reaction types for synthetic feasibility. The goal was to modify a known Cbl-b modulator at a synthetically accessible reaction vector (with the caveat of steep structure-activity relationships) to identify a distinct chemical series. The first cycle optimised the ROCS⁶ shape similarity, QED,⁷ and molecular weight of *de novo* molecules. This resulted in the synthesis of two molecules with pIC₅₀'s of 5.18 and 5.92. The second cycle, which included free-energy perturbation scoring, discovered two more molecules with different scaffolds. Finally, a third cycle discovered a third chemical series. Molecular dynamics was used to rank design proposals resulting in more potent selections. The end result was lead molecules with activities of 170 nM and 7 nM. Overall, 23 compounds were made and there was a 600-fold improvement in potency during design iterations. Kathryn noted the importance of several design iterations per cycle to acquire and implement knowledge.

Shifting to discuss more recent developments, Kathryn showcased how LinkINVENT can be used to connect two warheads using a similar encoder-decoder model followed by property optimisation by a Transformer trained on matched molecular pairs. She further emphasised that different runs under reinforcement learning can be divergent, and so beam search is employed to explore the surrounding chemical space. Looking ahead,

Kathryn highlighted the need for smarter and faster approaches in generative AI, particularly in the scoring process.

References

- (1) Olivecrona, M. et al. Molecular de-novo design through deep reinforcement learning. *J. Cheminform.* **2017**, *9*, 48. DOI: 10.1186/s13321-017-0235-x
- (2) Blashcke, T., et al. REINVENT 2.0: An AI Tool for De Novo Drug Design. *J. Chem. Inf. Model.* **2020**, *60*(12), 5918–5922. DOI: 10.1021/acs.jcim.0c00915
- (3) Fialková, F. et al. LibINVENT: Reaction-based Generative Scaffold Decoration for in Silico Library Design. *J. Chem. Inf. Model.* **2022**, *62*(9), 2046–2063. DOI: 10.1021/acs.jcim.1c00469
- (4) Guo, J. et al. Link-INVENT: generative linker design with reinforcement learning. *Digital Discovery.* **2023**, *2*, 392–408. DOI: 10.1039/D2DD00115B
- (5) He, J. et al. Molecular optimization by capturing chemist's intuition using deep neural networks. *J. Cheminform.* **2016**, *13*, 26. DOI: 10.1186/s13321-021-00497-0
- (6) Grant, J. A. et al. A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape. *J. Comput. Chem.* **1996**, *17*(4), 1653–1666. DOI: 10.1002/(SICI)1096-987X(19961115)17:14%3C1653::AID-JCC7%3E3.0.CO;2-K
- (7) Bickerton, G. et al. Quantifying the chemical beauty of drugs. *Nature Chem.* **2021**, *4*, 90–98. DOI: 10.1038/nchem.1243

NVIDIA BioNeMo: A framework and service for generative AI in drug discovery (Keynote 4)

Michelle Gill, NVIDIA, USA

Michelle presented the fourth keynote talk and final talk of the conference. She initiated her presentation by highlighting how language models are revolutionising drug discovery in various areas such as biomedical natural language processing, generative chemistry, protein structure prediction, and virtual screening. She emphasised the significance of symmetric-aware representations in addressing biological problems (and cited work by Ferruz et al.¹ as an example). Michelle also mentioned the success of combining language and symmetry, for example, as done in the RosettaFold² model. However, Michelle stated that models often have a finite lifespan due to the rate of development, and that the biggest value is in the learnings achievable. Therefore, it is important to develop internal research that is scalable and accelerates different factors such as by tackling efficiency of new models by addressing bottlenecks, which further drives software and even hardware design. This is true also for the work Michelle conducts on BioNeMo including MolMIM³ and DiffDock⁴ optimisation.

Michelle then introduced BioNeMo, an AI platform at NVIDIA to help accelerate drug discovery. This two-part platform consists of an inter-connected inference service and underlying framework.

The inference service collates and implements models from the literature that altogether offer protein and molecular representations, folding, generation, and docking capabilities available in the cloud. Michelle clarified that no inference model is implemented exactly as-is, usually performance or efficiency is tweaked to try and achieve improvements before deployment, also fostering learning within the team for future projects. A Python interface is also available, making it accessible through requests or pip installation. Michelle then illustrated the landing page and user interface showing how it can be used to conduct several tasks (such as, running structure folding prediction models from sequence, docking, or molecule generation) or to extract the corresponding Python code to be run from the Python API.

With regards to the BioNeMe framework, it contains capabilities for data processing, model training, and fine-tuning. It includes various large language models (LLMs) and equivariant models like MolMIM and DiffDock. Michelle briefly described that NeMo is the technology stack with associated library for accelerating the development of LLMs on multiple GPUs. This stack allows development from the application layer down to

hardware. She highlighted the use of the BioNeMo framework to design a variational autoencoder called ProT-VAE for generating modified proteins with enhanced functionality.

Michelle then moved on to examples of model development, implementation, and optimisation. Firstly MolMIM,³ a small-molecule foundation model consisting of representation and translation models. MegaMolBART was the first molecular representation model, built in collaboration with AstraZeneca to generate molecule embeddings.⁵ Although not too useful for molecular generation, it has utility but for translation tasks and molecular representation. A main caveat was that the decoding time increased as it generated very long molecules. To address variable sequence length a perceiver model was used to down-sample the hidden dimension⁶ which reduced runtime complexity to $\mathcal{O}(Sk + k^2)$ from $\mathcal{O}(S^2)$. However, this did not solve the organisation of the latent space. Therefore, a mutual information machine that maximises mutual information and minimises marginal entropy was implemented. This clustered the latent space as demonstrated by a dimensionality reduction plot Michelle presented. This new model, MolMIM, was then compared to MegaMolBART and CDDD.⁷ It showed a significant reduction in test runtime with high effective novelty, validity, novelty, and uniqueness. Looking at the latent space qualitatively, small distances resulted in small molecular structure perturbations, while larger distances resulted in larger chemical perturbations – as one would expect. This was also shown quantitatively as Tanimoto similarity correlated well with interpolation steps in the latent space. Next, to demonstrate *de novo* molecule generation and optimisation, a simple covariance matrix adaptation evolutionary strategy was implemented to optimise QED⁸ and Penalised LogP⁹ constrained by Tanimoto similarity. This resulted in state-of-the-art performance compared to a small selection of alternative generative models. However, they observed that an exploitation hack is sometimes used involving a repeated use of particular functional groups. Repeating this with multiple objectives (i.e., QED,⁸ SAScore,¹⁰ JNK3¹¹ and GSK3 β ¹¹) also resulted in comparable performance to other generative models. Now, MolMIM is undergoing productionisation which is a different but equally important challenge.

The second case study involved the acceleration of DiffDock.⁴ Michelle shared how they conducted GPU specific optimisation by reducing numerical precision to TensorFloat32 (TF32). As opposed to the more standard FloatingPoint32 (FP32), TF32 decreases the precision without negatively affecting the available dynamic range like FP32. Note TF32 is currently only available on NVIDIA A100 GPUs. This already accelerated the DiffDock model by 1.8-fold without impacting numerical stability or performance. Moreover, Michelle stated they are also working to improve tensor operations required to achieve symmetry in the equivariant network (using the e3nn library¹²) by utilising CUDA parallelism.

Michelle concluded by outlining future interests in molecular dynamics assisted refinement of docked poses, as well as machine learning potentials. Lastly, the [BioNeMo](#) will enter open beta soon and is open to enrolment.

References

- (1) Ferruz, N. et al. ProtGPT2 is a deep unsupervised language model for protein design. *Nat. Commun.* **2022**, *13*, 4348. DOI: 10.1038/s41467-022-32007-7
- (2) Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*. **2021**, *373*, 871-876. DOI:10.1126/science.abj8754
- (3) Reidenback, D. et al. Improving Small Molecule Generation using Mutual Information Machine. *arXiv*. **2022**. DOI: 10.48550/arXiv.2208.09016
- (4) Corse, G. et al. DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking. *arXiv*. **2022**. DOI: 10.48550/arXiv.2210.01776
- (5) Irwin, R. et al. Chemformer: a pre-trained transformer for computational chemistry. *Mach. Learn.: Sci. Technol.* **2022**, *3*, 015022. DOI: 10.1088/2632-2153/ac3ffb
- (6) Jaegle, A. et al. Perceiver: General Perception with Iterative Attention. *arXiv*. **2021**. DOI: 10.48550/arXiv.2103.03206

- (7) Winter, R. et al. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chem. Sci.* **2019**, *10*, 1692-1701. DOI: 10.1039/C8SC04175J
- (8) Bickerton, G. et al. Quantifying the chemical beauty of drugs. *Nature Chem.* **2021**, *4*, 90–98. DOI: 10.1038/nchem.1243
- (9) Jin, W. et al. Junction tree variational autoencoder for molecular graph generation. *arXiv.* **2018**. DOI: /10.48550/arXiv.1802.04364
- (10) Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform.* **2009**, *1*, 8. DOI: 10.1186/1758-2946-1-8
- (11) Li, Y. et al. Multi-objective de novo drug design with conditional graph generative model. *J. Cheminform.* **2018**, *10*, 33. DOI: 10.1186/s13321-018-0287-6
- (12) Geiger, M. et al. Euclidean neural networks: e3nn. DOI: 10.5281/zenodo.6459381

Poster prizes

The **Industry prize** sponsored by the *Reaction Chemistry and Engineering* journal was awarded to **Jon Paul Janet** for his poster titled “Active learning for reinforcement learning based de novo drug design”

The **Academia prize** sponsored by both the *Molecular Systems Design & Engineering* (MSDE) and *Physical Chemistry Chemical Physics* (PCCP) journals was awarded to **Guy Durant** for his poster titled “Robustly interrogating machine learning scoring functions: do they learn any biophysics?”

The **People’s prize** sponsored by *Digital Discovery* journal was awarded to **Molly Bartlett** for her poster titled “Uncovering structure activity relationships using machine learning techniques on experimental data generated in high throughput”

Congratulations to all of the Poster Prize winners.

The Summer of Science Festival at Nantwich Museum

Contribution from Dr Helen Cooke, RSC CICAG Newsletter Editor and Trustee at Nantwich Museum, email:

helen.cooke100@gmail.com

Although a local-history museum, Nantwich Museum incorporates STEM subjects within its events and exhibitions wherever appropriate and is fortunate to have a number of scientists and engineers amongst its volunteers, providing a strong foundation for scoping, planning and delivering such activities.

For example, in 2017, in partnership with Keele University, school visits and family workshops were organised which involved analysis of the quality of the water in the local River Weaver, and in 2019 during the International Year of the Periodic Table the Museum’s Research Group developed the “From Nantwich to Oxygen” exhibition and associated events, exploring the life of Joseph Priestley (discoverer of oxygen and Nantwich resident 1758-61).

In our experience, discussing science in a context to which people young and old can relate can help to reduce apprehension about science. At Nantwich Museum building connections between science and local history provides this opportunity and this was a key motivation for the Summer of Science festival which ran from 20 July-30 September 2023.

The festival was also an opportunity to engage with relevant community groups, other external organisations and individuals, both local and from further afield, which enhanced our offerings and strengthened relationships.

We were delighted to welcome Professor Mark Ormerod (PhD chemist), Deputy Vice Chancellor and Provost at Keele University, to open the festival.

Bringing objects to life

Explaining how objects displayed at the Museum were made or how they worked (in the process revealing the science and technology embedded within them) became a key objective. For example, many visitors look at the Museum's cannon balls and muskets dating back to the Civil War, not realising that it's the chemistry of the gunpowder used when weapons were fired which makes them work. This reinterpretation should have long-term benefits for visitors of all ages.

We created new roller banners featuring a variety of objects and local people with connections to science. They proved extremely popular with visitors, and have the added benefit of being portable and thus suitable for events outside the Museum. The banners featured: fire pump, mangle, weighing chair, salt, muskets, cannon balls, herbalist John Gerard, Joseph Priestley,

and a Nantwich apothecary – more are in the pipeline.



Family drop-in workshops

As the festival fell mainly in the school holidays, family drop-in events with a science theme were developed. Many were delivered by our volunteers, for example exploration of the salinity of local brine on "Salty Saturday", but some were developed by external organisations, including a "Making Medicines" workshop led by pharmacists from the local hospital and "Shocking Saturday" delivered by the Institute of Physics North West branch. Workshops were free of charge, to help families and ensure inclusivity.

Experience gained from the children's workshops will enable the Museum to include more science in its educational offerings for schools and its holiday family drop-in workshops.

Talks and demonstrations

We wanted to offer an engaging programme of talks to inspire adults as well as children, so that parents and grandparents might feel less intimidated by science and more confident to discuss it with children.

Topics included a local apothecary, herbalist John Gerard, Nantwich gas works, the local salt industry, Joseph Priestley, sustainable power from the local River Weaver, the science of brewing, genetics, and the geology of the area. Some were delivered by members of our Research Group, others by external contributors.

nantwich MUSEUM

PILLS, POTIONS & POISONS!

EXPLORING THE CONTENTS OF A 17TH CENTURY APOTHECARY

Apothecaries were predecessors to modern-day pharmacists. The probate inventory of Nantwich apothecary Raphe Walley's (1625-1661) home and shop provides insights into his life and the services he provided. His apothecary shop is believed to have been in the building later occupied by Grice's chemist until 1978, in Nantwich High Street. Raphe practised as an apothecary during the Civil War, possibly providing medicines to treat soldiers.

The inventory lists over 200 herbs, spices, animal products, chemicals and items of equipment.

Would you use any of these?

- Nutmegs, cloves, saffron, turmeric, white pepper, oregano, hellebore, tamarind, cummin and capers. Raphe would have used these to make medicines and also to sell to customers for cooking.
- Earthworms, to make oil of earthworms for treating bruises, arthritis, gunshot wounds and other complaints.
- Powdered mummified remains to cure abscesses, fractures and more!

Selection of items in the inventory of Raphe Walley's shop

- Arsenic, antimony, mercury, borax, saltpetre. Some of these are deadly poisons, though this wouldn't have been known in Raphe's time. Mercury was used to treat ulcers, sores and worms in children!
- Pestles, mortars, lancets and incision knives, possibly used by Raphe for performing minor surgery like lancing abscesses or blood-letting.

Some of the above are still used today, but for different purposes.

PLEASE DO NOT TRY THESE AT HOME!

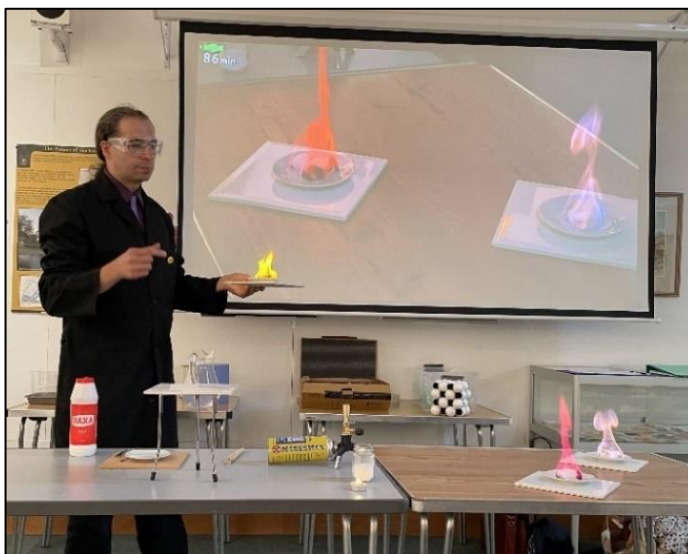
18th century syringe jar for oil of earthworms
(Image reproduced courtesy of the Nantwich Group Collection, Creative Commons License)

18th century mortar and pestle
Pestle and mortar used to crush and grind ingredients into a fine paste or powder.
(Image reproduced courtesy of the Nantwich Group Collection, Creative Commons License)

Advertisement for Grice's Chemist
impounded from Johnson's Directory, 1910

ROYAL SOCIETY OF CHEMISTRY

The festival ran in parallel with the Museum's "Nantwich Illuminated" exhibition on the history of the town's gas works site, which included the production of gas from coal and the manufacture of by-products. This provided considerable scope for chemistry demonstrations and the Museum is especially grateful to Professor Fabio Parmeggiani (Politecnico di Milano) for the bespoke demonstrations he created involving coal, gases, how gas holders and Davy lamps worked, the chemistry of sodium, chlorine, salt, lead, and more. There were plenty of flames and bangs to keep everyone entertained! A [summary video](#) is available on the Museum's YouTube channel.



Dr Mark Whalley (University of Chester and Institute of Physics) also entertained and informed the audience with demonstrations involving static electricity, the historic connection being Joseph Priestley's explorations of electricity.

In addition, the historic method of salt-making was re-enacted in the town's main square, showing how salt was made from local brine during Roman times, and for many centuries afterwards. This proved very popular with passers-by and will be repeated.

Guided walking tours

The Museum schedules walking tours throughout the year, and for the festival new walks which included some introductory science were introduced: "Science around Nantwich", "Riverside Nature" (especially suitable for children) and "Exploring Green Spaces" (led by the Sustainable Nantwich community group). These were repeated several times throughout the festival, though unfortunately sometimes affected by poor weather.

Other events and displays

- Sustainability and climate change themed coffee mornings
- Embroidered periodic table (made by the Museum's Craft Group)
- Live photosynthesis experiment, showing how oxygen is produced by plants
- Chester's Grosvenor Museum's "Getting Drastic with Plastic" travelling exhibition
- Screening of the film "2040", exploring potential solutions to climate change

Feedback from visitors, surveys and social media

"I had no idea what the periodic table was when my grandson mentioned it recently – now I know it lists all the known elements."

"The best chemistry demo I have ever seen...if I had seen this 60 years ago I would have studied chemistry instead of...!"

Acknowledgements

Nantwich Museum is extremely grateful for financial support from the RSC's Outreach Fund, the RSC's North Staffordshire & South Cheshire Local Section, and Museum Development North West's Sustainable Improvement Fund.

Cheminformatics: A Digital History – Part 4. Ladies First

Contribution from Dr Wendy Warr, Wendy Warr & Associates, email: wendy@warr.com



*Dr Wendy A. Warr, MA, DPhil (Oxon.), CChem,
FRSC, FCLIP.*

In Part 4 of this historical series of articles, I am honoured to follow in the distinguished company of Peter Willett, Henry Rzepa and Johnny Gasteiger, all of whom had a significant influence on me during my progress in the path of cheminformatics. I started out as a synthetic organic chemist. When I matriculated at Oxford in Michaelmas 1964 the ratio of women to men Oxford undergrads in all subjects was 1:5. (At Cambridge it was 1:7, but Cambridge had only 2-3 women's colleges then, while Oxford had five.) I was one of seven women in total from the five Oxford colleges who studied for chemistry Part I in 1964-1967. I don't know how many men there were, but the number was certainly over 200. A great deal of progress has since been made as regards equality. On the Oxford undergraduate course in 2023 there are 421 men (57%) and 317 women (43%). There is also a [web page about women in chemistry](#). Jonathan Goodman tells me that third years reading chemistry at Cambridge in Michaelmas 2023 are about 40% female.

I was at Lady Margaret Hall, Oxford, whose undergraduates were always addressed as 'ladies': hence the title of this article. One of my sons says that women have been the dark matter of science: they have always been there and are immensely important, but they have only recently been recognised. In 1964, when I went up to Oxford, Dorothy Crowfoot Hodgkin won the Nobel Prize in Chemistry, the first woman winner since the Curies early in the century. Historically, eight out of 191 chemistry Nobel Laureates have been women, but things are improving: five of the eight winning women were recognised in 2009 or later.

There was no such as thing as 'cheminformatics' in 1964, but in 1968, I was offered a job in the newly founded Experimental Information Unit at Oxford University. I spent the summer learning to code structures into Wiswesser Line Notation (WLN) as part of a project to make ISI's Index Chemicus available to UK universities. After a few weeks, I admitted that I had really wanted to work for a DPhil but had been put off by geographic and financial hurdles. I had married very young and after a year we bought a house in North Oxfordshire. The travelling and other costs were formidable in the bad old days when married women were given reduced grants (only £400 instead of £500). To cut a long story short, I did do the DPhil but I went on coding WLN's part time. My earnings were expenses for commuting, so income tax was not deducted from my £400 grant.

So, when I qualified, I had the option of two alternative careers. For rather more than a year I did process chemistry in industry, before Ernie Hyde persuaded me to join his team at ICI Pharmaceuticals at Alderley Park in Cheshire. ICI were earlier adopters of WLN, and I was one of the few people in the UK who knew the lingo. At that time ICI applied the standard '80% rule': women were not allowed to earn more than 80% of the maximum salary that a man could achieve in the same role. Fortunately, that rule was soon abolished, and I can honestly say that ICI gave me all sorts of opportunities and even allowed to me to take three months off (albeit unpaid) in an era when pregnancy always meant job loss. Thus, for better or for worse, I was able to get promoted at regular intervals.

From 1972 to 1976 I was an information scientist. At that time, Information Services Section managed external information and Data Services Section (where I worked) managed internal information. All searches were mediated by an expert. In data services, chemical structure databases were searched using CROSSBOW (Computerized Retrieval of Organic Structures Based on Wiswesser). Computers were huge mainframes operated by IT staff in big, air-conditioned facilities. My lifelong interest in chemical reactions moved from the bench to the computer when I was involved in an early project on coding reactions in WLN. In the Information Services Section in the early 1970s, experts dialled in and used an acoustic coupler to do online searches for external information; later on, they dialled in at maybe 300 baud and used a teleprinter to search Dialog or Orbit.

In 1976 to 1979 I was a senior information scientist and systems analyst, and in 1979-1984 a senior systems analyst and project manager. I wrote an early program in COBOL to handle the Commercially Available Organic Chemicals Index (CAOCI). CAOCI later became the Available Chemicals Directory. As a systems analyst, I learned about database design, and I worked on the European Inventory of Existing Commercial Chemical Substances (EINECS) and a system for Home Office licences for radiochemicals. I attended my first ACS national meeting in 1977. I soon became involved in all sorts of ACS committees and ACS CINF and COMP activities. I was an Associate Editor of the *Journal of Chemical Information and Computer Sciences* and the *Journal of Chemical Information and Modeling* for 24 years (1989–2013). I continue to attend almost every ACS national meeting, two meetings every year, but recently I have been able to experiment with one in-person and one virtual meeting each year.

By about 1980, the world was moving away from mediated searches and WLN. MDL was founded in 1979 and MACCS eventually replaced CROSSBOW. I oversaw the conversion of the ICI Pharmaceuticals database from WLN to MACCS format. CAS ONLINE appeared in 1980 but had no proper substructure search at first. DARC was used to search EURECAS on Télésystèmes Questel in 1981 before CAS ONLINE could do substructure search.

MACCS was originally designed for a PRIME minicomputer, but ICI was determined to use it on the new VAX (virtual address extension) machines which replaced the PDP/11. I specified our 1984 VAX 11/750 machine which had a clock speed of 6 MHz, 2 MB memory, 134 MB fixed disk, two 67 MB exchangeable disk drives, and shared peripherals. It cost £100,000 (£320,000 in 2023 terms).

Computer graphics go back to the 1960s. The Imlac company was founded in 1969. Initially, MDL's MACCS could only be used on expensive Imlac terminals. A light pen was used to draw structures. Later, lower-priced VT640 terminals made feasible the use of MACCS by end users in ICI. The IBM PC 5150 was introduced August 1981. It cost \$1565–\$3000 (\$5060–\$9700 in 2023). It had 16–256 kB memory, and a 4.77 MHz CPU. Software used was IBM BASIC and PC DOS. The first Macintosh appeared in January 1984, but Macs were not initially popular in pharma.

When I became manager of research information in 1984, internal and external information services were finally united in one group in an enlarged section under Angela Haygarth-Jackson. Angela was a true pioneer in information science. She was the first woman President of the Institute of Information Scientists, in 1983-1984, and was awarded an OBE. She was not a chemist, but she served the RSC and was appointed FRSC in 1993. When she retired in 1986, I became manager of information services, managing three libraries, literature and patent searching, the reports collection, the research notebook collection, and the compound store, in addition to chemical and biological research information.

Through almost all my years at ICI Pharmaceuticals, the various divisions of ICI collaborated on many information issues, not just on the united strength of our buying power for services. I was on many interdivisional committees and projects. I was on a committee on chemical reactions which supported Johnny Gasteiger's team in Munich. ICI (like many other companies) also had a close relationship with the University of Sheffield and greatly appreciated the fundamental research done there. I am giving no literature references in this article, partly because the earlier articles in this series by [Peter Willett](#) and [Johnny Gasteiger](#) tell it all. Those who wish to see some of my own contributions will find references [here](#).

The Information School at Sheffield has been at the forefront of developments in the information field for more than 50 years. It is recognised as the leading school of its kind in the United Kingdom, with an international reputation for the quality of teaching and research, and for the achievements of its graduates. As professor there, Val Gillet's primary research area is cheminformatics and the application of machine learning methods to the discovery of new pharmaceuticals. As a cheminformatician at the peak of her research, she is both a worthy successor to Peter Willett and an example to future women cheminformaticians.

My relationship with Sheffield, and visiting lectures at many other universities, continued after I left ICI and launched my own business. My PR expert (and younger son) puts it this way:

“Since January 1992, Wendy Warr & Associates has been supplying business and competitive intelligence services to a broad spectrum of clients in the United States, Europe, Australia, the Middle East, and Asia. Our success stems from our extensive network and our specialised knowledge of cheminformatics and ‘high throughput chemistry’. Pharmaceutical companies, venture capitalists, publishers, software companies, and scientific database producers have benefited from our expert counsel and services in recent years.”

More recently I have become very interested in AI (who hasn't?) and especially AI in reactions. I joke sometimes that robots (or, rather, the lack of them) drove me out of ICI, but the wheel has turned full circle as I now look at robots and lab of the future, but all this is another story. During the pandemic I got involved in the subject of ultralarge databases and co-edited a special issue of the *Journal of Chemical Information and Modeling* on reaction informatics. I worked on conferences with the [AI4SD network](#) and the NIH National Cancer Institute. I had already been very interested in structure search; now that interest has turned to ultrafast searching.

I was very proud to win the ACS CINF Herman Skolnik Award in 2020, though I didn't get my symposium and reception until 2022 (see photo).

Only two women out of 47 awardees have won the Herman Skolnik Award. Yvonne Martin won it in 2009. She was a true pioneer of computational chemistry. She also won the prestigious ACS Award for Computers in Chemical and Pharmaceutical Research in 2017 and next year she will receive the ACS Division of Medicinal Chemistry Alfred Burger Award in Medicinal Chemistry.

Three of the 38 winners of the ACS Award for Computers in Chemical and Pharmaceutical Research have been women. For her symposium, Yvonne chose a dazzling array of female speakers one of whom was Katharine (Kate) Holloway (see photo), who won the award in 2021. Kate began her career at Merck in 1985 and was one of the chemists who developed protease inhibitors to inactivate the HIV virus, greatly extending the lives of AIDS patients. Over the years she has contributed to the invention of ligands for over 60 different therapeutic targets. Successful design examples include Crixivan, one of the first marketed HIV protease inhibitors, and a second-generation hepatitis C protease inhibitor, grazoprevir, which was combined with the HCV NS5A

inhibitor elbasvir in Zepatier. She is a former chair of the Gordon Research Conference on Computer-Aided Drug Design and is still actively involved in the structure-based design community.



Wendy Warr and Sue Cardinal (Chair of ACS CINF)

was chair of the Computer Aided Drug Design Gordon Research Conference in 2021-2023, is an editorial board member of more than one journal and has served on the Scientific Advisory Board of the Cambridge Crystallographic Data Centre.

You may wonder why I have not yet mentioned Olga Kennard OBE, the eminent crystallographer, and winner of many awards, who founded the Cambridge Structural Database (CSD) in 1965 and was leader of the Cambridge Crystallographic Data Centre until 1997. Crystallographers are not usually 'cheminformaticians', but Olga well

deserves credit here because of her contributions to databases and data science, and the value that CSD has added to our community. I first met her in the 1980s when I used to negotiate ICI's licence for CSD. I also collaborated with her when Greg Paris and I converted CSD structures to 3D MACCS ones.

I will mention a few other names. Christine Humblet was an early molecular modeller who built the foundation and blueprint for the software Sybyl, later commercialised by Tripos. From 1984 she spent 19 years at Warner Lambert/Parke Davis and later worked for Abbott, Wyeth and Eli Lilly. She pioneered the use of structure-based design, 3D molecular visualisation, and integrated platforms to analyse target families with ligand-based

Man-Ling Lee of Genentech (also pictured in the photo I took of Yvonne's speakers) is a powerhouse on scientific data management software and has been instrumental at companies small and large. Hanneke Jansen (also pictured) was introduced to the field of 3D-QSAR by Yvonne Martin who was visiting the lab in Groningen where Hanneke was a PhD student (around 1990). Yvonne's insights and passion for the field were inspiring and Yvonne pointed Hanneke to the emerging technology of 3D-QSAR.

Georgia McGaughey (pictured) has co-invented or 'co-authored' on molecules which have entered clinical trials and in some cases progressed to marketed medicines. Her successes have been in diverse therapeutic areas and in more than one company. In addition to her outstanding contributions in leading the Data and Computational Sciences team at Vertex, Georgia has been a relentless advocate for underrepresented groups in research. She was the only woman scientist in her Vertex research group when she was hired but the group now has 50% diversity. Georgia



L to R: Georgia McGaughey, Johanna (Hanneke) Jansen, Yvonne Martin, Kate Holloway, Man-Ling Lee, Katrina Lexa, Maricel Torrent

structure-activity data. Lisa Balbes also drew my attention to Michelle Francl-Donnay, a quantum chemist who is on a list of the 1000 most cited chemists, but I am omitting quantum chemists and theoretical chemists in order to reduce the length of this article.

Thus far, I have covered mainly computational chemists, but we should not forget women who led progress in research informatics and information science. Angela Haygarth Jackson was one, obviously. Sandra Ward was also a President of the Institute of Information Scientists (in 1997-1998). In 1985, Janet Ash, Pamela Chubb, Sandra, Peter Willett and Steve Welford wrote one of the seminal books in chemical information: *Communication, Storage and Retrieval of Chemical information*. Diana Leitch MBE is of course well-known to RSC CICAG. Anne Girard of the Institut Français du Pétrole and Ursula Schoch Grüber of BASF were both experts on patent searching and were internationally recognised.

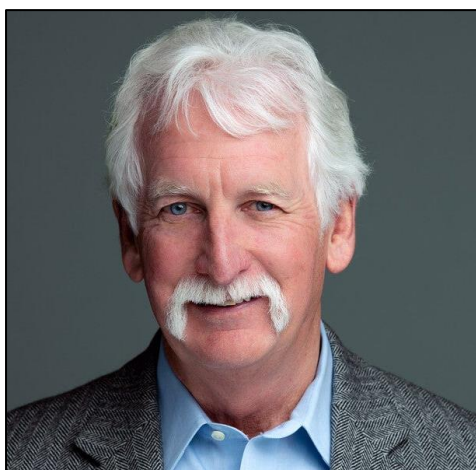
I have known Bonnie Lawlor in person since the 1970s, and I value her as a friend as well as a respected business colleague. Bonnie has served the American Chemical Society, the Chemical Notation Association, the Chemical Structure Association, and the CSA Trust for many, many years. After Thomson acquired the Institute for Scientific Information (ISI), she was Executive Vice President, Database Publishing at Thomson Scientific-ISI, and later, Executive Director of NFAIS.

I could go on and on producing ever longer lists of women who have made contributions to our profession. I am conscious that I have presented an arbitrary selection; regard it as exemplary (in all respects). I am sure that readers will immediately think of people whom I should have mentioned. Please forgive me for errors and omissions.

I was asked to write about myself so I must conclude with comments about yours truly. I have had a wonderful career in cheminformatics, and it has been a pleasure meeting so many influential and inspirational people on multiple continents. As my business statement says, the success of my current business has in many ways been built on my extensive network. I would like to thank the people in the network who have helped me with ideas for this article. Onward and upward 'ladies'!

William Jorgensen Wins 2024 Arthur C. Cope Award for Organic Chemistry

[Article by Jim Shelton](#), 13 September 2023, reproduced with permission from Yale University, USA.



Jorgensen, a pioneering computational chemist, is the first Yale faculty member to win the Cope Award since 1988.

William Jorgensen, Sterling Professor of Chemistry in Yale's Faculty of Arts and Sciences, has been named the recipient of the American Chemical Society's 2024 Arthur C. Cope Award for his ongoing achievements in organic chemistry.

The Cope Award, established in 1972, is one of the most highly regarded honors in organic chemistry worldwide. In announcing the award, the ACS lauded Jorgensen for his "pioneering computational studies of organic chemistry in solution,

development of free-energy methods, and demonstration of their utility in lead optimization for discovery of drugs.”

Jorgensen’s seminal research in molecular design and computational chemistry includes simulations of organic and enzymatic reactions, computer-aided drug design, and the synthesis and development of drug agents that combat HIV, inflammation, and cancer. In 2021, he and colleagues at Yale rapidly developed a new class of antiviral drug agents with the potential to create new therapies for COVID-19 and future coronaviruses.

“For me, the Cope Award is particularly meaningful,” Jorgensen said. “It is always special to receive an award that has been received by others whom you greatly admire, in this case, including my PhD advisor, E. J. Corey at Harvard, and Yale emeritus professor Ken Wiberg. The Cope Award predominantly has been awarded to centrist organic chemists, while my research has spanned between organic and theoretical/computational chemistry. Perhaps, our computationally guided discoveries of anti-HIV agents and inhibitors of the SARS-CoV-2 main protease were sufficient to pass the organic test.”

Jorgensen, who joined the Yale faculty in 1990, is the first Yale faculty member to win the Cope Award since Wiberg won the honor in 1988. The award comes with a \$25,000 prize, a medallion, and an unrestricted grant-in-aid of \$150,000 for research in organic chemistry.

Among Jorgensen’s other honors, he is the recipient of the prestigious Tetrahedron Prize for Creativity in Bioorganic & Medicinal Chemistry. He has been elected to the National Academy of Sciences and he is a fellow of the American Association for the Advancement of Science, the American Academy of Arts and Sciences, the American Chemical Society, and the International Academy of Quantum Molecular Sciences.

In 2021, he was selected as a Citation Laureate for his influential contributions to chemistry. According to the Web of Science, Jorgensen’s scientific publications have been cited more than 100,000 times.

Personal memories from Dr Willem van Hoorn, CICAG Committee Member and former postdoctoral researcher in the Jorgensen group 1997-1999, email: rsc@vanhoorn.co.uk

The first time I met Bill Jorgensen was in 1993 when he was a speaker at a conference the Dutch group where I was a PhD student had organised. Multiple American speakers had spoken before him, and a surprisingly large number had started their talk expressing how happy they were to visit the land of their Dutch ancestors. Bill started his talk saying, with a completely straight face, that his ancestry was actually not Dutch but Danish, but they had visited the Netherlands regularly and brought home many souvenirs. It took a while for the joke to be noticed, and he had already started by showing some overheads (remember those) he had scribbled in the preceding break. The last speaker before the break had argued something like all hydrogen bonds are equal strength. Bill disagreed and presented a few highlights of his work showing the opposite. With numbers and references. This made a big impression on me and the other PhD students, scientific fireworks! He told me later that normally he wouldn’t take someone down like that but he thought that there were too many young impressionable minds to let it slide unchallenged. And that he has a little book with key findings from his papers.

He visited our group towards the end of my PhD and I got to spend some time showing my work. Of course I also inquired about doing a postdoc in his group, and he answered along the lines that yes maybe he had a slot,

which I interpreted as a polite way of saying no. But when I finally finished my PhD I was asked when I could start! At least one other group member had a similar story. Bill means what he says.

Although not publishing often, I still use Bill's guide to scientific writing. One other scientific lesson: in statistical mechanics there is only one way to do the science right. So be a hawk when looking at your results. This also meant him looking at your results like a hawk. In complete silence he would go through your draft paper page by page while you were sitting in front of his desk. For minutes that seemed like hours he would occasionally hum before turning a page. And rub his chin or forehead. The first was generally a good sign, the latter not so much.

Bill's incredible list of accolades and scientific impact is well-deserved, but they were not achieved at the cost of his staff, quite the opposite. Teaching undergrads, and making sure PhDs and postdocs learn what they need for a future scientific career is equally important. All of the above was combined with a very dry sense of humour.

RSC Databases Update: ChemSpider

Contribution from Richard Kidd, Royal Society of Chemistry, email:

KiddR@rsc.org



ChemSpider has been running a series of webinars on chemistry data – focusing on how data is enabling research – the current challenges and examples and how a better future can be created using chemistry data. It has showcased current and planned initiatives to develop standards and tools, research infrastructures, and developing cultures to support Findable Accessible Interoperable Reusable (FAIR) chemistry data preparation, publication and reuse.

The webinars have featured:

Leah McEwen – on standards and notation

Kevin Jablonka – on chemistry LLMs

Pierre Morieux – on the Revvity Signals product suite (as our series sponsor)

Lynn Kamerlin – on the explosion of chemistry data

Simon Coles – on AI in research

May Copsey and Anna Rulka – on data sharing in RSC journals

Sonja Herres-Pawlis – on cultural change for digital chemistry

Samantha Pearman-Kanza – on heterogenous, unfair and disparate data

Guy Jones – on data journals supporting sharing and discovery

The [webinars are available online](#).

Hopefully by the time you read this: a new ChemSpider beta! We have been moving ChemSpider over to a cloud architecture, so expect to see more communications about our data clean-ups and simplification that we have done on the way. Not all functionality, or all the data, is there yet, but take a look at <https://beta.chemspider.com>.

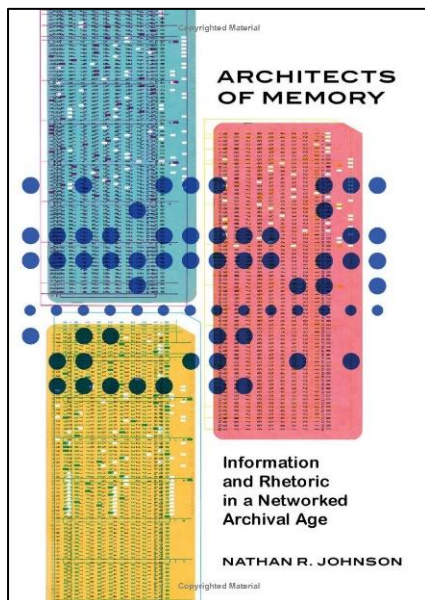
Book Review: Architects of Memory

Contribution from Robert E. (Bob) Bunrock, Bunrock Associates, Orono, ME, USA, email:

bunrock16@roadrunner.com

Architects of memory: information and rhetoric in a networked archival age

Nathan R. Johnson. Publisher: Univ. of Alabama Press, Tuscaloosa, USA, 2020. 224pp, hardcover ISBN 978-0-8173-2060-7, \$49.95 (~£38.66).



“In this strange but important book, Johnson frames 20th-century information management practices with the rhetorical principle of mnemonic techne, as expressed in classical retellings of the myth of Simonides, who was able to identify dinner guests after a roof collapse by picturing their seating arrangement during a banquet. Reading this history alongside the evolution of the Simonides parable induces one to reflect deeply on values underpinning the mechanics systems use to organise knowledge. Recommended.”

— [CHOICE](#)

Intrigued by a chance notice and this review, I acquired a copy of this three year old book for a possible review for this and other audiences. (In the absence of prior knowledge – even used copies on Amazon cost about as much as new – I obtained a copy on loan via the Maine Library System). Strange indeed, important? Maybe to some. This portion of a terse review – describing classical rhetoric – illustrates the intended

audience: rhetoricians. The author is a professor of rhetoric so the choice of audience is not too surprising. I was unfamiliar with rhetoric other than when I was in college, engineers took rhetoric, we scientists took freshman English. After a cram course on rhetoric, I’m still of the opinion that this book has a limited audience.

The memory of concern is public or collective memory, not personal. The history of the development of information – public memory – is mostly of US developments, many of which were spawned from US WW II and Cold War information efforts, so maybe even of less interest to British readers. I did find some interesting aspects of the author’s historical descriptions, including of some key organisations – The American Library Association (ALA), The American Society for Information Science (ASIS, now ASIST) – but I’ll challenge technical librarians and information specialists to review these histories, written by an ‘outsider’, for veracity in their opinion. The author also describes librarians and information scientists – devisers of the information technology used – as competitors without acknowledging the crew of SciTech information specialists (like myself) who were trained as scientists (especially chemists) and picked up information technology and techniques on the job.

A longer version of the review of this book will appear in the Winter Issue of the Chemical Information Bulletin (CIB) of the Chemical Information Division (CINF) of the ACS. In it I encourage the readers to obtain a copy for their own review and if in a library, for sharing and discussion with any rhetoricians on site, if any. (Actually, to my surprise, several US colleges and universities have Departments of Rhetoric.) So, I’ll make the same recommendation to British readers. Recommended to rhetoricians, those interested in the history of science and technology, and philosophy.

Cheminformatics and Chemical Information Books

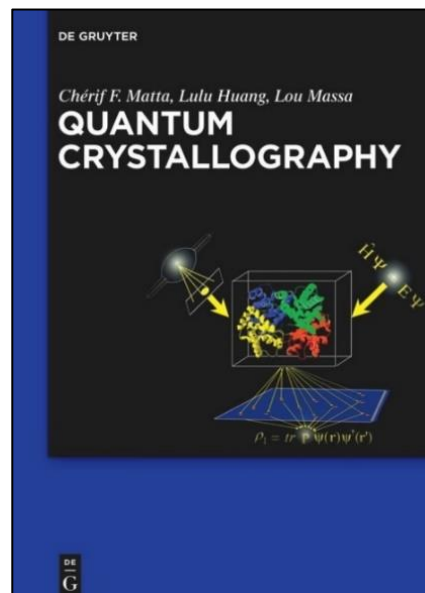
Contribution from Helen Cooke, CICAG Newsletter Editor, email: helen.cooke100@gmail.com

Descriptions are as provided by the publishers and not necessarily the view of the contributor or CICAG.

Quantum Crystallography

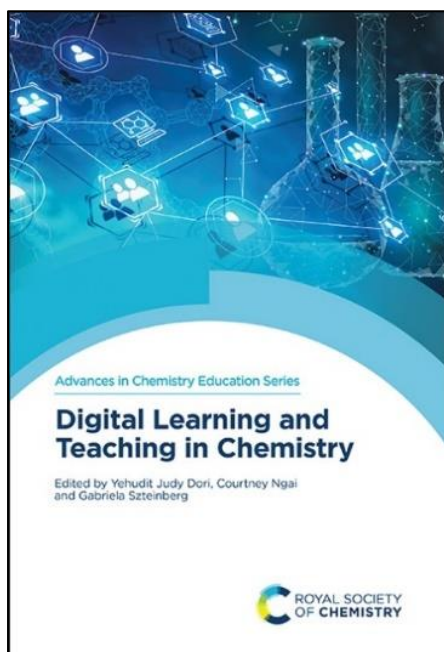
Quantum crystallography (QCr) is a novel scientific discipline combining quantum chemistry methods and crystal structure determination. Written by leading experts in the field, this book describes original quantum-mechanical approaches to obtain crystallographic data of enhanced value and explains how they correlate with real diffraction and scattering experiments. In particular, the book covers quantum N-representability, Clinton equations, kernel energy method (KEM), and quantum theory of atoms in molecules (QTAIM) methods and their applications in crystallographic studies. Readers will be interested in the Foreword written by Nobel Laureate Ada Yonath and the Epilogue by noted science philosopher Olimpia Lombardi.

- Outlines the latest achievements and developments of this novel area of research.
- Written by the leading scientists of the field.
- Bridges together theoretical methods and experimental studies.



Chérif F. Matta, Lulu Huang, Louis J. Massa. De Gruyter, September 2023, £107.50. DOI

<https://doi.org/10.1515/9783110566673>. eBook ISBN: 9783110566673, hardback ISBN: 9783110565669.



Digital Learning and Teaching in Chemistry

Education is always evolving, and most recently has shifted to increased online or remote learning. *Digital Learning and Teaching in Chemistry* compiles the established and emerging trends in this field, specifically within the context of learning and teaching in chemistry. This book shares insights about five major themes: best practices for teaching and learning digitally, digital learning platforms, virtual visualisation and laboratory to promote learning in science, digital assessment, and building communities of learners and educators. The authors are chemistry instructors and researchers from nine countries, contributing an international perspective on digital learning and teaching in chemistry.

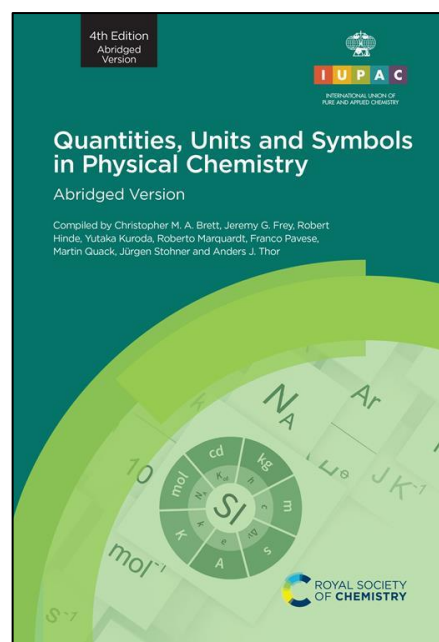
Edited by Yehudit Dori, Courtney Ngai, Gabriela Szeinberg. Royal Society of Chemistry, July 2023, £99.99. DOI

<https://doi.org/10.1039/9781839167942>. Hardback ISBN: 978-1-83916-523-8, PDF ISBN: 978-1-83916-794-2, EPUB ISBN: 978-1-83916-795-9.

Quantities, Units and Symbols in Physical Chemistry

The first IUPAC Manual of Symbols and Terminology for Physicochemical Quantities and Units was published in 1969 with the objective of “securing clarity and precision, and wider agreement in the use of symbols, by chemists in different countries, among physicists, chemists and engineers, and by editors of scientific journals”. Subsequent revisions have taken account of many developments in the field and were also substantially expanded and improved in presentation in several new editions of what is now widely known as the ‘Green Book of IUPAC’. This abridged version of the forthcoming 4th edition reflects the experience of the contributors and users of the previous editions.

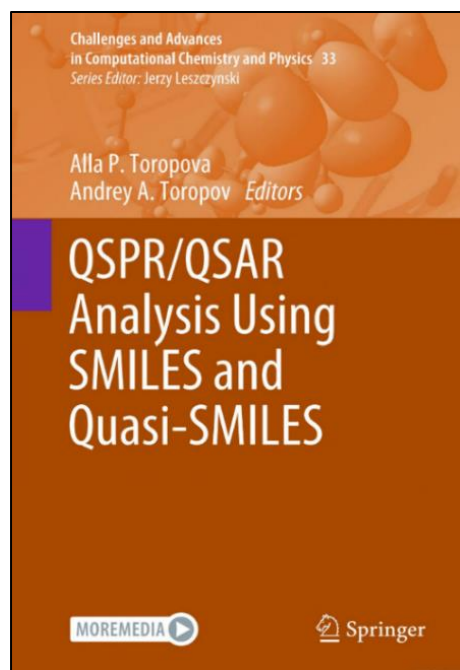
Edited by Christopher M A Brett, Jeremy G Frey, Robert Hinde, Yutaka Kuroda, Roberto Marquardt, Franco Pavese, Martin Quack, Juergen Stohner, Anders J Thor. Royal Society of Chemistry, November 2023. £30.99. DOI: <https://doi.org/10.1039/9781839163180>. Paperback ISBN: 978-1-83916-150-6, PDF ISBN: 978-1-83916-318-0.



QSPR/QSAR Analysis Using SMILES and Quasi-SMILES

This contributed volume overviews recently presented approaches for carrying out QSPR/QSAR analysis by using a simplifying molecular input-line entry system (SMILES) to represent the molecular structure. In contrast to traditional SMILES, quasi-SMILES is a sequence of special symbols-codes that reflect molecular features and codes of experimental conditions. SMILES and quasi-SMILES serve as a basis to develop QSPR/QSAR as well Nano-QSPR/QSAR via the Monte Carlo calculation that provides the so-called optimal descriptors for QSPR/QSAR models. The book presents a reliable technology for developing Nano-QSPR/QSAR while it also includes the description of the algorithms of the Monte Carlo optimisation. It discusses the theory and practice of the technique of variational autoencoders (VAEs) based on SMILES and analyses in detail the index of ideality of correlation (IIC) and the correlation intensity index (CII) which are new criteria for the predictive potential of the model. The mathematical apparatus used is simple so that students of relevant specialisations can easily follow. This volume is a valuable contribution to the field and will be of great interest to developers of models of physicochemical properties and biological activity, chemical technologists, and toxicologists involved in the area of drug design.

Alla P. Toropova, Andrey A. Toropov. Springer, June 2023. DOI <https://doi.org/10.1007/978-3-031-28401-4>. eBook 978-3-031-28401-4 £111.50, hardcover ISBN 978-3-031-28400-7 £139.99.



Cambridge Structural Database (CSD) Updates

Contribution from Michael Francis, Marketing Manager at CCDC;
Ana Machado, Marketing Executive at CCDC, email: hello@ccdc.cam.ac.uk



Update on our ongoing investigations into structures associated with a pre-print on a papermill in crystallography

In April 2023 we [provided an update](#) to confirm 209 of the implicated structures had been retracted from the CSD following the retraction of 125 associated publications. These retractions left 783 implicated structures under investigation.

Working closely with publishers and our community, we are following [COPE guidelines](#) as our investigations progress. We continue to add editorial comments to entries to highlight information that may be relevant so users can select fit-for-purpose data. [Read more](#).

CSD data release

In November we launched a new version of the entire Cambridge Structural Database (CSD) adding 10,837 structures (11,411 new entries). This took the total size of the database to 1,254,560 structures (1,284,316 entries).

This full release included improvements to existing entries and new entries. Knowledge bases such as Mogul and IsoStar were updated. [Read more](#).

CSD software release

Our software has been updated. The latest release included new drug discovery and materials science improvements.

- The torsions sampled by GOLD during docking have been improved including re-built and enhanced torsion distributions using latest Cambridge Structural Database (CSD) and Mogul knowledge. SMARTS can be used to define torsion patterns. [Read more on the improved CSD torsion data in GOLD here](#).
- CSD-Materials API Extensions. To support solid-form reporting and research, this update extends the CSD Python API to cover more of the CSD-Materials suite, including parts of the Solvate Analyser and Hydrate Analyser components, plus improvements to Hydrogen Bond Propensity, and Hydrogen Bond Statistics in the CSD Python API. [Read more on unlocking solid form innovation here](#).

Stay informed of new updates by [signing up to the CCDC email alerts](#) or join one of our regular free webinars to get live demos of new features, see www.ccdc.cam.ac.uk for the schedule.

News from the American Chemical Society Chemical Information Division (CINF)

Contribution from Ye Li, CINF Chair 2023, Massachusetts Institute of Technology, email: yel@mit.edu

Greetings from the American Chemical Society Chemical Information Division (ACS CINF)! I am delighted to share a recap of our experiences at the ACS Fall 2023 Meeting and provide updates on our division's leadership transition.

ACS Fall 2023 Meeting highlights

The ACS Fall 2023 Meeting, themed “Harnessing the Power of Data”, held in San Francisco, CA, was a resounding success as a hybrid conference. We extend our gratitude to RSC CICAG members who participated in CINF programs, whether in person or virtually. Special thanks to those who organised sessions and presented. We hope you felt the excitement as we explored the transformative potential of chemical data. The CINF Data Summit program, featuring insightful symposiums and engaging receptions, provided a platform for the exchange of innovative ideas. The Herman Skolnik Award celebration honouring Dr Patrick Walters and the recognition of Val Metanomski Meritorious Service award winners, Judith Currano, Sue Cardinal and Ye Li, were touching demonstrations of the profound impact of professional networks in our field. Further details about the symposia and reports can be found in the Fall 2023 Vol. 75(3) and Winter 2023 Vol. 75(4) of our [Chemical Information Bulletin](#).

Newly elected CINF officers

The 2023 election for CINF officers concluded on 20 September 2023. Please join me in congratulating our newly elected officers, who will assume their roles on 1 January 2024:

- Chair-elect: Nick Ruhs
- Secretary: Rachele Bienstock
- Councilor: Sue Cardinal (3-year term), Andrea Twiss-Brooks (2-year term)
- Alternate Councilor: Stuart Chalk (3-year term), Svetlana Korolev (2-year term)

Our gratitude goes out to all nominated candidates, the nomination committee, and CINF members who participated in the election. The newly elected officers, alongside the existing [CINF Exec team](#), will lead our division forward. If you would like to contribute or provide suggestions for CINF operations, please reach out to us or any [CINF Exec](#) members. We would love to work with you more. We welcome collaboration and are actively seeking volunteers for leadership positions, especially on communications and membership committees. Visit our [website](#) for more details.

Invitation to ACS Spring 2024

Join us for the ACS Spring 2024 Meeting with the theme “Many Flavors of Chemistry” in New Orleans, LA, and Hybrid. The [CINF program](#) is designed to inspire cross-discipline and cross-sector collaborations. [Registration for the ACS Spring 2024](#) opens in mid-December, and the full conference program will soon be available.

As data, machine learning, and AI increasingly integrate into the professional landscape of chemists worldwide, ACS CINF eagerly anticipates forging stronger collaborations with international organisations in these domains. If you have project ideas for collaboration between CICAG and CINF, please reach out to us. Your contributions and insights are invaluable as we navigate this evolving landscape together.

Call for Nominations for the 2025 Skolnik Award

Contribution from Rajarshi Guha, Chair, Awards Committee, ACS Division of Chemical Information, email:
rajarshi.guha@gmail.com

The ACS Division of Chemical Information established this Award to recognise outstanding contributions to and achievements in the theory and practice of chemical information science. The Award is named in honour of the first recipient, Herman Skolnik.

By this Award, the Division of Chemical Information is committed to encouraging the continuing preparation, dissemination and advancement of chemical information science and related disciplines through individual and team efforts. Examples of such advancement include, but are not limited to, the following:

- Design of new and unique computerised information systems
- Algorithmic advances in manipulating chemical information
- Preparation and dissemination of chemical information
- Editorial innovations
- Design of new indexing, classification, and notation systems
- Chemical nomenclature
- Structure-activity relationships
- Applications of chemical information in novel domains

The Award consists of a \$3000 honorarium and a plaque. The recipient is expected to give an address at the time of the Award presentation. In recent years, an Award Symposium has been organised by the recipient.

Nominations for the Herman Skolnik Award should describe the nominee's contributions to the field of chemical information and should include supportive materials such as a biographical sketch and a list of publications and presentations. Three seconding letters are also required. CINF aims to serve a diverse and inclusive worldwide community of scientists and professionals working with chemical information. Thus we encourage nominations of qualified women, members of underrepresented minority groups, and scientists from around the world.

Nominations and supporting material should be sent by email to awards@acscinf.org. Paper submissions will not be accepted.

The deadline for nominations for the 2025 Herman Skolnik Award is 1 June 2024.

Physical Sciences Data Infrastructure (PSDI) News

*Contribution from Dr Samantha Kanza, email: S.Kanza@soton.ac.uk, Dr Nicola Knight, email: n.knight@soton.ac.uk,
Professor Jeremy Frey and Professor Simon Coles, University of Southampton*

Introduction to PSDI

PSDI is a project, funded under the EPSRC DRI, looking to develop a roadmap for future investment in a UK Physical Sciences Data Infrastructure. The overall long-term vision of this project is to provide a data infrastructure that facilitates connections between existing experimental and computational facilities within the physical sciences. PSDI will accelerate and enhance research in the physical sciences by empowering digitally driven research. More information about PSDI can be found on the [PSDI website](#).

PSDI details



www.psdi.ac.uk



[@PSDI_UK](https://twitter.com/PSDI_UK)



[linkedin.com/company/psdiuk](https://www.linkedin.com/company/psdiuk)

Mailing List: <https://www.jiscmail.ac.uk/PSDI>

There are lots of different ways you can engage with the PSDI project, read about some of our recent and upcoming activities below.

PSDI webinars

Our PSDI webinar series is now in full swing with several recordings already available on our YouTube channel. We have held 6 webinars so far, and are planning many more in 2024!

If you are interested in presenting your work at a PSDI webinar do [get in contact](#) via our website.

Catch up on our previous webinar recordings on [YouTube](#) (with more published as recordings become available).

- [Introduction to PSDI](#)
- PSDI Pathfinders: [PF2 Process Recording](#)
- PSDI Pathfinders: [PF1 Experimental Data Capture in Catalysis](#)

Recordings will be available soon for:

- PSDI Pathfinders: PF4 - FAIR Data for the Biomolecular Simulation Community
- Making the intangible tangible: The journey from lab notebook to digital insight
- PSDI Pathfinders: PF5 – Data to Knowledge (to data)

Recent events and presentations

The PSDI team has been busy over the last couple of months attending a variety of conferences, exhibitions and events. We are updating our [zenodo community](#) with many of the presentations and publications made by the team.

Highlights from the last few months include:

Machine Learning School

In September PSDI, in collaboration with [PSDS](#), [AI4SD](#), [STFC-SCD](#) and [CCP5](#) ran the first iteration of the “Machine Learning for Atomistic Modelling Autumn School” at Daresbury Laboratory. This three-day training course, which was targeted predominantly at PhD students in Materials and Molecular simulations, was extremely popular with more than five applications for each place available.



Over the three days attendees learnt about the basics of machine learning, machine learning interatomic potentials and graph neural networks. Alongside lectures there were also a variety of practical sessions utilising a training cloud environment where students could write and run their own code.

The feedback from the students was very positive about the course and we are currently in planning and discussion about future iterations, so watch this space! If you are interested in contributing to this training course please contact Nicola Knight, n.knight@soton.ac.uk

International Data Week 2023 in Salzburg

Several of the teams attended [IDW](#) (Including SciDataCon & RDA Plenary) last month, both in person and online. This vibrant event had many wide-ranging sessions from discipline- and nation-specific initiatives, through to technical working group sessions. The PSDI team took part in many of these sessions, including presenting as part of the SciDataCon session “Beyond FAIR: Reusing Chemical Data Across-disciplines with Care, Trust and Openness”.

Lab Innovations 2023

Samantha Pearman-Kanza was invited to speak at the laboratory industry exhibition [Lab Innovations](#) in Birmingham in November. In the conference track at the exhibition Samantha presented about FAIR data and making research reproducible/reusable in her talk "To the well organised FAIR dataset, re-use is but the next great adventure". Alongside presenting her work Samantha also got the opportunity to speak to a wide variety of companies about their process recording tools and ELNs. Read more about it on our [website](#).

Annual Digital Catalysis & Catalysis-Related Sciences Conference 2023 (ADCR 2023) – Abraham Nieva de la Hidalgo

Abraham attended [ADCR 2023](#), organised by NFDI4Cat in Frankfurt. At this catalysis community event there was the opportunity for Abraham to present work on creating Galaxy workflow tools for processing and analysis of catalysis data. The slides from the presentation and a flash pitch are available on [zenodo](#).

Collaborations

ELNFinder

[ELNFinder](#) is an online tool comprising over 40 Electronic Lab Notebooks (ELNs) designed to help researchers select the right tool for them using a range of detailed filter criteria. The full metadata schema for this criteria can be found [here](#). We urge all ELN companies who aren't already on ELNFinder to get in touch with us so we can onboard you into the system.

ELN Finder

The ELN Finder helps you to search and select a suitable Electronic Lab Notebook (ELN) for your purposes.

- More than 40 filter criteria available.
- Filter criteria clearly divided into categories.
- Result list of the identified ELN tools displayed in an overview.
- Brief descriptions of the individual tools included.

 Find ELNs

AI4Green

AI4Green is an ELN developed by Professor Jonathan Hirst's Group at the University of Nottingham. It is web based and open source, and designed to foster collaboration, good data management and above all sustainability in the lab. Its key features include: automatic calculations, hazard lookup and CAS database linkage, solvent selection guide, and reaction summary including colour coding of solvent sustainability and hazards. PSDI will be working with the AI4Green team to help integrate the ELN into the student laboratories and perform qualitative research on the integration process. Any queries on AI4Green can be addressed to ai4green@nottingham.ac.uk.

News from CAS

Contribution from Dr Anne Jones, Senior Customer Success Specialist, email: ajones2@acs-i.org and Zornitsa Ivanova, CAS Communications Manager



CAS SciFinder Discovery Platform™

The CAS SciFinder Discovery Platform consists of multiple solutions that connect you with the world's scientific knowledge to find the answers you need to advance your research.

These solutions are enhanced regularly, providing easier-to-use tools and more information to improve your work. Here are several notable updates for CAS SciFinder[®], CAS Formulus[®], and CAS Analytical Methods™:

- **An enhanced 'All' search results page** – The results displayed when a search is conducted using the 'All' search option minimises scrolling and improves your ability to scan and interact with results quickly.
- **Coloured heteroatoms in substance displays** – These new colourations conform to CPK conventions and make it easier for you to quickly scan and assess your substance results.
- **View similar reactions** – A single click in a reaction search result allows you to easily explore similar reactions with varying scopes related to the specific reaction centre.
- **Enhanced Knowledge Graph** – Enhancements, including an updated content key, sticky preferences, a search feature, and a clearer colour scheme, make the Knowledge Graph easier to use for understanding the competitive landscape.
- **Expanded Experimental Properties** – Additional experimental properties, such as phase diagram point, thermodynamic properties, and dissociation constants, are now available in substance details for easier access to more information.
- **New Formulation Designer in CAS Formulus** – The new format lays out selections visually and displays top choices for the following criteria, making it simple to adjust your selections.
- **Group by Document in CAS Analytical Methods** – This enables you to view results as a collection of documents with related methods, as well as view all methods from a given document as a unique results set, providing the opportunity to filter, compare, export, and save the methods from a single document.

More details on the latest enhancements can be found by viewing the "What's New?" section available within CAS SciFinder[®]. You can also contact us for more information when using the CAS SciFinder Discovery Platform.

STN IP Protection Suite™

The STN IP Protection Suite consists of multiple solutions, including CAS STNext[®], designed to help IP searchers uncover comprehensive insights and minimize risk. CAS continues to enhance these solutions to meet the growing search needs of our users. Recent notable enhancements for CAS STNext include:

- **Interactive claim viewer** – Users can quickly discover the relationships between claims with a new graphical tree display, which provides an easy-to-navigate overview of the relationships between the different claims within a patent application.
- **Prior Art Analysis tool** – With this AI-powered tool, searchers can quickly generate a set of relevant patent and non-patent literature references published before starting a patent document.
- **Full coverage of EP Unitary Patent information** – Full coverage of Unitary Patent information is now available in databases covering EP patent publications.

- **Ultimate owner information** – Assess current patent ownership with new information to help searchers determine the current IP owner.
- **Identification of substances from claims simplified in CAS PatentPak®** – The CAS PatentPak workflow solution gives you easy access to substances in the full text of patent documents. Because the substances mentioned in the claims section are often of interest due to their legal importance, CAS added claim tags to help you find and jump to the claims for a more detailed review.

More details on these enhancements can be found in the “What’s New” section within CAS STNext.

CAS Insights™

CAS Insights is an open resource for actionable perspectives on the latest developments across science, technology, and innovation powered by CAS human-curated data collection and the expertise of our science team. [Subscribe](#) to get new insights delivered to your email.

Content spans scientific disciplines and industries, including the following recent publications relevant to CICAG members:

- **The impact of AI in R&D:** With the rise of [large language models](#) and ChatGPT, the AI revolution has only begun in chemistry and R&D. From [drug repurposing](#) to improving protein function [predictions](#), the use cases for [AI in chemistry](#) that you can explore in CAS Insights are wide-ranging.
- **Emerging trends in immuno-oncology:** While investments have grown in this area, what should leaders be prioritising? From [covalent inhibitors](#) to [antibody-drug conjugates](#) – discover new [modalities](#), checkpoint inhibitors, biomarkers, and targets in recent CAS Insights articles.
- **Sustainable gains for green chemistry:** Read CAS Insights to learn about new, sustainable [approaches](#) being used across industries to [reduce emissions](#), advance [biomaterials](#), and increase agricultural [efficiency](#).

CAS Future Leaders™

The CAS Future Leaders program supports the growth of science leadership among early-career scientists. Since 2010, the program has awarded PhD students and postdoctoral scholars with opportunities to learn leadership skills, engage in scientific discourse, and connect with peer scientists and innovators worldwide. [Applications for the 2024 program](#) are now open through 28 January 2024.

The Development of the Chemist’s Notebook – Meeting Announcement

Contribution from Dr Helen Cooke, RSC CICAG Newsletter Editor, email: helen.cooke100@gmail.com

This one-day in-person meeting organised by the RSC’s Historical Group will take place on Wednesday 13 March 2024, 10.30-17.00, at Burlington House, Piccadilly, London W1J 0BA.

For many centuries chemists have used notebooks to record their experiments, results, literature research and thoughts. This meeting will feature analysis of the notebook practices of some famous chemists starting from the time of Robert Boyle and consider their evolution until their most recent manifestation in electronic form.

For more information and to book please go to the [RSC Events page](#) for the event or email Peter Morris, Historical Group Secretary, directly at doctor@peterjtmorris.plus.com, giving your name, email address and

any special requirements. The event is free of charge. Coffee and tea will be available, but lunch is not included, although there are plenty of cafes nearby in Piccadilly and adjoining streets.

Programme

- 10.30 Coffee
- 10.50 Welcome
- 11.00 Michael Hunter (Birkbeck, University of London): The Workdiaries of Robert Boyle
- 11.40 Matthew Eddy (University of Durham): What was a Scientific Notebook? Amelie Kier, Chemistry and the Power of Annotation during the 1790s
- 12.20 Lunch (not supplied)
- 1.40 Sharon Ruston (Lancaster University): Protean Poetics in Humphry Davy's Notebooks
- 2.20 Frank James (University College London): How Michael Faraday's Laboratory Notebooks Developed into a Diary
- 3.00 Tea
- 3.30 Kostas Gavroglu (University of Athens): Notebooks as Laboratories: The case of Linus Pauling
- 4.10 Samantha Pearman-Kanza (University of Southampton): Electronic Lab Notebooks and Beyond
- 4.50 Closing remarks
- 5.00 Meeting ends

AI in Drug Discovery 2023 – A Highly Opinionated Literature Review (Part I)

Contribution from Pat Walters, Relay Therapeutics, email: pwalters@relaytx.com

Reproduced from Pat Walters's [Practical Cheminformatics blog post](#). Parts 2 and 3 will be reproduced in the Summer 2024 edition of the CICAG Newsletter.

Here's the first part of my review of some interesting machine learning (ML) papers I read in 2023. As with the previous editions, this shouldn't be considered a comprehensive review. The papers covered here reflect my research interests and biases, and I've certainly overlooked areas that others consider vital. This post is pretty long, so I've split it into three parts, with parts II and III to be posted in the next couple of weeks at <https://practicalcheminformatics.blogspot.com/>.

- I. Docking, protein structure prediction, and benchmarking
- II. Large Language Models, active learning, federated learning, generative models, and explainable AI
- III. Review articles



2023 was a bit of a mixed bag for AI in drug discovery. Several groups reported that the deep learning methods for protein-ligand docking weren't quite what they were initially cracked up to be. AlphaFold2 became pervasive, and people started to investigate, with mixed success, the utility of predicted protein structures.

There were reports of significant advances in protein-ligand docking, but no code or supporting methodology was provided. Finally, several benchmarking studies cast doubt on earlier claims that deep learning and foundation models outperformed more traditional ML methods. For the impatient, here's the structure of Part I.

1. Are Deep Learning Methods Useful for Docking?
 - 1.1 Are the Comparisons Fair?
 - 1.2 Training/test Set Bias
 - 1.3 Structure Quality
 - 1.4 Reporting Scientific Advances in Press Releases
2. Can We Use AlphaFold2 Structures for Ligand Discovery and Design?
 - 2.1 Experimentally Evaluating AlphaFold2 Structures
 - 2.2 Generating Multiple Protein Conformations with AlphaFold2
 - 2.3 Docking into AlphaFold2 Structures
3. Can We Build Better Benchmarks?
 - 3.1 Overviews
 - 3.2 Benchmark comparisons
 - 3.3 Dataset splitting
 - 3.4 New datasets

1. Are deep learning methods useful for docking?

2022 saw the emergence of deep learning (DL) methods for docking. These methods, trained on data from the PDB, learned to predict the poses of ligands based on interactions in known protein-ligand complexes. There were papers on DiffDock, Eqibind, TANKBind, and more. In 2023, these methods underwent additional scrutiny, and it turned out that they weren't quite as good as originally reported. Criticism of DL docking methods fell into three categories: the methods used for comparison, biases in the datasets used for evaluation, and the quality of the generated structures.

1.1 Are the comparisons fair?

One potential advantage of DL docking programs is their ability to perform "blind docking". Unlike conventional docking programs, the DL methods don't require the specification of a binding site. The DL programs use training data to infer the protein binding site and the ligand pose. In earlier comparative studies, conventional docking programs were simply given an entire protein structure without binding site specifications. Since this is not how they were designed to operate, the conventional methods were slow and inaccurate. A preprint by Yu and coworkers at DP Technologies decomposed blind docking into two problems: pocket finding and docking into a predefined pocket. The authors found that DL docking programs excelled at pocket finding but didn't perform as well as conventional methods when pockets are predefined.

Do Deep Learning Models Really Outperform Traditional Approaches in Molecular Docking
<https://arxiv.org/abs/2302.07134>

1.2 Training/test set bias

Most DL docking programs were trained and tested on time splits from the PDB. For instance, DiffDock was trained on structures deposited in the PDB before 2019 and tested on structures deposited in 2019 and later. Quite a few structures in the test set are similar to those in the training set. In these cases, prediction becomes a

simple table lookup. One way to address this bias is to create train/test splits that don't contain similar structures.

A paper by Kanakala and coworkers from IIT analyzed several datasets commonly used for affinity prediction, including [PDBBind](#) and [KIBA](#), and found that typical splitting methods overestimate model performance. The authors propose a clustered cross-validation strategy that provides more realistic estimates of model performance.

Latent Biases in Machine Learning Models for Predicting Binding Affinities Using Popular Data Sets
<https://pubs.acs.org/doi/10.1021/acsomega.2c06781>

A preprint by Li and coworkers from UC Berkeley described a similar effort. The authors cleaned the PDBBind dataset and divided it into segments that minimized leakage between the training and test sets. This new dataset was then used to retrain and evaluate several widely used scoring functions.

Leak Proof PDBBind: A Reorganized Dataset of Protein-Ligand Complexes for More Generalizable Binding Affinity Prediction
<https://arxiv.org/abs/2308.09639>

1.3 Structure quality

The third problem with many DL docking programs is the quality of the generated structures. To put it technically, the structures were really messed up. Bond lengths and angles were off, and there were often steric clashes with the protein. To address these challenges, Buttenschoen and colleagues from Oxford University developed PoseBusters, a Python package for evaluating the quality of docked poses. PoseBusters performs a series of geometry checks on docked poses and also evaluates intra and inter-molecular interactions. The authors used the [Astex Diverse Set](#) and a newly developed [PoseBusters benchmark](#) set to evaluate five popular deep learning docking programs and two conventional docking approaches. The conventional docking programs dramatically outperformed the deep learning methods on both datasets. In most cases, more than half of the solutions generated by the DL docking programs failed the PoseBusters validity tests. In contrast, with the conventional docking programs, only 2-3% of the docked poses failed to validate.

PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences
<https://pubs.rsc.org/en/content/articlepdf/2024/sc/d3sc04185a>

Many of the same problems encountered with DL methods for docking can also impact generative models that produce structures in the context of a protein binding site. A paper by Harris and coworkers from the University of Cambridge describes PoseCheck, a tool similar to PoseBusters, for identifying unrealistic structures. PoseCheck evaluates steric clashes, ligand strain energy, and intramolecular interactions to identify problematic structures. In addition, structures are redocked with AutoDock Vina to confirm the validity of the proposed binding mode. In evaluating several recently published generative models, the authors identify failure modes that will hopefully influence future work on structure-based generative design.

Benchmarking Generated Poses: How Rational is Structure-based Drug Design with Generative Models
<https://arxiv.org/abs/2308.07413>

1.4 Reporting scientific advances in press releases

The other (potentially) significant docking developments in 2023 weren't reported in preprints or papers; they were published in what can best be described as press releases. In early October, the Baker group at the

University of Washington published a short preprint that previews RoseTTAFold All-Atom, the latest incarnation of their RoseTTAFold software for protein structure prediction. In a brief section entitled “Predicting Protein-Small Molecule Complexes”, the authors mention their efforts to generate structures of bound non-covalent and covalent small molecule ligands. On benchmark structures from the [CAMEO](#) blind docking competition, RoseTTAFold All-Atom generated high-quality structures (<2Å RMSD) in 32% of cases. This compared favorably to an 8% success rate for the conventional docking program [AutoDock Vina](#).

Generalized Biomolecular Modeling and Design with RoseTTAFold All-Atom

<https://www.biorxiv.org/content/10.1101/2023.10.09.561603v1.full.pdf>

In late October, the DeepMind group published a blog post entitled “A glimpse of the next generation of AlphaFold,” where, among other things, they made this statement.

“Our latest model sets a new bar for protein-ligand structure prediction by outperforming the best reported docking methods, without requiring a reference protein structure or the location of the ligand pocket — allowing predictions for completely novel proteins that have not been structurally characterized before.”

The accompanying whitepaper provided impressive performance statistics for the PoseBusters set described above. The AlphaFold method achieved a 73.6% success rate compared to 52.3% for the conventional docking program AutoDock Vina. The AlphaFold performance was even more impressive when considering how the comparison was performed. While Vina was provided protein coordinates and a binding site as input, AlphaFold was only given the protein sequence and a SMILES string for the ligand.

A glimpse of the next generation of AlphaFold

<https://deepmind.google/discover/blog/a-glimpse-of-the-next-generation-of-alphafold/>

Performance and structural coverage of the latest, in-development AlphaFold model

https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/a-glimpse-of-the-next-generation-of-alphafold/alphafold_latest_oct2023.pdf

Unfortunately, neither the RoseTTAFold All-Atom preprint nor the DeepMind whitepaper contained any details on the methodology. In addition, at the time I’m writing this, neither group has released the code for their methods. Hopefully, papers with details on the methods will appear soon, along with public code releases. It’s safe to assume that others, such as the OpenFold consortium, who tend to be more forthcoming with their code and methods, are probably working on similar ideas.

Perspective: Like many other areas in AI, DL docking programs began with an initial period of exuberance. The community was excited, and everyone thought the next revolution was imminent. As people started using these methods, they discovered multiple issues that needed to be resolved. We’re not necessarily in the valley of despair, but this is definitely a “hey, wait a second” moment. I’m confident that, with time, these methods will improve. I wouldn’t be surprised to see DL docking methods incorporating ideas from more traditional, physics-inspired approaches. Hopefully, newly developed, unbiased training and test sets and tools like PoseBusters will enable a more rigorous evaluation of docking and scoring methods. With the co-folding approaches in RosettaFold All-Atom and AlphaFold, we’ll have to wait hope for the code to be released so that the community can evaluate the practical utility of these methods.

2. Can we use AlphaFold2 structures for ligand discovery and design?

2.1 Experimentally evaluating AlphaFold2 structures

Since it took the CASP14 competition by storm in 2020, AlphaFold2 has greatly interested people involved in drug discovery and numerous other fields. In addition to benchmark comparisons with the PDB, there have been several other efforts to evaluate the structural models generated by AF2 experimentally. Rather than simply comparing the atomic coordinates of AF2 structures with corresponding PDB structures, a paper by Terwilliger and colleagues from Lawrence Livermore National Labs compares AF2 structures with reported crystallographic electron density maps. The authors argue that this approach puts less weight on loops and sidechains that are poorly resolved experimentally. They found that prediction accuracy varied across individual structures, and regions with prediction score (pLDDT) > 90 varied by less than 0.6Å from the deposited model. They suggest that even inaccurate regions of AF2 structures can provide plausible hypotheses for experimental refinement.

AlphaFold predictions are valuable hypotheses, and accelerate but do not replace experimental structure determination

<https://www.nature.com/articles/s41592-023-02087-4>

A paper by McCafferty and coworkers from UT Austin used mass spec data from protein cross-linking experiments to evaluate the ability of AF2 to model intracellular protein conformations. The authors compared experimentally observed distances in cross-linked proteins from eukaryotic cilia with corresponding distances from AF2 structures and found an 86% concordance. In 42% of cases, all distances within the predicted structure were consistent with those observed in cross-linking experiments.

Does AlphaFold2 model proteins' intracellular conformations? An experimental test using cross-linking mass spectrometry of endogenous ciliary proteins

<https://www.nature.com/articles/s42003-023-04773-7>

2.2 Generating multiple protein conformations with AlphaFold2

In 2023, there was great interest in AF2's ability to generate multiple relevant protein conformations. A paper by Wayment-Steele and coworkers showed that clustering the multiple sequence alignment (MSA) used by AF2 enabled the program to generate multiple relevant protein conformations.

Predicting multiple conformations via sequence clustering and AlphaFold2

<https://www.nature.com/articles/s41586-023-06832-9>

These ideas have spurred additional investigations and stirred up a bit of controversy. A paper by Chakravarty and coworkers from NCBI and NIH examined the performance of AF2 on 93 fold-switching proteins. The authors found that AF2 only identified the switched conformation in 25% of the proteins in the AF2 training set and 14% of proteins not in the training set.

AlphaFold2 has more to learn about protein energy landscapes

<https://www.biorxiv.org/content/10.1101/2023.12.12.571380v1>

Wayment-Steele and coworkers proposed that their clustering of the MSAs captured the coevolution of related proteins. A subsequent preprint from Porter and coworkers at NCBI challenged this assumption and demonstrated that multiple protein conformations could be generated from single sequences.

ColabFold predicts alternative protein structures from single sequences, coevolution unnecessary for AF-cluster
<https://www.biorxiv.org/content/10.1101/2023.11.21.567977v2>

2.3 Docking into AlphaFold2 structures

After the publication of the AF2 paper and the subsequent release of the code, many groups began experiments to determine whether structures generated by AF2 and related methods could be used for ligand design. The initial results weren't promising. Díaz-Rovira and coworkers from the Barcelona Supercomputing Center compared virtual screens using protein-crystal structures and structures predicted by AF2 for 11 proteins. The authors found that the average enrichment factor at 1% for the x-ray structures was double that of the AF2 structures.

Are Deep Learning Structural Models Sufficiently Accurate for Virtual Screening? Application of Docking Algorithms to AlphaFold2 Predicted Structures
<https://pubs.acs.org/doi/10.1021/acs.jcim.2c01270>

Holcomb and coworkers from Scripps took a different approach and compared the performance of AutoDockGPU on AF2 structures with the performance on corresponding crystal structures from the PDBBind set. The authors noted a significant loss in docking accuracy with the AF2 structures. AutoDockGPU generated poses within 2Å of the experiment in 41% of the cases when docking into the crystal structures. This success rate dropped to 17% for the AF2 structures. On a brighter note, the authors reported that the docking success rate for AF2 structures was better than with corresponding apo structures.

Evaluation of AlphaFold2 structures as docking targets
<https://onlinelibrary.wiley.com/doi/full/10.1002/pro.4530>

A paper by Karelina and coworkers from Stanford examined the utility of AF2 for modeling the structures of GPCRs. While the authors found that AF2 could model structures and binding pockets with high fidelity, the docking performance of the models was poor. The results of this study were consistent with those in the papers described above. In this case, the success rate for docking into AF2 structures (16%) was less than half of that for experimentally determined structures (48%). As mentioned above, it was encouraging that the docking performance of the AF2 structures was better than that of structures with other ligands bound.

How accurately can one predict drug binding modes using AlphaFold models?
<https://www.biorxiv.org/content/10.1101/2023.05.18.541346v2>

While the results in the papers above aren't encouraging, all hope may not be lost. In the last week of 2023, there was a paper from Brian Shoichet, Bryan Roth, and coworkers that reported successful prospective virtual screening results with AF2 structures of the sigma2 and 5-HT2A receptors. The odd bit here is that while the AF2 model performed well prospectively, its retrospective performance on prior screens of the same targets wasn't good. To demonstrate that they got the right answer for the right reason, the authors solved a cryoEM structure of one of the 5-HT2A agonists and found that the docked pose was consistent with the experimental structure. The authors suggest that AF2 structures may sample the underlying manifold of conformations and posit that retrospective screening studies such as those described above may not predict prospective performance.

AlphaFold2 structures template ligand discovery
<https://www.biorxiv.org/content/10.1101/2023.12.20.572662v1>

Many earlier papers describing the use of AF2 structures for docking suggested that performance could be improved by refining the predicted structures. Zhang and coworkers at Schrödinger compared virtual screening performance using holo structures, apo structures, and AF2 structural models. The authors compared virtual screening performance across 27 targets from the DUD-E set and found that the enrichment factor at 1% (EF1%) on AF2 structures (13%) was similar to that for apo structures (11%). However, EF1% increased to 18% when the AF2 structures were refined using induced fit docking.

Benchmarking Refined and Unrefined AlphaFold2 Structures for Hit Discovery

<https://pubs.acs.org/doi/10.1021/acs.jcim.2c01219>

Perspective: The publication of the AF2 paper and subsequent release of the code has sparked work in numerous areas. There are already more than 17,000 references to the original AF2 paper. Protein structure prediction has become an integral component of experimental structural biology. Programs like Phenix can generate AF2 structures that can subsequently be fit to experimental data. While there is still work to be done, AF2 may be capable of generating ensembles of relevant protein conformations. It's exciting to think about how this work will progress as we achieve tighter integration between protein structure prediction and physics-based modeling. It currently appears that the jury is still out on the utility of predicted protein structures for drug design. While the results of retrospective evaluations are somewhat disappointing, the recent prospective success from the Shoichet and Roth labs is encouraging.

3. Benchmarking

I spent a lot of time this year ranting about benchmarks. I [highlighted](#) severe flaws in commonly used datasets like [MoleculeNet](#) and the [Therapeutics Data Commons](#) (TDC). In a [second rant](#), I bemoaned the lack of statistical analysis in most papers comparing ML methods and molecular representations. Fortunately, there were some rays of sunlight within the dark clouds. Here are a few benchmarking papers pushing the field in the right direction.

3.1 Overviews

Two recent papers provide insight into some of the challenges associated with benchmarking. A preprint by Green and coworkers from DeepMirror, provides an excellent overview of the field and some factors that complicate current benchmarking efforts. The authors compared several molecular representations and ML algorithms in evaluating model accuracy and uncertainty. These evaluations highlighted the strengths of different QSAR modeling and ADME prediction methods. Consistent with other papers published in 2023, 2D descriptors performed best for ADME prediction, while Gaussian Process Regression with fingerprints was the method of choice when predicting biological activity.

Current Methods for Drug Property Prediction in the Real World

<https://arxiv.org/abs/2309.17161>

A paper by Janela and Bajorath outlines several limitations in current benchmarking strategies. The authors used sound statistical methodologies to examine the impact of compound potency value distributions on performance metrics associated with regression models. They found that across several different ML algorithms, there was a consistent relationship between model performance and the activity range of the dataset. These findings enabled the authors to define bounds for prediction accuracy. The method used in this paper should be informative to those designing future benchmarks.

Rationalizing general limitations in assessing and comparing methods for compound potency prediction
<https://www.nature.com/articles/s41598-023-45086-3>

3.2 Benchmark comparisons

Three benchmarking papers stood out for me in 2023. These papers check a couple of critical boxes.

- For the most part, the authors used high-quality datasets. In a couple of cases, the papers included some of the MoleculeNet and TDC datasets. However, when these datasets were used, the authors did additional curation to clean up some dataset errors. It was nice to see the paper by Deng and coworkers (see below) point out the folly of trying to predict endpoints in the MoleculeNet ClinTox and SIDER datasets based on chemical structures.
- The authors used statistical tests (cue hallelujah chorus) to determine where method performance was different and where it wasn't.

After numerous papers claiming learned representations and foundation models were the current state of the art, it was refreshing to see careful studies showing this is not the case. In all three papers, the best-performing methods used good old fingerprints and 2D descriptors coupled with gradient-boosting or support vector machines (SVM).

A paper from Fang and coworkers at Biogen introduced several new ADME datasets. Unlike most literature benchmarks, which contain data collected from dozens of papers, these experiments were consistently performed by the same people in the same lab. The authors provided prospective comparisons of several widely used ML methods, including random forest, SVM, XGBoost, LightGBM, and message-passing neural networks (MPNNs) on several relevant endpoints, including aqueous solubility, metabolic stability, membrane permeability, and plasma protein binding.

Prospective Validation of Machine Learning Algorithms for Absorption, Distribution, Metabolism, and Excretion Prediction: An Industrial Perspective
<https://pubs.acs.org/doi/10.1021/acs.jcim.3c00160>

One of my favorite papers of 2023 provided a tour de force in method comparison. Deng and coworkers from Stony Brook University compared many popular ML algorithms and representations, curated new datasets, and performed statistical analysis on the results. My only complaint about this paper is that there may have been too much information presented. Each figure contains dozens of subfigures, and it's easy to get overwhelmed and miss the overall message. The authors used some of the MoleculeNet datasets that I'm not fond of, but they also point out some of the limitations of these datasets. Ultimately, this paper provides one of the best comparisons of ML methods published to date. The authors compare fixed representations, such as molecular fingerprints, with representations learned from SMILES strings and molecular graphs and conclude that, in most cases, the fixed representations provide the best performance. Another interesting aspect of this paper was an attempt to establish a relationship between dataset size and the performance of different molecular representations. While fixed representations performed well on smaller datasets, learned representations didn't become competitive until between 6K and 100K datapoints were available.

A systematic study of key elements underlying molecular property prediction
<https://www.nature.com/articles/s41467-023-41948-6>

A preprint by Kamuntavičius and coworkers at Ro5 uses several widely used ADME datasets to examine the performance of molecular representations, both individually and in combination. The authors found that when

tested individually, the RDKit 2D descriptors outperformed fingerprints and representations derived from language models. When examining feature combinations, they found that performance was highly dataset dependant.

Benchmarking ML in ADMET predictions: A Focus on Hypothesis Testing Practices

<https://chemrxiv.org/engage/chemrxiv/article-details/6578c39fbec7913d2774d6e6>

3.3 Dataset splitting

In many cases, cross-validation of ML models is performed by randomly splitting a dataset into training and test sets. As many have argued, these random splits can provide overly optimistic estimates of model performance. More recently, many groups have moved away from random splits and are using splits that avoid putting the same scaffold in the training and test sets. While this is an improvement, several subtle issues can confound scaffold splits. To better predict prospective performance, several groups have developed alternate methods for splitting molecular datasets.

A preprint by Tossou and coworkers at Valence Labs (now Recursion) examined several approaches to estimating the performance of ML models in a real-world deployment. The authors evaluated the impact of molecular representations, algorithms, and splitting strategies on the generalization abilities of ML models. In another victory for “classic” algorithms, the authors found that Random Forest was the best option for out-of-domain (OOD) classification, regression, and uncertainty calibration. When comparing representations, they found that 2D and 3D descriptors provided the best uncertainty estimates, while fingerprints provided the best generalization. In a comparison of splitting methods, the authors found that scaffold splits provided the best estimates of generalization, while maximum dissimilarity and random splits provided the best uncertainty estimates.

Real-World Molecular Out-Of-Distribution: Specification and Investigation

<https://chemrxiv.org/engage/chemrxiv/article-details/64c012a1b053dad33ae21932>

Diverse datasets used for hit identification differ significantly from the congeneric datasets encountered during lead optimization. As such, we should tailor our splitting strategies to the task. A preprint from Steshin discusses the differences in hit identification (Hi) and lead optimization (Lo) datasets and proposes different benchmarking strategies for each. When employing a scaffold split on a Hi dataset, the author found that many test set molecules had a neighbor in the training set with an ECFP Tanimoto similarity greater than 0.4. To address this limitation, Integer Linear Programming was used to develop a specific method for splitting. For the Lo benchmarks, a clustering strategy was used to divide the data into training and test sets. The author also provides a GitHub repository with the software for generating the splits and preprocessed versions of common (and unfortunately flawed) benchmarks.

Lo-Hi: Practical ML Drug Discovery Benchmark

<https://arxiv.org/abs/2310.06399>

As mentioned above, many commonly used dataset-splitting methods allow information from the training set to leak into the test set. When benchmarking docking or activity prediction models, it has been common to split sets of protein-ligand structures from the PDB based on the structure deposition date. Structures deposited before a specific date are used for training, and those deposited after that date are used for validation and/or testing. Unfortunately, this typically results in a test set that contains many structures that are almost identical to those in the training set. Similar issues can impact datasets used for QSAR or ADME modeling. To overcome some of these issues, Joeres and coworkers at the Helmholtz Institute for Pharmaceutical Research Saarland

developed Data Splitting Against Information Leakage (DataSAIL). This data-splitting method uses Binary Linear Programming to minimize the overlap between training and test sets. The authors demonstrate that their method scales better than the LoHi splitter, which can bog down when the dataset size approaches 100K.

DataSAIL: Data Splitting Against Information Leakage
<https://www.biorxiv.org/content/10.1101/2023.11.15.566305v1.abstract>

As seen above, several methods and software packages exist for molecular dataset splitting. Keeping up with work in the field and installing and learning to apply new methods can be time-consuming. To simplify this process, Burns and coworkers at MIT developed *astartes*, a Python package that aspires to be the Swiss army knife of dataset splitting. The *astartes* package currently supports more than a dozen splitting methods using a simple syntax that will be familiar to scikit-learn users.

Machine Learning Validation via Rational Dataset Sampling with *astartes*
<https://joss.theoj.org/papers/10.21105/joss.05996>

3.4 New datasets

2023 saw the appearance of a few valuable new benchmark datasets. As mentioned above, the Biogen ADME dataset provides a high-quality, consistently measured collection of ADME datasets that will hopefully become a standard for the field.

Prospective Validation of Machine Learning Algorithms for Absorption, Distribution, Metabolism, and Excretion Prediction: An Industrial Perspective
<https://pubs.acs.org/doi/10.1021/acs.jcim.3c00160>

While activity cliffs, where small changes in chemical structure bring about large changes in properties or biological activity, are frequently encountered in drug discovery, they are rarely present in benchmark datasets. A paper by van Tilborg and coworkers from the Eindhoven University of Technology set out to remedy this by creating MoleculeACE, a series of datasets designed to evaluate the performance of ML models on data containing activity cliffs. The authors evaluated a range of ML models and representations and found a high correlation between overall performance and the performance on activity cliffs in 25 of the 30 datasets studied.

Exposing the Limitations of Molecular Machine Learning with Activity Cliffs
<https://pubs.acs.org/doi/10.1021/acs.jcim.2c01073>

Perspective: While there were a few benchmarking papers published in 2023 that followed best practices, there were a lot more that didn't. As a field, we must reach a consensus on appropriate datasets and statistical tests for method comparisons. It's great to see so many groups looking at topics like dataset splitting. In the coming year, I hope statistical approaches to comparing methods receive equal attention.

Other Chemical Information News

Contribution from Stuart Newbold, email: stuart@psandim.com

Opinions and Tips on AI & Chemistry— Chemistry Advent Calendar

The intersection of chemistry and artificial intelligence (AI) is a fascinating area that attracts a lot of attention in both research and industry. We talked to people working in the field about the potential of AI to revolutionise chemical research, but also concerns, (current) limitations, and ethical implications for chemical applications. We also asked for ideas to try or experiment with, as well as useful articles and videos for beginners and advanced users. These new mini interviews will be available until 24th December 2023.

<https://www.chemistryviews.org/opinions-on-ai-chemistry/>

Source: *ChemistryViews*

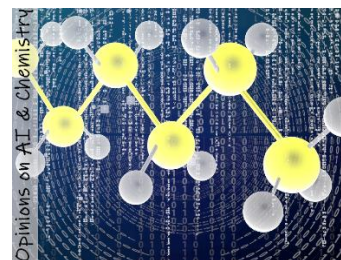


Image courtesy of
ChemistryViews

5 Challenges Derailing Research in the Chemical Industry (and how to fix them!)

Chemical-industry employees need to stay on top of emerging market demands with access to the latest scientific literature. Without a strategy to manage this information, barriers tend to emerge. Here are some of the most common challenges related to content that we hear among chemical industry researchers, with our tips for how to solve them.

https://www.copyright.com/wp-content/uploads/2022/03/CCC_5-Challenges_Chemical-Industry_Tip-Sheet.pdf

Source: *CCC*

AI Programs spat out known data and hardly learned specific Chemical Interactions when predicting Drug Potency

Take a look behind the scenes of machine learning in drug research.

<https://www.sciencedaily.com/releases/2023/11/231113141325.htm>

Source: *Science Daily*

Springer Nature introduces Curie, the AI-powered Scientific Writing Assistant

Springer Nature has announced a new AI-powered in-house writing assistant to support researchers, particularly those whose first language is not English, in their scientific writing. Global research shows that it takes non-native English-speaking scientists 51% more time to write a paper. This creates an unequal divide in research, limiting the advancement of knowledge and impacting the submission of high-quality research from across the globe.

Curie has been specifically trained in academic literature, spanning 447+ areas of study, more than 2,000 field-specific topics, and over 1 million edits on papers published including those in leading Nature journals. It combines the power of large language models (LLMs) with specialised AI digital editing developed in-house and designed specifically for scientific writing. Unlike generalist AI writing apps, Curie focuses on the unique pain points of researchers in their professional writing, including translation to English and English language editing to address grammatical errors and improve phrasing and word choice.

<https://www.knowledgespeak.com/news/springer-nature-introduces-curie-the-ai-powered-scientific-writing-assistant/>

Source: *Knowledgespeak*

Detector for AI-Generated Text in Chemistry Papers

Text generators based on artificial intelligence (AI), such as ChatGPT, pose new challenges for scientific journals. Many journals require disclosures of the use of ChatGPT in writing the manuscript, and often, they ban naming ChatGPT as a non-human author. So far, such AI text generators often make up facts that are not true, which can be an issue for the integrity of the scientific record if they are used without proper care. Methods for detecting if AI has been used in preparing a scientific paper are, thus, important. Existing tools for this can have drawbacks such as a bias against writers who are non-native English speakers.

<https://www.chemistryviews.org/detector-for-ai-generated-text-in-chemistry-papers/>

Source: *ChemistryViews*



Image courtesy of
ChemistryViews

ACS Publications provides a new option to support Zero-Embargo Green Open Access

From 1st October, the Publications Division of ACS has begun to provide authors with a new option to satisfy funder requirements for zero-embargo green open access. Through this pathway, authors are able to post accepted manuscripts with a CC BY license in open access repositories immediately upon acceptance. To ensure a sustainable model of delivering services from submission to final editorial decision, ACS Publications has introduced an article development charge (ADC) as part of this new zero-embargo green open access option.

<https://www.stm-publishing.com/acs-publications-provides-a-new-option-to-support-zero-embargo-green-open-access/>

Source: *STM Publishing*

Hypotheses devised by AI could find 'Blind Spots' in Research

Artificial intelligence is asking questions that humans hope to answer.

<https://www.nature.com/articles/d41586-023-03596-0>

Source: *Nature*

Elsevier takes Scopus to the Next Level with Generative AI

New generative AI feature will power Scopus, world's leading abstract and citation database of peer-reviewed literature, so researchers can get deeper insights faster.

<https://www.elsevier.com/about/press-releases/elsevier-takes-scopus-to-the-next-level-with-generative-ai>

Source: *Elsevier*

A Cautionary Tale for AI in Small Molecule Drug Discovery

Despite the buzz around artificial intelligence (AI), most industry insiders know that the use of machine learning (ML) in drug discovery is nothing new. For more than a decade, researchers have used computational techniques for many purposes, such as finding hits, modelling drug-protein interactions, and predicting reaction rates. A new webcast supported by Dotmatics explains more.

<https://www.scientific-computing.com/webcast/cautionary-tale-ai-small-molecule-drug-discovery>

Source: *Scientific Computing World*



Image courtesy of SCW

Chat-based AI “not enough for Scientific Research”

Study says there is a clear demand to move away from chat AI towards “interactive AI” from scientific researchers.

<https://www.researchinformation.info/news/chat-based-ai-not-enough-scientific-research>

Source: *Research Information*

Landmark Intellectual Property Ruling could offer new opportunities for Chemists working with AI

The UK intellectual property office has passed a judgement declaring an artificial neural network (ANN) patentable for the first time. This decision could set a powerful precedent enabling greater protection of AI technologies, which could provide both new commercial opportunities for chemistry AIs and attract investment into this growing UK sector.

<https://www.chemistryworld.com/news/landmark-intellectual-property-ruling-could-offer-new-opportunities-for-chemists-working-with-ai/4018633.article>

Source: *Chemistry World*

Springer Nature introduces AI-powered Scientific Writing Assistant

Tool to support researchers, particularly those whose first language is not English, in their scientific writing.

<https://www.researchinformation.info/news/springer-nature-introduces-ai-powered-scientific-writing-assistant>

Source: *Research Information*

ANSRs to Hard AI Questions

DARPA selected the following teams to explore diverse, hybrid architectures that integrate data-driven machine learning with symbolic reasoning, a problem-solving method that uses symbols or abstract representations to understand information or follow rules to reach conclusions.

<https://www.darpa.mil/news-events/2023-09-25>

Source: *DARPA News*

How a Methods Database increases Lab Efficiency

Pistoia's Dr Birthe Nielsen and Merck's Pankaj Aggarwal discuss the Methods Hub project and its impact on scientific research.

<https://www.scientific-computing.com/article/how-methods-database-increases-lab-efficiency>

Source: *Scientific Computing World*

In Musk Era, Pharma Companies move Ads away from X

Even before its billionaire owner made recent headlines by endorsing an antisemitic post and hurling an expletive at advertisers, biopharma companies had begun to find different uses for their advertisement budgets.

<https://www.biospace.com/article/in-musk-era-pharma-companies-move-ads-away-from-x/>

Source: *Biospace*

The Chemists creating Knowledge-sharing Websites

Speeding up scientific progress by sharing organic chemistry techniques, lab safety resources and computational procedures.

<https://www.chemistryworld.com/careers/the-chemists-creating-knowledge-sharing-websites/4018415.article>

Source: *Chemistry World*

A Map of every Conceivable Molecule could be possible with AI

A map of all chemicals that places compounds with similar properties next to each other could speed up the process of discovery for everything from drugs to materials.

<https://www.newscientist.com/article/2388562-a-map-of-every-conceivable-molecule-could-be-possible-with-ai/>

Source: *New Scientist*

ResearchGate and AAAS announce new Journal Home partnership for Science Partner Journals

ResearchGate and AAAS have announced a new partnership that will see all AAAS Science Partner Journals benefit from enhanced visibility and reach through ResearchGate's new Journal Home offering. AAAS, renowned for its Science family of journals, launched its Science Partner Journal (SPJ) program in 2017. Consisting of 14 high-quality, fully open-access journals produced in collaboration with international research institutions, foundations, funders, and societies, the SPJ program will now expand its reach through Journal Home on ResearchGate. With Journal Home, all version-of-record content from the 14 SPJs will be available to researchers on ResearchGate, including all archive content, and all new articles on publication. Reading usage data will be consistently provided to AAAS via COUNTER-compliant reporting that can be included in institutional usage reporting, providing increased value for institutional customers.

<https://www.knowledgespeak.com/news/researchgate-and-aaas-announce-new-journal-home-partnership-for-science-partner-journals/>

Source: *Knowledgespeak*

Enabling Remote Operations

Many routine operations and maintenance tasks can be performed remotely.

<https://www.scientific-computing.com/white-paper/enabling-remote-operations>

Source: *Scientific Computing World*

£300 Million to launch First Phase of new AI Research Resource

AIRR, a cluster of advanced computers for AI research, has received a £300 million investment, to include a new Cambridge-based supercomputer.

<https://www.ukri.org/news/300-million-to-launch-first-phase-of-new-ai-research-resource/>

Source: *UKRI*

Sage launches Literacy Information Microsite for Combatting Mis-, Dis-, and Malinformation

Sage has unveiled a literary resource hub aimed at countering the rise of online misinformation, disinformation, and deceptive content.

<https://www.infotoday.eu/Articles/News/Featured-News/Sage-Launches-Literacy-Information-Microsite-for-Combatting-Mis--Dis--and-Malinformation-161274.aspx>

Source: *Information Today*

Introducing the Hybrid Open Access Dashboard (HOAD)

In a significant stride towards advancing open access in academia, the Hybrid Open Access Dashboard (HOAD) has been unveiled as an innovative data analytics tool, aimed at empowering academic libraries and consortia. Developed by the State and University Library Göttingen and supported by funding from the German Research Foundation (DFG), HOAD ingeniously integrates open data from Crossref, OpenAlex, and the cOAlition S Journal Checker Tool to visually demonstrate the evolving shift of hybrid journal portfolios within transformative agreements towards full open access.

<https://www.knowledgespeak.com/news/introducing-the-hybrid-open-access-dashboard-hoad/>

Source: *Knowledgespeak*

Creating better Treatments with AI

The increasing use of AI promises to revolutionise drug discovery, but only if laboratories can organise their data, writes Robert Roe.

<https://www.scientific-computing.com/article/creating-better-treatments-ai>

Source: *Scientific Computing World*

Springer Nature strengthens AI capabilities with the Acquisition of Slimmer AI's Science Division

In a bid to fortify its technological prowess and enhance scientific publishing, Springer Nature has formally sealed an agreement to acquire the Science division of Slimmer AI (S-AI), a Netherlands-based AI company, in a landmark deal that underscores the growing significance of AI in academia.

<https://www.knowledgespeak.com/news/springer-nature-strengthens-ai-capabilities-with-the-acquisition-of-slimmer-ais-science-division/>

Source: *Knowledgespeak*

ACS Publications joins SDG Publishers Compact

The Publications Division of the American Chemical Society (ACS) has joined the [Sustainable Development Goals](#) (SDGs) Publishers Compact, and will invest up to \$50 million over the span of five years on four initiatives.

<https://www.researchinformation.info/news/acs-publications-joins-sdg-publishers-compact>

Source: *Research Information*

Farewell to Static Print and PDF Articles?

Individual sections of scientific publications, such as introductions, consistently follow the same narrative structure. AI tools like ChatGPT are capable of generating such texts. So, Michael Bojdys, Humboldt-Universität Berlin, Germany, asks: why do we persist in creating, publishing, and reviewing such narratives? And who actually reads all these texts?

<https://www.chemistryviews.org/farewell-to-static-print-and-pdf-articles/>

Source: *Chemistry Views*

PubHive launches Local Literature Dashboard & Tailoring Workflows

PubHive Ltd has announced the launch of its latest feature, the Local Literature Dashboard and Tailoring Workflows. This new functionality gives drug safety and pharmacovigilance local literature users a more personalised and efficient way to access and manage their local literature resources.

<https://www.stm-publishing.com/pubhive-launches-local-literature-dashboard-tailoring-workflows/>

Source: *STM Publishing*

Nearly 400,000 new Compounds added to Open-Access Materials Database

The contribution grows the open-access resource that scientists use to invent new materials for future technologies.

<https://www.sciencedaily.com/releases/2023/11/231129112351.htm>

Source: *Science Daily*

STEM employers pay Scientists with Disabilities up to \$14,360 less than those without Disabilities

Ableist attitudes in STEM responsible for pay disparities, researcher says.

<https://cen.acs.org/careers/salaries/STEM-employers-pay-scientists-disabilities/101/i40>

Source: *Chemical & Engineering News*

Intersection of AI & Copyright

This page serves as a resource for information on the responsible development and use of AI technologies with copyright-protected content.

<https://www.copyright.com/resource-library/insights/intersection-ai-copyright/>

Source: CCC

Crossref acquires Retraction Watch database, unveiling a new era of Research Integrity

In a significant development for the scientific community, [Crossref](#), the global infrastructure supporting research communications, has acquired the [Retraction Watch](#) database, a renowned resource for tracking retractions in academic publishing. The announcement, made jointly by Crossref and the Center for Scientific Integrity, the organisation responsible for the Retraction Watch blog and database, signals a collaborative effort to enhance transparency and trustworthiness in scholarly outputs.

<https://www.knowledgespeak.com/news/crossref-acquires-retraction-watch-database-unveiling-a-new-era-of-research-integrity/>

Source: Knowledgespeak

Springer Nature Developing new Peer Review Platform

Snapp 'a key investment in the future of publishing', which has today reached the milestone of one million submissions

<https://www.researchinformation.info/news/springer-nature-developing-new-peer-review-platform>

Source: Research Information

Scientists Use Quantum Biology, AI to Sharpen Genome Editing Tool

Scientists used their expertise in quantum biology, artificial intelligence and bioengineering to improve how CRISPR Cas9 genome editing tools work on organisms.

<https://www.sciencedaily.com/releases/2023/11/231109141444.htm>

Source: Science Daily

Research Integrity at your Fingertips with new World-leading Dimensions App

Digital Science company Dimensions has launched its new Dimensions Research Integrity app, enabling users to ensure the highest standards of research integrity and helping to build global trust in research. The Dimensions Research Integrity app uses AI to track the presence of several Trust Markers across tens of millions of publications worldwide.

<https://www.stm-publishing.com/research-integrity-at-your-fingertips-with-new-world-leading-dimensions-app/>

Source: STM Publishing

Revealed: the 50 new Technologies that could Shape the Future

A real-life invisibility cloak and worm-like robotics for disaster recovery are just some of the 50 emerging technologies that could shape our future.

<https://www.ukri.org/news/revealed-the-50-new-technologies-that-could-shape-the-future/>

Source: UKRI

Why Postdocs need Entrepreneurship Training

Landing tenure is a pipe dream for most postdoctoral researchers. They need business skills to help them thrive outside academia.

<https://www.nature.com/articles/d41586-023-03411-w>

Source: Nature

EBSCO Information Services pursues Generative Artificial Intelligence Opportunities

Company embarks on projects to determine how GenAI can enhance search discovery and content creation while avoiding “hallucinations” and spurious information.

<https://www.infotoday.eu/Articles/News/Featured-News/EBSCO-Information-Services-pursues-Generative-Artificial-Intelligence-opportunities-160767.aspx>

Source: *Information Today*

Cambridge hits £1 Billion Revenue for first time

More than 114 million research papers, book chapters and scholarly materials were downloaded.

<https://www.researchinformation.info/news/cambridge-hits-1-billion-revenue-first-time>

Source: *Research Information*

Elements of Responsible AI for the Library Community

Beth Rudden delivered an inspiring keynote talk about moving responsible AI forward in our communities at Internet Librarian Connect (ILC), the virtual conference held on 16th-19th October 2023. She concentrated on describing the AI revolution, giving a practical guide to AI experimentation, and stressing the need for librarians to be courageous stewards.

<https://www.infotoday.eu/Articles/Editorial/Featured-Articles/Elements-of-responsible-AI-for-the-library-community-161284.aspx>

Source: *Information Today*

How to Switch Research Fields Successfully

Four researchers offer tips for excelling in interdisciplinary training.

<https://www.nature.com/articles/d41586-023-03337-3>

Source: *Nature*

An artificial Nose that sniffs like a Wine Taster

Precisely copying the capabilities of a biological nose with an artificial one is a lofty but potentially world-changing goal.

<https://www.advancedsciencenews.com/an-artificial-nose-that-sniffs-like-a-wine-taster/>

Source: *Advanced Science News*

Cutting-Edge Chemistry will help new Laundry Detergents Clean-up

New tools harnessing advances in fundamental chemistry are aiding the development of detergents that will deliver cleaner clothes with a lower carbon footprint.

<https://www.ukri.org/news/cutting-edge-chemistry-will-help-new-laundry-detergents-clean-up/>

Source: *UKRI*

“Heartbreaking” to read how Researchers feel – Octopus Founder

New report questions impact of publishing on research culture.

<https://www.researchinformation.info/news/heartbreaking-read-how-researchers-feel-octopus-founder>

Source: *Research Information*

Five years of Plan S: A Journey towards Full and Immediate Open Access

In September 2018, a transformative initiative known as Plan S took shape, driven by a coalition of national research funding organisations and backed by the European Commission. This visionary collective, cOAlition S, embarked on a mission to make research publications openly accessible to all, with the adoption of ten key principles designed to catalyse the rapid transition to full and immediate Open Access. While the policies and

tools to support Plan S were primarily implemented in 2021, the impact of these initiatives will continue to unfold over several years.

<https://www.knowledgespeak.com/news/five-years-of-plan-s-a-journey-towards-full-and-immediate-open-access/>

Source: Knowledgespeak

How Will AI in Hiring be Regulated?

With 88% of life sciences organisations using or planning to use AI in recruitment and/or hiring, AI regulation is a priority for the industry.

<https://www.biospace.com/article/how-will-ai-in-hiring-be-regulated-/?s=120>

Source: Biospace

Over 40% of Cell Press Papers now include an Inclusion and Diversity Statement

<https://www.stm-publishing.com/over-40-of-cell-press-papers-now-include-an-inclusion-and-diversity-statement/>

Source: STM Publishing

Two-Dimensional Compounds can Capture Carbon from the Air

MXene and MBene compounds hold promise for new technologies to combat climate change.

<https://www.sciencedaily.com/releases/2023/10/231004201937.htm>

Source: Science Daily

Rethinking Chemistry – Talking with Presidents of Chemical Societies

Compilation of video statements by Presidents of chemical societies (ACS, GÖCH, RACI, RSC, SCI) on the GDCh's motto "Rethinking Chemistry".

<https://www.chemistryviews.org/rethinking-chemistry-talking-with-presidents-of-chemical-societies/>

Source: Chemistry Views

Clarivate Accurately forecasts four new 2023 Nobel Prize Laureates

Clarivate correctly predicted the four new Nobel Laureates who were Nobel Prize recipients. Ferenc Krausz, Mounqi G. Bawendi, Louis E. Braus (both for chemistry) and Claudia Goldin were named to the Citation Laureates™ list from Clarivate™ several years before being recognised by the Nobel Assembly.

<https://www.stm-publishing.com/clarivate-accurately-forecasts-four-new-2023-nobel-prize-laureates/>

Source: STM Publishing

2023 Information Seeking and Consumption Study

Using the latest data from Outsell, Inc., this study offers insights into how people think and behave in the context of copyrighted content consumption, use, and sharing, and presents analysis and recommendations to organisations that depend on published content.

<https://www.copyright.com/resource-library/insights/outsell/>

Source: CCC

Integrating AI in Life Sciences to change employee behavior with Microsoft and IQVIA

In this episode, you can hear from senior leaders at Microsoft and IQVIA to get their take on how generative AI is impacting productivity, employee engagement and how to mitigate risks.

<https://www.biospace.com/article/integrating-ai-in-life-sciences-to-change-employee-behavior-with-microsoft-and-iqvia/?s=120>

Source: Biospace

EBSCO Information Services pursues Generative Artificial Intelligence (AI) Opportunities

EBSCO is embracing the power of generative AI. Recognising the transformative potential of generative AI in the realm of academic research and libraries, EBSCO is making proactive strides to incorporate AI into the company's products, undertaking AI pilot projects in specific environments with the goal of amplifying the effectiveness of research.

<https://www.stm-publishing.com/ebsco-information-services-pursues-generative-artificial-intelligence-ai-opportunities/>

Source: *STM Publishing*

So, what is (and Isn't) protected by Copyright?

Understanding the extent to which materials are copyright protected can help you minimise the risk of infringement by well-intentioned employees.

<https://www.copyright.com/wp-content/uploads/2023/03/What-is-isnt-Protected-by-Copyright.pdf>

Source: CCC

Science Funding Agencies reject AI for Peer Review

In a recent turn of events, science funding agencies, including the NIH and the Australian Research Council (ARC), have decided to prohibit the use of artificial intelligence tools for peer review. The decision came after concerns were raised about the potential risks and drawbacks associated with employing AI-generated critiques for evaluating research proposals.

<https://www.knowledgespeak.com/news/science-funding-agencies-reject-ai-for-peer-review/>

Source: *Knowledgespeak*

Digital Science announces Brand Redesign for ReadCube and Papers

Digital Science is excited to announce that ReadCube, a leader in Literature Management, is unveiling a comprehensive repositioning of the brand.

<https://www.stm-publishing.com/digital-science-announces-brand-redesign-for-readcube-and-papers/>

Source: *STM Publishing*

Diamond Light Source launches £500m Upgrade Programme

Diamond-II is a £519 million investment by government, predominantly from the UK Research and Innovation (UKRI) Infrastructure Fund and the Wellcome Trust.

<https://www.ukri.org/news/visit-to-diamond-light-source-launches-500m-upgrade-programme/>

Source: *UKRI*

Chemists are teaching GPT-4 to do Chemistry and control Lab Robots

Augmenting the artificial intelligence GPT-4 with extra chemistry knowledge made it much better at planning chemistry experiments, but it refused to make heroin or sarin gas.

<https://www.newscientist.com/article/2370923-chemists-are-teaching-gpt-4-to-do-chemistry-and-control-lab-robots/>

Source: *New Scientist*

ACS launches new Grants and Awards to support Sustainability practices

The ACS has announced two new grants and three new awards that align with the ACS Board of Directors' Campaign for a Sustainable Future, which was announced in 2022. The campaign aims to advance chemistry innovations to address the challenges articulated in the U.N. Sustainable Development Goals.

<https://www.stm-publishing.com/acs-launches-new-grants-and-awards-to-support-sustainability-practices/>

Source: *STM Publishing*

AI Forward Recap Q&A

Earlier this year, DARPA's Information Innovation Office (I2O) announced plans for AI Forward – the agency's latest initiative to reimagine the future of AI research that will result in trustworthy systems for national security missions. Approximately 200 participants from across the commercial sector, academia, and government attended two workshops that generated ideas that will inform DARPA's next phase of AI exploratory projects.
<https://www.darpa.mil/news-events/2023-10-23>

Source: *DARPA News*

Clarivate Enriches Web of Science Platform with Integration of ProQuest Dissertations and Theses Global
Clarivate has announced the integration of ProQuest™ Dissertations & Theses Global with its renowned Web of Science™ platform.

<https://www.stm-publishing.com/clarivate-enriches-web-of-science-platform-with-integration-of-proquest-dissertations-and-theses-global/>

Source: *STM Publishing*

Improving University Workflows

Free librarian time, share resources, add value and reduce administration costs by consolidating and automating workflows.

<https://about.proquest.com/en/products-services/Rialto/>

Source: *Proquest*

Thomson Reuters unveils generative AI strategy designed to transform the future of Professionals

The strategy update coincides with generative AI enhancements to its flagship product, Westlaw Precision.

<https://www.infotoday.eu/Articles/News/Featured-News/Thomson-Reuters-unveils-generative-AI-strategy-designed-to-transform-the-future-of-professionals-161626.aspx>

Source: *Information Today Europe*

Research Solutions acquires ResoluteAI

[Research Solutions](https://www.knowledgespeak.com/news/research-solutions-acquires-resoluteai/) has announced its acquisition of [ResoluteAI](https://www.knowledgespeak.com/news/research-solutions-acquires-resoluteai/), an advanced search platform aimed at equipping organisations with search, discovery, analysis, and knowledge management tools powered by AI and NLP technologies.

<https://www.knowledgespeak.com/news/research-solutions-acquires-resoluteai/>

Source: *Knowledgespeak*

New ScioWire Widget: the short-cut to Research Teasers

Gain knowledge and time: embed the [ScioWirebeta](https://www.stm-publishing.com/new-sciowire-widget-the-short-cut-to-research-teasers/) Widget in your website for a custom feed of new research summaries. SciencePOD has launched the beta version of its research newsfeed ScioWirebeta now available in its Widget edition.

<https://www.stm-publishing.com/new-sciowire-widget-the-short-cut-to-research-teasers/>

Source: *STM Publishing*

New research shows global divide in Pharmaceutical Research is significant – but closing

Traditional disparities in funding and collaboration in global north and global south are changing.

<https://www.digital-science.com/news/new-research-shows-global-divide-in-pharmaceutical-research-is-significant-but-closing/>

Source: *Digital Science*

CAS and Molecule.one to Collaborate for Developing AI-based Solutions

CAS and [Molecule.one](https://www.moleculeone.com/), a tech-bio leader in AI-based solutions for pharmaceutical chemistry, have established a strategic collaboration focused on the joint development of computer-aided synthesis design technologies to accelerate scientific breakthroughs in early-stage drug discovery and aid chemists in the discovery of novel small molecules. Leveraging their existing technologies and expertise, including Molecule.one's proprietary generative deep learning models and synthesis planning platform with chemist-first UI, and CAS' world-class chemical reactions content collection and deep industry knowledge, the two organisations will work together to enhance and develop solutions for efficient and innovative chemical synthesis planning. Beyond their complementary capabilities, the organisations' collaboration is fuelled by their shared goal to empower scientists and accelerate breakthroughs.

<https://www.knowledgespeak.com/news/cas-and-molecule-one-to-collaborate-for-developing-ai-based-solutions/>

Source: *Knowledgespeak*

Robotic AI Chemist Creates Catalyst for O₂ Production Using Martian Ores

Sustaining human life on Mars is a common dream and a frequent topic in fiction. To achieve this in reality, a reliable oxygen supply is required. However, synthesising oxygen from local Martian resources can be challenging. Weiwei Shang, Jun Jiang, Yi Luo, University of Science and Technology of China, Hefei, China, and colleagues have developed a robotic system that uses artificial intelligence to optimise catalysts for the oxygen evolution reaction (OER) from water driven by solar energy, using Martian ores as materials. The process is completely automated with no need for human intervention, including ore pretreatment, catalyst synthesis, testing, and intelligent optimisation.

<https://www.chemistryviews.org/robotic-ai-chemist-creates-catalyst-for-o2-production-using-martian-ores/>

Source: *Chemistry Views*

Wolters Kluwer acquires AI-enabled Drug Diversion Detection Software, expands Clinical Surveillance capabilities

Wolters Kluwer Health has announced it has signed and completed the acquisition of Invistics Corporation (Invistics), a U.S.-based provider of cloud-based, AI-enabled software for drug diversion detection and controlled substance compliance. Invistics will join the company's Clinical Surveillance, Compliance & Data Solutions unit, part of Clinical Solutions.

<https://www.stm-publishing.com/wolters-kluwer-acquires-ai-enabled-drug-diversion-detection-software-expands-clinical-surveillance-capabilities/>

Source: *STM Publishing*

Research Integrity at your fingertips with new world-leading Dimensions app

App uses AI to track Trust Markers across tens of millions of publications.

<https://www.digital-science.com/news/new-world-leading-dimensions-research-integrity-app/>

Source: *Digital Science*

Clarivate Report Finds Significant Adoption of AI in IP

The report explores the effects the rapid advancements in artificial intelligence (AI) will have on IP law, practice and processes, and the legal sector. It analyses the responses of intellectual property (IP) and R&D professionals to understand their attitudes towards AI.

<https://clarivate.com/news/clarivate-report-finds-significant-adoption-of-ai-in-ip/>

Source: *Clarivate*

AI for Academia: Digital Science acquires Writefull to empower Researchers and Publishers

Digital Science has today announced it has fully acquired the AI-based academic language service Writefull, which assists users worldwide with all aspects of their scholarly writing.

<https://www.stm-publishing.com/ai-for-academia-digital-science-acquires-writefull-to-empower-researchers-and-publishers/>

Source: STM Publishing

Digital Science announces brand redesign for ReadCube and Papers

Change reflects expansion and innovation across research literature management.

<https://www.digital-science.com/news/digital-science-announces-brand-redesign-for-readcube-and-papers/>

Source: Digital Science

Research Solutions Acquires Search and Discovery Platform, scite

Research Solutions, Inc., a provider of cloud-based workflow solutions for R&D-driven organisations, has announced its strategic acquisition of scite, an award-winning search and discovery platform that employs AI to enhance the discoverability and evaluation of research. The acquisition adds a powerful layer to Research Solutions' product portfolio, providing a significant opportunity for cross-selling to scite's extensive B2C customer base of around 21,000 active subscribers and diverse B2B customer base, including corporate entities, academic institutions, and government agencies.

<https://www.knowledgespeak.com/news/research-solutions-acquires-search-and-discovery-platform-scite/>

Source: Knowledgespeak

Fair Global Pricing: Consultation

cOAlition S commissioned [Information Power](#) to explore how a [globally fair pricing framework for academic publishing](#) could be devised and implemented. The key objective of this project was to identify ways in which readers and producers of scholarly publications or their proxies – research funders and universities – can financially contribute to supporting academic publishing services in a globally equitable and sustainable manner.

<https://www.stm-publishing.com/fair-global-pricing-consultation/>

Source: STM Publishing

Chemists make Breakthrough in Drug Discovery Chemistry

Chemists offer two new methods to develop a way to easily replace a carbon atom with a nitrogen atom in a molecule. The findings could make it easier to develop new drugs.

<https://www.sciencedaily.com/releases/2023/11/231101180614.htm>

Source: Science Daily

Artificial Intelligence could 'revolutionise' chemistry but Researchers warn of Hype

Artificial Intelligence can revolutionise science by making it faster, more efficient and more accurate, according to a survey of European Research Council (ERC) grant winners. And while the report looks at the impact of AI on all scientific fields, the field of chemistry, in particular, can be expected to benefit greatly from the revolution, say researchers. But there are also warnings that AI is being overhyped, and avowals of the importance of human experts in chemical research.

<https://www.chemistryworld.com/news/artificial-intelligence-could-revolutionise-chemistry-but-researchers-warn-of-hype/4018645.article>

Source: Chemistry World

Springer Nature acquires protocols.io

Researchers will now have the option to make their protocols openly available on the fully OA platform.

<https://www.researchinformation.info/news/springer-nature-acquires-protocolsio>

Source: *Research Information*

Publishing Open Access with Springer Nature delivers highest Global usage and benefits for Researchers

Publisher's latest open access report shows the growth, impact and value of its OA portfolio and how the company is delivering greater transparency around its publishing activities. Authors publishing open access (OA) with Springer Nature see their work used more than if they had published with other mixed model or pure OA publishers.

<https://www.stm-publishing.com/publishing-open-access-with-springer-nature-delivers-highest-global-usage-and-benefits-for-researchers/>

Source: *STM Publishing*

New Whitepaper explores the future of Research Publishing, addresses the Challenges and Opportunities in Open-Access Publishing

The MIT Press has announced the release of a whitepaper titled "Access to Science and Scholarship: Key Questions about the Future of Research Publishing." The project, featuring contributions from notable figures including MIT Press's Director and Publisher Amy Brand and Director of Journals and Open Access Nick Lindsay, delves into the current state of the research enterprise and explores potential future trajectories for academic publishing.

<https://www.knowledgespeak.com/news/new-whitepaper-explores-the-future-of-research-publishing-addresses-the-challenges-and-opportunities-in-open-access-publishing/>

Source: *Knowledgespeak*

Understanding Compound Selectivity with Data-Driven Drug Design

Selectivity is a crucial property in the development of new active pharmaceutical ingredients (APIs).

<https://www.scientific-computing.com/white-paper/compound-selectivity-data-driven-drug-design>

Source: *Scientific Computing World*

RSC Signs up for Chronoshub Journal Guide

The RSC is proud to have signed Read & Publishing agreements with institutions in 32 countries. And while more choice can be a positive for researchers, the RSC understands that having agreements with different publishers can mean it is increasingly difficult for authors to keep track of which journals their article processing charges are covered in.

<https://www.stm-publishing.com/royal-society-of-chemistry-signs-up-for-chronoshub-journal-guide/>

Source: *STM Publishing*

Wiley launches new Database - Wiley Database of Predicted IR Spectra

Wiley, one of the world's largest publishers and a global leader in research and learning, has announced the release of the new Wiley Database of Predicted IR Spectra. The database combines over 60 years of expertise in infrared (IR) spectroscopy and spectral data curation with the most current machine-learning techniques to significantly expand the number of IR spectral data available for spectral analysis.

<https://www.knowledgespeak.com/news/wiley-launches-new-database-wiley-database-of-predicted-ir-spectra/>

Source: *Knowledgespeak*

Computational Model captures the elusive Transition States of chemical Reactions

Using generative AI, MIT chemists created a model that can predict the structures formed when a chemical reaction reaches its point of no return.

<https://www.eurekalert.org/news-releases/1011250>

Source: EurekaAlert

ARL and CNI form joint Task Force to envision AI and Machine-Learning futures in Research

A joint task force, representing the membership of the Association of Research Libraries (ARL) and the Coalition for Networked Information (CNI), has begun working on a six-month initiative to develop a set of possible future scenarios examining how AI and ML might transform the research enterprise.

<https://www.knowledgespeak.com/news/arl-and-cni-form-joint-task-force-to-envision-ai-and-machine-learning-futures-in-research/>

Source: Knowledgespeak

Managing Scientific Literature access and Copyright Compliance in a Remote Workforce

Learn how to manage compliance from home, from the perspective of knowledge and information management consultant, Heather Desmarais.

https://www.copyright.com/wp-content/uploads/2021/02/CCCRD_Remote-Workforce-Tip-Sheet.pdf

Source: CCC

Clarivate Expands Partnership with VeriSIM Life to Accelerate and De-risk Research and Drug Development

New AI-enabled, integrated workflow provides pharma and biotech companies with comprehensive R&D insights to help minimise late-stage failures during clinical trials.

<https://clarivate.com/news/clarivate-expands-partnership-with-verisim-life-to-accelerate-and-de-risk-research-and-drug-development/>

Source: Clarivate

Digital Science announces exclusive rollout of Dimensions AI Assistant beta version

Digital Science has announced a limited and exclusive beta launch of Dimensions AI Assistant, a new research tool designed to enhance how users engage with the wealth of knowledge available on Dimensions, among the world's largest linked research databases.

<https://www.stm-publishing.com/digital-science-announces-exclusive-rollout-of-dimensions-ai-assistant-beta-version/>

Source: STM Publishing

Chemists image basic blocks of Synthetic Polymers

Researchers have developed a new method to image polymerisation catalysis reactions one monomer at a time.

<https://www.sciencedaily.com/releases/2023/11/231109141457.htm>

Source: Science Daily

Report: Recruiting with AI: Trends and Challenges in Life Sciences

BioSpace surveyed life sciences employers to understand attitudes and current trends on AI usage in recruiting. This report explores the benefits of using AI tools in recruitment and provides practical recommendations for HR and talent acquisition professionals to leverage AI effectively.

<https://www.biospace.com/article/report-recruiting-with-ai-trends-and-challenges-in-life-sciences/?s=120>

Source: Biospace

UKRI announces Open Access Policy for Books and Monographs, including £3.5 million funding support

Starting from January 1, 2024, [UKRI](#) open access policy will extend to include monographs, book chapters, and edited collections that acknowledge UKRI funding. The policy is designed to ensure that research findings funded by UKRI, using public money, are freely accessible, fostering collaboration and innovation across the research and innovation community.

<https://www.knowledgespeak.com/news/ukri-announces-open-access-policy-for-books-and-monographs-including-3-5-million-funding-support/>

Source: *Knowledgespeak*

Artificial Intelligence paves way for new Medicines

Researchers have developed an AI model that can predict where a drug molecule can be chemically altered.

<https://www.sciencedaily.com/releases/2023/11/231129112519.htm>

Source: *Science Daily*

Biochemical Society advances Open Access with innovative Subscribe to Open Model

The [Biochemical Society](#), along with its trading arm [Portland Press](#), has announced an innovative Subscribe to Open (S2O) model for five of its esteemed research and review journals. This move exemplifies the Society's ongoing dedication to providing accessible research while upholding the highest standards of quality.

<https://www.knowledgespeak.com/news/biochemical-society-advances-open-access-with-innovative-subscribe-to-open-model/>

Source: *Knowledgespeak*

10 Questions to ask when Searching for a Corporate Literature Management Solution

Finding the right content at the right time is essential for any R&D-intensive company, but it is equally important to consider how that content is acquired and managed, and when the time is right to consider a literature management tool.

<https://www.copyright.com/wp-content/uploads/2023/04/Tip-Sheet-10-Questions-for-Corp-Lit-Management-Tool.pdf>

Source: *CCC*

Fourteen things you need to know about Collaborating with Data Scientists

Experimentalists often need help to analyse data. Here's how to ensure your collaboration is productive.

<https://www.nature.com/articles/d41586-023-02291-4>

Source: *Nature*

Biomedical Publisher Future Science Group Joins Taylor & Francis

As well as bringing a portfolio of cutting-edge journals and digital hubs, FSG's leading publishing solutions program will enable Taylor & Francis to offer researchers and medical communication planners a host of additional services.

<https://newsroom.taylorandfrancisgroup.com/future-science-group-joins-taylor-and-francis/>

Source: *Taylor & Francis*

UKRI invests in the Next Generation of AI Innovators

UKRI has announced investment in 12 UKRI Centres for Doctoral Training (CDTs) in artificial intelligence (AI) based at 16 universities.

<https://www.ukri.org/news/ukri-invests-in-the-next-generation-of-ai-innovators/>

Source: *UKRI*

ChatGPT Used for Text Mining of MOF Syntheses in the Literature

Large language models, such as the GPT series of models used in ChatGPT, are trained using large amounts of text and can predict the probabilities of series of words in a given language. This can be used for a variety of applications, e.g., to generate a probable text output based on a user input. The chemical literature also contains vast amounts of text, and performing a comprehensive literature review and extracting useful data and insights for a specific application quickly can be challenging. Large language models could help with this issue.

<https://www.chemistryviews.org/chatgpt-used-for-text-mining-of-mof-syntheses-in-the-literature/>

Source: *Chemistry Views*

Sage acquires IOS Press, expands Research Portfolio

Global independent academic publisher [Sage](#) has acquired [IOS Press](#), an independent publisher founded in Amsterdam in 1987 that specialises in health, life, and computer sciences. With this move, Sage acquires nearly 100 journals and a frontlist of 70 plus books each year covering subjects such as neuroscience, medical informatics, cancer research, AI, data science, and the semantic web.

<https://www.knowledgespeak.com/news/sage-acquires-ios-press-expands-research-portfolio/>

Source: *Knowledgespeak*

The Global Flourishing Study Launches Open Access of Sample Research Data with the Center...

The first sample dataset from the Global Flourishing Study (GFS) initiative is now available to researchers, with the project's initial full dataset scheduled for release in the coming months through the Center for Open Science (COS).

<https://www.stm-publishing.com/the-global-flourishing-study-launches-open-access-of-sample-research-data-with-the-center-for-open-science/>

Source: *STM Publishing*

Future Science Group launches new OA Journal: Future Medicine AI

[Future Science Group](#) has launched Future Medicine AI, a new open access, peer-reviewed journal committed to advancing the application of artificial intelligence in medicine. It is essential to build digital healthcare technologies upon a foundation encompassing safety and responsibility, and as such, Future Medicine AI aims to cover these challenges and ethical issues to provide a critically reliable source of information for health regulators and policymakers. Key topics covered by Future Medicine AI include: Virtual reality; Precision medicine; Ethics and regulation; Medical imaging and biomedical diagnostics; Multi-omics research; Drug discovery and development; Next-generation clinical trials; Health management/optimisation; and Real world evidence.

<https://www.knowledgespeak.com/news/future-science-group-launches-new-oa-journal-future-medicine-ai/>

Source: *Knowledgespeak*

5 Content Challenges derailing your Biotech's Research Process

Establishing a company information centre is the ideal approach, but many growing companies may not be able to invest time and money in that kind of initiative. If developing a centralised information centre isn't viable, here are a few ideas to solve the challenges that may be present for researchers.

<https://www.copyright.com/wp-content/uploads/2021/04/Tips-Sheet-5-Content-Challenges-Derailing-Your-Research-Process-and-How-to-Fix-Them-Life-Sciences.pdf>

Source: *CCC*

Elsevier unveils Scientific Datasets to drive Innovation across Industries

Elsevier has announced an initiative to empower R&D across various industries. The company's latest offering, enriched and authoritative scientific Datasets, aims to fuel innovation and support critical decision-making in

fields such as life sciences, chemicals, and other research-intensive industries. The introduction of [Elsevier's Datasets](#) provides researchers, data scientists, and industry leaders with a powerful tool to expedite R&D processes and enhance precision across a spectrum of sectors, including life sciences, energy, chemicals and materials, and technology. Use cases for these Datasets encompass a wide range of data science and analytical projects, from disease target identification using natural language processing to predicting molecule efficacy and toxicity through neural networks. Other applications include predictive modelling, Key Opinion Leader (KOL) analysis, and more.

<https://www.knowledgespeak.com/news/elsevier-unveils-scientific-datasets-to-drive-innovation-across-industries/>

Source: *Knowledgespeak*

Reducing Risk & Improving Synthesis Outcomes with SYNTHIA Retrosynthesis Software

Addressing the many intrinsic and external challenges to chemical synthesis has led to a highly complex system of regulations that involve numerous academic and governmental organisations in modern times. Existing methods have now been bolstered by powerful computer software with advanced algorithms to support the synthetic chemist and this is leading to new and innovative drugs that can be developed faster and with less risk.

<https://www.scientific-computing.com/white-paper/reducing-risk-improving-synthesis-outcomes-synthia-retrosynthesis-software>

Source: *Scientific Computing World*

Clarivate Announces Partnership with AI21 Labs as part of its Generative AI Strategy to Drive Growth

Clarivate has announced a strategic partnership with AI21 Labs, a pioneer in generative AI. The collaboration will integrate large language models into solutions from Clarivate, to enable intuitive academic conversational search and discovery, specifically designed to foster researcher excellence and drive success for researchers and students, while adhering to core academic principles and values.

<https://www.stm-publishing.com/clarivate-announces-partnership-with-ai21-labs-as-part-of-its-generative-ai-strategy-to-drive-growth/>

Source: *STM Publishing*

Wiley launches Advanced zyLabs to help Computer Programming Students learn in Professional, real-world Environment

Wiley has announced the launch of Advanced [zyLabs](#), an addition to classic zyLabs in zyBooks computer and data science courses designed to help computer programming students learn coding skills in a realistic, professional environment. More than a million students across more than 1,200 institutions have used zyBooks.

<https://www.knowledgespeak.com/news/wileys-launches-advanced-zylabs-to-help-computer-programming-students-learn-in-professional-real-world-environment/>

Source: *Knowledgespeak*

Denmark puts its Money where its Life-Sciences Strategy is

Building on the success of blockbuster drugs, the country's focus on reinvestment is feeding a stream of discovery.

<https://www.nature.com/articles/d41586-023-03446-z>

Source: *Nature*

A Vision for Science Culture

The RSC has launched its vision for a great science culture. The vision builds on in-depth engagement with the chemical sciences community and sets the direction for changes that can benefit everyone across the chemical sciences community.

<https://www.rsc.org/news-events/articles/2023/09-september/a-vision-for-science-culture/>

Source: RSC News

UKRI Funding set to transform EMBL-EBI Bioinformatics Capability

UKRI is set to invest more than £80 million in the European Bioinformatics Institute (EMBL-EBI), part of the European Molecular Biology Laboratory (EMBL).

<https://www.ukri.org/news/ukri-funding-set-to-transform-embl-ebi-bioinformatics-capability/>

Source: UKRI

Clarivate unveils Search Platform enhanced by Generative AI

Clarivate has launched its new enhanced search platform leveraging generative artificial intelligence (GenAI). The new Clarivate offering enables drug discovery, preclinical, clinical, regulatory affairs and portfolio strategy teams to interact with multiple complex datasets using natural language to obtain immediate and in-depth insights.

<https://www.knowledgespeak.com/news/clarivate-unveils-search-platform-enhanced-by-generative-ai/>

Source: Knowledgespeak

IOP Publishing is the winner of the new ALPSP Impact Award 2023

IOP Publishing (IOPP) was confirmed as the winner of the new ALPSP Impact Award 2023 with its entry as the first society publisher to combine double anonymous and transparent peer review.

<https://www.stm-publishing.com/iop-publishing-is-the-winner-of-the-new-alpsp-impact-award-2023/>

Source: STM Publishing

Springer Nature uses Generative AI to Publish Academic Book

Book took less than five months from inception to publication – about half the time normally taken.

<https://www.researchinformation.info/news/springer-nature-uses-generative-ai-publish-academic-book>

Source: Research Information

Elsevier releases Alpha Version of Scopus AI for Testing

Elsevier has announced an alpha version of Scopus AI for researcher testing – a next generation tool that combines generative AI with Scopus' trusted content and data to help researchers get deeper insights faster, support collaboration and societal impact of research.

<https://www.knowledgespeak.com/news/elsevier-releases-alpha-version-of-scopus-ai-for-testing/>

Source: Knowledgespeak

Seven AI/ML for Life Sciences Companies Identified as Innovators in New Clarivate Companies to Watch Report

Rapid acceleration in IP creation and scientific knowledge drives growing acceptance of AI and ML in drug discovery and development.

<https://clarivate.com/news/seven-ai-ml-for-life-sciences-companies-identified-as-innovators-in-new-clarivate-companies-to-watch-report/>

Source: Clarivate

PubHive announces Real-Time access to Clinical Trial Data

PubHive Ltd, a leading provider of scientific literature & safety information workflows provider, has announced the seamless integration of clinical trial data sources, providing clinical researchers and healthcare professionals with unparalleled real-time access to critical information and enabling them to conduct research more efficiently and effectively.

<https://www.stm-publishing.com/pubhive-announces-real-time-access-to-clinical-trial-data/>

Source: STM Publishing

India's rise as a Centre of Research Excellence: Joint seminar highlights academic potential

India has cemented its position as the world's third-largest producer of research output, surpassing the UK, according to data presented at a joint seminar organised by the Ministry of Education and Elsevier. The seminar, held in preparation for the 4th G20 Education Working Group meeting, emphasised the critical role of accessible science and collaboration in driving sustainable development. Global academic leaders, policy experts, and researchers gathered to discuss best practices and explore avenues for utilising science to support global progress.

<https://www.knowledgespeak.com/news/indias-rise-as-a-center-of-research-excellence-joint-seminar-highlights-academic-potential/>

Source: Knowledgespeak

Polyethylene Waste could be a thing of the past

Experts have developed a way of using polyethylene waste (PE) as a feedstock and converted it into valuable chemicals, via light-driven photocatalysis.

<https://www.sciencedaily.com/releases/2023/12/231208190002.htm>

Source: Science Daily

Clarivate Partners with EveryLibrary to support and nurture Libraries in the U.S.

Clarivate has announced a partnership with EveryLibrary, the non-partisan advocacy organisation for libraries.

<https://www.stm-publishing.com/clarivate-partners-with-everylibrary-to-support-and-nurture-libraries-in-the-u-s/>

Source: STM Publishing