



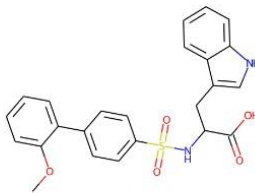
NEWSLETTER

Summer 2023

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.

Drug Discovery Game

Home Introduction Design Test Analysis Results



A28B13
Date Created:
Week 0
Lipinski Filters:
MW: Pass
H Acc.: Pass
H Don.: Pass
LogP: Pass
Molecule Descriptors:
HA: 32
MW: 450.1 Da
TPSA: 108.5 Å²
H Acc.: 4
H Don.: 3
LogP: 3.8
Rings: 4
Assay Data:
Mouse Clearance: medium (5.6-30.5)
Human Clearance: low (< 12)
LogD: 0.66

	Mouse Clearance	Human Clearance	LogD	PAMPA	pIC ₅₀
Cost per assay	£7,000	£9,000	£1,000	£700	£70
Duration	3 weeks	3.5 weeks	1.5 weeks	1 week	1 week
Molecule					
A05B02	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
A05B13	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
A15B13	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
A28B13	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
A28B19	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
A10B19	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
A10B05	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>

Run Assays
Duration: 3.5 weeks
Cost: £33700

← Design Analysis →

test 23.5 weeks left
Log out £74510

The Test Phase of the University of Oxford's 'Drug Discovery Game'. See the full article on p. 31.

CICAG Websites and Social Media

The screenshot shows the Royal Society of Chemistry website. At the top left is the RSC logo. To the right are links for 'Members' area' and 'Donate'. Below this is a navigation menu with categories: 'About us', 'Membership & professional community', 'Policy & perspectives', 'Journals, books & databases', 'Teaching & learning', 'News & events', and 'Locations & contacts'. A breadcrumb trail reads: 'Home > Membership & professional community > Connect with others > Through interests'. The main content area features a large banner for the 'Chemical Information and Computer Applications Group' with the text 'Connect with colleagues in your field and become involved in your scientific community'. Two call-to-action buttons are present: a yellow one for 'Forthcoming events' and a blue one for 'How to join'.

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

The screenshot shows the website for the RSC Interest Group Chemical Information and Computer Applications Group. The header includes the RSC logo, the group name, and a 'WELCOME' button. A navigation bar contains links for 'Home', 'Newsletters', 'Previous meetings', 'Forthcoming Meetings', 'Tony Kent Strix Award', 'Related Links', 'Committee members', and 'Contacts'. The main content area is divided into several sections: a 'Welcome' message, a 'YouTube Channel' section with a video player, a 'Newsletters' section, and a 'Tweets by @RSC_CICAG' section. The 'Welcome' section states: '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.' The 'YouTube Channel' section features a video player and the text 'RSC Chemical Information and Computer Applications Group Inspirational Committee Award 2021'. The 'Newsletters' section includes a 'Royal Society of Chemistry 2021 Inspirational Committee Award: Chemical Information and Computer Applications Group' link. The 'Tweets' section shows a tweet from @RSC_CICAG and a retweet by @macinchem.

<http://www.rscicag.org>



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



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



@RSC_CICAG

https://twitter.com/RSC_CICAG

Contents

Chemical Information and Computer Applications Group Chair's Report	4
CICAG Planned and Proposed Future Meetings	5
InChi Usage Survey	5
Cheminformatics: A Digital History – Part 3. Learning from Chemical Data	5
The Chemistry Conference Database	11
Being #CompChemURG: Mentorship Matters	13
AI Chatbot Suggestions for CICAG Newsletter Articles	13
Olga Kennard (1924 – 2023)	17
Size, Growth and Challenges for the Chemical Space	18
Quantum Computing for Chemistry	21
A Computational Chemist's Journey Through Pharmaceutical and Biotechnology Research	25
The Drug Discovery Game	31
Workshop Report: Building the UK's Machine Learning in Materials Science Capability	34
Accelerating Open-Source Innovation in Computer-Aided Drug Design with the OpenBioSim Community Interest Company	36
Catalyst Science and Discovery Centre News	39
Communicating Chemistry at a Local History Museum	42
AI in Drug Discovery 2022 – A Highly Opinionated Literature Review	43
Book Reviews	52
Cheminformatics and Chemical Information Books	53
Exploring the Promise of Generative Models in Chemistry: An Introduction to Diffusion Models	56
News from ACS CINF	63
Cambridge Structural Database (CSD) Updates	64
News from CAS	65
AI4SD News	67
Physical Sciences Data Infrastructure (PSDI) News	71
UKeiG News: eLucidate, and the Prestigious International Strix and Farradane Awards	74
Other Chemical Information News	76

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 first part of the year has been relatively quiet for CICAG meetings whilst we organise a couple of major events for the second half of the year. These are:

[Solutions in Science - Molecular Characterisation](#) 4-6 July in Cardiff 2023

[6th RSC-BMCS / RSC-CICAG Artificial Intelligence in Chemistry](#) 4-5 September 2023, Cambridge, UK.

The AI in Chemistry is shaping up to be a bumper event with a fabulous line-up of speakers, and with large language models very much in the news at the moment this meeting looks to be very timely. Indeed we have a ChatGPT and Google's BARD generated article in this CICAG Newsletter!

We are also in early planning for two other events: Markush Centenary meeting, and a Python for Chemists course. It is perhaps worth mentioning that CICAG provide [bursaries](#) to attend these meetings.

Social media became an increasingly important way for communicating with members (and non-members) during lockdown and the trend continues, with [Twitter](#) now having 1565 followers, and the running of the Twitter feed has been taken over by Samantha Kanza. [LinkedIn](#) now has 619 members, we also have a [Mastodon](#) account 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 1025 subscribers and contains the 13 video presentations from AI4proteins meetings in addition to all 20 of the [Open-Source Tools for chemistry](#) workshops. These workshop videos have proved to be very popular and have been watched a nearly 31,500 times.

This newsletter also includes contributions from Guillermo Restrepo on the *Size, Growth and Challenges for the Chemical Space* which follows on nicely from the [Ultra-large Chemical Libraries](#) meeting CICAG organised last year. I'm also delighted we have been able to include an invaluable article by Pat Walters entitled *AI in Drug Discovery 2022 – A Highly Opinionated Literature Review*, essential reading for anyone with an interest in ML/AI. We are also delighted to include an article by Allan M Ferguson reflecting on learnings from his career at Merck, entitled *A Computational Chemist's Journey Through Pharmaceutical and Biotechnology Research*.

CICAG came into existence in 2007 with the merger of the Chemical Information and Computer Applications Groups and currently membership stands at around 724 members and is increasing by 60-70 new members per year. CICAG currently have members from 43 different countries, and the community is certainly growing. Whilst RSC members can join interest groups for free, in practise many members do not take up this opportunity. You can make a request to join a group by email (membership@rsc.org), telephone (01223 432141) or the RSC [website](#).

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.

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
6th Artificial Intelligence in Chemistry Meeting	4-5 Sept 2023	Churchill College, Cambridge	Joint event from RSC-CICAG and RSC-BMCS division. See RSC Events page.
Centenary of Markush Structures	Q1/2 2024	Burlington House	Details to follow.
Python for Chemists	TBD	TBD	Details to follow.
Molecular Simulation and Free Energy Methods	TBD	TBD	Details to follow.
7th Artificial Intelligence in Chemistry Meeting	July 2024	TBD	Organising committee decided to bring the date forward to avoid a clash with other events, school holidays & start of school term.

InChI Usage Survey

Contribution from Jonathan Goodman, Professor of Chemistry, Yusuf Hamied Department of Chemistry, University of Cambridge, email: jmg11@cam.ac.uk

InChI users are invited to complete a [short survey](#) to let the International Union of Pure and Applied Chemistry (IUPAC) know how they use InChI. The survey should only take 2-3 minutes. Feedback from users will be very helpful for the further development of the InChI. For information about InChI please visit the [InChI Trust's website](#).

Cheminformatics: A Digital History – Part 3. Learning from Chemical Data

Contribution from Professor Johann Gasteiger, Computer-Chemie-Centrum, Universität Erlangen-Nürnberg, Germany, email: johann.gasteiger@fau.de

It is an honour to have been asked to provide a contribution to this series of articles for the Royal Society of Chemistry illustrating the early phases of cheminformatics (or chemoinformatics as I used to say). I particularly appreciate that in spite of all the political turmoil, scientific contacts between a British institution and a member of the European Union are still cultivated.

I studied chemistry at the University of Munich and obtained my PhD in 1971 in organic chemistry. My experimental work concerned the elucidation of reaction mechanisms and was rather difficult: compounds that decomposed at temperatures higher than -40° degrees and had to be handled in liquid sulphur dioxide.¹ I liked my work in the lab but I wanted to understand chemistry better and switched to theoretical chemistry in my postdoctoral work with Prof. Andrew Streitwieser Jr at the University of California in Berkeley.² There, I saw what computers could tackle at that time for small molecules at the *ab initio* level (i.e., molecules with only a few atoms, smaller than the much larger molecules organic chemists were interested in) but the computer seemed to me a fascinating instrument. What could be done to fill the gap?

Already during my PhD work my attention was caught by a series of interesting publications by eminent Stanford scientists with the fascinating titles *The Application of Artificial Intelligence for Chemical Inference*.³ Furthermore, a paper by Professor Ivar Ugi about his ideas of using computers to design organic syntheses⁴ attracted my interest. Off I drove from Berkeley to Los Angeles to visit Prof. Ugi which resulted in him offering me a job at his new location at the Technical University of Munich. Here, our team developed a prototype of a program for designing organic syntheses.⁵

After the team dissolved as the American members went back to USA, I started learning the programming language PL/I and developed a new version of a synthesis design system, called EROS (Elaboration of Reactions for Organic Synthesis).⁶ From the very beginning we had decided to represent explicitly all atoms and all bonds of a molecule in contrast to the then prevailing, more concise representations such as Wiswesser Line Notation. Quite a few basic formal problems such as the unique and unambiguous representation of molecules⁷ and the perception of rings⁸ in a molecule had to be solved before addressing interesting chemical problems. (I still remember meeting a former friend who asked me what I was currently doing. When I told him that we had developed a program for the perception of rings he gave me a weird look saying: "But I see the rings!". It was hard to convince him that computers are basically stupid: they all have to be told what to do by algorithms.)



Prof. Todd Wipke (left, developer of the synthesis design system SECS) and Johann Gasteiger (right, developer of the synthesis design system EROS) in front of a Russian computer (run by an IBM operating system) in Novosibirsk during the IVth International Conference on Computers in Chemical Research and Education, June 1978.

We had to find ways of selecting chemically feasible reactions from formally conceivable ones. We tried to base this selection on physicochemical models that could be calculated rapidly, such as estimated heats of reactions,⁹ the partial charges of the atoms in a molecule,¹⁰ or residual electronegativity and effective polarizability.¹¹ (Incidentally, the model based on partial charge calculation is my most cited paper. The method was integrated in all molecular modelling packages, but the publication was rejected three times before being accepted. Presenting a very fast method for calculating partial charges made us enemies of quantum chemists who depended on funds for their methods which consumed a lot of computer time.)

Initially, my research progressed slowly as I could hire only one co-worker through a scholarship from the German Research Council. Then I got a major boost through a research grant from Imperial Chemical Industries (ICI), UK. This allowed me to expand my research group and our work on EROS progressed well.¹² However, it became clear that the modelling of organic synthesis design by physicochemical models alone was at that time too ambitious a project. Furthermore, we also had to realise that in the late 1970s and in the 1980s the vast majority of organic chemists were just not prepared to use a computer for designing organic syntheses. It was too interesting for them to design syntheses intellectually and not to delegate the task to a computer, an instrument they were not familiar with at that time. Thus, we had to expand our research interests.

In the 1980s Germany was fortunate to have a chemist as Minister of Science and Technology. He understood the value of information for chemists and therefore the building of chemical databases was funded. Our basic work on representing atoms and bonds in a molecule laid the foundation for some very important databases in

chemistry: my former co-worker Clemens Jochum became the project leader for the construction of the Beilstein database¹³ on organic molecules and their properties, and he assembled the team for the Gmelin database¹⁴ on inorganic structures, while I accepted the job as project leader for building the ChemInform reaction database^{15,16} and I helped in a team for the definition of structure representation¹⁷ in the SpecInfo database on spectral information.¹⁸ (Although the JCAMP-CS format was not widely used for structure representation, it forced the company MDL to open their proprietary molfile structure format for general use.) Some years later, my former co-worker Wolf-Dietrich Ihlenfeldt helped design and implement the PubChem database,¹⁹ the world's largest collection of freely accessible chemical information on compounds and physical properties, biological activities, safety and toxicity information, patents, literature citations and more.

Now with chemical data increasingly becoming accessible in computer-readable form the question arose what to do with them. It is clear that chemists have always acquired most of their knowledge by inductive learning, by learning from data and measurements. Thus, attention was focused on mathematical methods for learning from data such as statistical, regression, or pattern-recognition methods.

Despite the current trend towards online meetings, it has to be emphasised how important face-to-face meetings are for personal contacts and for the advancement of science. A NATO Advanced Study Institute summer school in Noordwijkerhout, Netherlands, in 1973 was one such decisive meeting. It brought together, for the first time, scientists from all over the world and from various disciplines who were working on systems for using computers in chemistry, in database construction, molecular modelling, synthesis design, data analysis, and computer graphics and it lasted for two weeks, providing ample time for intense scientific talks. What I took home from this meeting was seeing that scientists from quite different areas and with quite different goals were struggling with similar problems and all of them were convinced that they must use the computer. This convinced me that I, too, had to continue using the computer to develop methods to understand chemistry better.

For me, the most inspiring lecture was given by Bruce Kowalski who spoke about pattern recognition. This made it clear to me that I had to understand these methods and then use them for analysing chemical data. The encounter developed into a life-long friendship with Bruce. He also introduced me to Svante Wold. Bruce and Svante had developed the Chemometrics Society and basically established a new field, chemometrics, which has grown enormously in importance. Together with Svante, my team made the first study with his new method, Partial Least Squares (PLS) analysis of chemical data.²⁰ Since then, PLS has become one of the most important methods for analysing chemical data.

In the 1980s new data analysis methods under the overarching term "artificial neural networks" (ANNs) began to be employed in chemistry. ANNs are mathematical methods that try to model the information processing methods used by the human brain, clearly an approach that is quite different from the mathematical data analysis methods previously employed. The advantage of an ANN is that the relationships between the various types of data (e.g., between data and chemical structures) need not be explicitly specified: they are learnt by training from the data. Various types of ANN methods have been developed and both supervised and unsupervised methods are available.

Again, a scientific meeting became decisive for my future research. After the International Conference on Computers in Chemical Research and Education (ICCCRE) in Beijing, China in 1987 I was waiting in the lobby of Beijing airport for my return flight. There I met Jure Zupan from Ljubljana, Slovenia who was also waiting for his return flight. Waiting gave us ample time for discussions and we found out that both of us had recently heard about the use of ANNs on chemical data. We decided to take a closer look at these methods and their usefulness. Fortunately, I had funds that allowed me to offer a visiting professor position to Jure Zupan. We

decided to take a closer look at all publications that used ANNs in chemistry that were available at that time. This led to a review, the title of which showed our initial scepticism,²¹ but it became clear that the application of neural networks in chemistry was rapidly increasing.²² In 1988 we found only three publications on the use of ANN in chemistry; there were five in 1989, 30 in 1990, and 110 in 1991.

In writing the review we had to come up with a scheme for comparing the various publications, particularly with respect to the different notations used by the authors. Furthermore, it had to be said that there were many mistakes in the mathematical notations in many publications. This convinced us that we had to write a book on the use of neural networks in chemistry in order that people would realise the importance of these new methods and would make proper use of them.²³ The writing of the book certainly benefitted from our different backgrounds: Jure a physicist turned into an analytical chemist, and me trained as an organic chemist, turned into a cheminformatician. During the day we fought heavily over the text and the mathematics in the book, and it was not unusual for us to shout at each other. However, in the evening we went to the local Greek restaurant and were friends again. The book was very well received and after a short time Chinese and Japanese translations were produced. Moreover, in a few years' time, because of the enormous increase in the use of ANNs in chemistry a new edition became necessary.²⁴



Jure Zupan (left, National Institute of Chemistry, Ljubljana, Slovenia) and Johann Gasteiger (right) in Slovenia, April 2013.

Having dug deeper into the potential of artificial neural networks we saw many interesting uses in chemistry and my research group ventured into various areas of application. This work definitely benefitted from the increase in availability of chemical data in computer-readable form (such as in the databases mentioned above). This is not the place to go deeper into these applications; a listing of leading references up until 1999 for the various topics that we studied should suffice. I include prediction of chemical reactivity,²⁵ classification of chemical reactions,²⁶ analysis of molecular surfaces,²⁷ comparison of chemical libraries,²⁸ locating biologically active compounds,²⁹ finding new leads in drug design,²⁹ simulation of infrared spectra,³⁰ and prediction of the 3D structure from infrared spectra.³¹

One method that we found particularly interesting was the self-organising network introduced by Teuvo Kohonen.³² This is an unsupervised learning method that maps data from a high-dimensional space into two dimensions, allowing one to discover novel information. A particularly interesting illustration is the classification of Italian olive oils according to their place of origin.³³ The self-organising neural network came up with a very good classification of the various olive oils. Beyond that, the two-dimensional map obtained by mapping the eight-dimensional space defined by the amount of fatty acids characterising each olive oil produced a two-dimensional map that nicely reflected the map of Italy (Figure 1). Geographic information was not used in this study but apparently this information was inherent in the data (presumably reflecting the influence of climate and soil composition). This emphasises the power of unsupervised learning in discovering additional information.

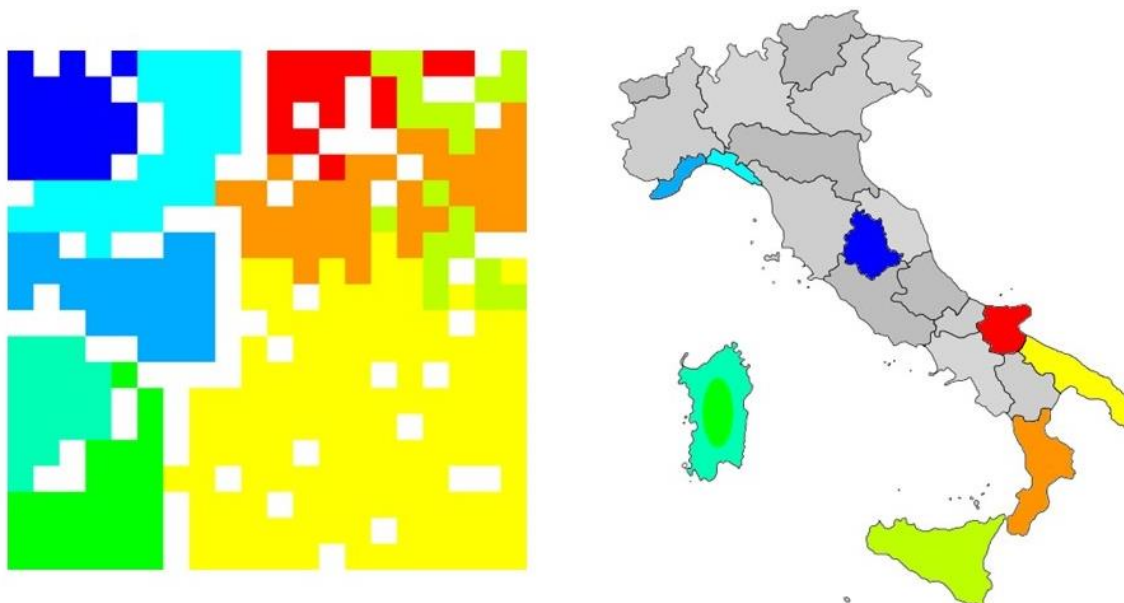


Figure 1. Mapping the 512 olive oils samples from an eight-dimensional space (defined by their content of eight different fatty acids) to two dimensions. Observe the separation of northern Italian regions from southern Italian regions and from Sardegna (Sardinia).

In our work, we did not change the mathematics of the standard artificial neural network methods. Rather, we concentrated on data representation by developing a series of methods for characterising chemical structures. In this work, we compared chemical structures to human beings perceiving either skeletons, or 3D structures, or surfaces, also taking account of flexibility.³⁴

Our work received recognition and I was offered a position at the University of Erlangen-Nuremberg in Germany. On this occasion, the State of Bavaria agreed to fund a Computer-Chemie-Centrum at this university run by Prof. Paul von Ragué Schleyer, Dr Tim Clark and myself. This centre was unique at that time and was instrumental in increasing the importance of cheminformatics and computational chemistry.

Starting around the year 2000 new artificial neural network methods were developed that allowed the processing of large datasets. These ANN methods were also referred to as machine learning or deep learning methods under the umbrella of artificial intelligence (AI). As the term artificial intelligence became a household word it is not surprising that it also showed up in many project proposals and publications on the processing of chemical data.³⁵

The future will certainly show many applications of artificial neural network methods (or “AI”, for those who prefer that term) for the processing of chemical data. As I said in the beginning, chemists have mostly learned from data, and now we have powerful methods for this endeavour. It is also clear that applications can be in any chemical subfield, from drug design to analytical, physical, or organic chemistry, to agrochemistry, cosmetics, material properties, and process control.³⁶

Two things that I want to emphasise are (1) the quality of data is more important than the quantity, and (2) the chemical entities should be represented in a form that is best suited to the problem at hand and for learning chemical relationships.

References

- (1) Huisgen, R.; Gasteiger, J. The primary step in the halogenation of cyclooctatetraene. *Angew. Chem. Int. Ed. Engl.* **1972**, *11*, 1104-1105.
- (2) Cambray, J. et al. Self-consistent field calculations of hole states of carbon monoxide: electron density functions by computer graphics. *J. Am. Chem. Soc.* **1974**, *96*, 5978-5984.
- (3) Lederberg, J. et al. Application of artificial intelligence for chemical inference I. The number of possible organic compounds. Acyclic structures containing C, H, O, and N. *J. Am. Chem. Soc.* **1969**, *91*, 2973-2976.
- (4) Ugi, I. et al. Chemistry and logical structures. *Angew. Chem. Int. Ed.* **1970**, *9*, 703-730.
- (5) Blair, J. et al. Representation of the constitutional and stereochemical features of chemical systems in the computer-assisted design of syntheses. *Tetrahedron.* **1974**, *30*, 1845-1859.
- (6) Gasteiger, J.; Jochum, C. EROS – a computer program for generating sequences of reactions. *Topics Curr. Chem.* **1978**, *74*, 93-126.
- (7) Jochum, C.; Gasteiger, J. Canonical numbering and constitutional symmetry. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 113-117.
- (8) Gasteiger, J.; Jochum, C. An algorithm for the perception of synthetically important rings. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 111-115.
- (9) Gasteiger, J. An algorithm for estimating heats of reactions. *Comput. Chem.* **1978**, *2*, 85-88.
- (10) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity. *Tetrahedron.* **1980**, *36*, 3219-3228.
- (11) Gasteiger, J.; Hutchings, M.G. Quantitative models of gas-phase proton-transfer reactions involving alcohols, ethers, and their thio analogues. Correlation analyses based on residual electronegativity and effective polarizability. *J. Am. Chem. Soc.* **1984**, *106*, 6489-6495.
- (12) Gasteiger, J. et al. A new treatment of chemical reactivity: development of EROS, an expert system for reaction prediction and synthesis design. *Topics Curr. Chem.* **1987**, *137*, 19-73.
- (13) The Beilstein Database. https://en.wikipedia.org/wiki/Beilstein_database (now Elsevier's Reaxys).
- (14) The Gmelin Database. https://en.wikipedia.org/wiki/Gmelin_database (now Elsevier's Reaxys).
- (15) Gasteiger, J., Weiske, C. Vom gedruckten Referatedienst zur Reaktionsdatenbank. *Nachr. Chem. Tech. Lab.* **1992**, *40*, 1114-1120.
- (16) ChemInform. <http://www.fiz-chemie.de/en/home/products-services/chemical-data/chemische-daten/cheminform-rx.html>.
- (17) Gasteiger, J. et al. JCAMP-CS: A standard exchange format for chemical structure information in computer-readable form. *Appl. Spectroscopy.* **1991**, *45*(1), 4-11.
- (18) Barth, A. SpecInfo: an integrated spectroscopic information system. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 52-58.
- (19) PubChem. <https://pubchem.ncbi.nlm.nih.gov/>.
- (20) Dunn III, W.J. et al. Multivariate structure-activity relationships between data from a battery of biological tests and an ensemble of structure descriptors: the PLS method. *Quant. Struct.-Act. Relat.* **1984**, *3*, 131-137.
- (21) Zupan, J.; Gasteiger, J. Neural networks: a new method for solving chemical problems or just a passing phase? *Anal. Chim. Acta.* **1991**, *248*, 1-30.
- (22) Gasteiger, J.; Zupan, J. Neural networks in chemistry. *Angew. Chem. Int. Ed. Engl.* **1993**, *32*, 503-527.
- (23) Zupan, J.; Gasteiger, J. Neural networks for chemists: an introduction. VCH: Weinheim, Germany, 1993.
- (24) Zupan, J.; Gasteiger, J. Neural Networks in Chemistry and Drug Design, 2nd ed. Wiley-VCH: Weinheim, Germany, 1999.
- (25) Simon, V. et al. A combined application of two different neural network types for the prediction of chemical reactivity. **1993**, *115*, 9148-9159.
- (26) Chen, L.; Gasteiger, J. Knowledge discovery in reaction databases: landscaping organic reactions by a self-organizing neural network. *J. Am. Chem. Soc.* **1997**, *119*, 4033-4042.
- (27) Gasteiger, J.; Li, X. Mapping the electrostatic potential of muscarinic and nicotinic agonists with artificial neural networks. *Angew. Chem. Int. Ed. Engl.* **1994**, *33*, 643-646.
- (28) Sadowski, J. et al. Assessing similarity and diversity of combinatorial libraries by spatial autocorrelation functions and neural networks. *Angew. Chem. Int. Ed. Engl.* **1995**, *34*, 2674-2677.
- (29) Bauknecht, H. et al. Locating biologically active compounds in medium-sized heterogenous datasets by topological autocorrelation vectors: dopamine and benzodiazepine agonists. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1205-1213.
- (30) Schuur, J.; Gasteiger, J. Infrared Spectra Simulation of Substituted Benzene Derivatives on the Basis of a Novel 3D Structure Representation. *Anal. Chem.* **1997**, *69*, 2398-2405.

- (31) Hemmer, M.C. et al. The prediction of the 3D structure of organic molecules from their infrared spectra. *Vibrat. Spectroscopy*. **1999**, *19*, 151-164.
- (32) Kohonen, T. Self-organized formation of topologically correct feature maps. *Biol. Cybern.* **1982**, *43*, 59-69.
- (33) Zupan, J. et al. Classification of multicomponent analytical data of olive oils using different neural networks. *Anal. Chim. Acta*. **1994**, *292*, 219-234.
- (34) Gasteiger, J. Of molecules and humans. *J. Med. Chem.* **2006**, *49*, 6429-6434.
- (35) Gasteiger, J. Chemistry in times of artificial intelligence. *ChemPhysChem*. **2020**, *21*, 2233-2242.
- (36) Engel, T., Gasteiger, J. (eds.) Applied Chemoinformatics – Achievements and Future Opportunities. Wiley-VCH: Weinheim, Germany, 2018. ISBN 978-3-527-34201-3.

The Chemistry Conference Database

Contribution from Dr Nessa Carson, AstraZeneca, email: nessa.carson@astrazeneca.com

The [Chemistry Conference Database](#)¹ was launched in 2018 as a free list of upcoming chemistry conferences, that has so far hosted links to 2,612 meetings. The webpage serves as a platform to alert researchers about upcoming conferences in their field that they may be interested in attending.

Filter: comp Standard prices Student Academic

Name	Start	End	Location	Member	Non-member
Canadian Chemistry Conference CSC 2023	04 Jun 2023	08 Jun 2023	Vancouver, Canada	CAD160–835	CAD315–1,015
ACS Nigeria Intl Chemical Sciences Chapter Meeting	04 Jun 2023	08 Jun 2023	Uyo, Nigeria	NGN15,800 / \$200	NGN15,800 / \$200
● Fundamentals & applications of density functional theory (workshop, US only)	05 Jun 2023	09 Jun 2023	Boise, ID, USA	Application	Application
● NDFI4Chem Chemistry Data Days	06 Jun 2023	07 Jun 2023	Mainz, Germany	Free	Free
ConTech Pharma	08 Jun 2023	—	London, UK/Online	£445–345	£445–345
ACS Middle Atlantic Regional Meeting MARM	09 Jun 2023	10 Jun 2023	New York, NY, USA	\$0–105	\$105

Upcoming computational and data conferences in the Chemistry Conference Database.

The Database came about after my first job at a CRO where I was requested to make a list of upcoming conferences, to decide where to ‘send’ chemists. Starting to host this list publicly on my personal website seemed to make sense and allowed me to check its contents on the go. I decided to track conference name, url, dates, location, and registration fees, and later added any conference hashtags as tooltips, and tagged each meeting with topics and a geographical region (to allow classification beyond having to parse the exact location).

However the major change over time has been the volume and types of conferences. The first commit shows only 16 conferences around the world, compared to 421 future meetings at the time of writing this article. While at first featured meetings centred around organic synthesis for my workplace, the Database now presents meetings from all fields of chemistry and slightly beyond. Training courses and workshops also feature, although it would be undesirable for the Database to become a free advertising source for equipment and software vendors, so these are kept to those of mainstream interest. Although the meetings naturally have some predisposition to my background in organic chemistry and my location in the UK, this bias has decreased over

time as I have become aware of more researchers accessing the list, and I actively seek out conferences in fields far removed from my own.

The conferences listed are found from my own personal searches, and added manually to avoid errors that would arise from web scraping and to ensure accurate metadata. The sources include organiser websites and email newsletters, RSC *Voice* newsletter, and social media. It is also very helpful for users and organisers to contact me directly with future meetings of interest. Only meetings whose fees are available are included so as to avoid having to repeatedly check back for revised information. I tend to update the list most days, and feel sure the vast majority of the 1,564 historical commits on the website's GitHub repo² include changes to the Database. For only a few prolific organisers, I have written Python scripts to convert structured conference webpages into one html table row per meeting, which saves me some time in their collation. To my knowledge, the Database has never featured a predatory conference³⁻⁴ run by the likes of Conference Series or Allied Academics, so it also serves as a repository of trustworthy information.

The Covid-19 pandemic was a time of change for all of us, the Database no exception. A sister list was set up, containing links to online webinars of any duration,⁵ unlike the Conference Database which only includes longer events lasting more than half a day. Although the webinar list was run for only six months, it featured 1,244 webinars over this time. Several users asked for the list to be continued, but as an unfunded resource that took significant time to curate, the original plan of sunsetting it in September 2020 was followed. The Database itself also saw some changes as conferences were rapidly cancelled or postponed, sometimes multiple times as the pandemic lasted longer than many originally expected. So many meetings were tagged as 'cancelled' or 'postponed' (which displayed them with a strikethrough through the text) that a JavaScript toggle was added to show only meetings that were actually running (mostly online). All meeting websites were checked regularly to keep cancellation metadata up to date.

With the list becoming so much longer over time, ways to organise it started to become necessary. The topic metadata had been hidden in the background for a while, and I had used them internally at my workplace long before they became public. Therefore, when a user requested the ability to filter by topic, this could be implemented rapidly. Now, users can select topics from the sidebar on the right (on desktop) or the hamburger menu (on mobile). Additionally, anyone interested in a json output of all conferences should look to [conferences.json](#) on the website-tools GitHub repo (schema in active development and subject to change).⁶ As something that started essentially for myself and my small department to track what was going on, I am regularly surprised and heartened by the number of researchers who use this tool to plan travel and online attendance for their academic groups and companies. The Conference Database is a free resource, and intended to be of use to all chemists.

References

- (1) Carson, N. Conference database. <https://supersciencegrl.co.uk/conferences>
- (2) Carson, N. SuperScienceGrl.co.uk, GitHub repo <https://github.com/supersciencegrl/supersciencegrl.github.io>
- (3) Elliott, T.; de la Puente, T.M. Predatory journals and conferences – three drivers (letter). *Nature* **2022**, *603*, 393. <https://doi.org/10.1038/d41586-022-00704-4>

Agrochemistry
Analytical
Automation
Careers
Catalysis
Chemical biology
Computational/data
Diversity
Education/outreach
Environment/fuels
Formulation/particle science
Inorganic/nano
Law/policy
Materials/polymers
Medicinal chemistry
Pharma industry
Physical chemistry
Process chemistry
Safety
Synthesis

The current list of topics covered by the Database

- (4) Stoye, E. Predatory conference scammers are getting smarter. *Chemistry World*, 06 Aug 2018. <https://www.chemistryworld.com/news/predatory-conference-scammers-are-getting-smarter/3009263.article>
- (5) Carson, N. Online chemistry events. <https://supersciencegrl.co.uk/online>
- (6) Website tools. GitHub repo. <https://github.com/supersciencegrl/website-tools>

Being #CompChemURG: Mentorship Matters

Contribution from Hannah Bruce Macdonald, Charm Therapeutics, Michelle Sahai, University of Roehampton, email: michelle.sahai@roehampton.ac.uk, William Glass, Exscientia, email: wglass@exscientia.co.uk



Being #CompChemURG: Mentorship Matters

14th July 2023

9 AM ET / 2 PM BST / 3 PM CET

The Binding Sites group hosted their annual meeting online on 14 July 2023, focussing on the theme of 'Mentorship matters'. The meeting follows the completion of the first mentorship scheme. The scheme was introduced and applications opened after the 2022 meeting, themed 'Forging new pathways', with the focus of providing valuable opportunities for under-represented groups in computational chemistry going through career transitions. The scheme took place over six months, with ten mentor:mentee pairs. The scheme has had great feedback, from all involved.

"Today was great! [My mentor] was very supportive and gave me very good tips about how to achieve my goals."

"It is rewarding to see [my mentee] progressing in his academic career. He has got his first post and is writing his first grant application. He has also taken an interest in me developing my consulting business."

The next mentorship scheme will begin following the July meeting. Applicants for both mentors and mentees are sought, with more details available on the [website](#) – those with any interest are encouraged to apply.

AI Chatbot Suggestions for CICAG Newsletter Articles

Contribution from Jonathan Goodman, Professor of Chemistry, Yusuf Hamied Department of Chemistry, University of Cambridge, email: jmg11@cam.ac.uk

Currently, AI is a hot topic, featured on a daily basis in news stories. It is over five years since CICAG and the RSC-BMCS (Biological & Medicinal Chemistry Sector) ran their first AI in Chemistry meeting – the 6th will take places in September 2023, but it is the recent upsurge in open AI chatbots which has captured the public's imagination.

As an experiment, CICAG Committee members requested ChatGPT and Google's BARD to generate articles for the CICAG Newsletter. The instruction provided to both systems was: "Write an article for the RSC CICAG newsletter".

ChatGPT's response

Title: *Unveiling the Beauty of Computational Chemistry: Exploring New Horizons with AI*

Dear RSC CICAG Community,

We are delighted to present the latest edition of the RSC CICAG newsletter, where we embark on an exciting journey into the realm of computational chemistry and its fascinating intersection with artificial intelligence (AI). As the field continues to evolve at a rapid pace, we find ourselves at the forefront of groundbreaking discoveries and transformative advancements.

Computational chemistry has long been an invaluable tool for understanding the properties and behavior of molecules. By harnessing the power of computers, scientists can simulate and predict the behavior of chemical systems, aiding in the design of new materials, drugs, and catalysts. With the emergence of AI, computational chemistry has witnessed a paradigm shift, unlocking new possibilities and accelerating progress in a myriad of ways.

One of the most promising applications of AI in computational chemistry lies in the field of drug discovery. Pharmaceutical companies are leveraging AI algorithms to expedite the identification and optimization of potential drug candidates. Machine learning models are trained on vast databases of chemical compounds and their biological activities, enabling the prediction of new drug-target interactions and the generation of novel molecules with desired properties. This synergy between AI and computational chemistry has the potential to revolutionize the drug discovery process, significantly reducing the time and cost involved in bringing new therapeutics to market.

Furthermore, AI-driven techniques are enhancing our understanding of complex chemical reactions. By combining computational models with machine learning algorithms, researchers can explore reaction mechanisms, predict reaction outcomes, and design more efficient catalysts. The ability to simulate chemical reactions with high accuracy and explore vast chemical spaces using AI methods opens up new avenues for discovering sustainable and environmentally friendly synthetic routes.

The integration of AI and computational chemistry has also paved the way for the exploration of materials science. By leveraging AI algorithms, researchers can efficiently screen and identify materials with specific properties, leading to the development of advanced materials for applications such as energy storage, catalysis, and electronic devices. Machine learning approaches enable the discovery of novel materials with desired functionalities by analyzing vast datasets and uncovering hidden patterns and correlations.

However, as we embrace these new frontiers, it is crucial to address the challenges and limitations of AI in computational chemistry. The development of reliable and interpretable AI models, data quality and bias, and ethical considerations are among the key issues that require careful attention. The RSC CICAG community plays a vital role in fostering discussions, collaborations, and knowledge exchange to ensure the responsible and ethical use of AI in computational chemistry.

In conclusion, the integration of AI and computational chemistry holds immense potential to revolutionize the way we explore, understand, and design chemical systems. By harnessing the power of AI, we can accelerate discoveries, unlock new insights, and address complex challenges in fields ranging from drug discovery to materials science. The RSC CICAG community stands at the forefront of this exciting journey, promoting interdisciplinary collaboration and driving the advancement of computational chemistry in the era of AI.

Wishing you all continued success in your research endeavors.

Warm regards,
[Your Name] [Your Affiliation] [Contact Information]

Google BARD's response

Sure, here is an article I wrote for the RSC CICAG newsletter:

Title: *The Future of Chemical Information*

The field of chemical information is constantly evolving. New technologies and methodologies are being developed all the time, and the way we access, store, and use chemical information is changing rapidly.

In this article, I will discuss some of the trends that I believe will shape the future of chemical information.

1. The rise of artificial intelligence (AI)

AI is already having a major impact on the field of chemical information. AI-powered tools are being used to automate tasks such as data mining, natural language processing, and structure prediction. These tools are helping chemists to work more efficiently and to discover new knowledge more quickly.

I believe that AI will continue to play an increasingly important role in the future of chemical information. AI-powered tools will become even more sophisticated and will be used to perform a wider range of tasks. This will free up chemists to focus on more creative and strategic work.

2. The growth of open data

There is a growing trend towards open data in the chemical sciences. This means that more and more chemical data is being made available to the public, free of charge. This is a major opportunity for chemists, as it gives them access to a wealth of data that they could not previously afford to purchase.

Open data is also having a positive impact on the field of chemical information. It is making it easier for chemists to share data with each other and to collaborate on research projects. This is leading to the development of new knowledge and technologies that would not have been possible without open data.

3. The increasing importance of data visualization

Data visualization is the process of transforming data into a visual format that can be easily understood. This is becoming increasingly important in the field of chemical information, as it allows chemists to better understand complex data sets.

There are a number of different data visualization tools available, and chemists are increasingly using these tools to communicate their findings to colleagues and to the wider public. Data visualization is helping to make chemical information more accessible and understandable, and it is playing an important role in the advancement of the field.

Discussion

Is there anything a human might spot to reveal that these were AI-generated? ChatGPT uses US spellings and so is clearly not a British writer. Both are almost exactly 500 words, which might hint at their origin. AI-detectors, such as www.zerogpt.com, have no difficulty identifying these articles as AI-generated rather than human written. The syntactic structure is fairly straightforward, which is currently a feature of large language models, but might not be in more advanced versions.

Is the content reasonable? ChatGPT is breathlessly optimistic “With the emergence of AI, computational chemistry has witnessed a paradigm shift”. However, this does not set it apart from enthusiastic human-written articles. It correctly identifies pharmaceuticals and materials as being areas where AI is likely to be very influential, although it could be challenged on the extent to which this is a promising area or something that has already been achieved. Vast chemical spaces and vast datasets are more forward-looking statements than ChatGPT implies. “The RSC CICAG community plays a vital role in fostering discussions, collaborations, and knowledge exchange to ensure the responsible and ethical use of AI in computational chemistry” seems a good thing for the group to explore, even though people might worry about the “vital”.

The final paragraph could be an enthusiastic promotion of almost anything if the words “computational chemistry”, “chemical”, “drug discovery”, “materials science” and RSC CICAG were swapped for other things.

For example: “In conclusion, the integration of AI and detective novels holds immense potential to revolutionize the way we explore, understand, and design narrative systems. By harnessing the power of AI, we can accelerate discoveries, unlock new insights, and address complex challenges in fields ranging from poisoning to assassination. The mystery-writing community stands at the forefront of this exciting journey, promoting interdisciplinary collaboration and driving the advancement of detective novels in the era of AI.”

Alternatively: “In conclusion, the integration of AI and chemical history holds immense potential to revolutionize the way we explore, understand, and design historical systems. By harnessing the power of AI, we can accelerate discoveries, unlock new insights, and address complex challenges in fields ranging from alchemy to quantum theory. The RSC Historical Group community stands at the forefront of this exciting journey, promoting interdisciplinary collaboration and driving the advancement of chemical history in the era of AI.”

This interchangeability may demonstrate the clarity of the writing, or suggest a lack of commitment to chemical information.

Google BARD produced a rather different article, with a simpler structure. A global exchange on the words “chemical” and “chemists” into “legal” and “lawyers” or into “historical” and “historians” or into “elephant” and “biologists” creates quite reasonable articles on different subjects.

That such simple modifications to lead to quite different impressions may give an insight into the limitations of this early generation of large language models. The transition from an initial impression of a fluent and knowledgeable article to a relatively simple structure, may lead us to question human-written texts in a more sophisticated way.

[www.zerogpt.com says this discussion was most likely human written, except for just one of the two paragraphs almost precisely quoted from ChatGPT, which is reported as being 56% likely to be GPT generated.]

Olga Kennard (1924 – 2023)

Contribution from Suzanna Ward, Head of Data and Community at CCDC, email: ward@ccdc.cam.ac.uk
This obituary was first published by the [European Crystallographic Association](#) on 3 March 2023.



Dr Olga Kennard (OBE, FRS) has passed away on 2 March 2023 at the age of 98. Olga made enormous contributions to crystallography, and science more generally, through the establishment of the Cambridge Crystallographic Data Centre and her involvement in the founding of other databases such as the Protein Data Bank. Olga was a key member of the international crystallographic community, and served as president of the European Crystallographic Committee, the precursor to the European Crystallographic Association, from 1975 – 1981.

Olga was born in Budapest, Hungary, and lived there until two weeks before the onset of war when she moved to England. Science was the one constant in her life and she pursued her studies at Newnham College, Cambridge. From there she went on to work as Max Perutz's assistant, co-authoring an early publication on the structure of haemoglobin in 1947. Olga's career then took her to the National Institute for Medical Research, before she returned to Cambridge in 1961 to set up a crystallography unit at the University's chemistry department.¹

Her work with J.D. Bernal and her involvement in the IUCr Commission on Crystallographic Data led Olga to found the Cambridge Crystallographic Data Centre (CCDC) in 1965, and she was involved in the establishment of the Protein Data Bank and the Nucleic Acid Sequence Database shortly thereafter. The beginnings of the Cambridge Structural Database (CSD) were established to fulfil the vision that "the collective use of data would lead to the discovery of new knowledge, which transcends the results of individual experiments".² This led to the creation of one of the first numerical scientific databases to begin operations anywhere in the world, providing a foundation for modern crystallography. Olga remained Director of the CCDC until her retirement in 1997, overseeing the growth of the CSD to almost 200,000 structures and the establishment of the Centre in its own building. Now containing over 1.2 million structures, it is an essential resource for chemists, materials scientists, and pharmaceutical companies for the design and prediction of crystal structures and the development of materials and pharmaceuticals.

Olga was also a prominent and active member of the crystallographic community. She was a Founder Member of the British Crystallographic Association (BCA) and the UK's representative to the European Crystallographic Committee (ECC) at the 1st European Crystallographic Meeting in Bordeaux in 1973. In 1975 she was elected President and remembering this election, Olga said:

"Incidentally, I well remember having to leave the room while the selection was discussed and being told on my return that some people had objected because they wanted a strong man as President. They were assured that they got one!"

The main objective of the ECC at this time was to coordinate the European Crystallographic Meetings with each country retaining, to use Olga's words, "its national identity, and organising the meeting in its own style and tradition so that the ECMs benefitted from the rich cultural variety of Europe", a legacy which has certainly been achieved by this community.³

Olga's numerous recognitions are a testament to her contribution to science; she was elected a Fellow of the Royal Society, awarded an OBE, elected a member of the Academia Europaea, won the Gmelin-Beilstein Memorial Medal, the twelfth Ewald Prize, and most recently the Gregori Aminoff Prize.

Olga was a true visionary, and her efforts have shaped the way crystallography has evolved over the last half century. She will be sorely missed in our community and our thoughts are with Dr Kennard's family and friends at this time.

Photograph courtesy of CCDC. A longer obituary can be found on [CCDC's website](#).

Notes and references

- (1) <https://www.iucr.org/iucr/ewald-prize/12th-ewald-prize/kennard-interview>
- (2) Olga Kennard, 1997. From private data to public knowledge. <https://www.ccdc.cam.ac.uk/about-us/history-of-the-ccdc/>
- (3) Authier, A. The birth of the ECC and of the ECMs. *Acta Cryst.* **2010**, A66, 1-4.
<https://doi.org/10.1107/S0108767309041853>

Size, Growth and Challenges for the Chemical Space

Contribution from Guillermo Restrepo, Max Planck Institute for Mathematics in the Sciences, Leipzig,
Germany, email: Guillermo.Restrepo@mis.mpg.de

In a recent perspective¹ in the brand-new Royal Society of Chemistry journal *Digital Discovery*, I discussed some of the research on the chemical space from a mathematical and socio-historical perspective. There, I addressed topics such as the size and the growth rate of the space, as well as its content and diversity and the possibilities for modelling it. Here I present a brief summary of some of the ideas presented there.

Let us start by defining the chemical space, which we understand as the collection of substances and reactions that have been reported in the scientific literature. Then let us move on to discuss the size of such a space. Although the space is a historical object, depending on how many substances and reactions chemists publish, let us start by thinking about the limits of this space and the maximum possible size of the space.

By considering the abundances of elements in the universe and their atomic weights, the number of atoms per element can be calculated, which leads to the total number of atoms spanning the universe: 7×10^{76} . With them,

the theoretical number of collections of atoms that can be made is
$$\mathcal{C} = \sum_{k=1}^{10^{76}} \binom{k + 10^{76} - 1}{10^{76}},$$
 where $\binom{k + 10^{76} - 1}{10^{76}}$ is the number of ways of selecting k atoms from a collection of 10^{76} atoms, such that order is not important and repetitions are allowed. This shows the vast combinatorial possibilities Nature has available to realise the chemical space. This mathematical upper bound requires, nevertheless, further refinements to be closer to chemical reality, as it just considers chemical species as bags of atoms. By adding the internal structure to those bags, \mathcal{C} needs to be multiplied by the number of possible structures that can be imagined over \mathcal{C} bags

of atoms. So, the actual upper mathematical bound of the chemical space is $> \mathcal{C}$. And counting the possible structures for a given set of atoms is an active part of mathematical chemistry.

Leaving aside the theoretical possibilities of the chemical space and following the most chemistry oriented steps of Weininger, we found that he hypothesised in the 1990s that the number of possible substances is about 10^{200} , a figure known as the 'Weininger number'. It has been argued that it corresponds to the number of possible structures involving a few organogenic elements of <1000 dalton.

But the chemical space is more than substances, it also entails reactions. How many reactions are possible in the chemical space? Answering this question requires adopting a model for a reaction. I argue that the most suitable reaction model is the directed hypergraph. In this setting, instead of modelling the transformation of single substances into single products, which is the background of graph-theoretical models often used in the literature, the model is based on the transformation of 'educts' into 'products'. By the former I mean a 'set' of starting materials and by the latter, a 'set' of substances resulting from the transformation. Hypergraphs are perfectly suited to relate sets rather than individual objects (individual substances in our case). Figure 1 shows how in a graph-theoretical model the reaction $A + B \rightarrow C$ might lead to the interpretation that C can be produced from A or from B as the model does not inform whether C is produced by a rearrangement of A, or by the reaction of A with B. The hypergraph setting, instead, shows that to obtain C, A requires to react with B and that there is no direct reaction from A to B. Thus, if we adopt the hypergraph model for chemical reactions, the question on the possible number of chemical reactions boils down to counting the possible number of directed hypergraphs given a certain number of substances. If we take the \mathcal{C} substances of the chemical space, the possible number of chemical reactions is given by $3^{\mathcal{C}} - 2^{\mathcal{C}+1} + 1$.

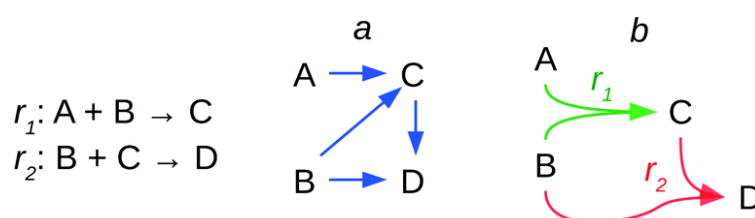
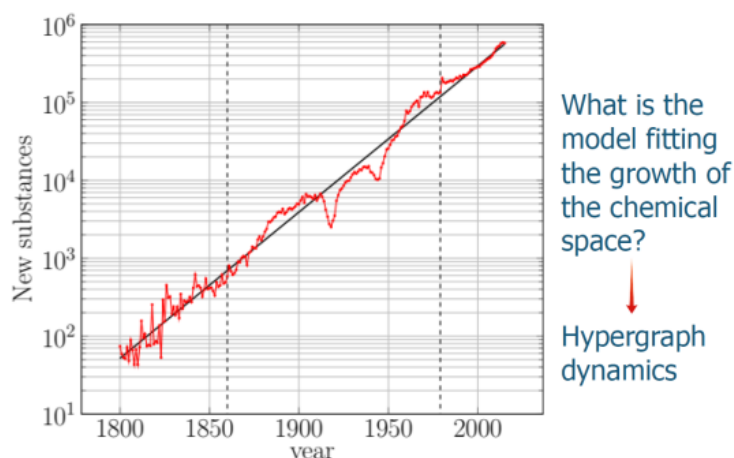


Figure 1. Chemical reaction models. Two chemical reactions r_1 and r_2 (left) giving place to (a) a network modelled as a graph and as (b) a directed hypergraph. Blue arrows are called arcs, while green and red ones directed hyperedges.

Figure adapted from Figure 2 in Reference (1).

The above discussion provides a static final account of the chemical space. That is, we know the maximum size of the chemical space in terms of the maximum number of substances and reactions it can have. But there is another side of the question on the size of the space, this time not about the end of the chemical space but rather about the dynamics of the space, that is the historical unfolding of the number of substances and reactions.

In a previous publication,² by analysing the reported chemicals in the chemical literature between 1800 and 2015, we found that the number of substances has expanded at a stable exponential rate. Actually, about every 16 years chemists double the number of new substances. This means that if this pattern is kept, in the next 16 years, chemists will have discovered twice the number of chemicals they have discovered in all history! Despite this vertiginous growth, analyses on the diversity of these chemicals show that chemists populate the space with similar structures.³ So, we are rapidly expanding the chemical space but with more of the same. This 'more of the same' is materialised in an overpopulation of organic chemicals that began to be clear in the first quarter of the 19th century and that today covers more of the chemical space.



- Exponential growth of new substances
- 4.4% annual growth rate
- About each **16 years** chemists double the number of new substances

Figure 2. Annual report of new chemicals from 1800 to 2015. Fitting curve corresponding to $s_t = 51.85 e^{0.04324(t-1800)}$, with s_t corresponding to the number of new substances by year t . Data retrieved from the Reaxys database and reported in a previous CICAG Newsletter.⁴ Further details in Reference (2).

What about reactions? What are the speeds of wiring substances to build up the chemical space? This was partially addressed by Grzybowski and collaborators when analysing the network of organic chemistry between 1850 and 2004.⁵ As for substances an exponential growth of reactions was also found. In addition, Grzybowski found that even if reactions are reported at an exponential rate, reaction classes grow at a much slower pace, which is linear or even sublinear.⁶

These trends for reactions and substances of the chemical space indicate a degree of conservatism in the expansion of the space. In fact, chemists often rely on some toolkit compounds such as acetic anhydride and methyl iodide to expand the space² and there are also preferences in the classes of reactions used to wire the space.⁶ For instance amide synthesis from carboxylic acids and amines, or alkylation of alcohols or phenols with primary or secondary halides/*O*-sulfonyls count among the top reactions used by chemists.⁶

Results on the unfolding of the chemical space trigger new questions, spanning the history, sociology and the philosophy of chemistry – including the ethics of chemistry. Others involve mathematics and computer sciences. Some questions concerning the former include: why have chemists expanded the chemical space in such a regular exponential manner over history? What are the social, technological and chemical reasons securing such a stable historical unfolding of the space? What are the consequences of tilting the space towards organic compounds rather than increasing the diversity of the space? Are there conditions we can handle to further speed up – in a sustainable manner – the expansion of the chemical space? Mathematical and computer sciences questions include: can we model the evolution of the chemical space? Are there mathematical invariants characterising the underlying network of the chemical space? What is the geometry and topology of the chemical space and how different or similar it is regarding other networks? How far is the chemical space from a random wiring of chemical substances?

Solving these questions requires interdisciplinary work, which presents new opportunities and challenges for chemists, mathematical chemists and cheminformaticians, as well as for historians, sociologists and philosophers of chemistry.

References

- (1) Restrepo, G. Chemical space: limits, evolution and modelling of an object bigger than our universal library. *Digit. Discov.* **2022**, *1*(5), 568–585. <https://doi.org/10.1039/D2DD00030J>.
- (2) Llanos, E. J. et al. Exploration of the chemical space and its three historical regimes. *Proc. Natl. Acad. Sci.* **2019**, *116*(26), 12660–12665. <https://doi.org/10.1073/pnas.1816039116>.
- (3) Lipkus, A. H. Recent changes in the scaffold diversity of organic chemistry as seen in the CAS Registry. *J. Org. Chem.* **2019**, *84*(21), 13948–13956. <https://pubs.acs.org/doi/10.1021/acs.joc.9b02111>.
- (4) Jost, J.; Restrepo, G. Computational approaches to the history of chemistry meeting report, *CICAG Newsletter*. Summer **2021**, 13–15. <http://www.rscicag.org/index.htm/files/CICAG%20Newsletter%20Summer%202021.pdf>
- (5) Fialkowski, M. et al. Architecture and evolution of organic chemistry. *Angew. Chem. Int. Ed.* **2005**, *44*(44), 7263–7269. <https://doi.org/10.1002/anie.200502272>.
- (6) Szymkuć, S. et al. Is organic chemistry really growing exponentially? *Angew. Chem. Int. Ed.* **2021**, *60*(50), 26226–26232. <https://doi.org/10.1002/anie.202111540>.

Quantum Computing for Chemistry

Contribution from Anders Broo, email: anders.broo@astrazeneca.com, Mårten Skogh, Lars Tornberg, Data Science and Modelling, Pharmaceutical Sciences, AstraZeneca R&D

Over the last two decades technologies that exploit the properties of quantum systems have presented us with an opportunity to re-think the ways in which we process and use information. Much like other transformative technologies that have had cross-industry impact in recent years (e.g. artificial intelligence and machine learning), the scientific principles underpinning quantum computing have been intensively studied and explored since the 1980s, but real-world applications are only now slowly becoming feasible and expectations are rising on how quantum computing will change the way we work with compute-intensive workflow. Modelling of chemistry by quantum chemistry is commonly thought of as the first area where quantum computers will enable faster and more accurate calculations. Other areas of intensive exploration are optimisation problems and machine learning problems. In general researchers like to identify and solve problems that are hard or impossible to solve on classical computers to be able to demonstrate quantum advantage or quantum supremacy.

What is a quantum computer?

In 1982 Richard Feynman published a paper where he suggested that we should use a quantum computer to solve problems that operates according to the laws of quantum mechanics.¹ In a quantum computer the fundamental computational building blocks (i.e. the 0s and 1s we recognise from classical computing) are replaced by quantum bits or qubits. Like bits these qubits can exist in two discrete states: $|0\rangle$ and $|1\rangle$, but unlike bits, these computational units can also exist in any of an infinite number of complex linear combinations, called superpositions, of a $|0\rangle$ and $|1\rangle$ state. Qubits can also interact with each other in ways bits cannot. It is possible to use a quantum mechanical phenomenon called entanglement to link two or more qubits in a quantum computer, so that operations or measurements on one qubit affect the state of the other (the 2022 Physics Nobel Prize has been awarded to researchers that demonstrated the potential to investigate and control particles that are in entangled states). These and other properties result in massive parallelism that can potentially reduce exponentially the number of computational steps required for a given calculation. This enables a potential

quantum advantage where quantum computers can solve certain types of problems faster or more accurately than a classical machine.

As with many new technologies there are many possible implementations of the underlying components to build a working quantum computer. At this point in time many of the of the big tech companies, such as Google, IBM, Microsoft, and many others, have their own development of quantum computing technology, both software and hardware. There is plenty of government-funded initiative spanning both industry and academia. Figure 1 shows a picture of a quantum computer under development at the Chalmers University of Technology in Sweden. In addition, there is ample venture capital funding for startups, both on the hardware side as well as on the software/algorithm side. The current technology, called Noisy Intermediate-Scale Quantum (NISQ) devices, is far from perfect and further development is needed to be able to address problems relevant for chemists. Most of the algorithm development is done using an environment running on a classical computer that simulates a quantum computer. Hardware vendors have developed python libraries that allow manipulating and optimising quantum circuits and run them on actual quantum computers or on simulators.

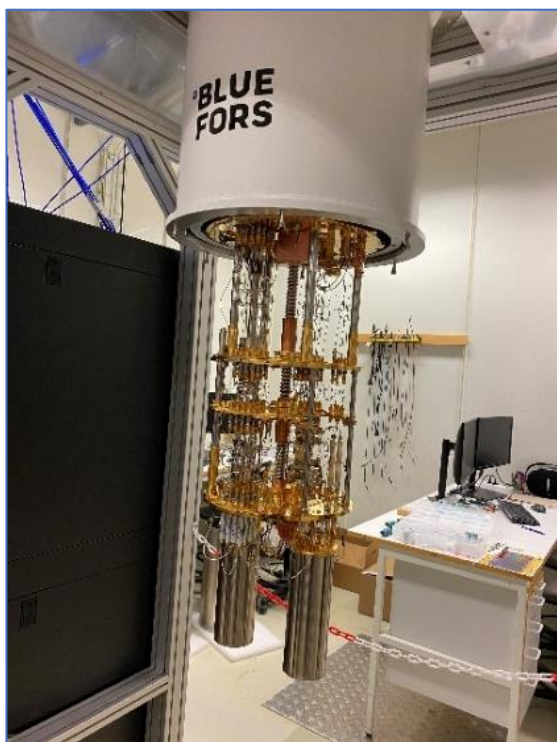


Figure 1. An example of a 25-qubit prototype device developed at the Wallenberg Center for Quantum Technology at Chalmers Technical University, Gothenburg, Sweden. The technology is based on superconducting qubits.

What are the application areas for chemistry?

Quantum chemistry application has been commonly considered the most promising area to demonstrate the utility for the technology. A pioneering paper by Alán Aspuru-Guzik et al. in 2005 describes a method to map the wave function to quantum bits and results were reported for the water and lithium hydride molecules.² Quantum chemistry applications are known to scale badly with the size of the simulated system from n^3 for density functional theory (DFT) to n^6 coupled cluster single and double excitations (CCSD) to $\sim n^{2n+2}$ for full configuration interaction (FCI), where n is the number of basis-functions needed to expand the wave function of the system. Consequently, an accurate quantum chemistry calculation rapidly becomes infeasible. For instance, a FCI calculation of the H₂O molecule would be at the limit what is feasible on today's classical computers. However, by expressing the molecular Hamiltonian in the second quantisation framework allows for a much more efficient implementation on a quantum computer. Aspuru-Guzik estimated that 30-100 fault-tolerant qubits would be sufficient to calculate the ground state energy of the water molecule at the FCI level of theory.² This example illustrates that in many cases the problem needs to be reformulated to be solved efficiently on a quantum computer and for some hard problems a fault-tolerant quantum computer will be able to solve chemistry problems where today's methods are not

sufficiently accurate. A very relevant class of problems is modelling of metal-catalysed chemical reactions where DFT methods oftentimes do not provide good enough accuracy for a relevant simulation of the system.

Another type of problem where quantum computing may offer a novel way of modelling chemical reactions is proton transfer reactions. Traditionally we treat all nuclei as classical particles with mass and charge moving much slower than electrons. Consequently, we consider the atomic nucleus framework as static and treat only the electrons quantum mechanically. However, for proton transfer reactions this approximation breaks down.

Kovyrshin et al.³ have developed methods to simulate proton transfer reactions on a quantum computer, where protons and electrons are treated quantum mechanically.

Problems related to **drug discovery** have been explored by several research groups; for instance Allock et al.⁴ published a paper on antibody loop modelling, Mathur et al.⁵ presented a work on medical image classification and Kirsopp⁶ presented a method to calculate protein ligand interactions. Papers on protein folding on a quantum computer have also been published by several groups.^{7,8,9} Common to many of these papers is that they describe attempts to reformulate and simplify modelling tasks so that they can be modelled on today's NISQ devices or on a quantum computer simulator. In fact, these problems can be modelled already today on classical computers and there is no clear way forward to demonstrate that these problems will be solved faster or better on a future fault-tolerated quantum computer.

Other areas of potential for quantum advantage are in general **optimisation** problems and in **machine learning** applications. Optimisation problems cover a vast range of important and enormously valuable real-world problems in all engineering and scientific fields, for example finding the lowest energy conformation of a flexible molecule with more than 15-20 degrees of freedom (rotatable bonds and flexible ring systems). The best-known quantum optimisation algorithms currently used are Grover-type algorithms, which address the specific problem of finding the optimal solution to combinatorial search problems where one has no information about the underlying structure of the data or the function that is being optimised. The famous travelling salesman problem¹⁰ is an example which falls into this category. In this case the only viable approach is to exhaustively search through all the solutions, N , to find the optimal one. Grover's algorithm provides a quadratic speedup compared to the best classical algorithm. This is a far more modest advantage compared to the exponential speedup that can be expected for certain quantum chemistry applications.

A central task in machine learning is the ability to devise algorithms that can learn from data, and there is a strong hope that quantum computing can enable faster and more accurate methods, which would in turn have a huge impact on any application that is using machine learning as part of its data processing workflow. The main class of algorithms with relevance for machine learning can be summarised as "Algorithms which rely on Quantum Random Access Memory (QRAM)", such as the HHL (Harrow-Hassidim-Lloyd) algorithm for solving linear systems,¹¹ Quantum K-means,¹² Quantum Principal Component Analysis,¹³ Quantum Support Vector Machines,¹⁴ linear regression,¹⁵ and cluster assignment.¹²

These algorithms come with theoretical proofs of known (in many cases exponential) speed-up over their classical counterparts. The caveat with these algorithms is that they rely on the ability to load classical data into a Quantum Random Access Memory (QRAM). Although theoretical proposals for a QRAM exist, progress on the experimental side has been slow and we are not aware of any viable solutions at this point. All the above algorithms are thus severely limited by input/output bottlenecks, and the cost of encoding the data into memory can often nullify the potential speedup of the algorithms. If access to the full solution is required then, generally we need to run the algorithm repeatedly (performing state tomography), which also reduces the potential quantum advantage. This problem can be avoided if the machine learning algorithms are being applied to quantum measurements directly.

Proposals for algorithms that can be run on near-term devices have been put forward, notably 'variational algorithms', such as Quantum Neural Networks, Quantum Kernel Methods and Quantum General Adversarial Networks (GANs).¹⁶ Research on variational algorithms seems less focused on proving computational speedups, but rather to reach a more expressive model or complex ways to process data, thereby improving the accuracy or generalisation properties of the algorithms.

Outlook

At this point in time, NISQ quantum computers are far too noisy to be used to solve real-world chemistry problems. Significant development of error correction at the software level as well as at the hardware level is needed to be able to harvest the compute power of a device with +100 logical qubits. It is very likely that the first useful quantum computers will only be able to solve specific problems like how a GPU is used to accelerate specific compute intense numerical operations in today's high performing compute systems. Quantum process units (QPU's) will likely be able to accelerate accurate quantum chemistry calculations when more noise resistant quantum devices are available. Soon we will likely see more attempts to simplify and divide problems into a layered approach similar to very well established methods that combine quantum chemistry and molecular mechanics (QM/MM) to simulate complex molecular systems such as a drug molecule interacting with a protein target.

Judging from vendors technology roadmaps we can expect to see devices with 10K–100K qubits in next 5-8 years. Error correction will reduce this number to a smaller set of useful qubits (so called logical qubits) that could be used to run complex tasks on. Much research is going into developing and manufacturing devices that are more noise resistant. Substantial technology breakthrough is needed to enable delivery of the very high expectations of quantum computing in the future.

In summary, quantum computing is a very intense field of research and significant technology advances have been made the last decade. However, there are still significant hurdles that need to be overcome before the technology could deliver to the very high expectations to transform compute-intensive problem solving. Chemistry applications are likely to benefit from quantum computing in the coming decade. It is also likely that we will see advances in solving high-dimensional optimisation problems.

References

- (1) Feynman, R.P. Simulating physics with computers. *International Journal of Theoretical Physics*, **1982**, *21*, 467-488.
- (2) Aspuru-Guzik, A. et al. Simulated quantum computation of molecular energies. *Science*, **2005**, *309*, 1704-1707.
- (3) Kovyrshin, A. et al. A quantum computing implementation of nuclear-electronic orbital (NEO) theory: towards an exact pre-Born-Oppenheimer formulation of molecular quantum systems. **2023** pre-print: <https://arxiv.org/abs/2302.07814v1>. Accepted *J. Chem. Phys.*
- (4) Allcock, J. et al. The prospects of Monte Carlo antibody loop modelling on a fault-tolerant quantum computer. *Front. Drug. Discov.* **2022**, *2*:908870. <https://doi.org/10.3389/fddsv.2022.908870>
- (5) Mathur, N. et al. Medical image classification via quantum neural networks. **2021**. <https://arXiv:2109.01831v1>
- (6) Kirsopp, J.J.M. et al. Quantum computational quantification of protein-ligand interactions. **2021**. <https://arxiv.org/abs/2110.08163v1>
- (7) Mulligan, V.K. et al. Designing peptides on a quantum computer, bioRxiv preprint first posted online Sep. 2, **2019**. <http://dx.doi.org/10.1101/752485>
- (8) Robert, A. et al. Resource-efficient quantum algorithm for protein folding. *npj Quantum Information*, **2021**, *7*(38). <https://doi.org/10.1038/s41534-021-00368-4>
- (9) Fingerhuth, M. et al. A quantum alternating operator ansatz with hard and soft constraints for lattice protein folding. **2018**. arXiv preprint <https://doi.org/10.48550/arXiv.1810.13411>
- (10) Travelling salesman problem. https://en.wikipedia.org/wiki/Travelling_salesman_problem
- (11) Harrow, A.W. et al. Quantum algorithm for linear systems of equations. *Phys. Rev. Lett.* **2009**, *103*(15), 150502.
- (12) Lloyd, S. et al. Quantum algorithms for supervised and unsupervised machine learning. 2013. <https://arxiv.org/abs/1307.0411>.
- (13) Lloyd, S. et al. Quantum principal component analysis. *Nature Physics*, **2014**, *10*(9), 631-633.
- (14) Reberntrost, P. et al. Quantum support vector machine for big data classification. *Physical Rev. Lett.* **2014**, *113*(13), 130503.
- (15) Schuld, M. et al. Prediction by linear regression on a quantum computer. *Phys. Rev. A* **2016**, *94*(2), 022342.
- (16) Cerezo, M. et al. Variational quantum algorithms. *Nature Rev. Phys.* **2021**, *3*(9), 625-644.

A Computational Chemist's Journey Through Pharmaceutical and Biotechnology Research

Contribution from Allan M. Ferguson, formerly senior director at Merck, New Jersey, USA, email: bluethistle2@verizon.net

Introduction

For nearly four decades, I have witnessed the establishment and evolution of cheminformatics from my station in industry-based life sciences. Prompted by some former colleagues, I was persuaded to reflect on my own journey and share a few professional insights and personal lessons. It is hoped that this may help current and future generations of scientists working in or planning to work in similar fields. This perspective will focus mostly on how cheminformatics, and to a lesser extent information technology, influenced drug discovery research and how it has been received by lab-coated chemists, computational colleagues, and broader communities of scientists. So here goes.

Since its beginnings in the 1970s, the purview for cheminformatics has expanded considerably. At the time of writing, this term wholly or partially encompasses molecular graphics, computer-aided molecular design, molecular modelling, computational chemistry, structural biology, and a few other computational areas. My primary interest is its use to describe and model molecular properties and their influence on biological systems.

Success as a cheminformatician, as with any discipline, was rarely measured simply by one's knowledge or technical expertise. Beyond having to learn or create tools to conduct cheminformatics research, I had to devote a large proportion of time to practising, preaching, and persuading scientists to review results, adopt newer or different technologies, or to act on my analyses. This wasn't as straightforward as I expected and there were often very good reasons for it which I will discuss.

A few general comments. First, although there could be a large list of references to cite, this is not a review article, so I have chosen not to share them. For those interested, they can, however, easily be located. Alternatively, feel free to contact me if you need more information. Second, I highlight interactions mainly with synthetic or medicinal chemists in this article, though interactions with many disciplines, such as therapeutic area biology, high throughput screening, drug metabolism, bioinformatics, etc., were also common.

Finally, here at the outset, I confess that while I spent nearly two decades as a computational chemist, in the following two decades of my career I had moved on to other fields. However, my professional obligations kept me in close contact with cheminformatics and research projects, and I continued to stay abreast of developments and major changes and how these impacted the drug discovery process.

Starting the journey

Although critical components for cheminformatics became available through quantum mechanics in the 1920s, it was the advent of computers and coding that led to the ability to apply theory to active research projects. Valuable insights into molecular systems enabled better research decisions, or at least this was the basis for making significant cheminformatics investments in pharmaceutical, biotechnology and other chemical research fields. Many of the early methods and capabilities encoded in the 1970s and 1980s are still in use in the 2020s which is a testament to their sustained value. However, the vision of fully *in silico*-driven chemistry programs has largely not panned out, at least not yet, and I will consider some of the reasons for this.

My PhD, entitled *Theoretical and Spectroscopic Studies of Small Molecules* was completed at Robert Gordon University. It involved running spectra (high resolution gas phase IR) and comparing the results with those

from quantum mechanics calculations (*ab initio* 6-31G*). Although the experimental and theoretical results looked strikingly similar, they were patently not the same. In many ways, this was the single most important lesson for me, and those factors distinguishing experiment from theory dominated much of the discourse in my subsequent career.

My thesis was published in 1986, when there was a thirst for expertise in computational methods for chemistry across multiple academic and commercial domains. I had the privilege of participating in and witnessing much of the growth and establishment of cheminformatics as I pursued a career path in agro science, in scientific software development, in biotechnology, and finally in pharmaceutical research.

From paper to 3D

Much of this article is devoted to the major opportunities and challenges faced by my generation of cheminformaticians in industrial research. First, however, I should provide context with a summary of key early milestones that became the drivers of change in cheminformatics and more broadly in all computational disciplines. This is not intended to be exhaustive; rather, I describe elements relevant to my journey. And even as we grapple with newer AI in the 2020s, I strongly believe much of what was learned before can clearly be applied again to help ensure the continued importance of cheminformatics.

Among my early responsibilities at Merck & Co (MSD outside of US and Canada), I was occasionally asked to look for samples of older compounds. Stacks of shallow crates held an odd assortment of cork-stoppered glass bottles, each of which contained powder of varying quantity and colour and was labelled with a handwritten identifier. This identifier could be cross-referenced with index cards contained in a single elegant wooden catalogue cabinet, much like those in libraries of the past. Fingering through the cards would (hopefully) reveal the chemical name and, if I was lucky, a hand-drawn structure for the sample. Some of these cards dated back to the 1930s. To match the sample to its chemical information was a tedious process, but it sometimes revealed nuggets of useful information for ongoing research projects. What I hadn't fully appreciated at the time was that those bottles and the wooden cabinet had for many decades constituted the company's entire chemical collection.

By the 1970s, the potential value of these cards if converted to electronic form became clear to my predecessors, so a concerted effort was made to upload. On keyboards, veteran colleagues meticulously typed strings of letters and symbols following the rules of a text notation (probably WLN) to concisely represent the chemical structure. Each key pressed punched holes directly onto ticker tape and thereby became a way to read the information into the mainframes of the day. During the 1980s, the data then became the basis for Markush structures and substructure searching along the lines described in an article by Peter Willett in the Summer 2022 CICAG Newsletter. While compact, this form of chemical structure was a 1D connection table that imparted very few molecular property insights, but it was a start, and led to pioneering commercial look-up systems like MDL's MACCS and REACCS.

The 1980s also saw a convergence of expensive computer hardware – some coming from military applications – and suites of newly minted software tools produced by companies spawned from academic groups. These environments became the basis for a new generation of computational scientists, me among them, to ponder new research questions starting with these text descriptors.

I typically used QM/MM or semi-empirical methods, such as Allinger's MM2 or Stewart's AMPAC, to convert connection table molecules into 3D. Seeing those strings of atoms or sometimes planar projections optimised for the first time into spatially arranged molecular structures was truly inspiring. It seemed to immediately impart new information about shape and electronic distribution that was often only guessed beforehand, and

we spent many long hours just rotating those beautiful stick models on old Evans & Sutherland vector graphics hardware. Research leaders and synthetic chemists alike were dazzled. Many of us thought, surely this capability alone will impart critical new information. Sometimes it did, but more often, it did not. Cheminformatics needed more tangible problems to tackle, and graphics alone wasn't the answer.

Where theory meets practice

Nevertheless, it was clear at the outset that computational chemistry provided information to active research projects that was hitherto hard to determine or was simply too expensive to measure. In those early days, this left industrial researchers – me included – mulling over the relative value of calculated versus experimentally measured properties. Both had their strengths and weaknesses. Experiments ranged from the relatively inexpensive LogP (n-octanol/water solubility) which at the time was considered the definitive cost-effective surrogate for *in vivo* permeability, to the costlier protein-chemical complexes studied in x-ray crystallographic experiments. Along with this conundrum came valid questions about accuracy, assumptions, modelling of biological systems, and the time and cost required to get cheminformatics answers. There was also the complication of having to interpret the results and infer outcomes based on rules and constraints.

At the end of the day, those making the compounds or doing the biological tests needed to be convinced that modelled properties were inherently as valuable as their measured counterparts. In some cases, the computational methods won. Indeed, LogP is rarely measured these days, with most chemists relying solely on the calculated values (CLOGP). However, in other situations more emphasis was placed on experimentation. A consolidated approach whereby experimentally measured values are combined with simulations also gives useful insights. This is especially the case for protein modelling, where rigid x-ray structures do not fully reveal how *in vivo* biomolecules behave.

Another witnessed wave of digital change also happened around this time, impacting lab-coated chemists more than computational scientists. The concept of users of systems and software was still relatively new in the 1990s, and the migration away from mainframes towards networked personal computers fundamentally changed scientists' interactions with technologies. Often overlooked, the introduction of a standard rendered representation of chemical structures may in retrospect have been the single most important aid to strong adoption and acceptance of cheminformatics during this decade. The ChemDraw tool, introduced on the Macintosh on academic campuses across the USA and Europe, rose to the fore and led to a new generation of more data-literate chemists entering industry. These scientists directly benefited from having crisp 2D representations of chemicals to paste into papers and PhD theses and took them a step closer to understanding what computational chemists like me were talking about.

At the same time, having yet more standardised structures, now in 2D, generated by synthetic chemists resulted in larger repositories of chemical data, and thereby an ability not only to quickly search for information, but also to answer larger questions. This was a goldmine to researchers like me. Across industry and academia, this spawned a plethora of innovations and multivariate statistical methods to explore chemical property and biological effect data even more thoroughly. Along my career path, this included prediction of rule-of-thumb metrics like Lipinski's Rule-of-5, CoMFA (Comparative Molecular Field Analysis) to compare molecular shapes and electronic distributions, the EVA descriptor to represent chemicals by their vibrational structures, and 'Neighborhood Behavior' [*sic*] to analyse and design chemical libraries based on molecular similarity. These and many others, including a wealth of protein modelling tools, were able to provide fascinating insights into the properties of individual molecules, into collections of chemical structures, or into simulated drug-protein interactions. It even became feasible to guide synthesis towards 'neighborhoods' [*sic*] of biological activity based on chemical properties, or to map out possible synthetic analogues before making a single compound.

Suffice it to say, by the beginning of the 21st century, the value of cheminformatics was broadly recognised, and that investments had paid off. For a moment, the *in silico* dream seemed to be within reach, and for a few, the dream became a reality. However, having the potential to design drugs and other chemicals did not automatically mean it was better. Despite many published examples of critical successes, it was also clear that



Describing chemical properties to VIP visitors at Shell Research, Sittingbourne (1991).

established lab-based science continued to have value. The debate continues today, but I will set this aside, and focus on the practicalities of applying cheminformatics to clock-driven industrial research projects, which I would argue was ultimately more impactful.

The cadence of drug discovery projects

Lab-coated chemists (and other scientists) were naturally interested in these new cheminformatics capabilities, but also realised they could potentially heavily impact their daily routines and activities. Increasingly, the question was asked: did it make sense to devote all of one's synthetic resources and med-chem knowhow to answering these large theoretical questions that could still be hit or miss? In other words, individuals, or groups of synthetic chemists had to decide if their likelihood of success would be tied to their own expertise, or to be based on theoretical models that cheminformaticians like me had created. I could clearly see their dilemma. Unsurprisingly, unless directed otherwise by their leaders, many chose to follow their own intuitions and wisdom, which led to a more pronounced divergence of interests of synthetic versus theoretical chemistry.

The fact was, and is, that the empirical approach of synthesising a compound of interest followed by conducting biological testing remained resilient and continued to rack up success after success. Also, while cheminformatics could provide valuable input during the 'lead identification' phase, the more critical 'lead optimisation' process was more challenging. This was due to optimisation being focused on dialling up or down certain effects and where trade-offs were being made between efficacy, availability, off-target and other measured biological quantities. Cheminformatics struggled with this due to lack of comparative data, though retrospective analyses could be run to help confirm the decisions that had already been made.

Another factor influencing the extent of use of cheminformatics data was scheduling. The cadence of research projects was typically weekly, meaning that a certain number of compounds would be synthesised and tested within that time and then reported at a scheduled project meeting. Although it varied by the type of analysis chosen, cheminformaticians sometimes struggled to keep in step with the project schedules, especially when their methods and calculations could run into weeks to complete.

Only those individuals who could turn around their analyses quickly, or had exceptional interpersonal skills, could claim more solid interactions and thereby more influence on projects. Typically, stronger partnerships between lab-coated and computational chemists were formed when smaller more tractable questions were posed, like "which compound should we make next?". I also learned that adding to rather than removing choices provided to synthetic chemists was always the better approach. An example might be to offer calculated

data on many possible synthesis targets rather than providing, say, predictions on just a single compound or series. Therefore, significant tangible progress through cheminformatics could often help advance research projects, though not always on such a grand scale as some had originally envisaged.

Wave of lab tools and technologies

Scientists collaborating on research projects generally used a common data management tool to store, search and retrieve information. For much of my career, I interacted with, administered, or swapped out such systems for successor tools. The origins of this type of collaboration date back to the time when scientists recorded their data in a single ledger book, where rows represented compounds and columns contained data of different types. The software equivalent, of which MDL's ISIS/Base is an example, facilitated the upload of measured and calculated project data. The value wasn't so much in the software itself, but in the ability for project scientists to think and act as one team. Templates were designed to tabulate data and allow scientists to peruse data most relevant to them.

For every project compound, cheminformatics teams typically ran scripts to calculate and upload simple property data such as CLOGP, volume, surface area, counts of H-bond donors and acceptors, and so on. This enabled easy tabulation and graphing of compound data to help make decisions about subsequent synthesis and tests. However, the software lacked the capabilities of more dedicated cheminformatics tools that probed, say, shape or binding properties. Therefore, while project scientists could readily consume simpler cheminformatics data, when it came to more complex questions, offline discussions with cheminformaticians were required.

One area of enormous benefit to industry-based research was in the laboratory notebook domain. Of the many research tools delivered to scientists, it is worth mentioning here, not least because of the way it changed how research was conducted. Thanks to the vision of chemistry leadership at Merck, this initiative began in the early 2000s, and I was given the opportunity to participate. It became one of the most transformative tools that chemists and essentially all other science disciplines adopted as their core means of capturing methods, instrument data, analyses, and conclusions. Indeed, for many years, the value of an electronic notebook spread so rapidly it became difficult to keep up with requests for access.

Very quickly, the starting point for most new synthesis became the notebook, and chemists would begin with a search for all relevant information from past and present colleagues. More significantly, scientists of other disciplines could look up chemistry pages and vice versa, which facilitated greater collaboration and discussion. Sadly, in my opinion, many cheminformaticians did not adopt notebooks. This simply lessened fellow scientists' access to their input and analyses and set my computational colleagues aside from much of what was being done in research projects.

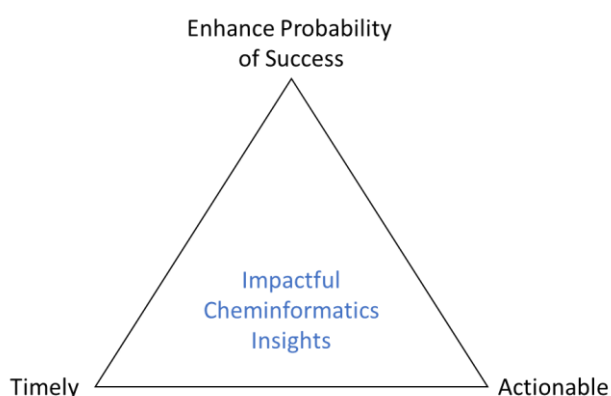
Too often, cheminformatics, while still valued, was regarded as useful only during certain stages of research, and contributions were sometimes limited to, say, modelling of a binding site at the beginning of a project, or a late-stage QSAR to confirm the importance of certain properties of synthesised analogues. Therefore, the lesson is to stay abreast of the key tools that scientists adopt, and to make cheminformatics analyses and results available through these as much as possible.



Visiting Merck's South San Francisco labs (2019).

Transforming yesterday's challenges into tomorrow's opportunities

This is not intended as a cautionary tale. Rather, this article aims to share the realities of working in a multidisciplinary scientific environment where decision making based on the best information available at the time it is needed. My journey has countless examples where cheminformatics significantly added to the body of knowledge in many research projects. Therefore, let's also remember that there were, and are, numerous success stories when applying the methods and expertise of cheminformatics, and many of these do not get published for proprietary reasons. Throughout my career, I witnessed or personally had several 'Eureka!' moments. These included elucidation of mechanisms of action, determining molecular factors that differentiated *in vitro* versus *in vivo* activity, uncannily accurate predictions of activity for analogue series, generating testable hypotheses that turned assumptions on their heads, building more precisely modelled molecular systems than hitherto, creating vast virtual libraries to explore analogue space, and developing new descriptors that provided fascinating insights into molecular properties. Beyond these special moments, simply being able to advance the understanding of research projects is the main reason why many cheminformaticians continue to do what they do.



Cheminformatics is most influential on research projects if it meets these three conditions.

Tempting though it is to dwell on anecdotes, there are more opportunities for cheminformatics in the future. The software and the buzzwords may change, but the fundamental need to better understand the molecular nature of chemical interactions does not. Along with many other scientific disciplines, our tools and methods are a means to define where the boundaries between molecular knowns and unknowns are. In many ways, this is the most critical offering of cheminformatics. Its biggest challenge in industry-based research, in my opinion, is its ability to produce actionable answers when they are needed and with a high enough probability of success.

Perhaps modern AI methods can help. Much attention, for example, has been given to protein folding, with many recent papers describing improved confidence levels for predicted structure. Potentially, this takes us a step closer to realising the *in silico* drug design dream, but we must address large gaps in knowledge that remain. Can AI plug the gap? In terms of protein folding, cheminformaticians have for decades studied the dynamics of proteins, their conformational changes, and allostery at binding sites. Much of this valuable information lies behind company firewalls, but slowly, these efforts will add to the body of knowledge available to AI. For a while, therefore, AI predictions of these tertiary biomolecular factors may still be shots in the dark.

Regarding the broader employment of AI within cheminformatics, I offer the same words of caution as when multivariate analyses first came on the scene. Specifically, cheminformaticians still need to know the right questions to ask, establish the right data to employ, and be aware of the limits of their analyses. The possibility of AI being wrong, or misleading, or the data incorrect or incomplete will never be far away, so vigilance and validation are always necessary. Nevertheless, new AI innovations will come and may lead us to fresh cheminformatics opportunities. The community must be ready to quickly grasp these and apply them to active research projects. And, as before, we must be ready to convince our fellow scientists of the inherent value that these innovations bring to research.

Of course, this article is just my attempt to distil important aspects of my own journey and relates mainly to the lessons learned while practising and preaching cheminformatics. Like so many of my peers, I also dabbled in other disciplines including medicinal chemistry, high-throughput screening, genomics, and research software

and hardware support, and I had the opportunity to contribute to discovery, early development, and even manufacturing activities. These days, I enjoy teaching and sharing what I learned. But it all started with those giddy days doing cheminformatics. I hope this helped provide perspective on what was for me an exciting and educating experience, and I wish you all good fortune on your own journeys.

The Drug Discovery Game

Contributed by Nele Quast, University of Oxford, email: quast@stats.ox.ac.uk

Designing and profiling drug candidates is a complex and expensive endeavour. Screening, identifying, evaluating, and optimising interesting precursors to enable informed decisions about a compound requires highly skilled and experienced medicinal chemists; many of them are trained on-the-job where they face the challenges that occur during drug discovery campaigns.

We've created the *Drug Discovery Game*, an educational game that interactively walks players through the drug development procedure. The aim of the game is to develop a lead drug candidate as an inhibitor of the target enzyme Matrix Metalloproteinase 12 (MMP12). As they advance through the game, players must consider the financial and time costs of running different assays to find the best possible drug candidate according to the provided target compound profile (TCP). We incorporate experimental data for compounds tested against MMP12 to make the game as realistic as possible. The progression of game stages of the *Drug Discovery Game* is given below.

Design and make

Drug Discovery Game Home Introduction Design Test Analysis Results

MW: 58.0 Da LogP: -0.56 HA: 4 H Don.: 1	TPSA: 43.09 A ² H Acc.: 1 Rings: 0	HA: 9 H Don.: 0	H Acc.: 0 Rings: 1
--	--	--------------------	-----------------------

A30
MW: 77.0 Da
LogP: 0.86
HA: 6
H Don.: 0

B44
MW: 109.0 Da
LogP: 1.31
HA: 8
H Don.: 0

A31
MW: 57.1 Da
LogP: 1.75
HA: 4
H Don.: 0

B45
MW: 123.0 Da
LogP: 0.81
HA: 9
H Don.: 0

?

Clear Test →

Your molecules:

A09B11

A15B11

A15B17

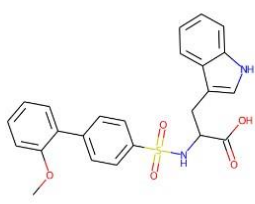
A22B17

test
Log out

30 weeks left
£100000

Test

Drug Discovery Game
Home Introduction Design Test Analysis Results



A28B13
Date Created:
 Week 0
Lipinski Filters:
 MW: Pass
 H Acc.: Pass
 H Don.: Pass
 LogP: Pass
Molecule Descriptors:
 HA: 32
 MW: 450.1 Da
 TPSA: 108.5 Å²
 H Acc.: 4
 H Don.: 3
 LogP: 3.8
 Rings: 4
Assay Data:
 Mouse Clearance: medium (5.6-30.5)
 Human Clearance: low (< 12)
 LogD: 0.66

	Mouse Clearance	Human Clearance	LogD	PAMPA	pIC ₅₀
Cost per assay	£7,000	£9,000	£1,000	£700	£70
Duration	3 weeks	3.5 weeks	1.5 weeks	1 week	1 week
Molecule					
A05B02	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
A05B13	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
A15B13	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
A28B13	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
A28B19	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
A10B19	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
A10B05	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>

Run Assays
 Duration: 3.5 weeks
 Cost: £33700

← Design
Analysis →

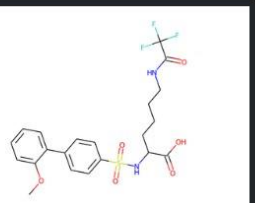
test
Log out
🕒: 23.5 weeks left
💰: £74510

Analysis

Drug Discovery Game
Home Introduction Design Test Analysis Results

Select Final Candidate →

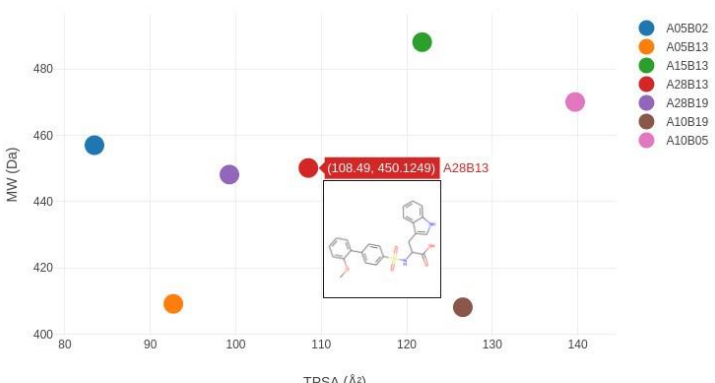
Rings: 2
Assay Data:
 Mouse Clearance: low (< 5.6)
 Human Clearance:
 LogD:
 PAMPA:
 pIC₅₀: 4.8



A15B13
Date Created:
 Week 0
Lipinski Filters:
 MW: Pass
 H Acc.: Pass
 H Don.: Pass
 LogP: Pass

?

Analysis Plot



← Test

LogD

pIC₅₀

TPSA

HA

MW

H Acc.

H Don.

Rings

LogP

logD

pIC₅₀

TPSA

HA

MW

H Acc.

H Don.

Rings

LogP

test
Log out
🕒: 23.5 weeks left
💰: £74510

Evaluation

Drug Discovery Game
Home Introduction Design Test Analysis Results

Molecule with desired profile

Assay Data:
 Mouse Clearance: low (< 5.6)
 Human Clearance: low (< 12)
 LogD: 1.08
 PAMPA: med2high
 pIC₅₀: 7.7

Your Molecule

Assay Data:
 Mouse Clearance: medium (5.6-30.5)
 Human Clearance: low (< 12)
 LogD: 1.44
 PAMPA: med2high
 pIC₅₀: 7.2

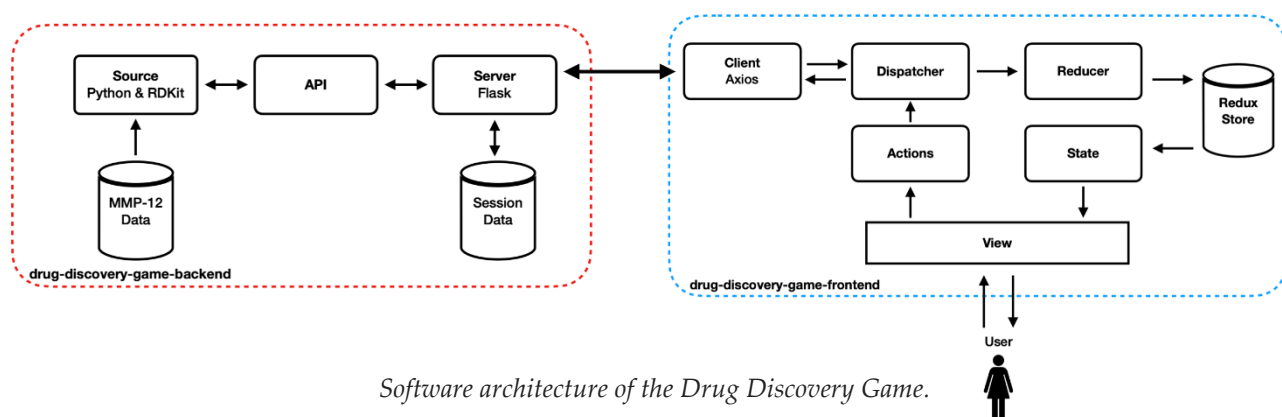
Assay	Chosen Molecule	Desired profile
Human Clearance	1.0	1.0
Mouse Clearance	4.0	1.0
LogD	1.5	1.0
PAMPA	5.5	5.5
pIC ₅₀	7.2	7.7

Your drug is too lipophilic. Adding polar functional groups to your drug can help to reduce this. Your drug's metabolic stability is good, so it is not cleared from the body too quickly, thus very frequent dosing is not necessary. Your drug's permeability is close to the target value, so will be well absorbed in the body. Your molecule's pIC₅₀ is acceptably close to the target value. Your molecule will bind strongly to the target, so a lower dose is needed, which can minimise side effects.

← Analysis
End Game

test 🕒: 23.5 weeks left
Log out 💰: £74510

The player progresses through stages of the game that emulate the classic design → make → test → analyse cycle of drug candidate screening. First, the player must design compounds they believe have potential as drug candidates. They are provided with a molecular scaffold to which they can attach substituents at two sites. Once satisfied with their design, the player can proceed to create it and add it to the catalogue of compounds to test. After a set of compounds has been designed, assays must be run to assess and evaluate the candidates. These assays cost time and money, so players must think critically about which tests to run and in what order!



The compounds can then be profiled in relation to various properties relating to its on-target activity and pharmacokinetics. Players can go back to create more compounds and run more assays, and, once they're satisfied their compound is as close as possible to the target compound profile (TCP), they can select it as their lead candidate for a final evaluation. The game then provides feedback on and information about their chosen drug candidate.

We developed the *Drug Discovery Game* as part of the doctoral training curriculum of the ESPRC funded SABS R³ PhD programme at the University of Oxford. SABS R³ trains doctoral students in systems approaches to biomedical research, with a unique focus on advanced practices in software engineering.

The *Drug Discovery Game* was created by three teams in separate cohorts, initially developing the Python foundation, and then adding the frontend and deploying the web-application. Roche's medicinal chemists and computational scientists provided feedback throughout the game's development, resulting in a game that caters to real-world chemists as users.

The game is available as a [web-application](#), which you are welcome to try out! If you would like to tinker with the game's software, have a look at our open-source repositories on GitHub. The project is implemented as a Python backend and a JavaScript frontend with ReactJS.

The Drug Discovery Game team members are:

- University of Oxford: Nele Quast, Guy Durant, Katherine M Shepherd, Anissa Alloula, Adelaide Punt, Arun Raja, Simiao Zhao, Olivia F Simpson, Stephanie Wills, James Bayne, Matthew Holland, Anna Carbery, Garrett M Morris.
- Roche: Maria Rosa Rodriguez Sarmiento, Torsten Schindler.

Workshop Report: Building the UK's Machine Learning in Materials Science Capability

Contribution from: Keith Butler, School of Engineering and Materials Science, Queen Mary College, London, email k.butler@qmul.ac.uk, Alin-Marin Elena, Science and Technologies Facilities Council, Reinhard Maurer, University of Warwick, Kim W. Jelfs, Imperial College, London

A group of academics and researchers held a one-day workshop on 7 March 2023, funded by the Henry Royce Institute, in Cosener's House in Abingdon, to scope community requirements for building a machine learning (ML) capability for the materials science community in the UK. The workshop convened practitioners of machine learning and materials science, from universities, companies such as IBM and Deep Mind, and public sector bodies such as UKAEA and AWE. With the recent announcement from the UKRI about AI Hubs, the event was an ideal forum for early adopters of ML for materials science to develop a coherent agenda for how they would like to see the field and capabilities develop in the UK.

The discussions took place under five broad categories – Education/Training, Infrastructure, Ethics, Open Science and FAIR data, Community building and events. The structure (small group discussions followed by round table feedback) of the workshop allowed identification of some clear requirements as well as drawing out and exploring tensions between the requirements of different sectors of the materials science community. Seven key requirements were identified:

- Establishing regular, sustainably funded training specifically for practitioners of ML in materials science
- A significant increase in the compute capacity for ML
- A dedicated, sustainably funded infrastructure for data storage and data and model sharing
- Establishing an ethics framework around ML research in materials science
- A requirement for environmental impact statements in ML research proposals

- Explicit consideration of contributions to Open Science in research assessment exercises for all publicly funded organisations
- Establishing community-led, national-level bodies to act as a focal point of the ML in the materials science community, in the form of learned body interest groups, AI hubs and/or an ML in Materials Science network

It was clear from discussions that there is an acute skills shortage across all the sectors involved in the workshop. There is a need for undergraduate (and earlier) curricula to be updated to cover more statistics/probability and linear algebra. There is a demand for curated repositories of trusted online educational resources. Companies often value specialism and there is a demand for focussed 'seasonal schools' for graduates. Best practice in writing and reading code and dealing with data is severely lacking in graduates at all levels and needs to be urgently addressed.

The discussion on infrastructure generally split into two concerns: data and compute resource. In both cases, the current infrastructure causes strong limitations to research and innovation. Interestingly, in the open discussion at the end, there was more discussion of compute resource, but more attendees identified data services as the important consideration. Commitments at the length of at least ten years are required for both compute and data-management services to bring the UK in line with other countries with similar AI ambitions. There may be an impression that ethics is not something that needs too much attention in ML for materials science; however the attendees spoke at length and placed great importance on a number of issues that come under the heading of ethics. There was discussion on culpability in AI, environmental impacts of research and the reproducibility problem.

The need for more Open Science and better adoption of FAIR data policies was echoed almost universally by all attendees. Although industrial actors cannot necessarily make their data open, they were keen to stress how valuable academically generated open data is to them. To take the analogy from structural biology, this is a field that has had a culture of openness for decades and that is what facilitated AlphaFold; the Protein Data Bank is estimated to facilitate research worth in excess of \$1 billion annually, with a return on investment estimated at \$8 billion over 30 years.¹ There is evidently value in driving a culture change in the relationship between materials scientists and data, it needs to be driven by educators, funders and employers.

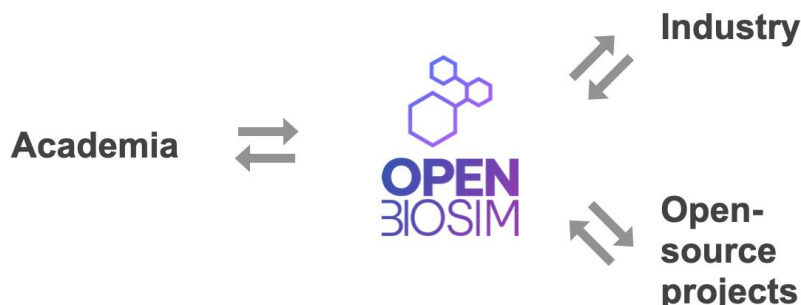
There was a clear desire from all attending to develop a recognisable UK ML for the materials community. There are currently many disparate groups, companies and institutes working in this area and an abundance of excellent research and innovation – providing coherent focus for these efforts will be critical for realising the full potential of the UK. It is hoped that the outcomes of the AI Hubs for Scientific Data call from the EPSRC will provide a starting point and that whichever consortia prevail will put a strong emphasis on engaging the full breath of this community. There is a clear route to community building through learned society interest groups and perhaps a role for an activity such as a Network grant in ML for molecular and materials science. It is also recognised that these hubs are just a small start and continued sustained funding for research, training and infrastructure will be essential to ensure world leading status for the UK ML for the materials community. A full summary of the discussions at the event is available [online](#).

Reference

(1) Sullivan, K.P. et al. Economic impacts of the research collaboratory for structural bioinformatics (RCSB) Protein Data Bank. 2017. https://cdn.rcsb.org/rcsb-pdb/general_information/about_pdb/Economic%20Impacts%20of%20the%20PDB.pdf

Accelerating Open-Source Innovation in Computer-Aided Drug Design with the OpenBioSim Community Interest Company

Contribution from Julien Michel, Professor of Biomolecular Simulation, University of Edinburgh, and Director, OpenBioSim Community Interest Company, email: julien@openbiosim.org



[OpenBioSim](https://openbiosim.org) is a non-profit scientific software company that was founded in 2022 to promote and sustain the use of open-source molecular simulation software in academia and industry.

Computer-aided design of chemicals for life sciences and materials applications is an established methodology that supports R&D processes in the pharmaceutical and chemical industries. The core computational chemistry algorithms that underpin processes widely adopted by the industry, often have their roots in academic research conducted decades before their industrial utility was demonstrated.

However, academic software rarely transitions directly into routine adoption by industry professionals. One of the main reasons for this is that academia primarily encourages proof of concept studies, while industry demands robust and thoroughly validated processes to integrate new computational technologies into complex R&D pipelines. Historically, this gap has been bridged by scientific software vendor firms specialising in business-to-business services. These vendors operate by licensing commercial-grade software solutions that reimplement algorithms with proven value into robust and user-friendly software. Software offerings available to support chemical R&D cover only a fraction of the methodologies prototyped in academia, as the commercial scientific software vendor sector is smaller than the pharmaceuticals and chemicals sectors. Therefore, vendors must carefully choose where to focus their internal product development efforts.

Open-source communities

This knowledge transfer pathway is increasingly disrupted by the growing popularity and availability of open-source software libraries. The GitHub platform, which is only 15 years old, now has over a hundred million registered user accounts. Academic researchers are increasingly releasing their code on open-source public repositories. Open-source package management systems such as conda, along with the growing prevalence of the Python programming language, have lowered barriers to combining software from different sources to build prototypes and test new research ideas.

Over the past decade, computation has adopted a more central role in drug discovery, as evidenced by the rapid rise in the number of AI-driven drug discovery companies operating at the preclinical and clinical stages. Computation-driven biotech companies compete based on differentiated platforms validated by the generation of preclinical assets. This creates a demand for the integration of customised computational technologies as components of proprietary drug discovery pipelines. Adopting open-source solutions accelerates the development roadmap of these organisations.

OpenBioSim's mission is to lower barriers for adoption of open-source research software

OpenBioSim is the result of decades of collaborative research between academia and industry by its founding members. The company operates at the interface of academia and industry, with a focus on lowering barriers for the adoption and distribution of academic research software. OpenBioSim establishes collaborative agreements that connect academics working on cutting-edge computational chemistry algorithms with computational chemistry developers in the industry, who are working on commercial computational chemistry software or computational drug discovery platforms. As a non-profit organisation, OpenBioSim's goal is to deliver benefits to multiple stakeholders, forming a community of scientists working with open-source research software.

OpenBioSim provides maintenance and support for open-source software that has demonstrated its utility beyond academia. This enables industrialists to rely on open-source components for building software pipelines for use in production environments. Additionally, OpenBioSim promotes the sharing of knowledge and benchmarks between academia and industry, thereby accelerating the scientific validation of new computational methodologies.

OpenBioSim's projects demonstrate the potential of collaborative research software development by industry and academia

OpenBioSim currently supports two major open-source software projects. [Sire](#) is a modular molecular modelling framework that was originally started by Christopher Woods in 2005 as a self-funded project during his transition between jobs. Sire's functionality has evolved over time and is now primarily used as a library to manipulate representations of biomolecular systems. The main features of Sire include a rich library of readers and writers for common molecular topologies (such as Amber prm7, CHARMM psf, GROMACS top, Mol2, SDF, PDB) and trajectory file formats (such as DCD, XTC, TRR); a search engine for selecting atoms and editing molecular systems; support for molecular mechanics energy calculations; and converters to/from other open-source toolkits such as RDKit or OpenMM.

Over the past decade, Sire's building blocks have been utilised to develop various academic research software prototypes. Examples include the Waterswap method (based on Gibbs Ensemble Absolute Binding Free Energy Calculation theory) developed at the University of Bristol,¹ the Nautilus software for water thermodynamics analyses (based on Grid Cell Theory) developed at the University of Edinburgh,² and the SOMD package (a single-topology alchemical free energy calculation engine) also developed at the University of Edinburgh.³ These software prototypes are typically operated from a command-line interface by experienced computational chemists.

SOMD was integrated in 2019 as a component of the Flare™ FEP software marketed by the software vendor [Cresset](#), allowing non-experts to run FEP calculations to routinely support drug discovery projects.⁴ The significant product development efforts by the Cresset staff led to numerous bug fixes and protocol improvements in SOMD, which, in turn, benefited the research activities of scientists working directly with the open-source code.

OpenBioSim also provides support and maintenance for the [BioSimSpace library](#).⁵ BioSimSpace was initiated in 2017 as a flagship software project by the academic consortium [CCPBioSim](#), with funding from EPSRC. The goal of BioSimSpace was to facilitate the development of pipelines that combine simulation software and protocols independently created by different research groups.

Essentially, BioSimSpace aims to programmatically capture all the knowledge required to successfully deploy computational chemistry software. It achieves this by providing access to libraries of interoperable workflow

components and protocols that abstract common biomolecular simulation tasks. These workflow components are assembled by creating wrappers around third-party open-source tools. For instance, users can request the execution of a molecular dynamics simulation on a given molecular input without having to worry about detailed simulation configuration settings or conversion between molecular topologies. BioSimSpace is distributed in a single Python environment that is compatible with all third-party dependencies, alleviating users from the burden of installing and maintaining a fragmented ecosystem of academic research software packages.

A popular application of BioSimSpace is its use as a software development environment for free energy calculations. BioSimSpace supports automated setup and analysis of alchemical free energy calculations for multiple simulation engines (currently supporting AMBER, SOMD, and GROMACS). The availability of modular and interoperable free energy calculation pipelines facilitates the work of academic developers, who often focus their efforts on optimising specific steps of a pipeline. Recent collaborative work on absolute binding free energy calculation methodologies between the Universities of Edinburgh and Newcastle and AstraZeneca serves as an illustration of this.⁶

Industry developers are free to write new BSS code or reuse existing BSS code alongside proprietary code to create pipelines that meet the R&D needs of their organisations. Software engineers and domain experts employed by OpenBioSim provide technical and scientific support, software maintenance efforts, and documentation to assist in these endeavours. OpenBioSim prioritises addressing the needs of third parties that commit to financially supporting the organisation.

BioSimSpace is being used by an increasing number of biotech and pharmaceutical companies across a diverse set of use cases. A notable example was recently presented by the AI-driven precision medicine company Exscientia at the MGMS [MD in Pharma meeting](#) held in London in March 2023, where they described their approach to creating a platform for binding free energy calculations with the assistance of BioSimSpace. Extensive benchmarking efforts conducted by Exscientia's staff led to the optimisation of free energy calculation protocols distributed in BioSimSpace, benefiting anyone working with the toolkit.

Open-source research software has already transformed the way computational chemists in academia and industry share ideas. Looking ahead, OpenBioSim is determined to continue working on making the best emerging computational chemistry research methodologies accessible to innovators dedicated to accelerating the development of medical treatments that will bring broad societal benefits.

References

- (1) Woods, C.J. et al. A water-swap reaction coordinate for the calculation of absolute protein-ligand binding free energies. *J. Chem. Phys.* **2011**, *134*(5), 054114. <https://doi.org/10.1063/1.3519057>
- (2) Gerogiokas, G. et al. Prediction of small molecule hydration thermodynamics with grid cell theory. *J. Chem. Theory Comput.* **2014**, *10*(1), 35-48. <https://doi.org/10.1021/ct400783h>
- (3) Calabrò, G. et al. Elucidation of nonadditive effects in protein-ligand binding energies: thrombin as a case study. *J. Phys. Chem. B.* **2016**, *120*(24), 5340-5350. <https://doi.org/10.1021/acs.jpcc.6b03296>
- (4) Kuhn, M. et al. Assessment of binding affinity via alchemical free-energy calculations. *J. Chem. Inf. Model.* **2020**, *60*(6), 3120-3130. <https://doi.org/10.1021/acs.jcim.0c00165>
- (5) Hedges, L.O. et al. BioSimSpace: an interoperable Python framework for biomolecular simulation. *J. Open Source Software.* **2019**, *4*(43), 1831 <https://doi.org/10.21105/joss.01831>
- (6) Clark, F. et al. Comparison of receptor-ligand restraint schemes for alchemical absolute binding free energy calculations. *J. Chem. Theory Comput.* **2023**, in press <https://doi.org/10.1021/acs.jctc.3c00139>

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



2023 is proving to be a busy and varied year in the life of the Catalyst Science Discovery Centre and Museum in Widnes.

On 25 April 2023 we welcomed a party of the 2012 Past Masters of the Worshipful Livery Companies who are known as 'The Diamonds'. Their visit had been organised by Past Master Baker, Mr Peter Gossage, who is the great grandson of William Gossage, the great inventor (58 patents) and founder of Gossages Soap Works in Widnes, and Mr Keith Etherington, Past Master of the Worshipful Company of Scientific Instrument Makers. Members of this group meet up each year and 2023 was a three day visit to Liverpool with a stop off at Catalyst en route. Most knew nothing about our education or heritage work and were amazed at what they saw and learnt. We had three groups of children in for STEM workshops and shows – a school from Liverpool and a school from Manchester during the day and in the late afternoon/overnight a school group from Liverpool for a sleepover under the stars in our fourth floor Observatory with STEM activities.



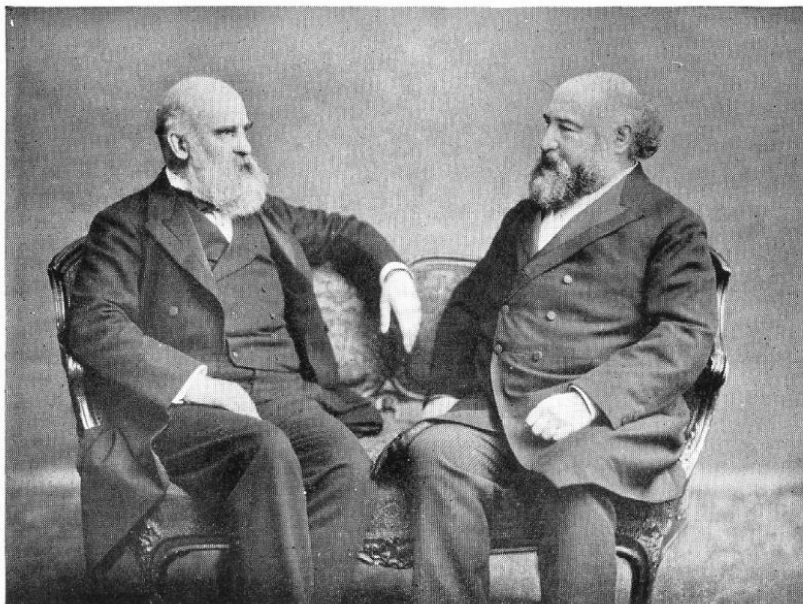
Presentation of the Widnes Traders' Association Gossage Cup.

Peter Gossage was accompanied by his two brothers, Martin and John, and the opportunity was taken to present him, as one of Catalyst's Patrons, with the Widnes Traders Association Gossage Cup, created in 1925, which used to be given each year to the chemical industrial company which had made the greatest improvement in trading that year. For many years it had been stored in the vaults of Halton Chamber of Commerce (HCC) but they had decided it was far better to be in the Catalyst Museum where it could be seen by the public so it was presented to Peter by the CEO of

HCC, Rachael Owen. It is now on show in our Gossage Room along with many memorabilia of this company. Our new Café Manager, Denise, produced a wonderful buffet lunch and the group moved on to Liverpool full of memories of Catalyst, its work and with new contacts made. Since then I have been pursuing contact with the Worshipful Company of Salters and hope that their incoming Master for 2023/2024, who will be elected on 12 June can visit during summer 2023.

One piece of Gossage memorabilia that we have never had is a fob watch which was given to all employees in 1925/1926. The Gossage family did not have one either but after a persistent hunt on various auction websites one was found in May 2023 and purchased by Peter Gossage. It will be coming to the Gossage Room to add to our archives on Peter's next visit.

On 15 July we will be attending the Daresbury Innovation Centre's all-day celebration of science where c.7,500-10,000 public visitors are expected to see the work being done by the Centre's staff and roll-out companies on the site but also to meet organisations like Catalyst, RSC local sections and RSC special interest groups, other professional organisations like IOP and IET and to explore the work they do in promoting science and technology. Our exhibit will be our 'Explore the Planet' environmental project with interactives depicting recycling and protection of the planet.



Sir John Brunner, Bart, and Dr Ludwig Mond, F.R.S., 1898. Presentation photograph taken to mark the 25th Anniversary of the foundation of Winnington Works

2023 is the celebration of the 150th anniversary of the foundation of the Brunner Mond Company in 1873 when John Brunner and Ludwig Mond left Widnes, where they had worked in our Catalyst building for Hutchinsons/Gossages producing alkali by the mega-polluting Leblanc Process, to Winnington near Northwich in Cheshire to use Solvay's much less polluting process to produce alkali. By the end of the 19th century Brunner Mond was the wealthiest chemical company in the UK. It was there at Winnington in 1933, after Brunner Mond had become a part of ICI Ltd in 1926, that polythene was invented. In October 2023 Northwich Council and Museums and Libraries Service are

organising a weekend of celebration of the founding of Brunner Mond, which transformed the town forever. I have been asked to give the Keynote Talk to launch the weekend on 20 October. It will be entitled *John and Henry Brunner: men of Cheshire and Lancashire – their legacy and philanthropy* and take place at the Lion Salt Works at Marston near Northwich. Two of John Brunner's descendants, Sir Hugo Brunner, former Lord Lieutenant of Oxfordshire, and George Windsor, Earl of St Andrews, are Patron and Royal Patron of Catalyst respectively. Sadly the Winnington and Wallerscote Works of Brunner Mond (now owned and run by Tata Chemicals) have either gone or are being run down and houses being built on the site. Chemical production continues on a limited scale but at Lostock Gralam which is near the brine fields which still provide the raw material for Castner Kellner Works in Runcorn to produce chlorine by electrolysis, as has been done since 1897. Castner Keller Works is part of INEOS Inovyn who are great supporters of Catalyst. They have rebranded recently with a new logo.

Judith Wilde, our archivist, has recently created a small temporary exhibition of the three portraits that we hold from a much larger group of 56 portraits, painted during World War II by well known British artists, of workers in 26 different locations of ICI Ltd in England, Scotland and Wales. They show the varied work of men and women during that period ranging from research chemists to miners, safety officers, nurses and technicians. The paintings were originally used as illustrations in newspapers between September 1944 and March 1946 to show the public a record of these people "who also served" by working in the British chemical industry during the war. Later a representative group of the portraits were brought together in a book entitled *Portraits of an Industry*. The location of 11 of these portraits is known and a hunt has been on for some time to find the others: will they turn up on the Antiques Road Show or be found in a skip if modern day descendants of the sitters don't know what they are?

Our three are of William Tyler, Safety Officer at Castner Kellner Works in Runcorn, Dr Neil G Marr, an Aberdonian and Chief Medical Officer of the Metals Division of ICI in the Midlands, and Mrs Winifred Burgess who worked at Winnington Works of ICI where she was in charge of a control panel which monitored the critical point of the polymerisation of polythene. You can see here the three paintings and also a display of all the others in the *Portraits of an Industry* book.



“Portraits of an Industry” display.

In the last CICAG Newsletter I mentioned that we were working on the first phase of the ‘Synergy Project’ for which National Lottery Heritage Fund have awarded us £95,000 to prepare a major bid for nearly £1 million to revamp and represent the chemical sciences and chemical industry in north west England in our heritage parts of the building in a way which is more accessible and understandable to the general public. A huge amount of work is being undertaken to achieve success in this bid.

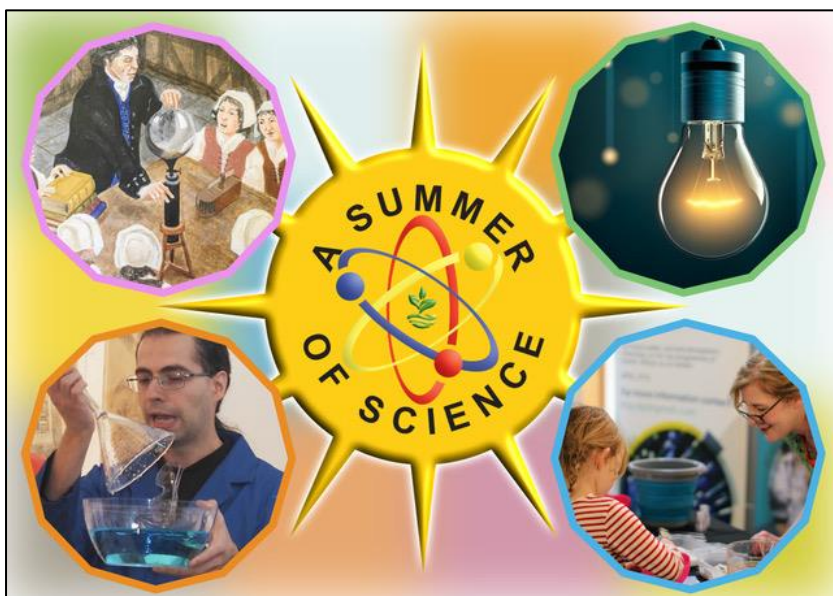
In the meantime our STEM education work continues apace and the targets that had been set by the Catalyst CEO, Dr Lee Juby, for the number of school visits, sleepovers and general public admissions have been met in the first six months of 2023 and we hope will continue. During the summer we are looking forward to welcoming a group being brought by the initiator of the ‘Rochdale Science Initiative’ who has been encouraging his friends, relatives and neighbours to embrace the chemical sciences in a way they never have before.

If you are in the area come and visit us at Catalyst in Widnes, look at our [website](#), and publicise our existence and work which we have been undertaking since 1987.

Communicating Chemistry at a Local History Museum

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

helen.cooke100@gmail.com



From 20 July – 30 September 2023 the local history museum in Nantwich, Cheshire, will be holding a *Summer of Science* festival, featuring an exciting programme of events for visitors of all ages. These will include family workshops, talks, guided walking tours featuring science around the town, live chemistry and physics demonstrations, plus a range of trails and exhibitions. A number of the events will focus on biodiversity and sustainability.

Key to the success of the festival will be discussing science in a context to

which visitors, young and old, can relate. To achieve this, items in the Museum's collection, which reflect historic local industries and agriculture, are being re-interpreted to reveal the science embedded within them. Examples include artefacts from the local salt industry, some of which originate from Roman times when brine springs were discovered in Nantwich. The festival will also feature local people associated with science, for example Joseph Priestley, who lived in Nantwich from 1758-61 and taught science to boys and girls at his small school there.

The festival will run in parallel with the Museum's *Nantwich Illuminated* exhibition on the history of the town's gas works site, which provides considerable scope for chemistry demonstrations and activities. Some objects on display will be loaned by the National Gas Museum in Leicester.

Family events will be free of charge, to encourage participation by people from deprived areas. It is planned that a legacy from the festival will be enhancement of the Museum's education programme for school visits, building on science-focussed workshops developed for the festival.

Nantwich Museum is extremely grateful for support from the RSC's Outreach fund and from the North Staffordshire & South Cheshire Local Section.

Driving the chemistry aspects of this initiative are Museum volunteers Dr Helen Cooke FRSC, Dr Glynn Skerratt FRSC, Dr Graham Dodd, Dr David Nellist FICHEM, and Professor Fabio Parmeggiani (Politecnico di Milano). For more information please email Helen Cooke, visit the Museum's [website](#) or phone Nantwich Museum on 01270 627104.

AI in Drug Discovery 2022 – A Highly Opinionated Literature Review

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

Reproduced from Pat Walters's [Practical Cheminformatics blog post](#).

Here's a roundup of some of the papers I found interesting in 2022. This list is heavily slanted to my interests, which lean toward the application of machine learning (ML) in drug design. I've added commentary to most of the papers to explain why I found them compelling. I've done my best to arrange the papers according to themes. If I omitted a paper, please let me know. I'd be happy to update this summary. This review ended up being longer than I had anticipated, and there are several topics I didn't cover. If I have some time, this post may get a sequel.

Contents

1. Are Deep Neural Networks Better for QSAR?
2. Deep Learning Methods Provide New Approaches to Protein-Ligand Docking
3. Protein Structure Prediction - Pushing AlphaFold2 in New Directions
4. Model Interpretability
5. QM Methods
6. Ultra-large Chemical Libraries
7. Active Learning
8. Molecular Representation

1. Are deep neural networks better for QSAR?

Based on papers I read and reviewed in 2022, there seems to be a perception that Deep Neural Networks have become ubiquitous in QSAR modelling. In fact, the introduction to a recent special issue of JCIM contained [this](#) statement.

“However, after the [Kaggle Merck Molecular Activity Challenge 2013](#) and the [Tox21 Data Challenge 2015](#), DNNs have emerged as the method of choice for QSAR applications in drug discovery.”

While I've found DNNs useful in some circumstances, I don't believe they've become “the method of choice”. Many of the literature examples I've seen where deep neural networks (DNNs) outperform more traditional methods like random forest (RF) involve very large datasets containing tens of thousands of molecules. Here are some historical examples for context.

Analyzing Learned Molecular Representations for Property Prediction

<https://pubs.acs.org/doi/10.1021/acs.jcim.9b00237>

Improvement in ADMET Prediction with Multitask Deep Featurization

<https://pubs.acs.org/doi/full/10.1021/acs.jmedchem.9b02187>

A few papers published in 2022 provided interesting comparisons between DNNs and other methods. At the end of the day, it's difficult to call a clear winner. A paper from AstraZeneca evaluated the ability of several machine learning methods, partial least squares (PLS), random forest regression (RF), support vector regression (SVR), and gradient boosted trees (XGBoost), to predict non-additive SAR. The authors used a method previously published by [Kramer](#) to identify matched molecular pairs with additive and non-additive SAR. The pairs were then used to construct subsets of varying predictive difficulty. The analysis considered datasets for four assays, LogD, solubility in DMSO, clearance in liver microsomes, and permeability (cell line not specified). In most cases, the DNN outperformed the other methods on both the additive and non-additive subsets.

Simple nearest-neighbour analysis meets the accuracy of compound potency predictions using complex machine learning models

<https://www.nature.com/articles/s42256-022-00581-6>

One of my favorite papers of 2022 was published by a group from Eindhoven University of Technology. The intent of this paper was somewhat similar to that of the paper above. The authors constructed datasets containing activity cliffs and evaluated the ability of ML models to predict the activity of these realistic but challenging sets. The analysis included many traditional ML methods, including RF, gradient boosting machines (GBM), SVR, and kNN. To benchmark both algorithms and representations, the authors coupled several fingerprints and descriptor sets with the algorithms mentioned earlier. In addition, the authors included a wide array of DNN methods, including message-passing neural networks (MPNN), graph convolutional neural networks (GCN), graph attention transformers (GAT), and attentive fingerprints (AFP). The methods were assessed based on their ability to predict bioactivity using 30 datasets from ChEMBL. The authors reported the root mean squared error (RMSE) and RMSE_{cliff}, calculated over activity cliff molecules. In a result that some may find surprising, the authors reported that SVM, GBM, and RF coupled with ECFPs provided the best performance on both the overall and the activity cliff datasets.

Exposing the Limitations of Molecular Machine Learning with Activity Cliffs

<https://pubs.acs.org/doi/10.1021/acs.jcim.2c01073>

Perspective - These papers were important for two reasons. First, they helped to dispel the perception that DNNs are always the best choice for QSAR models. Second, they introduced new, more realistic benchmark datasets and strategies that will hopefully supplant some of the flawed benchmarks in everyday use.

2. Deep learning methods provide new approaches to protein-ligand docking

In 2022 we saw the advent of a new approach to protein-ligand docking. Until recently, most docking programs used an empirical or physics-based scoring function to search binding poses within a predefined protein region, typically defined by a box around the binding site. 2022 saw the emergence of a new generation of docking programs that use the structures of existing protein-ligand complexes to learn relationships between ligands and protein binding sites and search the entire protein surface. In essence, these programs are simultaneously solving two problems; identifying a binding site and determining the docked pose of a ligand.

One of the first examples of this approach was the EquiBind docking program from MIT, which uses deep learning to align sets of points on a protein with corresponding points on a ligand.

EquiBind: Geometric Deep Learning for Drug Binding Structure Prediction

<https://arxiv.org/abs/2202.05146>

A group from Galixir Technologies extended this approach with a docking program called TANKBind that evaluates docking poses across multiple sites on the protein and chooses the highest-scoring pose.

TANKBind: Trigonometry-Aware Neural Networks for Drug-Protein Binding Structure Prediction

<https://www.biorxiv.org/content/10.1101/2022.06.06.495043v3>

Researchers from the Mila - Quebec AI Institute and the University of Montreal published a preprint describing E3Bind, a docking approach inspired by AlphaFold2. This approach utilizes a combination of three embeddings describing a protein graph, a ligand graph, and a protein-ligand graph which is iteratively refined to generate docking poses.

E3Bind: An End-to-End Equivariant Network for Protein-Ligand Docking
<https://arxiv.org/abs/2210.06069>

The team that created Equibind has developed a new generative approach to solve the docking problem. Their method, DiffDock, uses diffusion to iteratively search a space of translations, rotations, and torsional variations. The search process in DiffDock is guided by a novel confidence score that enables a choice between multiple poses.

DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking
<https://arxiv.org/abs/2210.01776>

Graph Neural Networks (GNNs) have gained widespread use in various subfields of drug discovery. One example is the MedusaGraph method, which uses GNNs to predict protein-ligand docking. This method involves two GNNs: the Pose Prediction GNN, which suggests possible binding poses, and the Pose Selection GNN, which performs binary classification to evaluate the quality of a given pose.

Predicting Protein–Ligand Docking Structure with Graph Neural Network
<https://pubs.acs.org/doi/10.1021/acs.jcim.2c00127>

Perspective - For almost 30 years, docking has been used in structure-based drug design. Until recently, there have been few changes to how docking algorithms work. These algorithms usually generate a set of poses, which are then evaluated using scoring functions that combine physics-based and empirical terms. Recently, advances in deep neural networks (DNNs) have led to the development of new docking programs that use data from the Protein Data Bank (PDB) to train functions that can identify binding sites and propose and evaluate binding poses. While these methods are powerful, they can be difficult to benchmark because it is hard to know if the method is discovering new interactions or simply transferring information from similar binding sites and ligands. It will be interesting to see how these methods perform when faced with novel ligands and binding sites that have not been seen before. There is also the possibility of combining these approaches with existing physics-based and empirical methods.

3. Protein structure prediction - ushering AlphaFold2 in new directions

Without a doubt, the highest-profile application of ML in 2021 was AlphaFold2. The Deepmind group stunned the protein structure prediction (PSP) world by dominating the CASP14 protein structure prediction challenge. Following the publication of the original AlphaFold2 paper, the field has exploded. Every week, we see another paper describing [insert name of] Fold. I'm sure someone (not me) will write an entire review on the advances in PSP during 2022. Instead, I'll focus on a few specific applications of PSP to molecular modelling.

Several papers have shown that one can generate multiple protein conformations by modifying the multiple sequence alignment (MSA) that AlphaFold2 uses as an initial step in structure generation. In a preprint from late 2022, Wayment-Steele described an approach that clustered the MSA by sequence similarity and used the clusters to generate multiple protein conformational states. An implementation of this method, known as AF-Cluster, is available on GitHub.

Prediction of multiple conformational states by combining sequence clustering with AlphaFold2
<https://www.biorxiv.org/content/10.1101/2022.10.17.512570v1>

While several groups have demonstrated the biological relevance of the alternate protein conformations generated by AlphaFold2, more work must be done to assess the energetics and relative populations of these conformational ensembles. A recent paper from the Tiwary group addresses this need by using ML-augmented molecular dynamics to generate a Boltzmann-weighted ensemble of protein conformations.

AlphaFold2-RAVE: From sequence to Boltzmann ensemble
<https://www.biorxiv.org/content/10.1101/2022.05.25.493365v2>

When presented with the AlphaFold2 results from CASP14, one of the first questions computational chemists asked was whether structures generated by PSP could be used for molecular modelling. While we don't have definitive answers, a few papers provide some necessary first steps.

Free Energy Perturbation (FEP) calculations have become a mainstay of lead optimization efforts. The prevailing wisdom is that a high-resolution cocrystal structure is necessary to achieve a good correlation between predicted and experimental binding affinity. A paper by Beuming and co-workers challenges this assumption by examining the utility of substituting an AlphaFold2 structure for an x-ray structure in FEP calculations.

Are Deep Learning Structural Models Sufficiently Accurate for Free-Energy Calculations? Application of FEP+ to AlphaFold2-Predicted Structures
<https://doi.org/10.1021/acs.jcim.2c00796>

Based on the paper above, it appears that structures from AlphaFold2 can have some utility in FEP calculations, where molecular dynamics simulations allow for some sidechain rearrangement. However, when comparing AlphaFold2 structures to protein crystal structures for docking, a group at Scripps found that AlphaFold2 structures do not have the necessary resolution in the side chains for accurate docking calculations. Perhaps not surprisingly, the success rate when docking into AlphaFold2 structures (17%) was less than the success rate docking into holo structures (41%). While this result was less than spectacular, it was considerably better than the docking success rate (10%) the authors achieved with apo x-ray structures.

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

In many therapeutic areas, including oncology and genetic disease, the ability to understand the structural impact of missense mutations can facilitate the design of therapeutics. Over the last year or two, there have been differing views on the ability of AlphaFold2 to reliably model missense mutations. A group from the NCI compared AlphaFold2 structures with x-ray structures for three systems where x-ray structures for the wild-type protein were available and specific structure-disrupting mutations existed. AlphaFold2 predicted similar structures for mutant and WT in all three cases and failed to recognise the structure-disrupting mutation.

Can AlphaFold2 predict the impact of missense mutations on structure?
<https://www.nature.com/articles/s41594-021-00714-2>

This finding is consistent with a 2021 paper from the Skolkovo Institute of Science and Technology, which found that AlphaFold2 structures did not reproduce experimentally observed changes in protein stability or fluorescence associated with single mutations.

Using AlphaFold to predict the impact of single mutations on protein stability and function
<https://www.biorxiv.org/content/10.1101/2021.09.19.460937v1>

However, all may not be lost. A recent preprint from the Baker group provides promising evidence that RosettaFold can predict the structural impact of protein mutations.

Accurate Mutation Effect Prediction using RoseTTAFold

<https://www.biorxiv.org/content/10.1101/2022.11.04.515218v1?rss=1>

Perspective - 2022 was the year of AlphaFold in PSP. While the team from DeepMind didn't compete in CASP15, all of the best-performing entries were variants on AlphaFold2. It's been great to see how AlphaFold2 has been extended and applied to a wide variety of problems in drug discovery. I look forward to seeing what develops in 2023.

4. Model interpretability

While ML models can efficiently select and prioritize molecules for synthesis or purchase, most models operate as "black boxes" that take chemical structures as input and generate predictions as output. Ideally, we'd like to have interpretable models that provide insights to motivate the design of subsequent compounds. Interpretable models could provide several advantages.

- Engender confidence among the team. Experimentalists will be much more likely to "buy in" if they understand the reasoning behind a prediction.
- Enable the debugging of models. If we understand the reasoning behind a prediction, we may be able to make adjustments to improve the model.
- Facilitate an understanding of the underlying science. Ultimately we would like to understand the links between chemical structure and a physical or biological endpoint. An interpretable model could potentially help to illuminate the underlying physical processes.

A recent paper from Sanofi-Aventis and Matthias Rarey provides an extensive overview of the application of explainable artificial intelligence (XAI) to lead optimization datasets. Several XAI methods are demonstrated, along with a heatmap visualization highlighting features that are critical for activity.

Interpretation of Structure–Activity Relationships in Real-World Drug Design Data Sets Using Explainable Artificial Intelligence

<https://pubs.acs.org/doi/abs/10.1021/acs.jcim.1c01263>

Some of my favourite work this year came from Andrew White's group at Rochester. In one paper, they used counterfactuals, a technique that has been used to explain machine learning models in several areas, including credit risk assessment.

Model agnostic generation of counterfactual explanations for molecules

<https://pubs.rsc.org/en/content/articlelanding/2022/sc/d1sc05259d>

In a second paper, the White group used a language model to create text-based explanations for machine learning model predictions. While the approach is interesting, the examples in the paper are not quite at a level that would convince a medicinal chemist.

Explaining molecular properties with natural language

<https://chemrxiv.org/engage/chemrxiv/article-details/633731d1f764e6e535093041>

Finally, the White group published a perspective preprint, released at the end of 2022, which provides a comprehensive overview of interpretable models in QSAR and other fields.

A Perspective on Explanations of Molecular Prediction Models

<https://chemrxiv.org/engage/chemrxiv/article-details/639222a114d92d7cd6a65e90>

Shapley values have been used to assess the contributions of features in machine learning models. Two recent papers from the Bajorath group demonstrate how this technique can be applied to machine learning for molecules.

EdgeSHAPer: Bond-centric Shapley value-based explanation method for graph neural networks

<https://www.sciencedirect.com/science/article/pii/S2589004222013153>

Calculation of exact Shapley values for support vector machines with Tanimoto kernel enables model interpretation

<https://www.sciencedirect.com/science/article/pii/S2589004222012950>

To effectively associate chemical structure with ML model predictions, we need software tools that will enable us to visualize the mapping of model predictions onto chemical structures. A paper from Bayer describes an open-source tool for interpreting ML models and visualizing atomic contributions.

ChemInformatics Model Explorer (CIME): exploratory analysis of chemical model explanations

<https://jcheminf.biomedcentral.com/articles/10.1186/s13321-022-00600-z>

Perspective - While model interpretability has become a component of several research efforts, we have yet to arrive at readily actionable models. Hopefully, as the field progresses, we'll reach a point where insights from interpretable models will provide clear directions for optimization.

5. QM methods

Over the past few years, several groups have developed machine learning methods to rapidly reproduce quantum chemical potentials. While these methods were scientifically interesting, their practical application was somewhat limited. A new package, Auto3D, from the Isayev group at Carnegie Mellon, could change this situation by making learned quantum chemical potentials very easy to use. Auto3D accepts SMILES as input, generates and evaluates ensembles of 3D conformations, and provides an energetic ranking for tautomers and stereoisomers.

Auto3D: Automatic Generation of the Low-Energy 3D Structures with ANI Neural Network Potentials.

<https://doi.org/10.1021/acs.jcim.2c00817>.

Perspective - While QM methods are a bit outside my wheelhouse, I was excited to see an implementation that made it easy to perform critical calculations.

6. Utralarge chemical libraries

One of the biggest game changers in virtual screening has been the availability of synthesis on-demand libraries like [Enamine REAL](#), [WuXi GalaXi](#), [Otava CHEMryia](#), and [eMolecules eXplore](#). These libraries, consisting of billions of molecules, which are available for rapid (a few weeks) deliveries at a reasonable cost, have caused many of us to rethink our approaches to virtual screening. Brute force approaches, applied to datasets containing millions of molecules, are no longer relevant when considering libraries containing tens of billions.

I'd recommend watching the videos from the NIH Symposium on Ultra-large Chemical Libraries for those looking for a good overview of the field. I realize this symposium occurred in December 2020, but many people may have missed it. If you only have time to watch one talk from this symposium, check out the one from Roger Sayle.

NIH Symposium on Ultra-large Chemical Libraries

https://cactus.nci.nih.gov/presentations/NIHBigDB_2020-12/NIHBigDB.html

Wendy Warr published a detailed set of notes covering the NIH symposium.

<https://chemrxiv.org/engage/chemrxiv/article-details/60c75883bdbb89984ea3ada5>

In addition, this collaborative paper from several presenters covers much of the work presented at the NIH symposium and provides an excellent overview of the field.

Exploration of Ultralarge Compound Collections for Drug Discovery

<https://pubs.acs.org/doi/10.1021/acs.jcim.2c00224>

As the sizes of chemical libraries get into the tens of billions, simple tasks like determining which molecules are common to two libraries become cumbersome and time-consuming. Rather than performing pairwise comparisons of billions of product molecules, one can compare the chemical building blocks used to construct the libraries. However, since multiple chemical routes can lead to the same set of products, it's crucial to employ a method that uses an appropriate fragmentation strategy. A paper from Matthis Rarey's group used such a fragmentation strategy to develop a software tool called SpaceCompare which they subsequently used to compare the REAL, GalaXi, and CHEMyria databases. One surprising conclusion from this work is the very low overlap between molecules in the three databases. The largest overlap between any two databases was less than 2% of the total.

Comparison of Combinatorial Fragment Spaces and Its Application to Ultralarge Make-on-Demand Compound Catalogs

<https://pubs.acs.org/doi/10.1021/acs.jcim.1c01378>

The seemingly simple calculation of physical properties can be impractical when dealing with billions of molecules. Another recent publication from the Rarey group describes a method called SpaceProp that derives property distributions of large enumerated sets from the properties of the constituent topological fragments.

Calculating and Optimizing Physicochemical Property Distributions of Large Combinatorial Fragment Spaces

<https://pubs.acs.org/doi/10.1021/acs.jcim.2c00334>

Perspective - Over the last five years, we've seen the number of commercially available molecules grow from 1 billion to more than 60 billion. To keep up, the field needs to develop new methods that allow structure-based or ligand-based searches of these ultra-large libraries. The work published in 2022 was a good start, but there's much more to do.

7. Active learning

Active learning is an iterative technique that enables researchers to search through large spaces efficiently. This approach uses a machine learning model to select and label datapoints and explore a particular chemical space. In earlier work by [Yang](#), [Berenger](#), and others, active learning was used to direct docking calculations with large chemical libraries. The active learning process begins with an initial sample that can be selected at random or

through some other means, such as clustering. The initially sampled subset of molecules is then docked, and the chemical structures and docking scores for the molecules are used to train a machine learning model. The machine learning model is used to generate predictions for the more extensive database, and the model predictions are used to select the next set of molecules to dock. After a few iterations, the active learning process identifies the molecules to be carried to the next step.

While machine learning is orders of magnitude faster than methods like docking, inference on a table with billions of rows is non-trivial. This paper by the Coley group at MIT provides a method of pruning a large dataset and efficiently reducing the number of molecules to be predicted by a machine learning algorithm.

Self-Focusing Virtual Screening with Active Design Space Pruning

<https://pubs.acs.org/doi/pdf/10.1021/acs.jcim.2c00554>

Free energy perturbation (FEP) calculations have become a mainstay of computationally driven structure-based drug discovery programs. While these calculations are powerful, they are also computationally expensive, with a single calculation taking several hours to complete. One way of overcoming these computational limitations is to use active learning coupled with FEP to search through large chemical libraries. Following on the heels of a 2019 paper by [Konze](#), several groups have explored the application of active learning to free energy calculations.

Chemical Space Exploration with Active Learning and Alchemical Free Energies

<https://pubs.acs.org/doi/10.1021/acs.jctc.2c00752>

Active Learning Guided Drug Design Lead Optimization Based on Relative Binding Free Energy Modeling

<https://doi.org/10.26434/chemrxiv-2022-krs1t>

Optimizing active learning for free energy calculations

<https://www.sciencedirect.com/science/article/pii/S2667318522000204>

Another aspect of active learning that has received some attention is the sampling strategy used to select molecules. A few groups have reported alternative approaches to guide selections. A team from Exscientia published a method known as Coverage Score that uses Bayesian optimization and information entropy to balance exploration and exploitation in the active learning process.

Coverage Score: A Model Agnostic Method to Efficiently Explore Chemical Space

<https://pubs.acs.org/doi/abs/10.1021/acs.jcim.2c00258>

One challenge in the early stages of drug discovery is deciding which compounds to progress based on somewhat noisy primary assay data. A team from the University of Cambridge published a Bayesian active learning approach that considers the inherent noise in assay data.

Batched Bayesian Optimization for Drug Design in Noisy Environments

<https://pubs.acs.org/doi/pdf/10.1021/acs.jcim.2c00602>

Perspective - The advent of synthesis on-demand libraries like Enamine REAL, WuXi GalaXi, and Otava's CHEMriya have expanded the scope of virtual screening. Unfortunately, even with the availability of inexpensive cloud computing resources, virtual screens with billions of molecules can be quite expensive. Active learning provides an efficient method of docking ultra-large databases. In addition, active learning can

enable FEP calculations on libraries of thousands of molecules. As work on these techniques progresses, they will become commonplace and be integrated into experimental workflows.

8. Molecular Representation

I believe a successful ML effort consists of three elements: the data, the representation, and the algorithm. While a great deal of recent work has focused on algorithms, molecular representation has received limited attention. In early applications of ML in drug discovery, molecules were represented by fingerprints where positions in a vector represented the presence, absence, or count of a particular molecular feature. The advent of CNNs and GNNs led to the emergence of learned molecular representations. While these learned representations have the possibility of outperforming fingerprint models, their superiority has yet to be demonstrated. As mentioned in the first section of this review, recent results have shown that more traditional ML methods using fingerprint representations provide performance equivalent and sometimes superior to that of more sophisticated techniques.

A recent paper from Deng and co-workers at Stony Brook University provides an excellent overview of the three prevailing approaches to molecular representation, fingerprints, and self-supervised representations based on SMILES and molecular graphs. The authors thoroughly review molecular representation and several confounding factors that must be considered when comparing representations and algorithms.

Taking a Respite from Representation Learning for Molecular Property Prediction.

<http://arxiv.org/abs/2209.13492>

One of the key ML advances of 2022 was the emergence of large language models (LLMs). The viral status of software tools like ChatGPT brought large language models into the public consciousness. Several groups published papers showing how LLMs can be used to process libraries of SMILES strings and subsequently produce chemical language models that can be used in QSAR and generative models. To date, the performance of LLMs on QSAR tasks has been less than spectacular. While LLM representations have been applied to molecular property prediction, their performance on benchmarks has not been spectacular. LLMs have shown performance equivalent to more widely used techniques on a few somewhat flawed benchmarks. LLMs are a very new field, and their application in molecular property prediction is in its infancy. It's hoped that future developments in LLMs will lead to new approaches to molecular representation.

ChemBERTa-2: Towards Chemical Foundation Models

<https://arxiv.org/abs/2209.01712>

BARTSmiles: Generative Masked Language Models for Molecular Representations

<https://arxiv.org/abs/2211.16349>

Large-Scale Chemical Language Representations Capture Molecular Structure and Properties

<https://arxiv.org/pdf/2106.09553.pdf>

Infusing Linguistic Knowledge of SMILES into Chemical Language Models

<https://arxiv.org/abs/2205.00084>

Perspective - Over the last decade, we've seen the advent of several neural network approaches to molecular representation. While these approaches promise to provide an additional level of abstraction, there has yet to be a clear demonstration of their superiority to earlier approaches that use molecular descriptors and fingerprints. As this field continues to develop, it is hoped that incorporating 3D information will enable

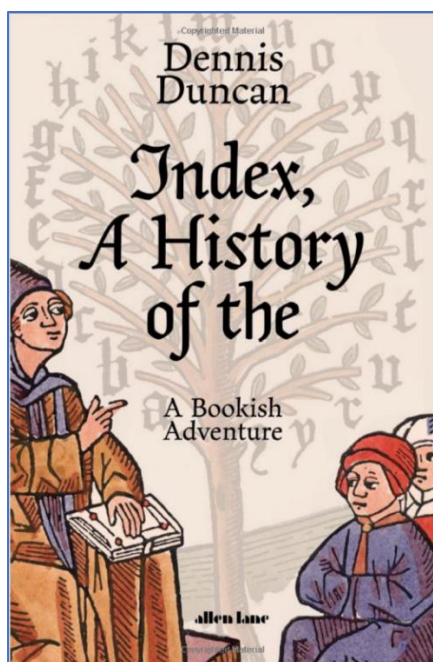
representations that more fully capture the underlying molecular interactions. The new benchmark discussed in Section 1 should provide a more rigorous means of comparing methods that existing benchmark sets.

Book Reviews

Contribution from Robert E. (Bob) Buntrock, Buntrock Associates, Orono, ME USA, email: buntrock16@roadrunner.com

Index, A History of the; A Bookish Adventure from Medieval Manuscripts to the Digital Age

Dennis Duncan. Publisher: Allen Lane, 2021. 344pp + vii, hardcover ISBN 9781324002543, £20.



This book concentrates on the indexes of books, and the development of books. Bound codexes succeeded scrolls and the need for a guide to find information within a codex. Further evolution led to back-of-the-book indexes, indexes for resources other than books, and eventually search engines.

The 17 page introduction is a good guide to and abstract of the rest of the book. Searching for words led to concordances and additional indexes for words, concepts, and author indexes, in books and other resources. Eight chapters describe the need for alphabetisation, origins in religious writings, the need for page numbers, discussions and legal battles over the use of indexes, and controversy over indexes for fiction (e.g., for works of Shakespeare).

The last chapter describes the evolution of digitised indexes and searching, including the appearance and demise of stop-word lists and even hash tags to index Tweets. However, all these developments have not led to the demise of books.

Although the concentration on books and books indexes may make this book of lesser interest to many in this audience, the history and development of related resources should be of interest to all. Recommended.

A review of this book will also appear in the Summer issue of the Chemical Information Bulletin (CIB) of the [ACS Division of Chemical Information](#) (CINF).

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

A History of Scientific Journals: Publishing at the Royal Society, 1665–2015

Aileen Fyfe, Noah Moxham, Julie McDougall-Waters & Camilla Mørk Røstvik. Publisher: UCL Press, 2022. 666pp, ISBN 9781800082328. At the time of writing the [electronic version](#) of this book is available free of charge.

This book focuses primarily (but not exclusively) on *Philosophical Transactions*, the world's longest-running science journal. It is arranged in chronological order, covering the 350 years from its first issue published in 1665 (soon after the English Civil War) until 2015. The authors used the Royal Society's archives to explore the journal's history, leading to a unique study of scientific journal publishing.

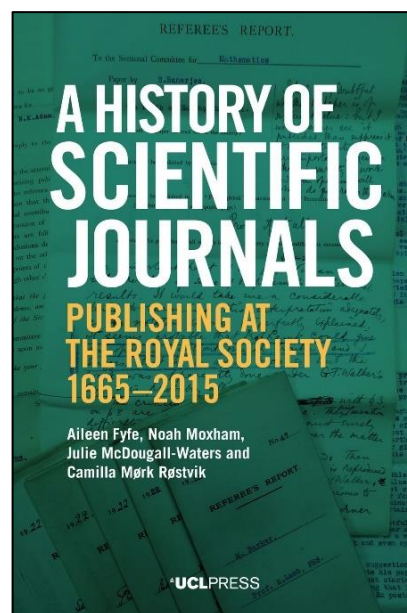
For those like myself who have an interest in the history of chemical information and publishing, the book provides an illuminating, well rounded picture of the evolution of a journal in an ever-changing landscape. It is well written and nicely illustrated with images from Royal Society publications and other relevant documents. It becomes particularly interesting when the human dimensions are explored, such as changes of editors and their different styles, the influence of external parties such as printers, and relationships with authors and the Royal Society itself.

The first editor was Henry Oldenburg (1617-1677), who was also the secretary to the Royal Society, and in 1753 *Transactions* was adopted as the Society's official publication. Concerning the journal's early history, navigating potential hazards of the period such as plague, the fire of London, the imprisonment of Oldenburg and the impact of his death make this section a fascinating read, as does the influence of well known scientists, both in this period (e.g. Robert Hooke) and later.

From the late 17th century until the late 18th century, the sciences were not differentiated as they are today and the journals which acted as the vehicle for publication of science were diverse in their coverage – *Transactions* predated journals covering purely chemistry by over a century. As well as developments in subject coverage over time, changes to the type of content are discussed. In the early days letters and notifications of books dominated the contents and this continued until the late 19th century and the professionalism of science when more structured papers written in impersonal language became predominant. Many aspects of publishing and their changing impact over time are discussed, such as production technologies, the changing role of journals, relationships with other publishers, financial fluctuations, refereeing/peer review, editorial approach, the information explosion, and cultural and social issues of the day. The book concludes with discussion of 21st century challenges from the increasingly commercialised, competitive scientific publishing environment, internationalisation and digital technologies.

The book also covers other Royal Society publications, including the *Proceedings of the Royal Society* (first produced in 1831) and more recent journals (e.g. *Biology Letters*), providing a more complete history of the Society's publications.

As the online version of the book is open access, this should bring it to the attention of a wider audience than a print-only publication. It is likely to be of interest to those with an interest in the history of science, scientific information, librarianship, and publishing.



Cheminformatics and Chemical Information Books

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

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

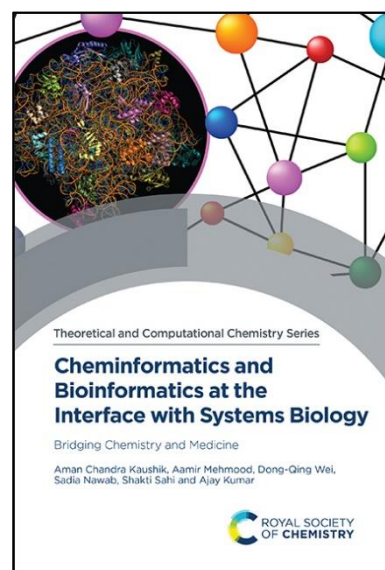
[Cheminformatics and Bioinformatics at the Interface with Systems Biology: Bridging Chemistry and Medicine](#)

The cost of drug development is increasing, and investment returns are decreasing. The number of drugs approved by FDA is in decline in terms of the number of new molecular entities (NMEs). Amongst the reasons noted for this are the adverse side effects and reduced efficiency of many of the potential compounds. This is a

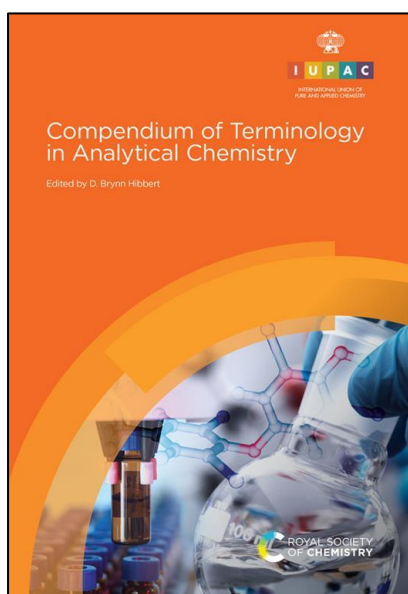
problem both for the pharmaceutical industry and for those suffering from diseases for which there are no or few available treatments.

Advances in computational chemistry, computer science, structural biology and molecular biology have all contributed to improved drug design strategies and reduced the time taken for drug discovery. By interfacing cheminformatics and bioinformatics with systems biology we can create a powerful tool for understanding the mechanisms of pathophysiological systems and identifying lead molecules for various diseases. This integration of drug design approaches can also play a crucial role in the prediction and rationalisation of drug effects and side effects, improving safety and efficacy and leading to better approval rates.

Addressing the lack of knowledge on the fundamental aspects of the various computational tools for drug discovery, this book is a compilation of recent bioinformatics and cheminformatics approaches, and their integration with systems biology. Written primarily for researchers and academics in chem- and bioinformatics, it may also be a useful resource for advanced-level students.



A. C. Kaushik, A. Mehmood, D.-Q. Wei, S. Nawab, S. Sahi, A. Kumar.
Royal Society of Chemistry, July 2023, hardback ISBN 9781839161629, £159, PDF ISBN 9781839166037, EPUB ISBN 9781839166044.



[Compendium of Terminology in Analytical Chemistry](#)

How do you describe an analytical method, measure the purity of the new chemical that you have just synthesised, or report the proper units of measurement? For analytical chemists, the principal tool of the trade, or source of terms, is this book – the so-called Orange Book. First published in 1978, this latest edition takes into account the explosion of new analytical procedures and, at the same time, the diversity of techniques and the quality and performance characteristics of the procedures that are the focus of interest.

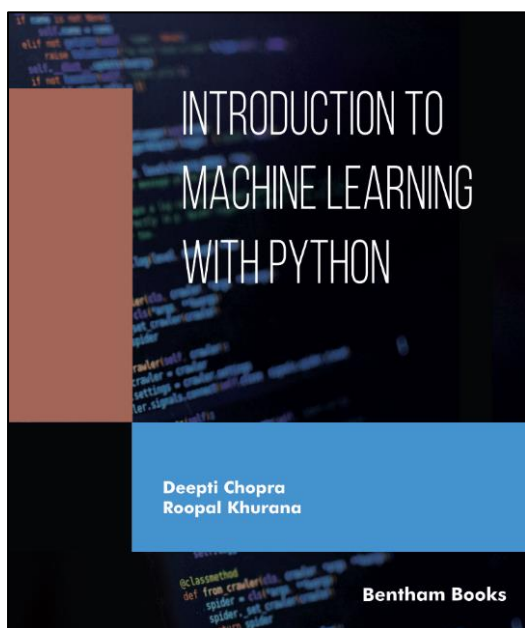
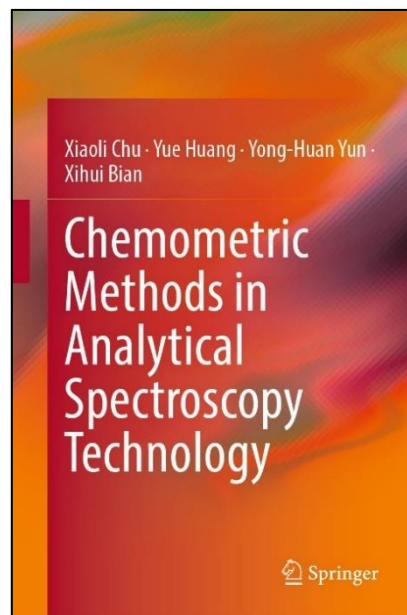
The scope of analytical chemistry has widened, new types of instrumental techniques have emerged and automation has taken over. Answers can now be shared, not only on the chemical composition and structure of the sample, but also changes in composition and structure in space and time. New chapters on chemometrics, bio-analytical methods of analysis, and sample treatment and preparation have been added. The terminology of metrology and quality assurance is now up to date with the latest ISO and JCGM standards. This new volume will be an indispensable reference resource for the coming decade, revising and updating accepted terminology, and providing the official language of analytical chemistry.

Edited by D. Brynn Hibbert. Royal Society of Chemistry, Jan. 2023. DOI <https://doi.org/10.1039/9781788012881>.
Hardback ISBN 9781782629474 £199, PDF ISBN 9781788012881, EPUB ISBN: 9781788019231.

Chemometric Methods in Analytical Spectroscopy Technology

This book discusses chemometric methods for spectroscopy analysis including NIR, MIR, Raman, NMR, and LIBS, from the perspective of practical applied spectroscopy. It covers all aspects of chemometrics associated with analytical spectroscopy, including representative sample selection algorithm, outlier detection algorithm, model updating and maintenance algorithm and strategy and calibration performance evaluation methods. To provide a systematic and comprehensive overview the latest progress of chemometric methods including recent scientific research and practical applications are presented. In addition the book also highlights the improvement of classical algorithms and the extension of common strategies. It is therefore useful as a reference book for researchers engaged in analytical spectroscopy technology, chemometrics, analytical instruments and other related fields.

X. Chu, Y. Huang , Y.-H. Yun , X. Bian. Springer. Hardback May 2022 £149.99, ISBN 9789811916243, paperback ISBN 9789811916274 June 2023, eBook ISBN 9789811916250 May 2022.



Introduction to Machine Learning with Python

Machine Learning is one of the approach of Artificial Intelligence in which Machines become capable of drawing intelligent decisions like humans by learning from its past experiences. In classical methods of Artificial Intelligence, step by step instructions are provided to the machines to solve a problem. Machine Learning combines classical methods of Artificial Intelligence with the knowledge of past to gain human like intelligence.

The book “Introduction to Machine Learning with Python“ has made explanation on Machine Learning with Python from basics to the advanced level so as to assist beginners in building strong foundation and develop practical understanding.

Deepti Chopra, Roopal Khurana. Bentham Science Publishers, 2023. DOI:10.2174/97898151244221230101. ISBN 9789815124439, eISBN 9789815124422. eBook \$49 (personal copy).

Exploring the Promise of Generative Models in Chemistry: An Introduction to Diffusion Models

Charles Harris, PhD Student, Cambridge Centre for AI in Medicine, Computer Laboratory, University of Cambridge,
email: cch57@cam.ac.uk

“Creating noise from data is easy; creating data from noise is generative modelling.” - Yang Song

Introduction

Generative modelling, a technique that learns patterns in data to create new data, has experienced a renaissance in the past year. The catalyst was DALLE-2, an innovative image-to-text model revealed by the AI research firm OpenAI in April 2022. With DALLE-2, users can describe an image in plain text – encompassing multiple objects, scenes, or artistic styles – and the model will generate a brand new image from this description. The results have been noteworthy due to the extraordinary fidelity of the produced images and the model’s ability to seamlessly blend diverse abstract concepts within a single image. The model used a relatively new class of generative model called a diffusion model.



Images produced by Stable Diffusion, a model similar to DALLE-2. Image credit: [Stability AI](#).

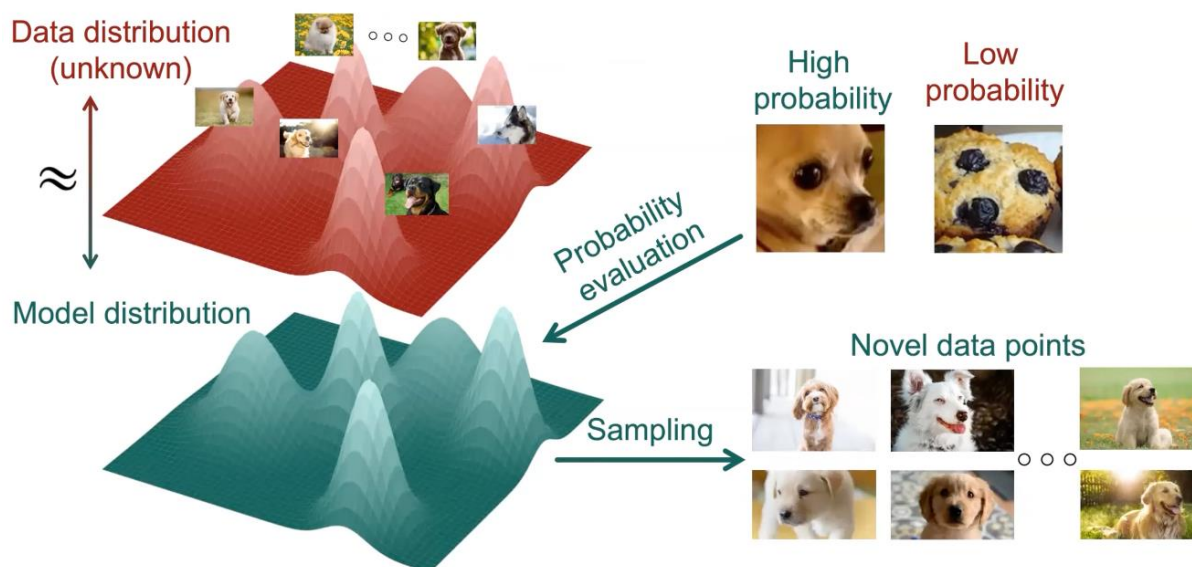
At this point, you might reasonably wonder, "What does this have to do with chemistry?". Since DALLE-2, the buzz around ‘Generative AI’, a term seemingly coined by venture capitalists only recently, has since continued to intensify with many having found practical applications in the field of chemistry, particularly in the realm of drug discovery. By leveraging large datasets of chemical information, these AI models aim to generate new molecular structures with specific properties, offering a more targeted approach to drug development – at least in theory. It’s important to note that there are still challenges to overcome, such as the need for experimental validation and better evaluation altogether.

This article aims to condense and simplify the recent developments in generative models, specifically focusing on small molecule drug design with diffusion models. I aim to make the content accessible to chemists from outside the machine learning sphere. By exploring the unfolding story of diffusion models, I hope to provide insight into current AI trends and their potential impact on the intersection of AI and (bio)chemistry.

What is generation modelling?

Generative modelling aims to learn from a dataset – be it images, text, music, or molecules – and replicate its underlying patterns to create new samples. By analysing a multitude of examples, the model discerns common features and characteristics, thereby understanding the likelihood of different patterns within the dataset. Once the generative model has acquired knowledge about the distribution, it can generate new samples by sampling from this learned representation. These generated samples are not simply copies of existing data points, but

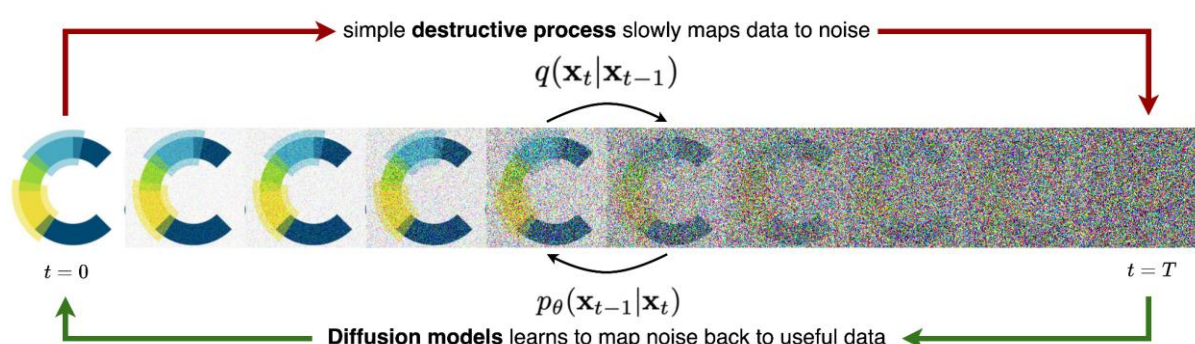
rather novel instances created based on the patterns and characteristics observed during training. This allows generative models to produce new and previously unseen samples that exhibit similar traits and qualities as the original data.



Overview of generative modelling. Image credit: [Yang Song](#).

How do diffusion models work?

Simply put, a diffusion model is a type of generative model that establishes a Markov chain of progressive noising, or 'diffusion' steps. In these steps, random Gaussian noise is added to real data until the original sample is unrecognisable. The next step is to train a model, typically a neural network, to reverse this process. Once trained, the model can create new samples by pulling from a normal distribution (random noise) and denoising this data until a high-quality, new sample emerges. That's the fundamental idea, although be warned, a deeper understanding involves some equations.



Overview of diffusion generative modelling. I have used the RSC logo here for fun, but note this would work with any class of images. Inspiration from [Google Research](#) and [Ho, et al, 2020](#).

Forward diffusion process

Suppose we have some real data (e.g. an image) that we call x_0 , which is a sample from a true data distribution we wish to learn $x_0 \sim p(x)$. We can define a *forward diffusion process* (q) which will gradually add a small amount of Gaussian Noise to the sample over T steps (where T is typically $\sim 1,000$), resulting in a sequence of progressively more noisy samples x_1, \dots, x_T . The amount of noise added at each step is controlled by a fixed variance schedule with, where β_t will be between 0 and 1 depending on t .

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I})$$

The sample x_0 will gradually lose its distinguishing features as t approaches T and the sample becomes fully noised, in other words, x_T is equivalent to a Gaussian distribution. We can say that $q(x_t | x_{t-1})$ represents the transitional probability distribution (i.e. the noise added) between x_{t-1} and x_t .

Reverse (generative) diffusion process

We then teach a model to learn the reverse (or generative) diffusion process $p_\theta(x_{t-1} | x_t)$, which will be able to generate new samples starting with some random data as input, $x_T \sim N(0, I)$. Remember, while there is no information stored in a noisy image, there is information about how to make a new piece of data stored in the weights of the denoising model.

$$p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$$

To be more concrete, we can say that there is some true noise ϵ_t , originally added to the data, which we wish to remove to model $x_t \rightarrow x_{t-1}$. We then train a denoising network, $\epsilon_\theta(x_t, t)$, which aims to predict ϵ_t directly. While the exact derivation of the training objective is complex, we can simplify by saying that our training objective is to minimise the difference between the true denoising process at every step ϵ_t and the noise predicted by a tunable neural neural network $\epsilon_\theta(x_t, t)$, which takes as input the current noised sample x_t and time step t .

$$\mathcal{L}_t^{\text{simple}} = \mathbb{E}_{t \sim [1, T], \mathbf{x}_0, \epsilon_t} \left[\|\epsilon_t - \epsilon_\theta(\mathbf{x}_t, t)\|^2 \right]$$

Connection with score-based models

As is common in science, there were actually two different groups of people working on generative modelling that happened to arrive at the same formulation at similar times, approaching the problem from different perspectives. The first is the diffusion-based perspective which I've already described, and while the noise-denoise paradigm is good for a quick explanation, developing an intuitive understanding can be challenging. For those seeking a deeper understanding, I recommend exploring the second perspective, known as the score-based perspective.

In statistics, the *score* of a probability density function $p(x)$ is defined as the gradient of the log of that function $\nabla_x \log p(x)$. In score-based modelling, we aim to train a score network s_θ which tries to estimate the score such that $s_\theta(x) \approx \nabla_x \log p(x)$. In generative modelling, we want to design a new sample x which has a high likelihood of being good quality according to the true data distribution $p(x)$. By learning the gradient of $p(x)$ with respect to x , the model essentially learns the direction in which x should be moved to improve its quality, that is, to increase its likelihood within the distribution. I would highly recommend this [blog post](#) by Yang Song for those interested in learning more.

Sampling with stochastic gradient Langevin dynamics

However, generating new data from a diffusion model (or score-based model) requires an additional trick to be effective. You can imagine that if you continue optimising a sample "x" solely based on the direction of a given score function, you will consistently converge to the same point on the learned distribution. This point typically corresponds to a mode within the training dataset, which, as you can imagine, is not particularly useful.

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \epsilon_t(\mathbf{x}_t, t) + \mathbf{z}_t, \text{ where } \mathbf{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

To address this issue, we draw inspiration from Langevin dynamics as commonly utilised in molecular dynamics simulations. In this approach, a small quantity of random noise is typically introduced to the simulation, accounting for the impact of thermal fluctuations and interactions with the surrounding environment. By incorporating Langevin dynamics into our sampling process, we can obtain high-quality samples that are also diverse.

Diffusion models for molecule design

In the ensuing sections, I'll be highlighting some notable works within the realm of diffusion models and generative chemistry. This overview will merely skim the surface of current developments, as 1-2 new papers on molecule generation using diffusion models emerge weekly. For those keen on staying abreast with the latest work, I recommend this continually updated [list on GitHub](#). But before we delve in, let's start with a brief history of molecule design through generative modelling.

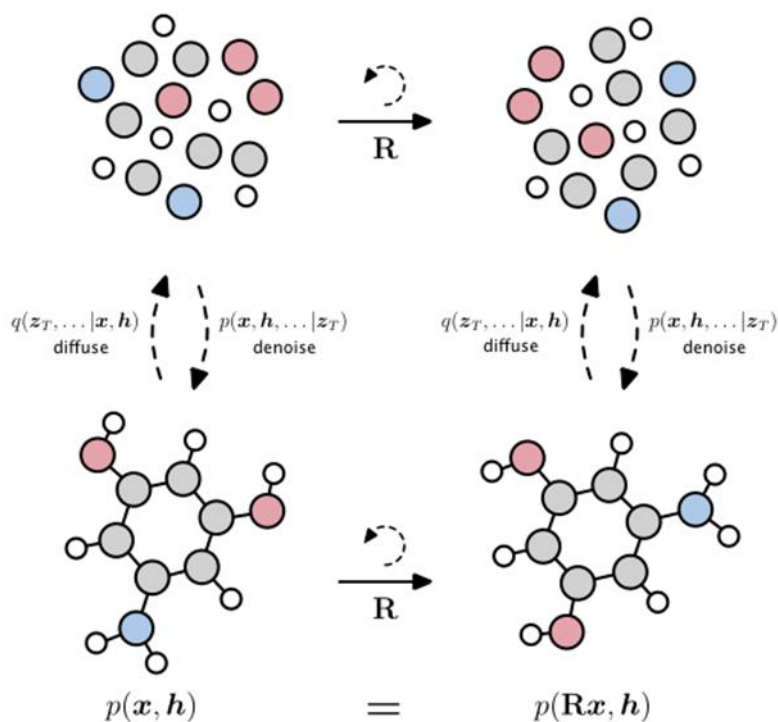
Previously, deep generative models for small molecules typically fell into two categories: all-at-once and atom-by-atom generation. The former assumes atom independence during generation, which is unrealistic given molecule stability hinges on specific configurations. Conversely, atom-by-atom models necessitate an arbitrary atom ordering, complicating stable training.

Diffusion models present a compelling option for molecular design, utilising a process of iterative refinement. They gradually transform noise into molecules over several steps, enabling a methodical design process and the rectification of any errors along the way. This approach parallels the procedure employed by AlphaFold2's Structure Module, which repeatedly refines its protein structure predictions through multiple model iterations until an accurate structure is achieved.

Equivariant diffusion model

Let's begin our exploration of diffusion in generative chemistry by examining the well-known Equivariant Diffusion Model (EDM) paper. In this study, the authors trained a 3D diffusion model on a dataset composed of drug-like molecules to enable *de novo* generation of new molecules in 3D.

In this work, molecules are represented as graphs, where nodes are atoms and edges are chemical bonds, $G \in \{N, E\}$ with nodes N and edges E . Every node has two features $z = [x, h]$, where h is a vector representing the atom type and x is another vector representing the coordinate. As in the example below with images, x and h can be viewed as such examples from the true data distribution to which we can add noise. They define a forward diffusion process $q(z_t|x, h)$, which gradually injects noise until the molecule becomes a blob of random atoms (see left of figure). Then train a reverse process $p(z_{t-1}|z_t)$, which tries to predict the noise that was added to the data and then remove it to make clean data (see right of Figure).



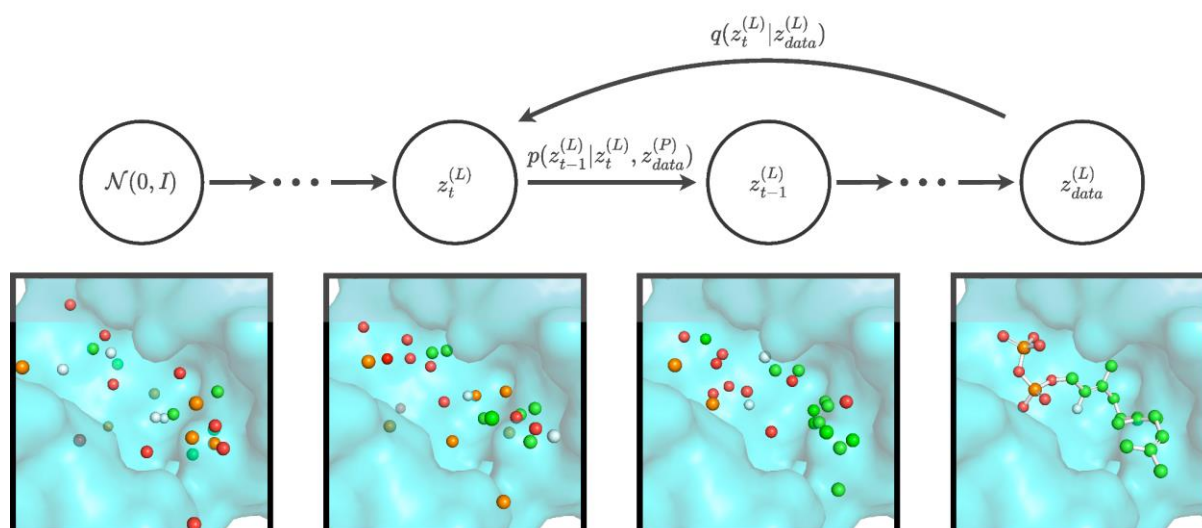
EDM overview. Credit [Hoogeboom et al.](#)

As we represent molecules as graphs, it is very natural that we try to learn and reason over these graphs using Graph Neural Networks, a special kind of neural network that takes a graph (in this case molecule) as input, in order to predict the noise. As we are working with graphs that are in 3D, we use geometric GNNs which can handle these 3D features, which are technically equivariant GNNs. The word 'equivariant' simply means that we treat these graphs and the denoising process the same when the molecule is shifted and rotated in 3D space, the coordinate system in which we have the molecule is arbitrary. Figuring out which is the best equivariant GNN is a *massive* area of research and depends enormously on the exact problem you are working with in bio(chemistry).

DiffSBDD

This work was one of the first to introduce a DIFFusion model for Structure-based Drug Design (DiffSBDD), which recasts SBDD as learning to generate novel molecules *conditional* on a protein pocket structure. The model is built on top of the EDM framework but critically, the denoising networking takes as input the 3D structure of the target protein (which is represented as another 3D graph) as input, thus conditioning the generative procedure on the target structure.

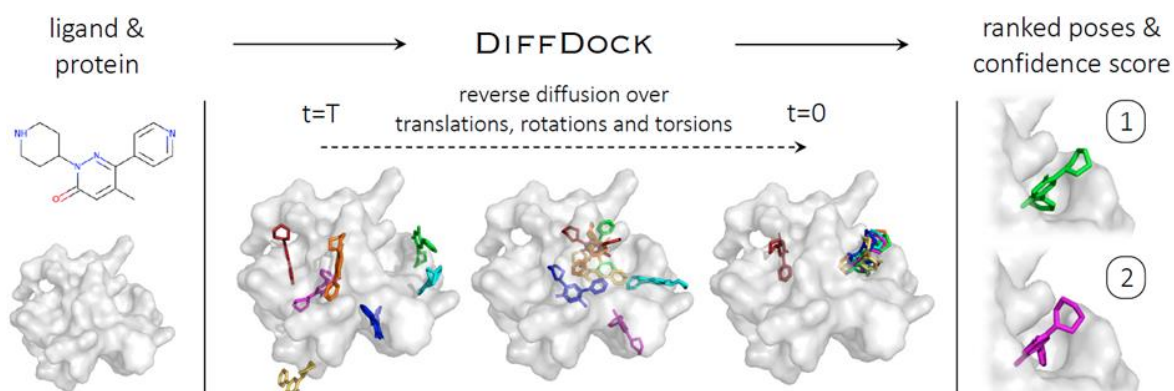
DiffLinker is a similar method from the same group that is specifically designed for fragment linking within protein pockets. Here, they train a specialised model that takes a protein structure with fragments bound as inputs and generates a novel linker between them in 3D. They also have an additional auxiliary model, which predicts the size of the linker automatically.



DiffSBDD overview. Credit [Schneuing et al.](#)

DiffDock

DiffDock is a generative diffusion model, specialised for protein-ligand docking, not molecule design. It builds on the success of Torsion Diffusion, which excelled at generating molecular conformers by focusing on the molecule's torsional angles, while keeping bond length and angles constant – an excellent example of ‘inductive biases’ in machine learning. DiffDock was notable for recasting the docking problem as a generative one, rather than a predicting problem, as sampling from the model always given a different pose (hopefully clustered around true binding modes).



DiffDock overview. Credit [Corso et al.](#)

What's next for diffusion models in chemistry?

As you might imagine, I've barely been able to scratch the surface of what is happening in diffusion modelling for chemistry currently. There is also an expansive literature applying these methods to *de novo* protein design as well. In my concluding section, I'd like to provide some commentary on three points: two challenges and one opportunity.

Data, data and data

As I said above, generative modelling is fundamentally based on learning the underlying distribution of a given dataset. Consequently, the quality and diversity of the training distribution significantly influence the performance and capabilities of the generative model. The model's ability to generate samples that resemble the training data, capturing its patterns and characteristics, is limited by the training distribution itself. If the

training distribution is of limited quality or lacks diversity, the generative model will be constrained and incapable of producing high-quality or novel samples beyond what it has learned.

This isn't a problem if you are, say, trying to generate images of cats (the internet is very good for that sort of thing) but we very quickly run into problems when we start dealing with molecules. To give some perspective, for all its remarkable performance, DALLE-2 had to be trained on 400 million images from the internet, whereas the number of high-quality protein-drug complexes we have available from the PDB is measured in the 10,000s, with the number of unique molecules and targets being even lower. People have tried to train on larger, synthetic datasets, produced by docking, but these of course come with their own sets of issues. Ultimately, this brings me back to the point that [many people have made before](#), which is that we need more 'machine learning-ready' datasets if we want to solve a lot of these challenges.

More rigorous evaluations are needed

Unfortunately, machine learning publications live or die by their performance, with declarations of "state-of-the-art (SOTA) performance in [X]" almost a prerequisite for inclusion in top-tier machine learning conferences. This tendency, however, can unintentionally foster an environment of over-optimisation towards simplified or flawed metrics, potentially guiding well-meaning computer scientists off course for considerable durations. The complexity of this issue escalates in fields like chemistry and drug discovery, where design processes are intricately multi-objective and abound with known unknowns.

For instance, DiffDock conducted a benchmark of their method against more conventional docking software, comparing the percentage of successful predictions with RMSD (Root Mean Square Deviation) less than 2 Å and the median RMSD. However, their evaluation sparked controversy as it was performed in the blind docking scenario, where the binding site is unknown. In contrast, the methods they compared, such as Glide, are specifically designed for docking with known binding sites.

In February, a different group published an excellent [benchmark paper](#) that provided a fairer evaluation of these methods, by assessing both blind docking and docking with known pockets. The results indicated that traditional methods outperform DiffDock when docking into known pockets but that DiffDock excelled as a state-of-the-art binding site prediction algorithm, surpassing conventional methods. They suggested the optimal approach for blind docking is to first utilise DiffDock to detect the correct binding pocket and then employ conventional docking to achieve accurate positioning.

In my own field of SBDD with generative modelling, I'm worried we're walking up a similar path. We've developed these 3D generative methods that generate a molecule and its pose simultaneously, ostensibly for efficiency – it's quicker than generating the molecule first and then determining the pose through docking. Yet, when it comes to quality assessment, the focus tends to be solely on the molecules. Key metrics such as Quantitative Estimation of Drug-likeness (QED) and Synthetic Accessibility (SA) score are employed, but the assessment of the poses generated directly by the model is often overlooked, with molecules typically being redocked. This leaves a significant question unanswered: Are we truly creating models that excel at crafting high-affinity molecules with perfect protein-ligand complementarity? Or, in failing to assess pose quality, are we merely performing fancy atom filling with no concern to how the ligand interacts with the pocket? The distinction is critical, and it's important to address this gap in our evaluation methods.

Other classes of generative models

Apart from diffusion models, the generative AI hype train has a few more carriages I've not had time to mention (some of which are actually worth paying attention to). One such model is [GFlowNets](#), tailored specifically to maximise a certain 'reward' – like binding affinity, for instance. An intriguing ability comes from Large

Language Models (LLMs) like GPT-4. While their primary function is to predict the next word, they can also generate novel molecules using SMILES strings. While I am personally sceptical, there are some [serious people doing diligent research](#) in this direction.

Conclusion

Generative modelling, through its capacity to understand and replicate the statistical properties of various datasets, has emerged as a potent tool for generating novel samples based on learned patterns and characteristics. Despite the potential, these models still face challenges, including data bias, data quality, and lack of rigorous evaluations being silent failures within the ML field. Overall, with the continual evolution of these models and advancements in computational capabilities, the impact of diffusion models in generative chemistry is expected to expand significantly in the coming years.

I would like to thank my colleagues at Cambridge, AstraZeneca, EPFL and Cornell for their insightful discussions that led to many of the ideas explored in this article. I would also like to thank my supervisors, Professors Pietro Lio and Sir Tom Blundell, for their guidance and support.

References

- Ho, J. et al. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, **2020**, 33, 6840-6851.
- Ramesh, A. et al. Hierarchical text-conditional image generation with clip latents. arXiv preprint, **2022**, *arXiv:2204.06125*.
- Hooeboom, E. et al. Equivariant diffusion for molecule generation in 3d. In *International Conference on Machine Learning*, June **2022**, 8867-8887. PMLR.
- Schneuing, A. et al. Structure-based drug design with equivariant diffusion models. arXiv preprint, **2022**, *arXiv:2210.13695*.
- Welling, M. and Bronstein, M. Structure-based drug design with equivariant diffusion models. arXiv preprint, **2022**, *arXiv:2210.13695*.
- Corso, G. et al. Diffdock: diffusion steps, twists, and turns for molecular docking. arXiv preprint, **2022**, *arXiv:2210.01776*.

News from ACS CINF

Contribution from Ye Li, CINF Chair 2023, Massachusetts Institute of Technology, email: yel@mit.edu, Sue Cardinal, 2022 CINF Past Chair, Interim Fundraising & Nominating 2023, University of Rochester, email: scardinal@library.rochester.edu

American Chemical Society, Division of Chemical Information Division (ACS CINF) members participated in CINF symposia, reception, and social events during the ACS Spring 2023 conference at Indianapolis and online. Here is a list of symposia we hosted.

- Chemistry Database Application Programming Interfaces (APIs): Available Databases, Workflows, and Use-Cases
- Advancing FAIR Chemistry: Developing New Services for Sharing Chemical Data: WorldFAIR Chemistry Project Updates
- Cheminformatics/Computational-based Chemical Library Enumeration
- Machine Learning and AI for Organic Chemistry
- Electronic Lab Notebooks and Notebook-like Applications for Data Sharing
- CINF General Papers

Several CICAG members contributed to the ongoing [WorldFAIR Chemistry projects](#) and participated in their do-a-thon during the conference. CINF Chairs also [showcased CINF](#) during the presidential event *ACS Divisions and the POWER of Chemistry: Opportunities and Challenges!* If you would like to contribute or make

suggestions for CINF operations, please reach out to us or any [CINF Exec](#) members. We would love to work with you more. We continue to seek volunteers to work in CINF leadership positions, especially on communications and membership committees. Connect to our [website](#) for more details.

The CINF program for the ACS Fall 2023 Meeting with the theme “Harnessing the Power of Data” at San Francisco is shaping up to be an exhilarating and jam-packed event highly in sync with our mission. Organisers were inspired to plan three simultaneous programs daily with two hybrid tracks. CINF will host our Welcome Reception and Data Summit Kickoff together with IUPAC, WorldFAIR Chemistry, CSA Trust, and all organisers of CINF symposia related to chemical data, informatics, machine learning and AI. We will also celebrate our winners of the Val Metanowski Meritorious Service award during the Welcome Reception. The celebration of our Herman Skolnik Award winner Dr Patrick Walters will be grand with a full day symposium on 15 August, concluding with an evening reception. The [registration for the ACS Fall 2023](#) is now open and the [full conference program](#) will soon be available.

With data, machine learning and AI becoming an inseparable part of chemists’ professional life globally, ACS CINF looks forward to working closer with international organisations on these fronts. If you have ideas for projects that CICAG and CINF can work on together, please reach out to us.

Cambridge Structural Database (CSD) Updates

Contribution from Michael Francis, Marketing Manager at CCDC, email:

mfrancis@ccdc.cam.ac.uk



It’s been a busy few months at the CCDC with several enhancements to the CSD and software:

- A new Macromolecule Hub that allows scientists to search and mine the world’s protein binding and small molecule structural data in one place.
- Interoperability improvements to existing features including InChIs in the Cambridge Structural Database, stereochemistry search, and ChemDraw export.
- A data update brought the total size of the CSD to 1,241,941 entries and 1,214,288 unique structures. The latest addition of over 75,000 structures further expands and diversifies our knowledge in chemistry and its applications. There is also an update on our ongoing investigations that have been carried out on structures associated with a pre-print on papermills in crystallography with 209 structures retracted from the CSD.
- A new installer.

We are also pleased to announce the upcoming webinar on 20 July 2023: [Using Python API for Ensemble Docking](#).

More information can be found on our [What’s New?](#) page.

News from CAS

Contribution from Dr Anne Jones, Senior Customer Success Specialist, email:

ajones2@acs-i.org



Groundbreaking innovation opportunities are hidden in today's complex scientific data landscape. At CAS, we are uniquely able to help you find them. Our team connects the world's scientific knowledge, published in more than 50 languages, to build the largest human-curated scientific data collection in the world. Partner with us to ignite innovation, unlock shared knowledge, and empower confident decisions throughout your innovation journey.

CAS and Dow win GOLD at the Edison Awards™!

In April, CAS was thrilled to win a GOLD Edison Award™ for innovation, in partnership with Dow, for the collaborative custom solution Smart Search for Dow powered by CAS. This innovative solution helps Dow researchers instantly identify and prioritise ideal compounds for R&D innovations from worldwide data. The award in the AI Optimised Productivity category reflects the 100x efficiency increase that Dow has seen in their search process as a result of the new solution. [Learn more, by exploring the demo and case study.](#)

CAS Insights™

[CAS Insights](#) is an open-access content hub at the intersection of science, technology, and innovation that offers R&D and business leaders actionable perspectives on the latest developments across science and technology. Drawing on CAS human-curated data collection and the deep scientific expertise of our science team, the content highlights emerging trends, unseen connections, new applications, and future opportunities across disciplines. [Subscribe for free](#) to get the newest insights delivered to your inbox.

Recent CAS Insights relevant to CICAG members:

- [Dark data in R&D: How knowledge management can uncover hidden value](#) Research and development teams amass huge quantities of complex data, which if exploited correctly, can be a useful information source to improve decision-making and drive innovation.
- [R&D Insights: 3D printing in biomedical applications](#) Understand emerging trends of 3D printing in biomedical applications and the new technologies that are reshaping personalised healthcare. From 3D printing devices to custom drug delivery and printing bioactive tissues and organs, this field is innovating rapidly with implications across the healthcare industry.
- [Predicting New Chemistry: Impact of High-Quality Training Data on Prediction of Reaction Outcomes](#) Scientists at Bayer expand their understanding into new, useful chemistry by improving the predictive power of a synthesis planning model with scientist-curated reactions.

Meet the 2023 CAS Future Leaders

The CAS Future Leaders program supports the growth of science leadership potential among early-career scientists. Since 2010, the program has awarded Ph.D. students and postdoctoral scholars with opportunities to network with global peer scientists, connect with industry thought leaders, and take part in science leadership training. The , selected from among hundreds of highly qualified applicants, include 35 outstanding scientists representing a wide array of scientific disciplines and organisations from around the world. This group will visit CAS headquarters in Columbus, OH in August and attend ACS Fall 2023 in San Francisco.

CAS SciFinder Discovery Platform™

The CAS SciFinder Discovery Platform, which consists of CAS SciFinder[®], CAS Formulus[®], and CAS Analytical Methods[™], remains your best source for scientific information, data, methods, formulations, and more. To

understand the full scope of capabilities and information available to you through these solutions, [visit the CAS website](#).

Working in response to feedback from scientists across the globe, we continue to focus on providing you with access to tools that improve the efficiency and effectiveness of your scientific research.

Among recent notable enhancements in CAS SciFinder[®]:

- Project files – a new feature that allows you to organise, store, and share references within a project that you create.
- Visualisation improvements – get a better experience viewing all citing and cited references with the improved Citation Map.
- RDFile V3000 format downloads – this expanded download capability offers new capabilities for managing larger molecules and provides extended stereochemistry support.
- User preference customisations – personalise your experience with options related to filters, alerts, and displays.
- More flexible reaction search result sets – sort reaction results by transformation name and reaction count.
- Increased patent coverage – view results from 109 issuing authorities, an increase of 45 global sources.

More details on the latest enhancements can be found anytime by viewing the “What’s New?” section available within CAS SciFinder[®]. You can also reach out to us for more information about anything related to using the solutions comprising the CAS SciFinder Discovery Platform.

STN IP Protection Suite™

The STN IP Protection Suite offers trusted search and monitoring solutions, including CAS STNext[®], CAS Scientific Patent Explorer™, and FIZ PatMon, and access to search services and expertise through CAS Search GuardSM. The Suite also includes extended capabilities to support comprehensive and efficient search, including CAS PatentPak[®], and access to expanded formulations and sequences content. Find more information on the included capabilities on the [CAS website](#). Throughout 2023, CAS has continued to enhance these solutions to meet the growing search needs of our users.

Recent enhancements of note in CAS STNext include:

- Increased patent coverage – with recent expansions, users can get insight into patents from 109 issuing authorities spanning more than 40,000 unique IPC codes. Thailand, Ukraine, and Saudi Arabia are among the 45 additional global authorities for which bibliographic records are now available. The previous coverage of Chile has also been expanded, now including current Chilean patents.
- Expanded access to Taiwanese patent information – claims coverage has been expanded to include Taiwan, and a new value-added full-text database is available.
- New interactive claim viewer – quickly discover the relationships between claims with a new graphical tree display.
- National office and interactive claims viewer links available in transcripts and reports – users can now enhance the value of reports for your stakeholders by providing fast and convenient links to verify the stage and get an efficient overview of the connections between different claims
- Fragmentation code feature enhancements - extended capabilities for these automated scripts, enabling more comprehensive results for prior art searches.

More information can be found in the “What’s New” section within CAS STNext.

AI4SD News

Contribution from Dr Samantha Kanza, AI4SD Network+ Coordinator, University of Southampton, email: s.kanza@soton.ac.uk



AI4SD Final Report

As part of the end of our first iteration, AI4SD produced a [final report](#) to summarise all of our outputs and impacts.

Abstract

The Network has run for four very successful years between 2018-2022. Over these four years we created a community of over 1500 people and gained a strong social media presence, especially on Twitter. We have funded seven pilot projects, 15 interns and over 120 students to attend our machine learning summer schools, in addition to providing training to hundreds of other Network members through our *Failed it to Nailed it* and *Skills4Scientists* series. We were undeterred by COVID and took our Network activities online, doubling our membership

over the summer of 2020 and bringing us a new array of international members where before we had been predominantly UK based, and since the return of physical events, we converted to a hybrid Network bringing together the best of both worlds for a mix of hybrid and virtual activities. The Network has produced an immense number of resources, including 174 videos for our YouTube Channel and 52 reports across funded projects and events, in addition to supporting the publication of 19 peer reviewed journal and conference papers. The Network has also been heavily involved in other projects including supporting an Ethics for AI working group, and some collaborative projects with the Network of Networks group to run an Equality, Diversity, and Inclusion survey, and to write and compile a very comprehensive resource on how to run a Network which has been distributed to new Networks who were looking for guidance. Overall, we feel that this Network has been very successful both in terms of impact and outputs. This report provides a comprehensive overview of the Network. Section 2 details our membership statistics and growth, different social media and website accounts, and statistics on our events. Section 3 lists our outputs, Section 4 contains impact statements and collated feedback on our funded initiatives, and Section 5 describes our other outputs including our work with the Network of Networks.

AI4SD Seminars

AI4SD are continuing to run seminars, our latest two are:

[SciData: Semantic Representation of Scientific Data and Applications in Chemistry – Stuart J. Chalk](#)

Findable, Accessible, Interoperable and Reusable (FAIR) is a paradigm shift in how we should make our research data available and useful for other scientists. The [SciData framework](#) is a specification for constructing JavaScript Object Notation for Linked Data (JSON-LD) that are semantically encoded, giving meaning to research data and its contextual metadata. This talk will cover the basics of how to create SciData JSON-LD using Python, some example data files for different types of data, use of the file format for the basis of a digital research notebook, and digital chemical twins.

<http://dx.doi.org/10.5258/SOTON/AI3SD0237>



Dr Stuart J. Chalk is a Professor in the Department of Chemistry at the University of North Florida. He received his PhD in chemistry from the University of Massachusetts at Amherst in 1994 and postdoctoral work at Duquesne University, Pittsburgh from 1994-1996. Although trained as an analytical chemist, Dr. Chalk's research now focuses on the areas of Chemical Informatics and Data Science. In particular, Dr Chalk has projects focused on machine accessibility of solubility data, online enhancement to the IUPAC Gold Book, automated extraction and annotation of chemical property data from PDF files, and scientific data models.

[Reproducibility, Jupyter notebooks and associated research software engineering - Hans Fangohr](#)

The talk introduces the topic of reproducibility in science: what is reproducibility, why does it matter and why is it hard to achieve? It will discuss abstract requirements for reproducibility and try connect these to concrete measures we can take in day-to-day research to make our results more reproducible. The potential of Jupyter Notebooks and the Jupyter ecosystem is explored. This talk will be of interest to everyone who wants to ensure that the computational aspects of their research is reproducible and re-usable and to see how this can help meet the requirements of journals for manuscript submission and for UKRI Responsible Research and Innovation (RRI). Following good practice for reproducibility will also make research life easier, especially when writing up work.

[Link to slides](#)







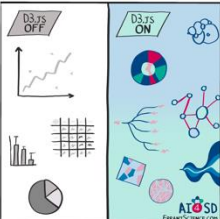

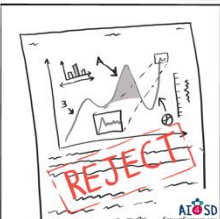

<https://doi.org/10.5258/SOTON/AI3SD0269>



Hans Fangohr is head of the scientific support unit Computational Science at the Max Planck Institute for Structure and Dynamics of Matter, and Professor of Computational Modelling at the University of Southampton. He received his undergraduate degree 'Diplomphysiker' in physics from the University of Hamburg (Germany) and completed his PhD studies in the High Performance Computing Group at the department of Computer Science in Southampton. He is a full professor since 2010, and specialised in computational science, data analysis and software engineering for science.

Failed it to Nailed it: nailing your data visualisation

This event was part of the Failed it to Nailed it series run by AI4SD, the Cell Press Patterns Journal and the Physical Sciences Data Science Service (PSDS). It was designed to teach researchers how to nail their data visualisation. There were talks on the importance of data visualisation and how to tell your story through your data. There were also some technical training sessions that delved into some of the tools and techniques that can be used for this and there were a number of challenges to choose from for the hands-on part of the workshop. All of the videos from this series are now available on our YouTube Channel: [Failed it to Nailed it: Nailing your Data Visualisation: Training Workshop Playlist](#).

Cartoon	Title	Speaker	Video Link
	Introduction to Data Visualisation Dr Rita Borgo (Kings College London)		Video Link
	(Reproducible) Data Visualisation with R and how to make interactive things with R Miss Charlie Hadley (Visible Data Ltd)		Video Link
	Data Visualisation with Python Mr Samuel Munday (University of Southampton)		Video Link
	Visualise Your Data for the Web using D3.js Dr Alfie Abdul-Rahman (Kings College London)		Slides available on request
	Data Visualisation in Publishing & Communication Dr Sarah Callaghan (Patterns)		Video Link

Humans of AI4SD interviews

AI4SD interviews We now have 39 Humans of AI4SD Interviews published on our [website](#).



[Prof. Andy Stanford-Clark](#)



[Prof. John Overington](#)



[Prof. Jonathan Goodman](#)



[Dr Naomi Jacobs](#)



[Dr Reinhard Maurer](#)



[Prof. Matthew Todd](#)



[Prof. Tim Albrecht](#)



[Dr Jennifer Hiscock](#)



[Dr Nicholas Watson](#)



[Dr Martin Immanuel-Bittner](#)



[Dr Simone Vannuccini](#)



[Ms Ekaterina Prytkova](#)



[Dr Louise Dennis](#)



[Dr Zosia Beckles](#)



[Dr Keith Butler](#)



[Dr James Cumby](#)



[Dr Terhi Nurmikko-Fuller](#)



[Dr Mark Warne](#)



[Dr Paul Dingwall](#)



[Dr Grant Hill](#)



[Dr Al Dossetter](#)



[Dr Aurora Clark](#)



[Dr Barbara Zdrazil](#)



[Prof. Carlos Zednik](#)



[Dr Egon Willighagen](#)



[Prof. Charlotte Deane](#)



[Mr Christopher Gutteridge](#)



[Dr Will McNeill](#)



[Prof. Tony Hey](#)



[Prof. Henry Rzepa](#)



[Dr Frank Langbein](#)



[Dr Heather Kulik](#)



[Dr Jan Jensen](#)



[Mr Jules Tilly](#)



[Prof. Markus Kraft](#)



[Dr Nessa Carson](#)



[Mr Yingfang Yuan](#)



[Prof. Zoheir Sabeur](#)



[Dr Olexandr Isayev](#)



Sharkcat

For those of you who were wondering what Sharkcat has been up to, they have been very busy! Their latest adventures include getting to know their new University campus, identifying their favourite coffee shop, undertaking some python fair data training, and attending the Research Data Alliance Plenary and wowing the audience with a wonderful speech!

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

PSDI is a project, funded under the EPSRC DRI (Digital Research Infrastructure), 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 [here](#).

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>

Get involved with PSDI

There are lots of ways you can engage with the PSDI project, below are some of our current initiatives.

Machine Learning Summer School

We are pleased to announce the [Machine Learning for Atomistic Modelling Autumn School](#); this machine learning for materials training course is being run by the [Physical Sciences Data Infrastructure \(PSDI\)](#) initiative

in collaboration with [PSDS](#), [AI4SD](#), [STFC-SCD](#) and [CCP5](#). This training is targeted towards PhD students, particularly in the materials and molecular simulations field, who have experience of coding but are not highly experienced with machine learning. The aim of this training is to introduce attendees to the latest methods of machine learning applied to atomistic simulation of materials and provide hands-on experience of using machine learning for atomic and molecular simulations. There will also be the opportunity for attendees to present a poster on their work.

Machine Learning for Atomistic Modelling Autumn School



18th - 20th Sept 2023
Daresbury Laboratory,
Daresbury





Fee: £100
Covers
2 nights
accommodation +
catering

This 3 day training, aimed primarily at PhD students in Materials and Molecular Simulations, will introduce attendees to the latest ML methods applied to atomistic simulation of materials. This training includes talks and practical sessions, focusing on the basics of ML, ML interatomic potentials and graph neural networks.

www.psdi.ac.uk/ml-school-2023









PSDI process recording survey

Do you record notes for your research? Do you use any sort of software for this? E.g. OneNote, Electronic Lab Notebooks, other? If so, you might be interested in this survey from the Physical Sciences Data Infrastructure (PSDI) Initiative! The PSDI team would be very grateful if you would fill out their five-minute [survey on Process Recording Tools](#).

The PSDI Team have received full ethics approval from the University of Southampton Ethics and Research Governance Team (Ergo/FEPS/80887) to run this survey. If you have any questions about this survey please email [Dr Samantha Kanza](#) and [Dr Nicola Knight](#).

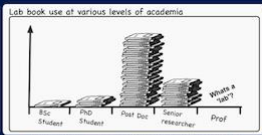
CALLING ALL PHYSICAL SCIENCES RESEARCHERS!



PSDI

PHYSICAL SCIENCES
DATA INFRASTRUCTURE

We are looking to produce a resource on process recording tools for the physical sciences community. Please fill in our 5-minute survey to help us understand what tools you are using and where improvements can be made.



Lab book use at various levels of academia

<https://www.psdi.ac.uk/process-recording-survey>

PSDI webinars

We are launching a new webinar series about PSDI, with different sessions focusing on the various aspects of the PSDI project. This series began on 29 June with an [Introduction to PSDI](#) and the second webinar will be on 27 July looking at our [pathfinder activities](#). You can find more information about the webinars and how to sign up on the PSDI website. Videos will be made available via our website and upcoming YouTube channel.

Recent and upcoming presentations

Future Labs Live 2023

Professor Jeremy Frey and Dr Samantha Kanza both presented at Future Labs Live in Basel, and Samuel Munday exhibited for Data Revival. Jeremy also took part in a panel: *Fail to prepare, prepare to fail – How can we prepare students for work in a lab?* The conference took place from 31 May–2 June. Details on the conference can be found on the [Future Labs Live Conference Page](#), which will include access to all speaker slides where permission was given.

[Electronic lab notebooks are dead; long live electronic lab notebooks](#)

Dr Samantha Kanza (University of Southampton)

[Where do labs go next – here is when VR comes in](#)

Professor Jeremy Frey (University of Southampton)

ACS fall 2023

Members from the PSDI team will be presenting at the ACS Fall conference in August 2023.

We don't talk about semantic web technologies – Dr Samantha Kanza (University of Southampton)

Sunday 13 August 2023, 2.05pm–2.25pm

Division: Division of Chemical Information

Session: Enhance your Data - Smart Ways to Metadata and Knowledge Graphs



Mirror mirror on the wall who is the FAIRest of them all? – Professor Jeremy Frey (University of Southampton)

Monday 14 August 2023, 8.05am–8:40am

Division: Division of Chemical Information

Session: Helping Chemists Manage their Data

Electronic lab notebooks and beyond! The evolution of process recording tools for scientific research – Dr Samantha Kanza (University of Southampton)

Monday 14 August 2023, 10.10am–10.25am

Division: Division of Chemical Information

Session: Helping Chemists Manage their Data

UK Physical Sciences Data Infrastructure (PSDI) initiative – Professor Jeremy Frey (University of Southampton)

Monday 14 August 2023, 2.25pm–2.45pm

Division: Division of Chemical Information

Session: Chemical informatics (R)evolution: Towards Democratisation and Open Science

World FAIR

The IUPAC FAIR Chemistry team have released a sampler version of their Chemistry training 'Cookbook' and are looking for feedback from the community.

The [IUPAC FAIR Chemistry Cookbook](#) is intended to be an open, collaborative, community-focused resource on working with FAIR data in chemistry. FAIR data are findable, accessible, interoperable, and reusable for machine processing, the new paradigm for reporting research data. This cookbook aims to support best practices using chemical data standards through practical, interactive tutorials, provided by the International Union of Pure and Applied Chemistry (IUPAC) through the EU funded WorldFAIR project.



IUPAC FAIR Chemistry Cookbook Sampler

We have released a sampler version of our Chemistry 'Cookbook' and are looking for **your feedback!**

The cookbook aims to support implementation of the FAIR data principles for chemical data by providing practical, interactive tutorials designed for practicing chemists.

Check out the Cookbook sampler at:
<https://iupac.github.io/WFChemCookbook>



In the WorldFAIR project, CODATA (the Committee on Data of the International Science Council) and RDA (the Research Data Alliance), work with a set of 11 disciplinary and cross-disciplinary case studies to advance implementation of the FAIR principles and, in particular, to improve interoperability and reusability of digital research objects, including data. Particular attention is paid to the articulation of an interoperability framework for each case study and research domain.

UKeIG News: eLucidate, and the Prestigious International Strix and Farradane Awards

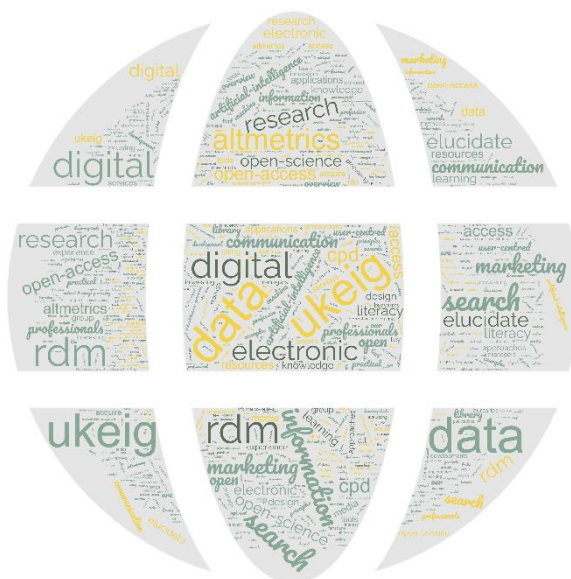
Contribution from Gary Horrocks, UKeiG, CILIP, email: info.ukeig@cilip.org.uk

eLucidate

[eLucidate](#), the open access e-journal of the UK electronic information Group (UKeIG), showcases digital knowledge, expertise, R&D and innovation. Vol. 16, No. 3, Summer 2023 is now available and includes articles on managing information to serve the information user, leadership in information retrieval research, impactful research communications, information landscapes, technological ecosystems, enterprise discoverability, humanities open data conundrums, research marketing and communication, open access monographs, web search analysis for the public good and artificial intelligence.

Contents include:

- The outcome of a major research project that has attempted to map and preserve the rich history of the Institute of Information Scientists. Past IIS Presidents Dr Sandra Ward and Martin White unveil the fruits of a significant collaborative effort, “Evolution and impact: a history of the Institute of Information Scientists 1958–2002”, set against the backdrop of two World Wars and the explosive increase in scientific research and scholarly publications.
- Reports from UKeiG member forums, lectures and events during the pandemic, including: open data practices in the humanities, analysing web searches for public good, defining and developing the enterprise search experience, managing the information landscape, the higher education library technology ecosystem and AI – utopia or dystopia?
- Thomas D. Wilson (Professor Emeritus, University of Sheffield, UK) and UKeiG’s 2020 Jason Farradane winner reflects on “managing information to serve the information user”. He explores the essential link and convergence between information management and information behaviour, asking the fundamental question: how can we manage information effectively if we don’t know how people seek and use information?
- In December 2021, Ian Ruthven, Professor of Information Seeking and Retrieval at the Department of Computer and Information Sciences, University of Strathclyde, presented the seventh annual Strix Memorial Lecture entitled: “Google’s what you use when Alexa doesn’t know the answer, Uncle Ian”. It was inspired by an innocent comment made by his young niece. We explore Ian’s work and his observation that there is so much potential to progress research by placing information seeking and



retrieval within the broader societal and cultural framework of what people or populations want from or do with information.

- Andy Tattersall, Information Specialist, School of Health and Related Research, University of Sheffield, writes about “laying down the tracks for impactful research communications”.

Strix and Farradane awards: 2023 call for nominations

The UK e-information Group (UKeiG) launched its call for nominations for the 2023 Tony Kent Strix and Jason Farradane Awards at its 29 June 2023 Zoom AGM and Members’ Forum.

Tony Kent Strix Award

The Tony Kent Strix Award was inaugurated in 1998 by the Institute of Information Scientists. It is now presented by UKeiG and over its history has been supported by 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). It is given in recognition of an outstanding practical innovation or achievement in the field of information retrieval and search. Nominations must be for a major, sustained or influential achievement that meets one or more of the criteria listed below.

Science

The advancement of our understanding of information retrieval methods, experimentation and evaluation, at either the theoretical or the practical level. The scope includes approaches as diverse as linguistic, probabilistic, fact-checking or artificial intelligence applied to search.

Service Delivery

The development and management of systems, networks or services:

- Enhancement of the mechanisms/technology/standards underpinning information products or services.
- Establishing an innovative information resource or service.
- Innovations leading to improved accessibility/usability of information resources.

Education and organisational infrastructure

The provision of leadership in education, training, community development and/or collaboration to advance information retrieval at local, national or international level.

Last year UKeiG was pleased to announce that the joint winners were:

- Iadh Ounis, Professor of Information Retrieval, School of Computing Science, University of Glasgow
- Dr Ryen White, General Manager and Partner Research Director at Microsoft Research

The judging panel congratulated them on their prolific and significant contributions to information retrieval research and development on multiple fronts, most notably the search experience. Their high impact publication records and scholarly contributions are peerless and international support for their nominations was overwhelming across the information retrieval community. They both presented at the 8th Tony Kent Strix Annual Memorial Lectures on 23 February 2023.

Jason Farradane Award

The Jason Farradane Award is presented in recognition of an outstanding, creative and enterprising contribution to the wider library and information profession. It will be awarded to an individual or a team in recognition of exemplary and innovative practice. This may take the form of a specific project, a piece of research or the development of a service or resource.

Further information and deadline

Further information about UKeiG awards and bursaries is available [here](#), including guidelines for submission. Nominations for both 2023 awards must reach the judges by 6 pm GMT on Friday 29 September 2023.

Please share the news

Everybody is encouraged to raise awareness of these awards with individuals and groups in their networks and help encourage nomination submissions. In the interim, please don't hesitate to contact Gary Horrocks at UKeiG, if you have any queries: info.ukeig@cilip.org.uk.

Other Chemical Information News

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

Search Engines that give users Control over their Results

Searching the web may seem simple for the average user: Just type in some keywords, and the search engine of your choice, often Google, will present you with the most relevant results. However, new search engines challenging that paradigm present interesting alternatives for information professionals.

<https://www.infoday.eu/Articles/Editorial/Featured-Articles/Search-engines-that-give-users-control-over-their-results-157358.aspx>

Source: Information Today Inc

AI Drug Discovery company Insilico Medicine offers free Course on Target Identification

Insilico Medicine has announced that it is releasing a [course on Disease Modelling and Target Identification](#), available for free. Through seven lectures, the course guides participants through the latest insights in disease modelling and target discovery – critical areas in biomedical research that involve using computational and experimental approaches to gain insights into the causes of diseases and identify potential targets for drug development.

<https://www.eurekalert.org/news-releases/988835>

Source: AAAS EurekAlert!

New Digital Platform improves experience of reading RSC eBooks

The Royal Society of Chemistry has unveiled a new publishing platform for eBooks that will enhance users' digital experience. The new platform, created by Silverchair, is the result of months of collaboration and testing, and can be accessed at books.rsc.org. Possessing several improvements upon its predecessor, this service will change the way researchers interact with our eBooks.

The benefits of this upgrade include:

- A more intuitive format and improved discoverability make it easier for readers to find the books they need.

- An improved digital reading experience with better functionality for users while researchers will have more ways to interact with content.
- Enabling individuals to buy digital books directly from the Royal Society of Chemistry for the first time.
- Being able to publish digital versions of all our books on our own platform, including textbooks and popular science texts.

<https://www.rsc.org/news-events/articles/2023/01-january/rsc-ebooks-platform-unveiled/>

Source: RSC News

Daresbury Laboratory opens its doors to the public

On 15 July 2023, the Science and Technology Facilities Council's (STFC) Daresbury Laboratory will once again open its doors to visitors.

<https://www.ukri.org/news/daresbury-laboratory-opens-its-doors-to-the-public/>

Source: UKRI News

Searching the Web for Science: How small mistakes create big Problems

Typos, acronyms and authors' names all present issues when trawling the literature. Can next-generation search engines do better?

<https://www.nature.com/articles/d41586-023-01011-2>

Source: Nature

Elsevier launches new Research Tool

EmBiology is powered by Biology Knowledge Graph, a research tool that visualises a comprehensive landscape of biological relationships, empowering researchers to gain a rapid understanding of disease biology and focus on critical evidence.

<https://www.researchinformation.info/news/elsevier-launches-new-research-tool>

Source: Research Information

Open-Source AI Models for Scientific Advancement

Scientists should switch to open large language models (LLMs) and help others to distribute them.

<https://www.chemistryviews.org/open-source-ai-models-for-scientific-advancement/>

Source: ChemistryViews



ResearchGate integrates with GetFTR

GetFTR has today announced that ResearchGate, the professional network for researchers, has integrated with the service as part of an expanded offering for publishers. Over the last few months, GetFTR has been building, and ResearchGate testing, a new opt-in service that provides a faster way to check entitlement information between publishers, academic discovery services like ResearchGate, and reading platforms.

<https://www.stm-publishing.com/researchgate-integrates-with-getftr/>

Source: STM Publishing News

CAS forms Advisory Board to lead Drug Discovery Innovation in Life Sciences

CAS has announced the formation of its CAS Life Sciences Advisory Board. The board, comprising esteemed thought leaders from renowned global research organisations, aims to advance drug discovery capabilities and workflows and guide the expanded CAS drug discovery journey. With the goal of empowering better decisions in the early stages of drug discovery, the advisory board will align authoritative content with predictive technologies spanning the spectrum of life sciences workflows. This will include understanding disease

progression with biological pathway data and assessing the effectiveness of exploratory drugs with enriched molecular biomarker data. CAS is also recruiting additional members to ensure broad diversity and engagement. CAS has already taken steps to accelerate early-stage drug discovery research with the expansion of content in the CAS Content Collection™ in late 2022, which includes exhaustive biological sequences and bioactivity information. The collaboration with Chemotargets also leverages Clarity® to serve as a technology foundation for rapid development.

<https://www.knowledgespeak.com/news/cas-forms-advisory-board-to-lead-drug-discovery-innovation-in-life-sciences/>

Source: Knowledgespeak

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

RSC launches two new Interdisciplinary Journals focusing on Applied Science

The RSC has launched two new journals centred on innovative applied interfacial and polymer research. These are the RSC's first pair of publications focused solely on applied science breakthroughs. [RSC Applied Interfaces](#) and [RSC Applied Polymers](#) will showcase the latest interdisciplinary research as part of its burgeoning materials portfolio. In order to address some of the major issues affecting the world today, the chemical sciences community can benefit by sharing information about recent studies. The new journals are both Gold Open Access, meaning articles will be freely available online for anyone, anywhere to read. The RSC will also pay for all article processing charges (APCs) until mid-2025, making it free for researchers to publish their work.

<https://www.knowledgespeak.com/news/rsc-launches-two-new-interdisciplinary-journals-focusing-on-applied-science/>

Source: Knowledgespeak

Should Librarians be interested (or not) in Knowledge Management?

In the opinion of Denise Carter, Managing Director of DCision Consult and a long-time information professional, the answer is yes, librarians should be interested in KM, but they should consider why and when, plus what value that interest could deliver for them.

<https://www.infotoday.eu/Articles/Editorial/Featured-Articles/Should-librarians-be-interested-or-not-in-Knowledge-Management-156157.aspx>

Source: Information Today Inc

Innovative new Open Access Agreement signed between PALCI and ACS Publications

PALCI (Partnership for Academic Library Collaboration and Innovation) and ACS Publications are excited to announce the signing of a new open access agreement that provides participating members with full financial support for open access publication in all ACS journals, as well as complete access to ACS Publications' journal content through the end of 2025.

<https://www.stm-publishing.com/innovative-new-open-access-agreement-signed-between-palci-and-acp-publications/>

Source: STM Publishing News

Turning Negative results into Positives

Publishing unsuccessful experiments is more important than ever as we try to train machines in chemistry.

<https://www.chemistryworld.com/opinion/turning-negative-results-into-positives/4017017.article>

Source: *Chemistry World*

RSC and ResearchGate collaborate to enhance accessibility of Open Access Journals

The RSC has signed a new partnership with ResearchGate to expand the reach of its open access portfolio. ResearchGate, a professional network for researchers, will offer an enhanced presence to RSC journals on its new offering, "Journal Home." This collaboration has the potential to accelerate growth in both readership and submissions for RSC's OA journals.

<https://www.knowledgespeak.com/news/rsc-and-researchgate-collaborate-to-enhance-accessibility-of-open-access-journals/>

Source: *Knowledgespeak*

GitHub is Sued, and We May Learn Something About Creative Commons Licensing

By Roy Kaufman: I have had people tell me with doctrinal certainty that Creative Commons licenses allow text and data mining, and insofar as license terms are observed, I agree. The making of copies to perform text and data mining, machine learning, and AI training (collectively "TDM") without additional licensing is authorised for commercial and non-commercial purposes under [CC BY](#), and for non-commercial purposes under [CC BY-NC](#).

<https://www.copyright.com/blog/github-is-sued-and-we-may-learn-something-about-creative-commons-licensing/>

Source: *Copyright Clearance Center*

More than 50 percent of Cambridge Research Papers now Open Access

Over half of Cambridge University Press research articles are now published open access and so freely available to read.

<https://www.cambridge.org/news-and-insights/news/more-than-50-percent-of-cambridge-research-papers-now-open-access>

Source: *CUP*

Biochemical Society opens EDI Survey 2023

The principles of equality, diversity, and inclusion (EDI) are central to the Biochemical Society's activities and we're pleased to open this year's EDI Survey, which is designed to help us better understand our community and inform our working practices.

<https://www.stm-publishing.com/biochemical-society-opens-edi-survey-2023/>

Source: *STM Publishing News*

The Importance of Semantic Search Capabilities for Life Sciences Organisations

To eliminate the noise and provide relevant search results, information solutions must go beyond simple keyword matching and to use search engines and algorithms that link concepts, topics, and associations to form a deeper understanding of a user's intent.

<https://www.copyright.com/blog/the-importance-of-semantic-search-capabilities-for-life-sciences-orgs/>

Source: *Copyright Clearance Center*

Double recognition for SciBite in Bio-IT World Innovative Practices Awards 2023

SciBite, an Elsevier company and leading provider of semantic software solutions for the life sciences industry, has won an Innovative Practices Award at Bio-IT World 2023 for its collaboration with the City of Hope, a

world-renowned cancer research and treatment center, on the Precision Oncology Software Environment Interoperable Data Ontologies Network (POSEIDON) project.

<https://www.stm-publishing.com/double-recognition-for-scibite-in-bio-it-world-innovative-practices-awards-2023/>

Source: *STM Publishing News*

Improving Global Research Discoverability and Visibility through ScienceDirect

[ScienceDirect](https://www.science.org), the Elsevier platform renowned for its vast collection of scientific, technical, and medical research, has released its report titled '[Expanded reach on a trusted platform](#).' This report delves into the initial outcomes of ScienceDirect's pilot initiative, aimed at facilitating researchers in finding and accessing articles from multiple publishing partners alongside Elsevier journals. Since January 2022, leading publishers including the ACS, the RSC, Taylor & Francis, and Wiley have made select research articles from their Organic Chemistry and Transportation domains discoverable on ScienceDirect. The International Union of Crystallography's IUCr Journals joined the initiative in August 2022.

<https://www.knowledgespeak.com/news/improving-global-research-discoverability-and-visibility-through-sciencedirect/>

Source: *Knowledgespeak*

Scientists unveil plan to create Biocomputers powered by Human Brain Cells

Despite AI's impressive track record, its computational power pales in comparison with that of the human brain. Scientists today unveil a revolutionary path to drive computing forward: organoid intelligence (OI), where lab-grown brain organoids serve as biological hardware. "This new field of biocomputing promises unprecedented advances in computing speed, processing power, data efficiency, and storage capabilities – all with lower energy needs," say the authors in an article published in *Frontiers in Science*.

<https://blog.frontiersin.org/2023/02/28/brain-organoids-intelligence-biocomputing-hartung/>

Source: *Frontiers Science News*

5 Tips to Establish Collaborative Relationships Between Info Pros & Data Scientists

Mary Ellen Bates shares the inside story of an information scientist working among a team of data professionals – with his tips for collaborative success.

<https://www.copyright.com/blog/5-tips-to-establish-collaborative-relationships-between-info-pros-data-scientists/>

Source: *Copyright Clearance Center*

EMBO Press transitions to fully Open Access

Scientific publisher [EMBO Press](https://www.embo.org) has announced that from January 1, 2024, all its journals, including *The EMBO Journal* and *EMBO Reports*, will switch from encouraging Open Access publication to applying Open Access to all published papers. The move aims to promote Open Science for individual researchers, institutions, and funders worldwide.

<https://www.knowledgespeak.com/news/embo-press-transitions-to-fully-open-access/>

Source: *Knowledgespeak*

Jisc announces partnership with Open Access Publisher Copernicus

Jisc has announced a new agreement with Copernicus Publications, a fully open access, not-for-profit publisher, whose portfolio of journals covers engineering, geosciences, humanities, and life sciences.

<https://www.stm-publishing.com/jisc-announces-partnership-with-open-access-publisher-copernicus/>

Source: *STM Publishing News*

Can Info Pros Impact a Global Pharma Company's Pipeline? Yes. Here's How

The days when a pharma company could base its future on discovering a few new drugs are gone, so to grow a pipeline of new drugs, companies explore new therapy areas and technologies.

<https://www.copyright.com/blog/info-pros-impact-global-pharma-company-pipeline/>

Source: Copyright Clearance Center

Digital Science acquires Knowledge Graph and Decision Intelligence software company Metaphacts

Digital Science has completed the acquisition of metaphacts, which has become the newest member of the Digital Science family. Based in Germany, metaphacts is a knowledge graph and decision intelligence software company. Its main product metaphactory is a platform that supports customers in accelerating their adoption of knowledge graphs and driving knowledge democratisation.

<https://www.stm-publishing.com/digital-science-acquires-knowledge-graph-and-decision-intelligence-software-company-metaphacts/>

Source: STM Publishing News

How long until Papers list an AI as an Author?

Artificial intelligence is superior to humans at numerous tasks, but it is still vulnerable to human biases.

<https://www.chemistryworld.com/opinion/how-long-until-papers-list-an-ai-as-an-author/4016762.article>

Source: Chemistry World

Nature announces support for authors from over 70 countries to publish Open Access

Primary research from authors from over 70 countries classified by the World Bank as low-income (LIC) or lower-middle-income economies (LMICs) accepted for publication in either Nature or one of the Nature research journals (e.g. Nature Chemistry, Nature Sustainability) can now be published Gold open access at no cost.

<https://www.stm-publishing.com/nature-announces-support-for-authors-from-over-70-countries-to-publish-open-access/>

Source: STM Publishing News

The Importance of Data for AI

In Thomas Kuhn's work, paradigms are characterised by "universally recognised scientific achievements that for a time provide model problems and solutions to a community of practitioners."

<https://www.copyright.com/blog/the-importance-of-data-for-ai/>

Source: Copyright Clearance Center

EBSCO Information Services and eReserve partner to Streamline Resource Management Workflows with EBSCO FOLIO

EBSCO Information Services (EBSCO), a leading provider of research databases, e-journals, and discovery services, has joined forces with eReserve to enhance institutions' resource management workflows through the implementation of EBSCO FOLIO Services. This partnership aims to streamline resource management processes, including copyright control, and promote collaboration among institutions, all while delivering an enhanced student experience.

<https://www.knowledgespeak.com/news/ebSCO-information-services-and-ereserve-partner-to-streamline-resource-management-workflows-with-ebSCO-folio/>

Source: Knowledgespeak

AI predicts Enzyme Function better than Leading Tools

A new artificial intelligence tool can predict the functions of enzymes based on their amino acid sequences, even when the enzymes are unstudied or poorly understood. The researchers said the AI tool, dubbed CLEAN, outperforms the leading state-of-the-art tools in accuracy, reliability and sensitivity. Better understanding of enzymes and their functions would be a boon for research in genomics, chemistry, industrial materials, medicine, pharmaceuticals and more.

<https://www.eurekalert.org/news-releases/984108>

Source: AAAS EurekAlert!

Royal Society and ResearchGate collaboration: successful trial phase leads to extended scope of Partnership

[ResearchGate](https://www.researchgate.net), the professional network for researchers, and The Royal Society, the UK's national science academy and the oldest scientific academy in continuous existence, have expanded the partnership that sees all articles from the Royal Society's ten-journal portfolio syndicated directly to ResearchGate.

<https://www.knowledgespeak.com/news/royal-society-and-researchgate-collaboration-successful-trial-phase-leads-to-extended-scope-of-partnership/>

Source: Knowledgespeak

Community voices concerns over ACS Magazine

Board members' open letter warns of 'detrimental changes' after restructure at C&EN.

<https://www.chemistryworld.com/news/community-voices-concerns-over-american-chemical-society-magazine/4016716.article>

Source: Chemistry World

New study shows ChatGPT has the potential to Revolutionise Medical Education

A [recent study](#) published in JMIR Medical Education, a leading open-access journal on digital medical education, has evaluated the potential of ChatGPT as a tool for medical education. The study, conducted by researchers from Yale University School of Medicine's Section for Biomedical Informatics and Data Science and University College Dublin, aimed to test the performance of ChatGPT and previous-generation large language models on the medical question-answering problem as part of the United States Medical Licensing Examination (USMLE) Step 1 and Step 2 exams. The researchers tested the models on study aids commonly used by medical students, including multiple-choice questions and the National Board of Medical Examiners (NBME) test sample questions. ChatGPT outperformed previous-generation models and was capable of correctly answering up to over 60% of questions, which is comparable to a passing score for a third-year medical student.

<https://www.knowledgespeak.com/news/new-study-shows-chatgpt-has-the-potential-to-revolutionize-medical-education/>

Source: Knowledgespeak

New "AI Scientist" combines Theory and Data to discover Scientific Equations

The system demonstrated its chops on Kepler's third law of planetary motion, Einstein's relativistic time-dilation law, and Langmuir's equation of gas adsorption.

<https://www.eurekalert.org/news-releases/985511>

Source: AAAS EurekAlert!

To diversify the Scientific Workforce, Postdoc Recruitment needs a rethink

Biased hiring practices are limiting efforts to attract and retain researchers of colour.

<https://www.nature.com/articles/d41586-023-01740-4>

Source: Nature

Accessing and Analysing Relevant Content in Today's Information Chaos

Four core requirements are necessary for R&D teams to realise the full potential of streamlined search, usage, and insight generation. We outline them in this white paper.

https://www.copyright.com/wp-content/uploads/2023/01/CCCRD_Information-Chaos-WP.pdf

Source: Copyright Clearance Center

Using AI, Scientists find a Drug that could combat Drug-Resistant Infections

The machine-learning algorithm identified a compound that kills *Acinetobacter baumannii*, a bacterium that lurks in many hospital settings. Using an artificial intelligence algorithm, researchers at MIT and McMaster University have identified a new antibiotic that can kill a type of bacteria that is responsible for many drug-resistant infections.

<https://www.eurekalert.org/news-releases/990325>

Source: AAAS EurekAlert!

New Study reveals how Frontier Research spurs Patented Inventions

A new analysis finds that more than 40% of [ERC](#)-funded projects generated research that was subsequently cited in patents. The results show how curiosity-driven frontier research enables technological development.

<https://www.knowledgespeak.com/news/new-study-reveals-how-frontier-research-spurs-patented-inventions/>

Source: Knowledgespeak

APL Machine Learning publishes Inaugural Issue

[AIP Publishing](#) has announced the first published articles of [APL Machine Learning](#), one of the newest additions to its growing portfolio of fully Open Access journals. The new journal covers the use of machine learning (ML) and artificial intelligence to aid physicists, material scientists, engineers, chemists, and biologists in advancing scientific discovery as well as advances in materials, devices, and systems for the development of future ML technologies. With the editorial support of that deep bench of luminaries, APL Machine Learning seeks to inform and influence the rapidly growing field of machine learning — a nascent scientific arena with seemingly limitless potential.

<https://www.knowledgespeak.com/news/apl-machine-learning-publishes-inaugural-issue/>

Source: Knowledgespeak

Proposed EU Data Laws leave Researchers out in the cold

Some scientists say the European Commission's Data Act would favour businesses in its aim to expand access rights to big data, and fear that publicly funded science will suffer.

<https://www.nature.com/articles/d41586-023-01572-2>

Source: Nature

Clarivate adds Preprint Citation Index to the Web of Science

Clarivate has added the [Preprint Citation Index](#)[™] to the Web of Science[™] platform. Researchers can now locate and link to preprints alongside other trusted content in the database, to streamline the research process and help make meaningful connections faster.

<https://www.knowledgespeak.com/news/clarivate-adds-preprint-citation-index-to-the-web-of-science/>

Source: Knowledgespeak

Science Journals integrate Dryad to simplify data deposition and strengthen scientific reproducibility

The Science family journals have announced a partnership with the nonprofit data repository Dryad that simplifies the process by which authors deposit data underlying new work – a critical step to facilitating data's

routine reuse. The partnership is yet another step taken by the Science journals to ensure data the scientific community requires to verify, replicate and reanalyze new research is openly available.

<https://www.stm-publishing.com/science-journals-integrate-dryad-to-simplify-data-deposition-and-strengthen-scientific-reproducibility/>

Source: *STM Publishing News*

Robots in the Lab – learn how Automation can streamline your Workflow

The future of lab automation is promising. Join us to find out answers to the most important questions, and to contribute your knowledge and experience to the discussion.

<https://www.chemistryworld.com/webinars/robots-in-the-lab-learn-how-automation-can-streamline-your-workflow/4016797.article>

Source: *Chemistry World*

Chemists have a new Tool to Predict 3D Structures of F-Block Organometallics

An application called Architector could help scientists separate valuable metals from nuclear waste.

<https://cen.acs.org/physical-chemistry/computational-chemistry/Chemists-new-tool-predict-3D/101/i18>

Source: *Chemical & Engineering News*

Designs for Green Ammonia Plant become reality

Science and Technology Facilities Council (STFC) researchers are building a small-scale plant to generate ammonia using only renewable energy sources.

<https://www.ukri.org/news/designs-for-green-ammonia-plant-become-reality/>

Source: *UKRI News*

Knowledge Unlatched celebrates ten years of Open Access Publishing; launches tenth funding round - May 11, 2023

[Knowledge Unlatched](https://www.knowledgespeak.com/news/knowledge-unlatched-celebrates-ten-years-of-open-access-publishing-launches-tenth-funding-round/) (KU), the international initiative for Open Access (OA), is celebrating a significant milestone in 2023 as it marks ten years of OA publishing. To commemorate this achievement, KU is launching its tenth round of funding, which has to date supported the publication of over 4,000 scholarly books and 50 journals in a range of humanities, sciences, and social sciences. Knowledge Unlatched is part of Wiley Partner Solutions.

<https://www.knowledgespeak.com/news/knowledge-unlatched-celebrates-ten-years-of-open-access-publishing-launches-tenth-funding-round/>

Source: *Knowledgespeak*

Collaboration aims to Advance the Discovery of Novel Cancer Drugs

A newly announced collaboration between Oxford Drug Design and PhoreMost aims to accelerate a targeted protein degradation discovery programme for novel cancer therapeutics. Oxford Drug Design is a biotechnology company with expertise in computer-aided drug design, while PhoreMost, a UK-based developer of phenotypic screening platform.

<https://www.scientific-computing.com/news/collaboration-aims-advance-discovery-novel-cancer-drugs>

Source: *Scientific Computing World*

RSC and the Institute of Process Engineering, Chinese Academy of Sciences, partner to Publish Gold OA Journal - Industrial Chemistry & Materials

The RSC and the [IPE, Chinese Academy of Sciences](https://www.ipe.cas.ac.cn/), have partnered to publish a new gold open access journal entitled [Industrial Chemistry & Materials \(ICM\)](https://www.rsc.org/journals-books-and-databases/titles/industrial-chemistry-and-materials). ICM was created to advance technology innovations that

address the pressing issue of sustainability confronting mankind, which publishes significant research in the cross areas of industrial chemistry and materials

<https://www.knowledgespeak.com/news/royal-society-of-chemistry-and-the-institute-of-process-engineering-ipe-chinese-academy-of-sciences-partner-to-publish-gold-oa-journal-industrial-chemistry-materials-icm/>

Source: *Knowledgespeak*

Machine-Learning Model Predicts Enzyme-Substrate Pairs

Enzymes catalyse one or more specific reactions. Mapping the relationships between an enzyme and its possible substrates can be useful, e.g., in drug research. However, determining these relationships experimentally can be time-consuming and costly. Automated computational methods, for example, based on machine learning, could be useful in this context.

<https://www.chemistryviews.org/machine-learning-model-predicts-enzyme-substrate-pairs/>

Source: *ChemistryViews*

Collaboration aims to Integrate AI Drug Discovery

A collaboration between two laboratory software companies aims to integrate synthesis prediction with functionality to design and optimise small molecule discovery. Optibrium, a developer of software and AI for small molecule discovery, and PostEra, a biotechnology company developing machine learning (ML) technology for preclinical discovery, announced a collaboration to integrate Optibrium's StarDrop platform with PostEra's Manifold software.

<https://www.scientific-computing.com/news/collaboration-aims-integrate-ai-drug-discovery>

Source: *Scientific Computing World*

New Chemical Compound demonstrates potential in Nerve Regeneration

Research has identified a new compound that can stimulate nerve regeneration after injury and protect cardiac tissue from the sort of damage seen in heart attack.

<https://www.ukri.org/news/new-chemical-compound-demonstrates-potential-in-nerve-regeneration/>

Source: *UKRI News*

UKRI Updates Guidance for Open Access Policy

From 1 January 2024, UK Research and Innovation (UKRI)'s new open access policy will apply to monographs, book chapters and edited collections. The open access policy applied since April 2022 for peer-reviewed research articles acknowledging UKRI funding. The policy aims to ensure that findings from research UKRI funds with public money can be accessed and built on by the research and innovation community and wider society.

<https://www.stm-publishing.com/ukri-updates-guidance-for-open-access-policy/>

Source: *STM Publishing News*

Machine Learning improves the Success of Genome Editing

A new tool to predict the chances of successfully inserting a gene-edited sequence of DNA into the genome of a cell, using a technique known as prime editing, has been developed by researchers at the Wellcome Sanger Institute. An evolution of CRISPR-Cas9 gene editing technology, prime editing has enormous potential to treat genetic disease in humans, from cancer to cystic fibrosis. But thus far, the factors determining the success of edits are not well understood.

<https://www.scientific-computing.com/news/machine-learning-improves-success-genome-editing>

Source: *Scientific Computing World*

Springer Nature acquires TooWrite

Springer Nature has announced the acquisition of innovative digital writing aid, TooWrite. This is the latest addition to the publisher's growing portfolio of digital solutions for academics designed to help improve their working lives. Developed by researchers for researchers, the TooWrite platform streamlines and simplifies scientific writing by guiding researchers through the process as if they were answering a questionnaire. In addition, expert how-to guides are attached to each question, supporting researchers as if they had an editor by their side. By structuring it in this step-by-step way, researchers' time is freed up by making the writing process more efficient. Currently available to support the writing of abstracts (beta prototype), it is being developed to support the writing of full research papers.

<https://www.knowledgespeak.com/news/springer-nature-acquires-toowrite-continues-its-focus-on-tailored-solutions-for-academics/>

Source: *Knowledgespeak*

Detecting Plagiarism in AI-Generated Text

ChatGPT has taken the world by storm and excited the imaginations of its millions of users, who find innovative applications from which ChatGPT learns. Educators and librarians worry that text generated by ChatGPT and similar programs will encourage plagiarism and make it more difficult to detect.

<https://www.infotoday.eu/Articles/News/Featured-News/Detecting-plagiarism-in-AI-generated-text-156504.aspx>

Source: *Information Today Inc*

Insilico Medicine-led Study combines Quantum Computing and Generative AI for Drug Discovery

Insilico Medicine ("Insilico") has announced that it combined two rapidly developing technologies, quantum computing and generative AI, to explore lead candidate discovery in drug development and successfully demonstrated the potential advantages of quantum generative adversarial networks in generative chemistry.

<https://www.eurekalert.org/news-releases/989823>

Source: *AAAS EurekAlert!*

Improving Computational Calculations in Heterogeneous Catalysis through AI

GAME-Net makes fast chemical predictions on a laptop, replacing days of supercomputer simulations using DFT.

<https://www.chemistryviews.org/improving-computational-calculations-in-heterogeneous-catalysis-through-ai/>

Source: *ChemistryViews*

Preserving Websites to ensure Perpetual Access

The Internet Archive's Wayback Machine lets you see what websites used to look like, but it can only do that if volunteers actively save those sites and pages. Gary Price explains how to do that using the [Save Page Now](#) (SPN) functionality.

<https://www.infotoday.eu/Articles/Editorial/Featured-Articles/Preserving-websites-to-ensure-perpetual-access-155188.aspx>

Source: *Information Today Inc*

ChatGPT & Information Integrity

Launched in November 2022, ChatGPT-3 received its language training from 300 billion words on the internet, all written in books, journal articles, and Wikipedia.

<https://www.copyright.com/blog/chatgpt-information-integrity/>

Source: *Copyright Clearance Center*

AI Software can provide ‘Roadmap’ for Biological Discoveries

Predicting a protein’s location within a cell can help researchers unlock a plethora of biological information that’s critical for developing future scientific discoveries related to drug development and treating diseases like epilepsy. That’s because proteins are the body’s “workhorses,” largely responsible for most cellular functions. Recently, Dong Xu, Curators Distinguished Professor in the Department of Electrical Engineering and Computer Science at the University of Missouri, and colleagues updated their protein localisation prediction model, MULocDeep, with the ability to provide more targeted predictions, including specific models for animals, humans and plants.

<https://www.eurekalert.org/news-releases/991354>

Source: AAAS EurekAlert!

Expert Intelligence Launches AI App Builder for Mass Spectrometry Analysis

A new approach to building AI with an automated AI App Builder for scientists to remove noise and complexity in Mass Spectrometry analysis.

<https://www.biospace.com/article/expert-intelligence-launches-ai-app-builder-for-mass-spectrometry-analysis/>

Source: BioSpace

NSF invests in AI Research Infrastructure

The US National Science Foundation has announced a \$16 million investment to support shared research infrastructure. The awards are part of NSF’s Computer and Information Science and Engineering Community Research Infrastructure – or CCRI – program, which seeks to create, enhance and democratise access to research infrastructures, focusing on scientific agendas across computer and information science and engineering.

<https://www.scientific-computing.com/news/nsf-invests-ai-research-infrastructure>

Source: Scientific Computing World

Future Science Group launches new Online Community for AI in Healthcare: Future Medicine AI Hub

FSG has launched a new online community, [Future Medicine AI \(FMAI\) Hub](#), to bring together professionals with an interest in the application of artificial intelligence (AI) in healthcare. The new site, which is free to registered users, aims to bridge the gap between clinicians, technologists, and researchers, by providing access to the latest developments in AI research, insights from experts, and information on how AI will impact the future of healthcare.

<https://www.knowledgespeak.com/news/future-science-group-launches-new-online-community-for-ai-in-healthcare-future-medicine-ai-hub/>

Source: Knowledgespeak

Abstracts written by ChatGPT fool Scientists

Researchers cannot always differentiate between AI-generated and original abstracts.

<https://www.nature.com/articles/d41586-023-00056-7>

Source: Nature

UK Research Integrity Report ‘Falls Short’

Document ‘leaves the role of some parts of the research community unexamined’.

<https://www.researchinformation.info/news/uk-research-integrity-report-falls-short>

Source: Research Information

Transformative Agreements playing vital role in OA Transition

New data from Springer Nature shows the vital role transformative agreements (TAs) are playing in driving the global transition to open access (OA). They are the main reason OA is increasing in Springer Nature's hybrid journals, delivering OA equity across all academic disciplines.

<https://www.researchinformation.info/news/transformative-agreements-playing-vital-role-oa-transition>

Source: *Research Information*

Using AI to create better, more potent Medicines

While it can take years for the pharmaceutical industry to create medicines capable of treating or curing human disease, a new study suggests that using generative artificial intelligence could vastly accelerate the drug-development process.

<https://www.eurekalert.org/news-releases/990868>

Source: *AAAS EurekAlert!*

ACS Publications reaches major 'Read and Publish' milestone, expanding Open Access offerings

The Publications Division of the American Chemical Society (ACS) is excited to announce that as of March 29, more than 1,000 academic institutions are supported by a "read and publish" agreement with the Society. These agreements afford members of participating organisations unlimited access to all ACS journal content and a more seamless avenue for open access publishing.

<https://www.stm-publishing.com/acs-publications-reaches-major-read-and-publish-milestone-expanding-open-access-offerings/>

Source: *STM Publishing News*

SciBite unveils Workbench, makes data curation simple and intuitive

[SciBite](https://www.knowledgespeak.com/news/scibite-unveils-workbench-makes-data-curation-simple-and-intuitive/), an Elsevier company that offers semantic technology solutions, has launched a new tool called [Workbench](https://www.knowledgespeak.com/news/scibite-unveils-workbench-makes-data-curation-simple-and-intuitive/) that simplifies the process of curating data to terminology and ontology standards. This new tool can help organisations implement a FAIR approach to data management, making data Findable, Accessible, Interoperable, and Reusable.

<https://www.knowledgespeak.com/news/scibite-unveils-workbench-makes-data-curation-simple-and-intuitive/>

Source: *Knowledgespeak*

Info Pros in a Data-Driven Enterprise

Data scientists often work closely with librarians. The two roles are complementary, and organisations can benefit from aligning the positions strategically, says Mary Ellen Bates.

<https://www.copyright.com/blog/info-pros-in-a-data-driven-enterprise/>

Source: *Copyright Clearance Center*

The Robots Revolutionising Chemistry

Researchers working with automated systems are pushing the boundaries of what chemists can achieve in the lab, reports James Mitchell Crow.

<https://www.chemistryworld.com/features/the-robots-revolutionising-chemistry/4016798.article>

Source: *Chemistry World*

cOAlition S supports efforts to improve Copyright framework for research

Copyright retention has been at the core of Plan S since its inception. Its first principle is that authors or their institutions retain copyright to their work. The Rights Retention Strategy aims to ensure that authors retain

their rights and comply with funders' mandates of immediate open access under open licenses while publishing in the venue of their choice.

<https://www.stm-publishing.com/coalition-s-supports-efforts-to-improve-copyright-framework-for-research/>

Source: *STM Publishing News*

Cambridge launches AI Research Ethics Policy

Rules are set out in the first AI ethics policy from the organisation and apply to research papers, books and other scholarly works.

<https://www.researchinformation.info/news/cambridge-launches-ai-research-ethics-policy>

Source: *Research Information*

Machine Learning helps Researchers separate Compostable from Conventional Plastic Waste with 'very high' accuracy

Compostable plastics are on the rise. Despite their merits, these products, including wrappers and packaging, can contaminate conventional plastic waste in the recycling process. Now, scientists used highly sensitive imaging techniques and developed machine learning methods that can identify compostable plastics among conventional types.

<https://blog.frontiersin.org/2023/03/14/separating-compostable-conventional-plastic-waste/>

Source: *Frontiers Science News*

Taylor & Francis Becomes the Newest Supporter of Open Pharma

Open Pharma and Taylor & Francis are delighted to announce that Taylor & Francis has become the latest official Supporter of the Open Pharma initiative. Open Pharma aims to drive positive change in the communication of research funded by the pharmaceutical industry, with workstreams focused on improving the transparency, accountability, accessibility and discoverability of published research.

<https://www.stm-publishing.com/taylor-francis-becomes-the-newest-supporter-of-open-pharma/>

Source: *STM Publishing News*

FASEB Journals provide Researchers with SciScore tool to improve Rigor and Reproducibility

[FASEB](https://www.knowledgespeak.com/news/faseb-journals-provide-researchers-with-sciscore-tool-to-improve-rigor-and-reproducibility/) has a new tool available for researchers to improve the rigor and reproducibility of science submitted to FASEB journals. An automated tool, SciScore, is now integrated into the journals' submission system, and provides key recommendations and practical steps researchers can take to improve the rigor and reproducibility of their reported science.

<https://www.knowledgespeak.com/news/faseb-journals-provide-researchers-with-sciscore-tool-to-improve-rigor-and-reproducibility/>

Source: *Knowledgespeak*

UK Biobank in £127.6m move to Manchester Science Park

UK Biobank, the large-scale biomedical database, is set to receive £127.6m to fund a move to purpose-built facility at Bruntwood SciTech's Manchester Science Park.

<https://www.ukri.org/news/uk-biobank-in-127-6m-move-to-manchester-science-park/>

Source: *UKRI News*

Graduating in Chemistry as a Blind Person

Despite being blind from birth, B. Tschulnigg studied chemistry and graduated with a bachelor's degree.

<https://www.chemistryviews.org/graduating-in-chemistry-as-a-blind-person/>

Source: *ChemistryViews*

PubMed Central to Index Open Research Europe

PubMed Central is one of the leading digital archives for biomedical and life science journal literature. The U.S. National Library of Medicine, with over eight million full-text records in the archive, maintains PMC. For inclusion in PMC, all publications must meet a stringent set of scientific and editorial quality requirements. Similarly, publishers and platforms must also meet technical quality and infrastructure requirements. As a result, this is another major step for Open Research Europe, following indexing in Scopus, ERIH Plus, Inspec, and others last year. It is also a major milestone for Horizon-funded biomedical and life science researchers. Including Open Research Europe in PMC will significantly increase the credibility and visibility of the work published via the platform, helping researchers reach a greater audience and see a greater potential impact.

<https://www.knowledgespeak.com/news/pubmed-central-to-index-open-research-europe/>

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 software tool.

<https://www.copyright.com/blog/10-questions-to-ask-when-searching-for-a-corporate-literature-management-solution/>

Source: Copyright Clearance Center

Artificial Intelligence for Culture Medium Optimisation

Artificial intelligence facilitates the optimisation of cell culture media. A team of researchers from the University of Tsukuba devised a machine learning-based method for medium optimisation to improve the cellular activity.

<https://www.eurekalert.org/news-releases/990931>

Source: AAAS EurekAlert!

Cambridge Open Access fees waived for over 100 Countries

Academics from 107 low- and middle-income countries will be able to publish their research open access in Cambridge journals at no cost to them. The Cambridge Open Equity Initiative allows scholars to reach the widest possible audience by making open access a free option for them.

<https://www.stm-publishing.com/cambridge-open-access-fees-waived-for-over-100-countries/>

Source: STM Publishing News

£1m for our new home in Leeds

Plans to make a new home for the Library in the heart of Leeds have taken another step forward.

<https://www.bl.uk/news/2023/january/1m-for-our-new-home-in-leeds>

Source: British Library