RSC INTEREST GROUP
CHEMICAL INFORMATION
AND COMPUTER APPLICATIONS

WINTER 2024-25

DISTILLATE

Bringing you news, features and reports about cheminformatics, computational chemistry, chemical information & data, all in one place

Open-Source Cheminformatics Toolkits
See full article on p. 39-42

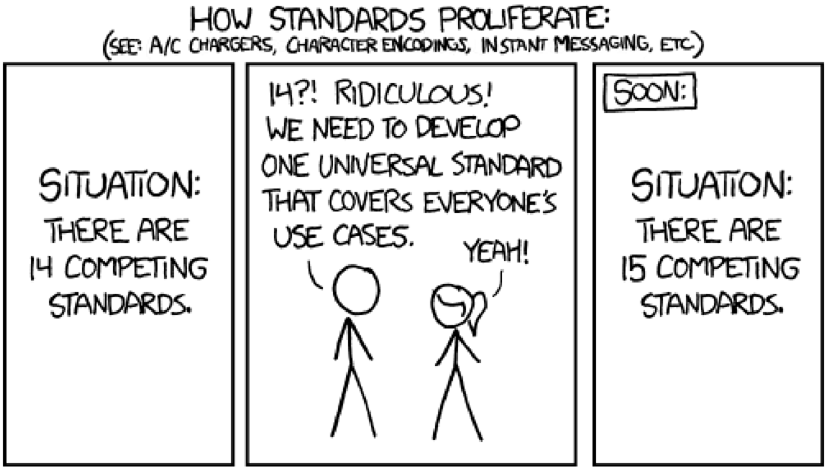
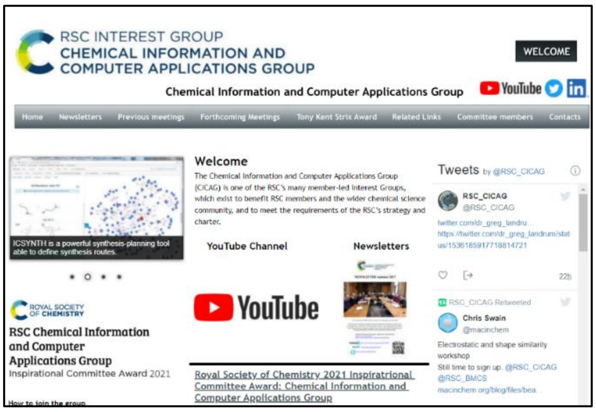

HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC.)
SITUATION: THERE ARE 14 COMPETING STANDARDS.
14?! RIDICULOUS! WE NEED TO DEVELOP ONE UNIVERSAL STANDARD THAT COVERS EVERYONE'S USE CASES.
YEAH!
SOON:
SITUATION: THERE ARE 15 COMPETING STANDARDS.

Image credit XKCD

CICAG Websites and Social Media



RSC INTEREST GROUP
CHEMICAL INFORMATION AND
COMPUTER APPLICATIONS GROUP

Chemical Information and Computer Applications Group

ICSYNTH is a powerful synthesis-planning tool able to define synthesis routes.

ROYAL SOCIETY OF CHEMISTRY

RSC Chemical Information and Computer Applications Group
Inspirational Committee Award 2021

How to join the group

Welcome

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.

YouTube Channel

Newsletters

Royal Society of Chemistry 2021 Inspiratrional Committee Award: Chemical Information and Computer Applications Group

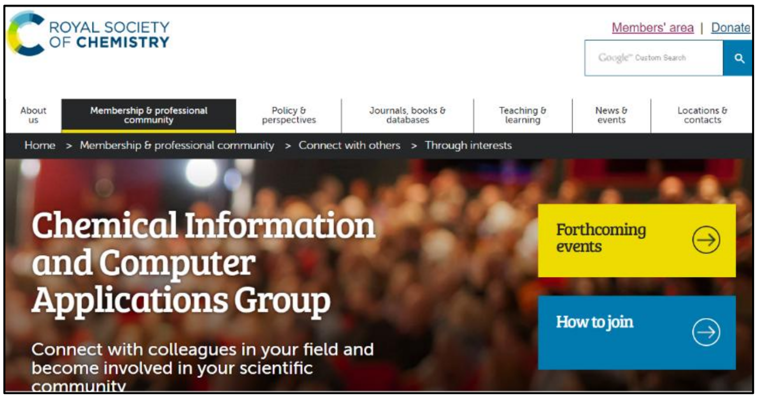
Tweets by @RSC_CICAG

RSC_CICAG
@RSC_CICAG
twitter.com/dr_greg_landru...
https://twitter.com/dr_greg_landrum/status/1536185917718814721

22h

RSC_CICAG Retweeted

Chris Swain
@macinchem
Electrostatic and shape similarity workshop
Still time to sign up. @RSC_CICAG @RSC_BMCS
macinchem.org/blog/files/bea...



ROYAL SOCIETY OF CHEMISTRY

Home > Membership & professional community > Connect with others > Through interests

Chemical Information and Computer Applications Group

Connect with colleagues in your field and become involved in your scientific community

http://www.rsccicag.org
http://www.rsc.org/CICAG

rsc-cicag page
rsc-cicag group

@RSC_CICAG

@rsc-cicag.bsky.social

@rsccicag

@RSCCICAG

Contents

Contributions to the CICAG Distillate are welcome from all sources – please contact the editor
Dr Helen Cooke FRSC: email helen.cooke100@gmail.com

Introducing the *CICAG Distillate*

Contribution from Helen Cooke, CICAG Distillate editor, email: helen.cooke100@gmail.com

RSC INTEREST GROUP
CHEMICAL INFORMATION
AND COMPUTER APPLICATIONS

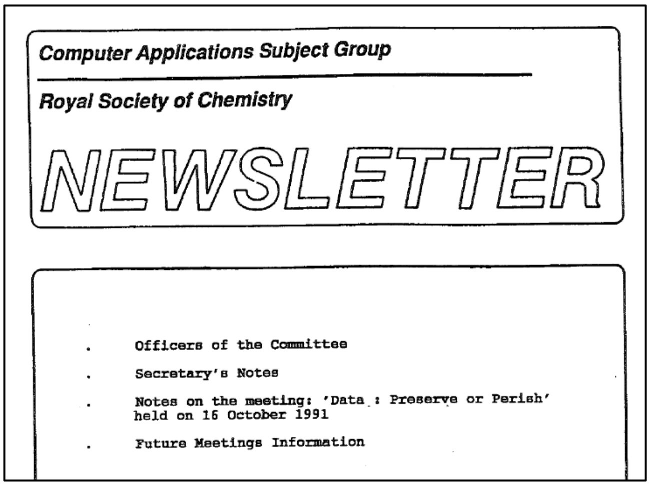
WINTER 2024-25

DISTILLATE

Bringing you news, features and reports about cheminformatics, computational chemistry, chemical information & data, all in one place

Earlier in 2024 the CICAG Committee decided that it was time for a change of title for the *CICAG Newsletter*, which had its origins in the 1990s (when CICAG was two groups, the 'Chemical Information Group' and the 'Computer Applications Subject Group').

When the two groups merged in 2006 and became the Chemical Information and Computer Applications Group, a single newsletter representing both preceding groups was started, going from strength to strength over the years as the *CICAG Newsletter*.

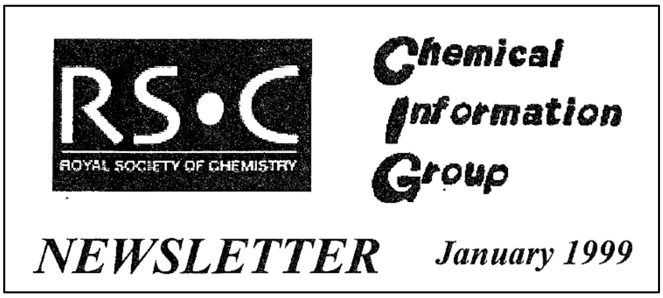


Computer Applications Subject Group

Royal Society of Chemistry

NEWSLETTER

- Officers of the Committee
- Secretary's Notes
- Notes on the meeting: 'Data : Preserve or Perish' held on 16 October 1991
- Future Meetings Information



RS•C ROYAL SOCIETY OF CHEMISTRY

Chemical Information Group

NEWSLETTER January 1999

In recent years the *Newsletter* has increased in size significantly. Recently, as well as news and meeting announcements, the *Newsletter* has included a more diverse range of articles, including reviews, in depth meeting reports, opinion pieces, and more, and it has become much longer than what would normally be thought of as a newsletter.

So the CICAG Committee decided to consider possible new titles for the publication, and after a democratic vote by CICAG Committee members on a range of possibilities, settled on the *CICAG Distillate* (the new cover page was kindly designed by Samantha Pearman-Kanza). It was felt that this better encapsulated the range of content reported, distilled into manageable articles which reflect the scope and activities of CICAG.

We hope you agree!

- -

Chemical Information and Computer Applications Group Chair's Report

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

You will have noticed the that the *CICAG Newsletter* has a new name, the *CICAG Distillate* and you can read more about the change in the article above.

We're very pleased that the membership of CICAG continues to grow and we currently have 823 members. The rise in membership is primarily due to the increasing importance of cheminformatics and computational chemistry, and the continued hard work of the CICAG Committee who organise conferences, workshops, webinars and reach out to the chemistry community. This is all undertaken voluntarily by Committee members in their own time. As Chair I'm enormously proud of their efforts.

In the second half of the year CICAG have been involved in organising and supporting several meetings:

- Molecular Simulation and Free Energies 14 June 2024
- 7th RSC-BMCS / RSC-CICAG Artificial Intelligence in Chemistry 16-18 September 2024
- Python Workshop for ChEMBL anniversary/EUbOPEN community event, 1-2 October 2024
- Hot topics: Robotics and Automation event 21 November 2024

Reports for some of the above are in this issue of the *Distillate*.

We are currently planning the programme of meetings for 2025 and beyond. We welcome suggestions from CICAG members.

CICAG is pleased to be able to provide bursaries to enable attendance at meetings which we organise. In 2024, CICAG awarded eight bursaries. Between them, CICAG and BMCS provided 11 bursaries for the 7th AI in Chemistry meeting, the most we have jointly given to any single event. With input from the RSC staff, we are currently exploring additional ways in which we can support the CICAG community and any updates will be provided in a future issue of the *Distillate*.

The ongoing 'Cheminformatics: A Digital History' series of articles has generated significant interest and we have included an item in this issue describing our rationale behind the series (page 7).

In recent years, social media has been an increasingly important means of communication with CICAG members (and non-members):

- The CICAG website is an important repository for information about future and past events, issues of the CICAG *Newsletter/Distillate*, information about bursaries, a list of committee members and more.
- Our LinkedIn group is growing and recently we have added a LinkedIn page, which allows extra flexibility in posting.
- BlueSky – recently added.
- Mastodon – not heavily used at present but which may become more important in the future.
- X has been an important channel, but its popularity appears to be waning.
- Our YouTube channel contains 29 videos including all of the Open-Source Tools for chemistry workshops. These have proved to be very popular and have been watched over 47,000 times.

An article with more information about CICAG's social medial channels is on page 77. We would be very interested to hear suggestions for additional content for all channels.

The *CICAG Distillate* continues to add unique and inspiring articles and reports, some provided by regular contributors others by new contributors. If you have an idea for an article please contact Helen Cooke, the *Distillate* editor, or me to discuss your ideas.

- -

President's Reception and Summer Party at Burlington House and the Royal Academy

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

The RSC Summer Party was held at Burlington House and the Royal Academy of Arts in July 2024. This annual event was an opportunity for the RSC to acknowledge contributions made by members and to recognise the election of six new honorary fellows. I was invited to attend to accept an Award for Exceptional Service in recognition of my work for CICAG and BMCS. I do think this award is entirely due to the fantastic people I've been able to interact with.

Chris Swain with Annette Doherty, RSC President.

- -

CICAG Planned and Proposed Future Meetings

The table below provides a summary of planned scientific and educational meetings which CICAG is currently organising or contributing to. For more information, please contact CICAG's Chair, Dr Chris Swain.

Meeting	Date	Location	Further Information
BMCS 2nd Conformational Design in Drug Discovery	3 March 2025	Discovery Centre, Cambridge Biomedical Campus, CB2 0AA	Conference web page
8th RSC-CICAG-BMCS AI in Chemistry meeting	22-24 Sept 2025	Churchill College, Cambridge	Meeting web page

CICAG is also planning to hold the following meetings and events in 2025, details to be confirmed:

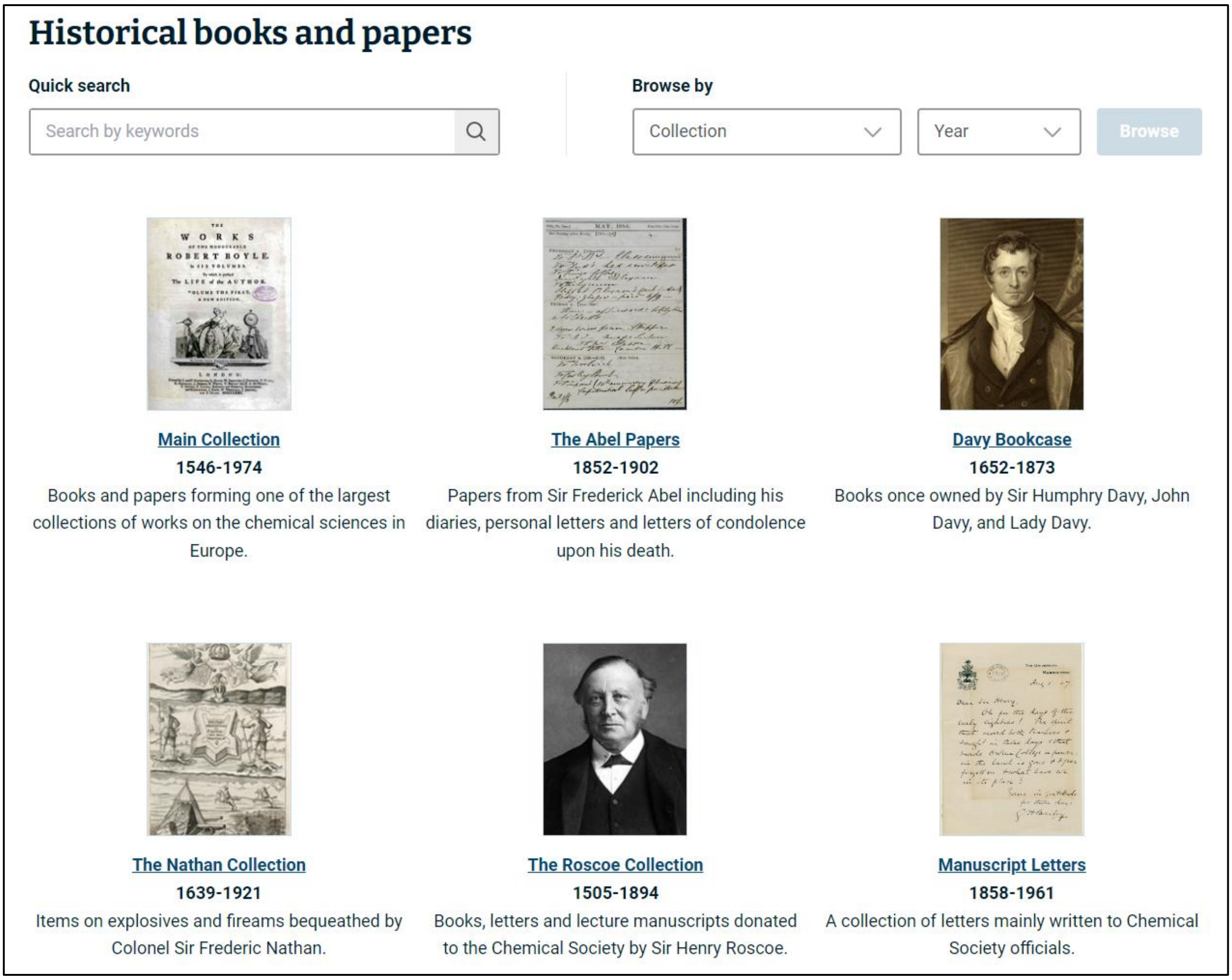
- Nomenclature for the Physical Sciences
- Centenary of Markush Structures
- CICAG AGM
- Python for Chemists
- Open Chemical Science

- -

RSC Library News: Royal Society of Chemistry Historical Collection Update

Contribution from David Allen, RSC Librarian, email: library@rsc.org

The newly re-designed Historical Collection went live on 14 June 2024. It can be accessed directly from the Library page or from the Members' area (Insight and Information). New content includes the Roscoe Letters (within the Roscoe Collection) and the Abel Papers.



Historical books and papers

Main Collection
1546-1974
Books and papers forming one of the largest collections of works on the chemical sciences in Europe.

The Abel Papers
1852-1902
Papers from Sir Frederick Abel including his diaries, personal letters and letters of condolence upon his death.

Davy Bookcase
1652-1873
Books once owned by Sir Humphry Davy, John Davy, and Lady Davy.

The Nathan Collection
1639-1921
Items on explosives and fireams bequeathed by Colonel Sir Frederic Nathan.

The Roscoe Collection
1505-1894
Books, letters and lecture manuscripts donated to the Chemical Society by Sir Henry Roscoe.

Manuscript Letters
1858-1961
A collection of letters mainly written to Chemical Society officials.

Screenshot of the historical books and papers search page.

The RSC Historical Collection is an extensive range of historical items including books, journals, letters, lecture notes, pamphlets, monographs and magazines. The collection covers the evolution of the chemical sciences from the 16th to the 20th century and includes publications from the Royal Society of Chemistry and its precursor societies. Society publications include *Chemistry in Britain*, and there are also Council minutes, lists of Fellows and annual reports. Historical books, papers and letters owned by the RSC include the Nathan Collection, the Davy Bookcase and a collection of manuscript letters written to Chemical Society officials.

- -

InChI News

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

The InChI project continues to progress steadily. The source code, which is available on GitHub, is becoming even more robust and secure as it is tested on more and more platforms and compilers.

We anticipate that a preliminary version of an extension of the InChI to improve the handling organometallic and molecular inorganics will be available for testing soon. In addition, a working group is being set up to address issues in generating InChI-based descriptors and identifiers for polymers.

Progress in the InChI project is now being discussed in monthly meetings. The next is on Wednesday 29 January 2025. You are welcome to attend. Enquiries to Jonathan Goodman (jonathan@inchi-trust.org).

We look forward to more InChI developments through 2025!

- -

Cheminformatics: a Digital History – the Thinking Behind the Series

Contribution from Chris Swain, CICAG Committee Chair, email: swain@mac.com and Helen Cooke, CICAG Distillate editor, email: helen.cooke100@gmail.com

About three years ago, Chris proposed that we initiate a series of articles for the *CICAG Newsletter* featuring the origins and early days of cheminformatics, as told by the people who were there at the time. We have invited one article per issue, starting in Summer 2022, and so far Peter Willett, Henry Rzepa, Johann Gasteiger, Wendy Warr and Gary Wiggins have written articles. This issue includes the latest in the series, Part 6, by Steve Heller. We are very grateful to all the contributors.

Our experience has been that when invited to contribute, at first people can be hesitant and modest about their achievements and contributions to the field. But once they have agreed and started to write, they enjoy the trip down memory lane, recording their recollections and reflections. In turn, Helen has found engaging with the authors as their articles have developed prior to publication to be very rewarding.

Between them, the authors, all of whom are pioneers in the field, have given illuminating insights into the evolution of cheminformatics and we are very grateful to have been able to record these in the *CICAG Newsletter/Distillate*. There are many lessons to be learned from their experiences, and we're pleased that their stories may inspire others, now and in the future.

The plan is to continue the series. At some point we may compile the articles into a book, along with others we have included from other entrepreneurs in the field of chemical information and computer applications.

To see the Digital History articles, go to the Newsletter page on the CICAG website (see issues from Summer 2022 onwards).

- -

Cheminformatics: a Digital History – Part 6. The Ups and Downs in the Development and Sustainability of Databases and Software While Working for the US Government

Contribution from Dr Steve Heller, email: steve@hellers.com

Wow! This is the sixth article in the *CICAG Newsletter/Distillate* series on those involved in the early days of cheminformatics – 'Cheminformatics: a Digital History' – and adding my voice to the first five contributors is really a very hard act to follow.

My initiation into cheminformatics started at the National Institutes of Health (NIH) in 1970 when I joined NIH as a Senior Staff Fellow in the (now defunct) Division of Computer Research and Technology (DCRT). Scotty Pratt, the head, wanted a chemist who did not want to work in a laboratory anymore. I fit his job description perfectly.

At NIH I was lucky and my first collaborations were with Hank Fales and Bill Milne at the NIH Heart Institute where Bill and I worked together, under Hank's guidance, to develop what was originally called the NIH mass spec search system (MSSS). The MSSS was housed on a research PDP-10 time sharing computer with dial up access. While very slow, it was still light years faster than looking up data in books. And, for my career, it was a great way to start with something very useful all over the world, which is what it became as the project evolved. Figure 1 shows me searching for a mass spectral hit and then displaying the result using a microfiche reader hooked up to the program. In those primitive days there were no hard drives available to store and quickly retrieve the mass spectrum.

In late 1972, NIH leadership felt it was inappropriate to support outside users of the DCRT/MSSS. The project became the NIH/EPA MSSS when I moved to EPA (Environmental Protection Agency) and began working with Bill Budde and John McGuire to use the database for identifying pollutants and hazardous materials. As for users and usage, I discovered it never hurts to create a product that a government regulatory agency used.

Finally, when EPA's new management interest in maintaining the database dwindled, I started a collaboration with the NIST (National Institute of Standards and Technology) office of standard reference data (OSRD) of Dave Lide and Lew Gevantman, which published five volumes of the spectra as well as adding to the database. This project was the first and only one of several chemistry projects I have undertaken over the years that is clearly sustainable. In fact, the database generates about $7 million dollars a year in income for NIST due a very unique piece of USA copyright law. Every, and I mean every, public product of a US Government employee cannot be copyrighted – except databases "pursuant to Section 290(e) of Title 15 of the United States Code". I never got a penny in royalties from NIST. But the best thing about this project is the yearly income to NIST which has made this a very sustainable project.

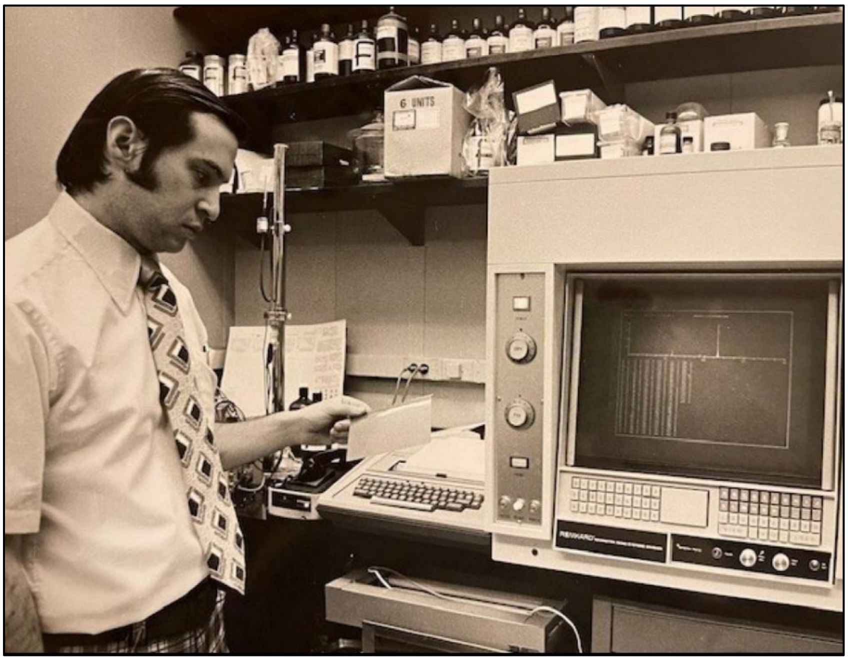


Figure 1. A microfiche reader connected to the mass spectral search program.

After the initial work on the mass spec database, Bill Milne and I went on to initiate, develop, collaborate with several databases and software packages to develop the NIH/EPA Chemical Information System (CIS) in collaboration with other US Government agencies including NBS, FDA and NIOSH.

As hard as we tried, in feature articles in *Science* magazine in 1977 and 1982 by me, Bill Milne, Richard Feldmann and Rudy Potenzone, to convince people that linking information was the future between politics, egos, and US Government agency budgets and mandates, the CIS lasted until only a few years ago.

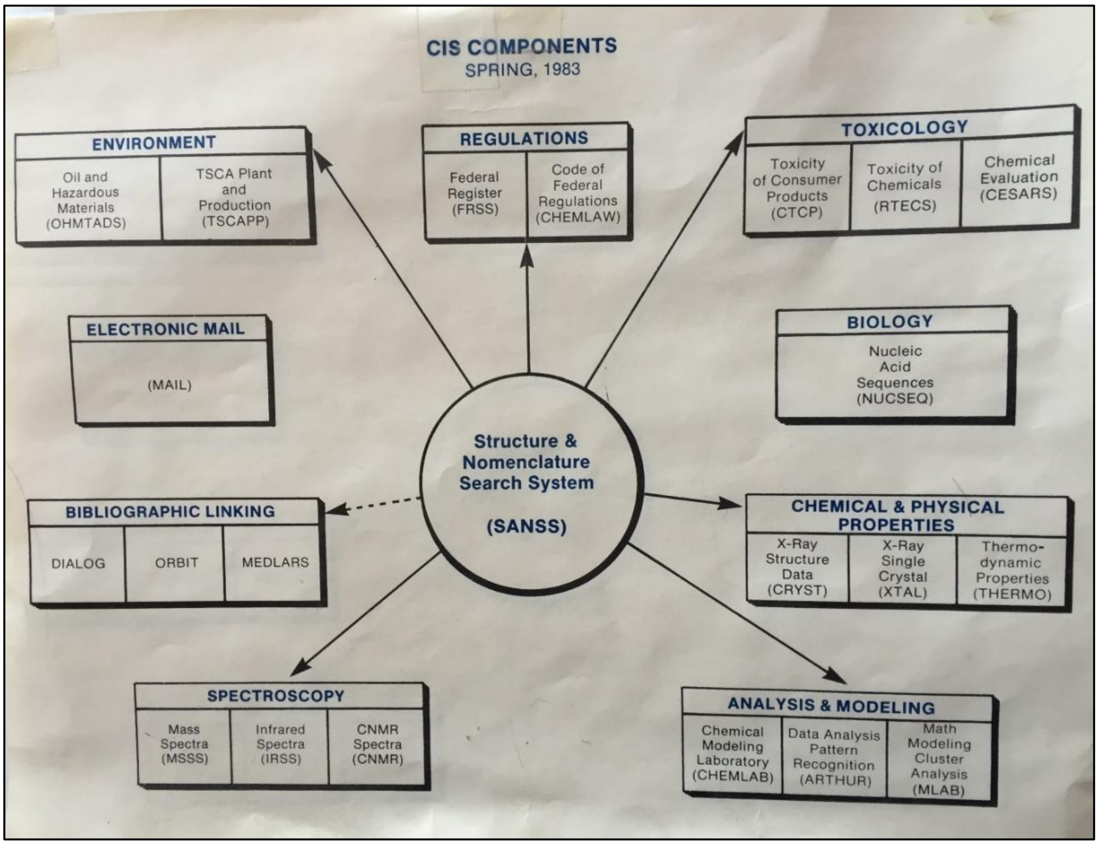

CIS COMPONENTS
SPRING, 1983
ENVIRONMENT
Oil and Hazardous Materials (OHMTADS)
TSCA Plant and Production (TSCAPP)
REGULATIONS
Federal Register (FRSS)
Code of Federal Regulations (CHEMLAW)
TOXICOLOGY
Toxicity of Consumer Products (CTCP)
Toxicity of Chemicals (RTECS)
Chemical Evaluation (CESARS)
ELECTRONIC MAIL
(MAIL)
BIOLOGY
Nucleic Acid Sequences (NUCSEQ)
Structure & Nomenclature Search System
(SANSS)
BIBLIOGRAPHIC LINKING
DIALOG
ORBIT
MEDLARS
CHEMICAL & PHYSICAL PROPERTIES
X-Ray Structure Data (CRYST)
X-Ray Single Crystal (XTAL)
Thermo-dynamic Properties (THERMO)
SPECTROSCOPY
Mass Spectra (MSSS)
Infrared Spectra (IRSS)
CNMR Spectra (CNMR)
ANALYSIS & MODELING
Chemical Modeling Laboratory (CHEMLAB)
Data Analysis Pattern Recognition (ARTHUR)
Math Modeling Cluster Analysis (MLAB)

Figure 2. A vision on how to connect, find and analyse chemical information.

While all these databases and software packages were being developed, tested, and used throughout the world we realised we needed to show off the resources we had created and/or collected. Being so new and different, we were invited very often to travel the world to show what could be done with this new and unique project we had undertaken. As part of this seeing the world I decided to take one of my three sons on just about every trip I took. One son (Matt) missed so much school that he was given the Pan Am travel award for doing homework around the world. On my third visit to Siberia (all in the summer) I even took my wife and all three boys to Siberia at the invitation of Valentin Koptyug, the President of the Siberian Academy of Sciences who I had visited previously and who had visited with me in the USA (Figure 3 & 3a). As a result of the growing interest in cheminformatics, I decided to organise meetings in this new and growing field starting in the 1970s, with at least one of these meetings still going strong. In particular, the Noordwijkerhout meetings were first organised by me, Todd Wikpe, Ernie Hyde, Charles Citroen, and Richard Feldmann in 1973. Figure 4 is a picture of Richard giving a talk in Noordwijkerhout and I was chairing that session.

Figure 3a (left): Visiting Valentin Koptyug and colleagues at their lab at the Siberian Academy of Sciences. Figure 3b (right): A dinner hosted by Valentin and his wife, pictured next to him, at their house, then (anticlockwise) my wife Shelly, our youngest son Matt and our two other sons across the table (faces not in the photo).

This first conference of what became a long-term series of Noordwijkerhout conferences was a two-week NATO/CNA (Chemical Notation Association) Advanced Study Institute on Computer Representation and Manipulation of Chemical Information. From 1998-2000 I organised the ChemInt series of conferences. From 1980-1992 I was a member of the organising committee for the International Conference on Computers in Chemical Research and Education (ICCCRE) conferences and chaired the 1982 conference. From the first meeting in 1973 to the meeting to be held in Noordwijkerhout in June 2025, this meeting brings together many of the leaders as well as those just starting out in cheminformatics.

Lastly, in the area of conferences and separate from all the chemistry meetings I organised with collaborators, in 1992 my supervisor, Jerry Miksche, at the US Department of Agriculture/Agricultural Research Service (USDA/ARS), asked me to think a of a way to publicise the agriculture genomics research activities he was heading up for ARS since the leader at ARS thought very little of genomics and preferred more ‘classical’ research areas on plants and animals. When I found a private company (Scherago International) “willing to give it try” and organise a meeting, we asked our supervisor for permission to provide technical input for the proposed conference with no personal financial support to either of us from USDA/ARS. As any good bureaucrat would do, rather than say yes or no, he just never replied to the request. We went ahead and the conference was a success with some 350 people attending. Now, 30+ years later the conference attracts some 3000+ attendees and is the premier conference in the field and I am still the chairman of the meeting – and still lacking approval.

Figure 4. Steve Heller and Richard Feldmann at the Noordwijkerhout meeting.

In another project I started with Bill Milne and Kay Pool we developed SciWords, a scientific and technical spell checker dictionary, for spell checking chemical scientific words in word processing programs like Word and WordPerfect. This turned out to be a short-lived for after a few years the database of terms was purchased by a company and integrated directly into word-processor programs.

As I retired from the US Government in 2000, I needed a project to keep me busy and thought about my initial work with adding ACS Registry numbers to the mass spec database to compare and/or weed out duplicates. In the 1970s at EPA and into the late 1990s at NIST, CAS contracted to provide the database with ACS Registry numbers for a reasonable cost. For whatever reason in the late 1990s CAS policy changed, and they no longer would provide CAS Registry numbers for the chemicals in the database which now had some 100,000+ entries. The need for a unique identifier remained after CAS would no longer provide one, so in November 1999 I proposed to the head of the mass spec lab at NIST, Steve Stein, to develop a unique identifier. From this discussion, in March 2000, a Strategy Roundtable meeting of some three-dozen people was held in Washington DC under the auspices of IUPAC and led by Ted Becker at NIH and Alan McNaught at the Royal Society of Chemistry (RSC). To create such a computer program IUPAC asked NIST to undertake the task as NIST was both the US Government standards agency and the mass spec lab needed this for their daily work. Steve agreed to develop the program and assigned Dmitrii Tchekhovskoi, a very bright programmer, to undertake the task. As Dmitrii finished the initial version of what we now call InChI the real issue of expansion, maintenance, and marketing/outreach arose. In the mid 2000s we were able to get Igor Pletnev, a visiting chemist/programmer at NIST, to take what Dimitri had done and carry on from there. Working with Alan McNaught of the RSC, we also set up a separate InChI Trust UK charity to develop, review and disseminate the InChI. Sadly, we lost Igor to COVID in 2022. Fortunately, a group at RWTH Aachen University, under the direction of dedicated InChI advocate Sonja Herres-Pawlis now provides the needed programming support and advances the developments towards the InChI for inorganic molecules.

I feel very pleased to have been part of the InChI project and to have seen the rules for the InChI standard expand and cover additional areas of chemistry, such as organometallics and inorganics. InChI is so useful already as it is with its current form and capabilities and has the potential to be more than just an IUPAC standard. I would expect that most of the future effort would be to add a few more enhancements, perhaps tautomer capability and extended stereochemistry.

Since 2004 I have been a member of the NIH/NLM/NCBI PubChem Advisory committee, working with Evan Bolton to add to and improve the capabilities of PubChem, more of which is mentioned at the end of this article.

It has been a really neat and fun ride to have been lucky enough to get into cheminformatics before there was cheminformatics. But what I most enjoyed and appreciated was the collaborators and friends I found along this journey. I have mentioned over a dozen people I have worked with on cheminformatics projects over the years and they have all been so very helpful in making these projects as successful as they have been. Without such coopetition I would never have been able to do most of what had been accomplished. I truly believe that this philosophy of working with others and giving credit to them has enabled me to accomplish what I have done. This cooperation is reflected in the awards I have received over the years. The ACS CINF Skolnik award was with Bill Milne. The Mike Lynch award was with Alan McNaught, Igor Pletnev, Steve Stein, and Dmitrii Tchekhovskoi.

Saving the best for last, I was pleased with my various activities in cheminformatics over the years and wanted to do something would survive me. Our family discussions led to the idea of endowing a chair in cheminformatics in the chemistry department at Stony Brook University on Long Island, New York, where my wife Shelly and I earned BS degrees. To further feel this was a good idea I discovered that there were chemistry department faculty for organic chemistry, analytical chemistry, physical chemistry, inorganic, and others, but no cheminformatics. We finalised this endowment in mid-2024 and look forward to this new faculty position being filled for 2025-2026 academic year. With the total package for this endowment reaching almost $5 million, roughly half for the chemistry department and the rest for scholarships and related support) I believe this will be a sustainable activity.

When I left my first US Government position, in my exit interview the reason I gave for leaving was “desire to do useful work”. And thanks to a good number of collaborators I think I have done some useful work over the past 55+ years. In conclusion what I have learned through the projects I have worked on I would like to be sustainable. First, and most foremost, when it comes to databases and computer software I learned that I needed to work with, and give credit to others and say thank you. Second, I learned that you need to find a first-rate solid organisation to work with a good public reputation and hope they can carry on once I go off to do something else.

So what about the future? I am only 82 and my father made it to 103, so assuming I was not an adopted child, I need a hobby or something to do for the next two or so decades. I seem to have found a link between my annual PAG (Plant and Animal Conference) meeting and chemistry. Thus, now I am off looking for data to enhance PubChem with agriculture information from the agriculture, plant and animal community. I was asked to look for databases/datasets linking to chemicals used in plant and animal work, including where in the animal or plant such data can be found, such as tissue/organ, pathways, and anything linking information to diseases.

Editor’s note: On 9 October 2024 an article was posted on Stony Brook University’s website, Motivated by the Matches: Alumni Establish New Endowed Chair in Cheminformatics, featuring Steve and Shelly Heller’s experiences and achievements while at the University and more information about the endowed chair in cheminformatics.

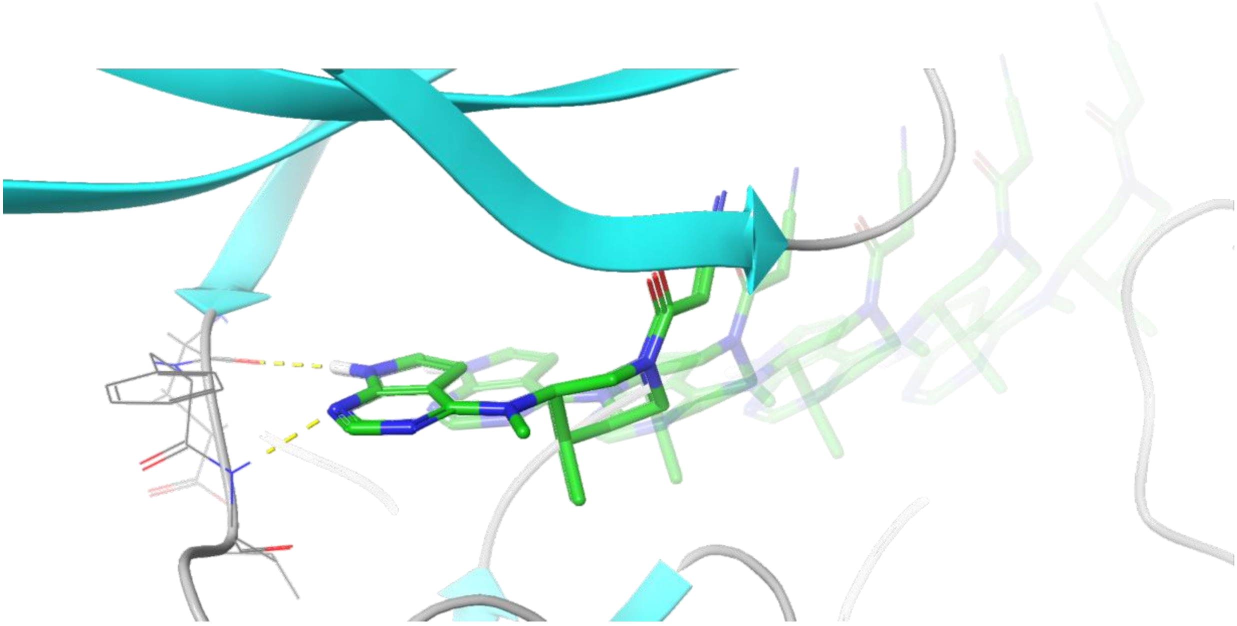
- -

Computational Chemistry Basics for Medicinal Chemists. Part 2: Ligand Binding Thermodynamics

Contribution from Andreas Ritzén, Principal Scientist at Novo Nordisk, email: auze@novonordisk.com

Part 1 of the CICAG Newsletter/Distillate series is in the Summer 2024 issue. The full series is available on LinkedIn.

Binding affinity



We usually want to use compchem methods to predict what design changes in our molecule will lead to better potency, or at least higher binding affinity. Alternatively, for an undesired off-target, we want to reduce affinity while keeping it for our target of interest. Even when detailed structural information on the ligand-target complex is available, these are not trivial tasks. Unfortunately, we can't know how important any given interaction is by just looking at the structure of the ligand-target complex. Maybe that hydrogen bond you see doesn't contribute to the binding affinity? In fact, by only looking at the ligand-target complex, we can make no prediction of binding affinity at all, because we are only looking at half the equation! There are two important factors, invisible in a static model of the ligand-target complex, that we must take into account: **solvation** and **conformational entropy**.

Solvation: don't forget that water is everywhere!

Remember that the binding affinity is determined from this equilibrium, where L is the ligand, T is the target, and LT is the ligand-target complex:

$$L + T \rightleftharpoons LT$$

It is tempting to reason that if we could make the ligand-target complex tighter we would move the equilibrium to the right and get a better affinity. This is unfortunately not always true! Why not? Because we are looking at this association in aqueous solutions, solvation/desolvation becomes important, but this is completely invisible in the ligand-target complex we are looking at. In the equilibrium above, solution behaviour is implicit. To remind us of this, let's add a '*solv*' suffix to explicitly specify that solvated species are involved:

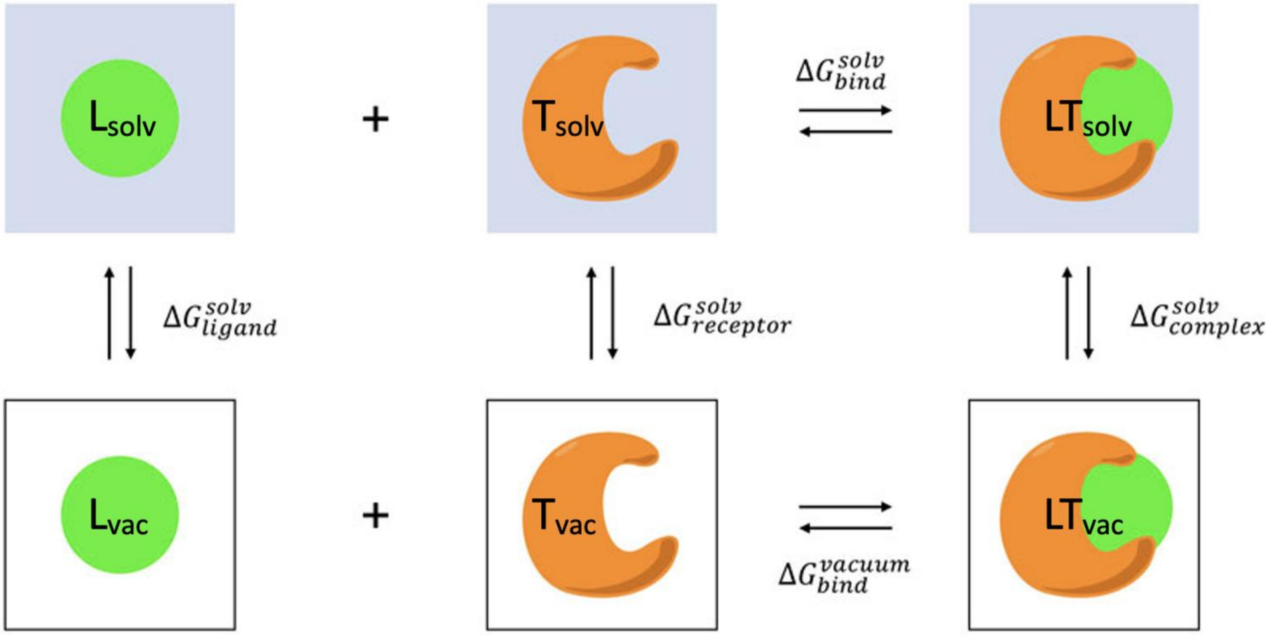
$$L_{solv} + T_{solv} \rightleftharpoons LT_{solv}$$

$$K_D = \frac{[L_{solv}][T_{solv}]}{[LT_{solv}]}$$

where *KD* is the dissociation constant, a measure of binding affinity. The relationship between *KD* and the standard-state Gibbs free energy of ligand binding in water is: [1]

$$\Delta G^{\circ}{}^{solv}_{bind} = RT \ln K_D$$

When thinking about this, it helps to consider solvation/desolvation explicitly as individual steps of the equilibrium. We can dissect solvation and binding using a series of coupled equilibria as shown in Figure 1.



L_{solv}
+
T_{solv}
ΔG^{solv}_{bind}
LT_{solv}
ΔG^{solv}_{ligand}
$\Delta G^{solv}_{receptor}$
$\Delta G^{solv}_{complex}$
L_{vac}
+
T_{vac}
LT_{vac}
ΔG^{vacuum}_{bind}

$$\Delta G^{solv}_{bind} = \Delta G^{vacuum}_{bind} + \Delta G^{solv}_{complex} - \left(\Delta G^{solv}_{ligand} + \Delta G^{solv}_{receptor}\right)$$

Figure 1. Thermodynamic cycles for ligand and target (receptor) solvation/desolvation and association/dissociation (i.e., *binding/unbinding). Adapted from King, et al. 2021*.

When we look at the ligand-target complex we only consider one term of that equation (Δ*G vacuum bind*) and ignore all the others. Most docking scores work this way. The problem is that solvation/desolvation losses can override the gains we make by design changes in the ligand-target complex. For example, it looks like we could make a new hydrogen bond from our ligand to our target by adding a hydroxyl group at a certain position. Docking looks great and the docking score predicts a stronger complex (i.e., a decrease in Δ*G vacuum bind* because more negative Δ*G* equals stronger binding). But when we make the molecule, the affinity is unchanged, even though X-ray crystallography shows the predicted hydrogen bond being formed. Why? Because the increased desolvation penalty of pulling off the water molecules from the ligand as it binds to the protein (Δ*G* *solv ligand*) costs as much energy as the gain from that new hydrogen bond (Δ*G vacuum bind*). Another situation can occur when we try to displace a buried water molecule from the target by replacing it with ligand atoms. If the increase in desolvation penalty of the target (Δ*G solv receptor*[2]) is greater than the gain from the new interaction (Δ*G vacuum bind*) we will lose affinity. There are more costly compchem methods to predict such changes, e.g., free energy perturbation (FEP), but they are beyond the scope of this guide. Figure 1 above was adapted from a recent review on such methods.

Conformational entropy: rigid is better – or is it?

Of relevance at the molecular level, the statistical definition of entropy is:

$$S = k_B \ln \Omega$$

where *kB* is the Boltzmann constant and Ω is the number of microstates or configurations of the system, assuming that all microstates are equally probable. For the following discussion, we can think of a microstate as a specific conformation selected from the conformational ensemble of a molecule. Because different conformers usually have different energies, their probabilities are not the same: the lower the energy of a conformer, the more probable it is that the molecule is found in that conformation. In this case, the entropy is calculated as:

$$S = -k_B(p_1 \ln p_1 + p_2 \ln p_2 + \cdots) = -k_B \sum_i p_i \ln p_i$$

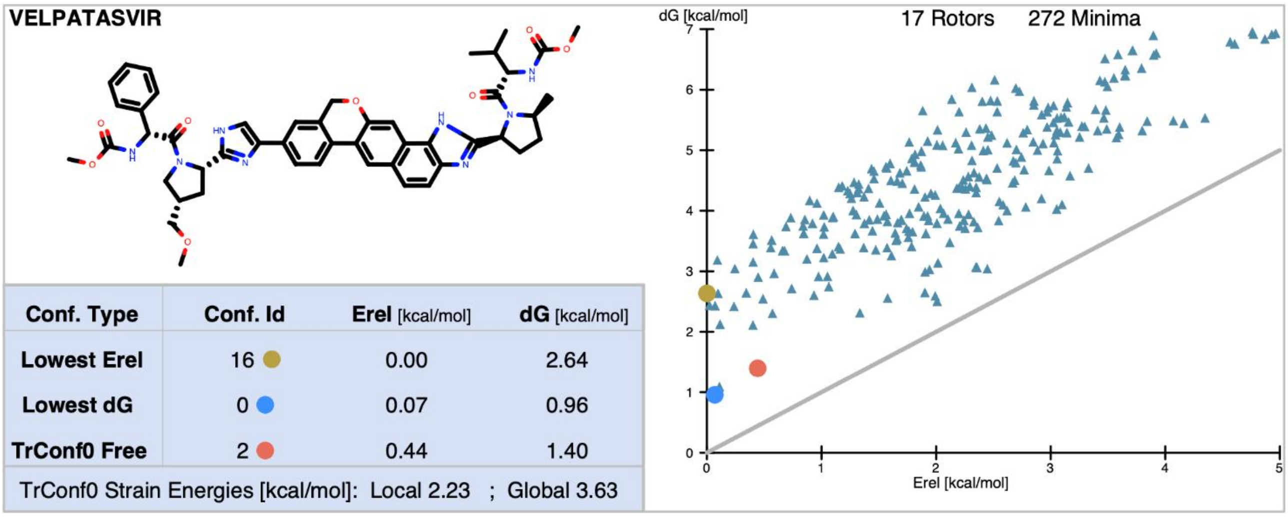
where *pi* is the probability of microstate *i*. This equation simplifies to the one above when all microstate probabilities *p* are equal. What this means is that the entropy of a system increases with increasing number of available microstates. The change in entropy is most pronounced for microstates with a relatively high probability. Adding or removing microstates with very low probabilities won't change the entropy very much. To summarise: if we go from fewer to more microstates, Δ*S* is positive and decreases Δ*G* (favourable), and if we go from more to fewer microstates, Δ*S* is negative and increases Δ*G* (unfavourable).

The entropy of the system of interest can be described as the sum of various entropy contributions. For example, consider the change in ligand conformational entropy upon binding. We can think of this as going from a freely moving ligand molecule that can access its entire conformational ensemble (many microstates) to a bound molecule in one single conformation in a ligand-target complex (one microstate). This would decrease the entropy of the ligand molecules in the system: we go from many microstates (conformers) to just one, so we get a negative Δ*S*. The ligand conformational entropy contribution to Δ*G vacuum bind* upon binding (see Figure 1) is therefore unfavourable: It increases Δ*G vacuum bind*.

Does this mean that flexible molecules should be avoided? This is a contentious topic, probably because the answer is: it depends! Solvation was not taken into account in the paragraph above, and solvation/desolvation can overwhelm the contribution of conformational entropy. For example, in solution, not all conformers may be populated because of e.g., hydrophobic collapse of a lipophilic, flexible molecule. And even upon binding, some flexibility may still be possible, i.e., more than one microstate may still be accessible. For a comprehensive discussion, see Geschwindner *et al.*, 2015, and for an example, see Wienen-Schmidt *et al.*, 2018. Claims that Δ*G* increases by 0.3-0.6 kcal/mol per rotatable bond are sometimes heard, but they derive from inappropriate assumptions, e.g., solvation effects are ignored, and all conformers are assumed to be equally populated. A computational study of the HIV protease inhibitor amprenavir showed that the conformational entropy decrease upon binding contributed only 2 kcal/mol to the Δ*G* of binding despite the molecule having 12 rotatable bonds. A conformational analysis showed that "...the single most stable conformation has a probability of 0.23, the 6 most stable conformations have a combined probability of 0.51, and the 45 most stable conformations have a combined probability of 0.91". In other words, most of the available conformations (microstates) have very low probabilities and excluding them won't increase Δ*S* very much. The vibrational entropy losses upon binding were much larger and contributed 25 kcal/mol to the Δ*G* of binding. Vibrational entropy losses are due to a restriction of the vibrational degrees of freedom upon binding.

In addition to ligand and target entropy changes, we must also take the entropy changes of the solvent (water) into account. The hydrophobic effect is thought to be largely due to changes in water entropy. Water molecules tend to form cage structures that surround hydrophobic molecules in solution. This restricts the freedom of movement of those water molecules compared to bulk water molecules. The entropy is therefore reduced and makes an unfavourable contribution to Δ*G*. Solvation is a complex phenomenon, because water molecules form hydrogen bonds to each other and to the solutes, and those bonds have a favourable enthalpic contribution to Δ*G*. Thus, entropic and enthalpic contributions are opposite and the balance between them determines if the net effect on the various Δ*G solv* terms in Figure 1 is positive or negative.

There are computational methods that can be used to estimate the conformational entropy and its effect on Δ*G*. For example, the OpenEye Freeform method combines a conformational analysis with a calculation of conformational entropy and can even take solvation effects into account. One example is shown in Figure 2. In the plot, every conformer is shown as a triangle. The horizontal axis shows the strain energy relative to the global minimum (tan dot) and the vertical axis shows the Δ*G* resulting from restricting the molecule to the respective single conformer. The conformer with the lowest Δ*G* is shown as a blue dot, and a specific conformer given as input (e.g., the bioactive conformer from an X-ray crystal structure) is shown as a red dot. In the example in Figure 2, the input conformer carries a penalty in Δ*G* = 1.40 kcal/mol. The molecule has 17 rotatable bonds, and 272 conformations were found within the specified energy window (5 kcal/mol).



Conf. Type	Conf. Id	Erel [kcal/mol]	dG [kcal/mol]
Lowest Erel	16 ●	0.00	2.64
Lowest dG	0 ●	0.07	0.96
TrConf0 Free	2 ●	0.44	1.40

TrConf0 Strain Energies [kcal/mol]: Local 2.23 ; Global 3.63

*Figure 2. Impact of conformational entropy loss on Δ*G *going from the entire conformational ensemble to a single conformation calculated using OpenEye Freeform. For an in-depth discussion of the theory behind this method, see the online documentation.*

Affinity is not the only consideration when thinking about flexible vs. rigid ligands. A flexible ligand can adopt many more shapes than a rigid one. This could potentially allow it to bind to undesired off-targets, i.e., be less selective than desired. It could also be more readily metabolised by CYP enzyme oxidation because it would be able to flex to fit the active site. It may also make the ligand more susceptible to be a substrate of efflux transporters, e.g., PGP and BCRP, which can be very problematic for compounds intended to reach the central nervous system. So even if affinity may not be the main issue with flexible ligands, it is worth considering the topic when designing molecules.

Summary

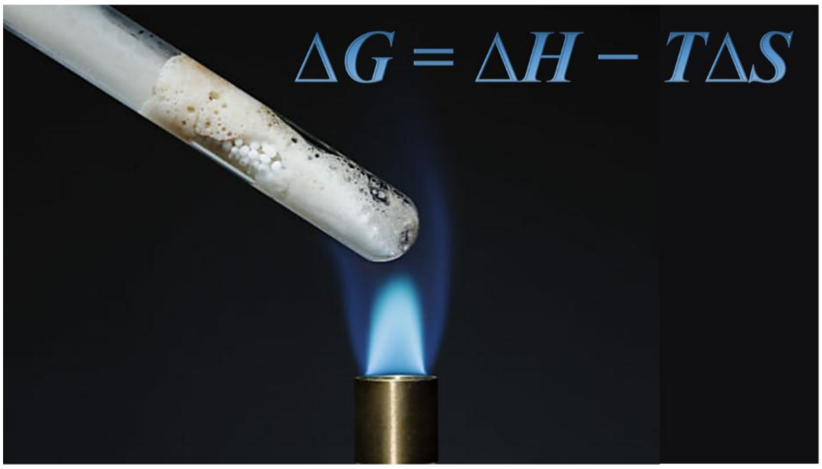
- Binding affinity reflects the equilibrium of solvated species in an aqueous environment.
- Changes in desolvation penalty of the ligand must be considered when comparing the affinity of a new design with an existing one.
- Changes in desolvation penalty of the target must also be considered.
- Flexibility, i.e., the number of rotatable bonds, may negatively impact binding, selectivity and ADME parameters.

Notes

[1] *Technically, the definition of KD in this equation is not entirely correct (a logarithm can only be taken for a dimensionless number). The correct definition uses the activity (a) of each species (ligand, target, and ligand-target complex) rather than the concentration in the calculation KD. For practical purposes (dilute solutions near room temperature at atmospheric pressure), activity can be approximated as a = c/c^{o}, where c is the concentration and c^{o} is an arbitrarily chosen standard-state concentration, usually 1 M. Thus, we can use the dimensionless number KD/c^{o} in this equation with c^{o} = 1 M to obtain the standard-state Gibbs free energy ΔG^{o}. For more information, see my article on ligand binding thermodynamics.*

[2] *In this situation, the buried water molecule is considered part of the receptor structure, because it is present even when the ligand is bound. This is why we account for its displacement by a different ligand in the "ΔG solv receptor" term.*

Ligand binding thermodynamics



Entropy and the Gibbs free energy

Let's admit it: entropy in thermodynamics is not easy to get your head around. It is often described as disorder, but that is not an ideal way to describe it. It is perhaps better to think of it as a means of defining in which direction a system evolves: Does the ligand bind or unbind given its affinity for the target and the concentrations of the species? The entropy of the universe always increases, but that doesn't mean that the entropy of a system does. We can use energy to decrease the entropy of a system. This is then counterbalanced by a greater increase in the entropy of the surroundings. The Wikipedia page on entropy has lots of background information, both regarding the classical thermodynamic and the statistical mechanics definitions. Because the entropy of the universe is impractical to consider (and not really of interest to us anyway), we define the Gibbs free energy *G* as a means to determine the direction a system evolves in without having to consider the surroundings. We are usually only interested in changes in this parameter, so we define:

$$\Delta G = \Delta H - T\Delta S$$

where ΔH is the change in enthalpy, *T* is the absolute temperature, and ΔS is the change in entropy. Although ΔG is called the Gibbs free energy, it is not actually an energy in the thermodynamic sense, but rather the potential to do work. We use it because at equilibrium, $\Delta G = 0$ and we don't need to consider the entropy of the surroundings. The Gibbs free energy is related to chemical potential, as discussed in Chen 2019.

How exactly does the Gibbs free energy let us determine the direction a system evolves in? Consider a ligand binding reversibly to a target with a 1:1 stoichiometry:

$$\mathrm{L + T \rightleftharpoons LT}$$

where L = ligand, T = target, and LT = ligand-target complex. The equilibrium constant for the forward direction is the association constant *KA* = [LT]/([L][T]), and the equilibrium constant for the reverse direction is the dissociation constant *KD* = [L][T]/[LT], which is a measure of binding affinity.

How is Gibbs free energy related to the equilibrium constant? For convenience, we define the standard-state Gibbs free energy, denoted ΔG^{o}, at a reference point (the standard state). This is arbitrarily chosen, usually with a pressure of 1 bar, a temperature of 298 K, and with all components at 1 M concentration. To find the actual ΔG at a set of given concentrations (i.e., not at the standard state) like in the example above, we only need to calculate the difference from the reference point. To do this, we use the actual concentrations to calculate the reaction quotient *Q* = [LT]/([L][T]). We can then use the following equation to find our ΔG at the given concentrations of ligand, target and ligand-target complex:[1]

$$\Delta G = \Delta G^\circ + RT\ln Q$$

At equilibrium, *Q* = *KA* (the equilibrium constant) and $\Delta G = 0$ so we have:[1]

$$0 = \Delta G^\circ + RT\ln K_A$$

Rearranging this, and remembering that *KA* is the inverse of the dissociation constant (*KD* = 1/*KA*), and that ln 1 = 0 and ln *a*/*b* = ln *a* - ln *b*, we get:[1]

$$\Delta G^\circ = -RT\ln K_A = RT\ln K_D$$

As an example, let’s consider a ligand with a binding affinity (*KD*) of 10 nM, an unbound target concentration of 1 nM, and a ligand-target complex concentration of 1 nM. In other words, we start out with a 50% target occupancy in this example. What happens when we change the ligand concentration? Will we get a higher or lower fraction bound? In Figure 3, a plot of ΔG vs. ligand concentration is shown. If the ligand concentration [L] < 10 nM, then $\Delta G > 0$. This tells us that the dissociation of the ligand from the already formed ligand-target complex is favoured, and unless any kinetic barriers prohibit it, the ligand will start to dissociate. There is insufficient ligand available to drive the association process to the right. On the other hand, if [L] > 10 nM, then $\Delta G < 0$ and the ligand will bind to the unbound target and increase the fraction bound.

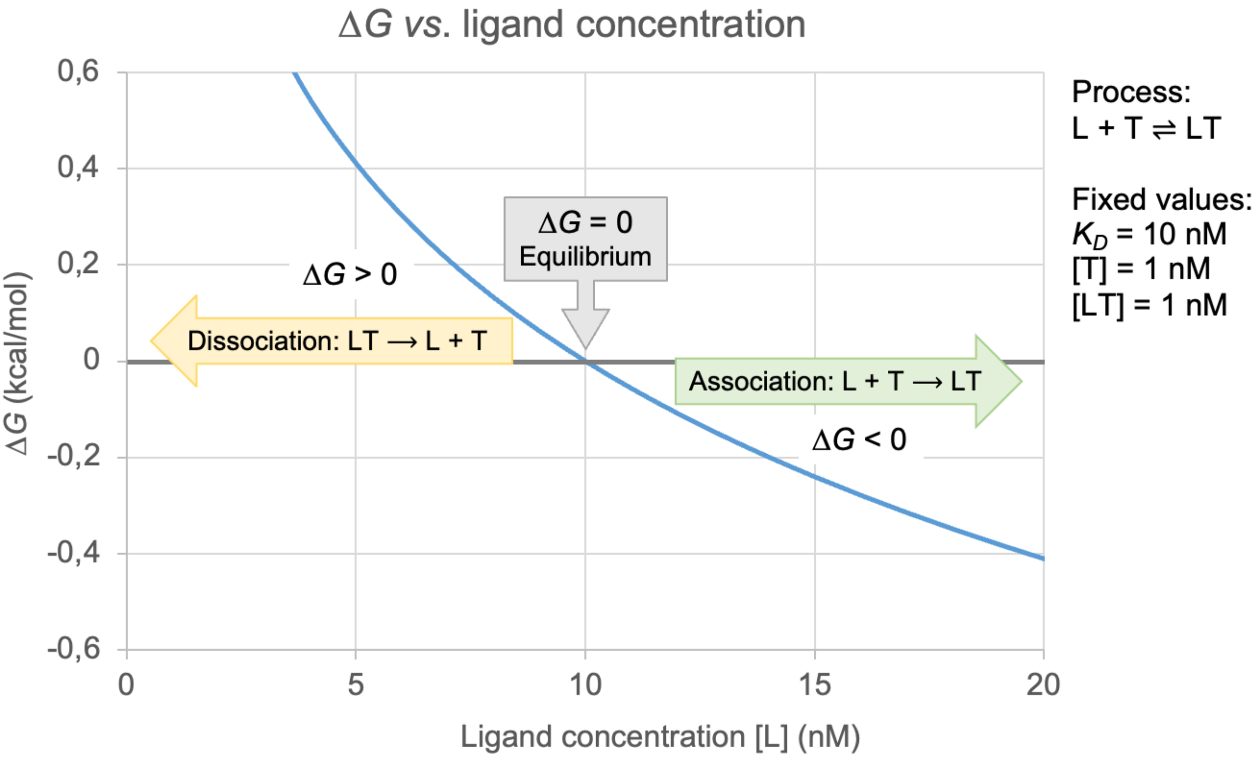


Figure 3. Plot of ΔG vs. ligand concentration under a set of given conditions.

Note

[1] *Technically, the definitions of Q, KA and KD in these equations are not entirely correct (a logarithm can only be taken for a dimensionless number). The correct definitions use the activity (a) of each species (ligand, target, and ligand-target complex) rather than the concentration in the calculations of Q, KA and KD. For practical purposes (dilute solutions near room temperature at atmospheric pressure), activity can be approximated as a = c/c^{o}, where c is the concentration and c^{o} is an arbitrarily chosen standard-state concentration, usually 1 M. Thus, we can use the dimensionless numbers Q/c^{o}, KA/c^{o} and KD/c^{o} in these equations with c^{o} = 1 M to obtain the standard-state Gibbs free energy ΔG^{o}.*

- -

Data-driven Discovery in the Chemical Sciences: Faraday Discussion

Contribution from Jeremy Frey, School of Chemistry and Chemical Engineering, University of Southampton, email: j.g.frey@soton.ac.uk

The Faraday Discussion series has over a century of history and more than 300 meetings at the forefront of the physical sciences and many Discussions have become landmark meetings in their field. This Discussion, held from 10-12 September 2024 in the lecture room of Trinity College, Oxford, focused on the increasingly central role of big data, machine learning (ML), and artificial intelligence (AI) in the chemical sciences. The meeting was very well attended, filling the lecture hall with more than 120 delegates.

For those unfamiliar with the Faraday discussion format, the papers accepted for a Faraday discussion are circulated to the attendees as pre-prints and the meeting itself centres around the discussions of these papers. The authors are given five minutes to highlight their paper, and in groups of typically three related papers, this is then followed by a discussion, with questions documented and will form part of the published proceedings.

The meeting brought together several different communities within chemistry, in particular the materials and molecular communities from both academic and industrial research institutions. Four central themes were part of the meeting: (a) Discovering chemical structure, (b) Discovering structure–property correlations, (c) Discovering trends in big data and (d) Discovering synthesis targets. The meeting started with an introductory talk by Alán Aspuru-Guzik setting out his quite dramatic view of the future of chemistry, which certainly engendered much discussion among the attendees.

There was a strong focus on AI/ML for materials discovery, and it was interesting to see where the use of AI/ML in drug discovery both overlapped with the materials discovery and where the two diverged. There were lively discussions when it came to predicting molecular and materials synthesis pathways and the issues of automatic extraction of data from the literature, and as frequently commented upon in the data driven science area – the need for negative data (or perhaps better put as less optimal pathways) which are essential to developing the best quality models; Heather Kulik's paper (***Leveraging natural language processing to curate the tmCAT, tmPHOTO, tmBIO, and tmSCO datasets of functional transition metal complexes)***, on curated datasets highlighted these aspects.

To highlight just a few others of the great papers, Jiayun Pang's talk (***Specialising and analysing instruction-tuned and byte-level language models for organic reaction prediction***) about the large language models for reaction prediction, making use of transfer learning from large language models (LLMs) trained on very large language datasets but with fine-tuning on much smaller reaction prediction datasets. Basita Das's paper (***Embedding human knowledge in material screening pipeline as filters to identify novel synthesisable inorganic materials***) showed some ways in which data and chemical knowledge can be combined in models.

There was also a nice discussion on how you explore space that you've not exposed experimentally, Chris Collins on the use of generative machine learning with crystal structure prediction. Veronika Juraskova's talk (***Modelling ligand exchange in metal complexes with machine learning potentials***) on using specially crafted algorithms on small datasets showed what can be done even if you don't have a huge dataset, which chimed nicely with Claudi Draxl's talk on ***How big is big data?***

The importance of making your work more generally and widely available was highlighted (along with making data available) by several speakers and Kim Jelfs' work ***Web-BO: towards increased accessibility of Bayesian optimisation (BO) for chemistry***, stood out.

To get a flavour of the exciting dialogue at the meeting I recommend looking at the articles, some of which are already available online. The printed version of the Discussion should be available in Spring 2025.

- -

Two Academic Librarians Open up on Open Access

Contribution from Sara Wallace, Librarian, Yusuf Hamied Department of Chemistry, University of Cambridge, email: sw2018@cam.ac.uk, and Sophy Arulanantham, Library & Information Manager, Department of Geography, University of Cambridge, email: sja60@cam.ac.uk

We approach this discussion from the viewpoint of librarians currently employed by Cambridge University Libraries, for the chemistry and geography departmental libraries respectively. Sophy also brings her knowledge of open access (OA) from her previous position in the Office of Scholarly Communication. The views expressed in this article are our own and are not necessarily those of the institution in which we are employed.

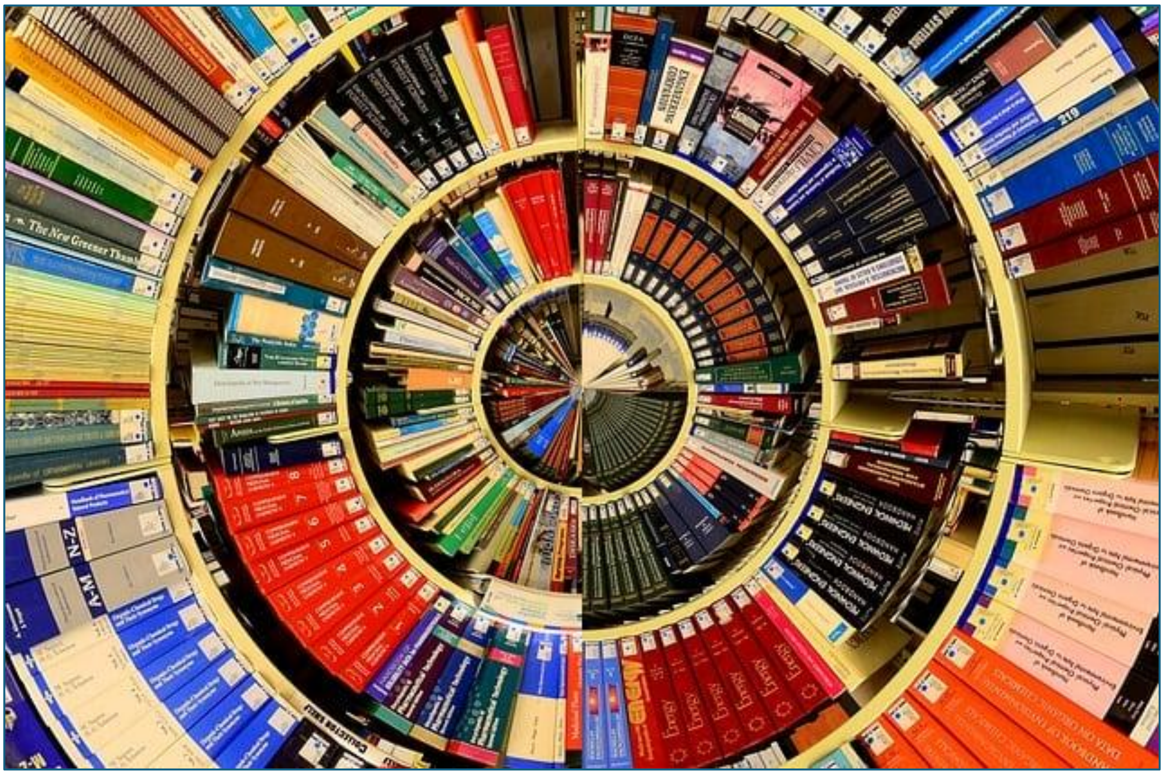


Image credit: Gerd Altmann/Pixabay.

In November 2020 the Royal Society of Chemistry held a five-day series of talks exploring open science. Clair Castle, the then Librarian at the Yusuf Hamied Department of Chemistry, University of Cambridge, spoke at the event on Open Access Publishing for Chemistry to present an academic librarian's perspective. She highlighted that OA publishing in chemistry was lagging behind other STEMM disciplines, and suggested strategies which could encourage chemists to be more open.[1] Over the last four years the OA movement has gained momentum across all disciplines with many academic institutions adopting self-archiving policies (SAP) and rights retention policies (RRP), along with many UK and non-UK research funders requiring immediate OA to research articles. It seemed timely to provide an update on the OA landscape, focusing on research articles, again from the viewpoint of academic librarians.

The current OA landscape – publishing at academic institutions

Institutional changes

Institutions are increasingly developing infrastructure to enable research to be shared more openly, whilst also ensuring that researchers can keep the rights to their manuscripts. They have developed policies to encourage or mandate self-archiving of research articles in open repositories to help researchers to comply with funder OA policies and the institutions' own open research statements. Similarly, the OA policy for the last Research Excellence Framework (REF) in 2021 required that journal articles and conference proceedings were made OA to be eligible for the REF submission. Cambridge is still following the REF 2021 policy guidance, however the final OA policy for REF 2029 is expected to be announced very soon and will be implemented from January 2026. Preparations are already underway at Cambridge to help support researchers through the potential changes.

Rights retention is a strategy that enables researchers to retain their rights over their manuscripts – it allows them to retain control over what they are used for and freely share their work with others. Institutional RRPs support researchers to exercise their rights and provide an extra layer of protection when dealing with publishers. They typically work by researchers granting the institution a licence to make the accepted manuscript (the final author-created version accepted following peer review but before any publisher formatting) available via the institution's repository with an open licence at the time of publication (also known as self-archiving or 'green' OA).

The number of UK institutional RRPs has grown considerably in recent years. The UK Institutional Rights Retention Policies website[2] lists 36 UK academic institutions with RRPs, of which four were piloted or fully adopted in 2022, 20 in 2023 and 10 in 2024 (the remaining two are due to be adopted in 2025). Across the world there are currently 141 universities and institutions with RRPs and eight more in preparation, as listed via the Open Access Directory. 54 of these have been adopted over the last two years.

The University of Cambridge adopted its Self-Archiving Policy (SAP) for research articles in April 2023. This followed a successful one-year pilot policy. The intention of the SAP is to disseminate Cambridge research as widely as possible and facilitate compliance with funder OA requirements. To support the policy, the University has notified most of the publishers that Cambridge researchers publish with. Researchers can opt out of the policy if they wish at the time they deposit their accepted manuscript in the University's repository Apollo. The SAP sits within the University's Open Access Publications Policy Framework.

Other recent developments at Cambridge include the introduction of the University's central OA fund, which provides a route to gold OA for researchers who want to publish their research article in a fully OA journal but do not have access to funding to cover the OA charges (article processing charges), the launch of a diamond OA journals platform pilot project, and a new preprints deposit service for Cambridge researchers.

Research funder changes

Many UK and non-UK research funders have joined cOAlition S and therefore shifted their OA policies for research articles to require immediate OA with open licences to maximise reuse. The Registry of Open Access Repositories Mandatory Archiving Policies ROARMap chart shows the growth in OA mandates from international universities, research institutions, and research funders.

UK Research & Innovation (UKRI - comprising AHRC, BBSRC, ESRC, EPSRC, MRC, NERC, STFC, Research England and Innovate UK) now requires peer-reviewed research articles to be OA immediately upon publication with a Creative Commons Attribution (CC BY) licence. Articles can be made OA either by making the published version OA on the journal's website ('gold' or 'diamond' OA) or by depositing the accepted manuscript in an open repository without an embargo ('green' OA).

UKRI-funded researchers can apply to have the OA charges paid from their institution's block grant when they publish their research article in a journal which meets UKRI's requirements (fully OA journals and Jisc-approved transformative journals only). Alternatively, they can publish their article gold OA if it is eligible for one of their institution's transitional read & publish agreements (TAs).

Transitional read & publish agreements (TAs) are contracts between publishers and institutions which are meant to encourage a transition from the traditional model of paying subscriptions to read journals, to paying for OA publishing in those journals. The Joint Information Systems Committee (Jisc) highlights that the agreements are intended to ensure that the transition to OA will have a minimal effect on institutional budgets.[3] Other funders mandate OA but require the costs to be budgeted into research grants or have other mechanisms via which OA charges are paid. The ERC's open science requirements depend on which framework programme funding has been obtained from. Researchers who are funded under Horizon 2020 are required to make research articles available in an open science repository within six months of publication for STEMM. Under

Horizon Europe, researchers are required to make all peer-reviewed publications (including monographs and book chapters) immediately OA upon publication with a CC BY licence, either by paying for gold OA or by using rights retention and self-archiving. In either case, the publication must be deposited in a 'trusted repository'. OA charges are eligible costs if they are incurred within the lifetime of the research project and within the provisions of the grant agreement.

Institutional RRPs and self-archiving are important because they provide a no-cost route to compliance for researchers whose funders mandate immediate OA. This is especially helpful in cases where the research will be published in a hybrid or subscription-based journal and it is not eligible for gold OA publication via other routes.

OA publishing in chemistry

In her presentation, Castle[1] highlighted that in October 2020 chemistry OA journals made up only 0.92% of all OA journals listed in the Directory of Open Access Journals (DOAJ) (142 of the 15,231 OA journals). Figures taken from DOAJ in November 2024[4] show that the number of chemistry OA journals has more than doubled to 342 and now make up 1.6% of all journals listed. The number of all OA journals listed in DOAJ, as of November 2024, is 21,134. Although the number of chemistry OA journals has increased in the last four years, there are still far fewer OA chemistry journals worldwide compared with other disciplines. The November 2024 figures from DOAJ show that medicine OA journals make up 22% of its listings, biology 3.5%, social sciences 17.4% and technology 12.9%.

It is important to state that DOAJ does not include hybrid journals, only fully OA journals are listed which meet the criteria stated in DOAJ's guide to applying. Therefore, the number of fully OA chemistry journals available, and the comparison with other disciplines, does not necessarily reflect the true prevalence of OA publishing in chemistry. Another method to investigate the uptake of OA publishing in chemistry, is by looking at the SCImago Journal and Country Rank which takes data from journals indexed in the Scopus database to calculate a SCImago Journal Rank (SJR). The data is provided from Scopus annually and the SJR is updated each June.[5] Filtering the search parameters to a subject area of chemistry, the results[6] show there are 1017 chemistry journals published worldwide which have been given an SJR for 2023, based on metrics provided by Scopus in March 2024. This is an increase of 18% on the figure obtained by Castle in 2020[1]. In the top 50 of these, 15 are listed as being OA, an increase of 12 from 2020 figures, or 400% compared against the increase in the number of chemistry journals published. The OA journals ranked here include hybrid journals (note that both JACS and JACS Au are listed), and the 2023 SJR ranking indicates an increase in chemistry OA publishing.

Recent figures shared by Jisc chart the worldwide growth of gold and hybrid OA publishing across all disciplines, including chemistry, for 2017-2022.[7] These figures show that in chemistry there has been a 20% growth in the uptake of gold OA publishing and 5% growth in the uptake of hybrid OA publishing, with a reduction of 10% in closed access or 'bronze' OA (free to read on the publisher's website but without an open licence). Data taken from Web of Science show a similar trend. In 2018, 36% of all research articles indexed in multidisciplinary chemistry journals were OA: 22% were gold OA and 14% were other forms of OA.[8] By 2021, this had increased to 57% of all chemistry research articles: 46% were gold OA and 11% were other forms of OA.[8]

There has also been a growing interest in preprints from funders and an increased uptake in chemistry. Preprint servers offer an easy way for researchers to share and find the latest research findings, prior to formal peer review, and get rapid feedback from the academic community. They also help to make science more inclusive and interconnected. However, some researchers may have concerns about making their research ineligible for publication in major journals on the grounds of prior publication. They may also be concerned about the absence of quality control and the dissemination of poor-quality research.

Chemists have come relatively late to preprinting. Researchers in subjects such as theoretical and computational chemistry and physical chemistry have been posting preprints for some time, however it has only become common practice for chemists in general in the last few years.[9] ChemRxiv was launched in 2017 with support from the Royal Society of Chemistry (RSC), the American Chemical Society (ACS), the Chemical Society of Japan, the Chinese Chemical Society and the German Chemical Society (GDCh). The number of new preprints posted to ChemRxiv has increased from 877 in 2018 to 4,051 in 2021.[10] While these numbers are small compared with the total number of chemistry research articles published each year (approx. 500,000), they are steadily growing. Many chemistry journals now have editorial policies on preprints and prior publication which help to mitigate the concerns of researchers.

What effects does the shift to zero-embargo OA and RRPs have on commercial publishers and transitional agreements?

Currently, chemistry researchers at academic institutions in the UK are often supported to publish gold or hybrid OA via funder block grants or institutional TAs, or as noted previously, some funding bodies will cover OA charges if they are budgeted into research grants. Particularly in the case of the TAs, researchers are often not fully aware of the huge costs involved for institutions, most often coming from library budgets. Ma[11] highlights that the gold OA model and TAs continue to stretch library budgets and warns that due to changes in funding and open research policies, gold and hybrid OA publishing will become more and more costly as commercial publishers continue to raise APCs. She also acknowledges that "gold open access is an attractive option because many hybrid journals are established with a track record, especially when funding is available. Nevertheless, the gold open access model heavily relies on commercial publishers and some argue that the pay-to-publish model is antithetical to bibliodiversity and equality in scholarly publications globally."[11]

The increase of RRPs and SAPs, whilst beneficial for encouraging and supporting zero-embargo green OA, may have knock-on effects to commercial publishing and gold OA publishing routes – and publishers are increasingly aware of the implications. Particularly relevant to chemistry, we only need to look at the rationale for the ACS's article development charge (ADC) to see the way publishers might respond to pressure for immediate OA. Whilst the ACS states that the ADC is only for those not covered by an institution's TA,[12] the increasing cost of TAs and the reduction in library budgets may mean that some institutions are unable to justify signing new agreements leaving their researchers open to the ADC and APC. The development of the ADC has been met with concern on both sides of the Atlantic. Rumsey[13] posits that a zero-embargo charge "perpetuates an increasingly out-of-touch and outdated position taken by some publishers, who aim to prevent researchers from retaining their rights to use their own work as they choose". In America, the Ivy Plus Libraries Confederation (IPLC) highlights that the ADC goes against the 'long-established' practice of authors sharing their research by depositing manuscripts in OA repositories and therefore "prevents universities from creating an accessible record of their scholarly output". The IPLC also emphasises that the ADC opposes open and inclusive scholarly communication by further increasing the cost of OA publishing for certain communities.[14]

In Europe, the ACS, along with Elsevier, Springer Nature, Wiley and Taylor & Francis are often known as 'The Big Deals' publishers due to their dominance of the academic publishing market. Ma highlights findings from The 2019 Big Deals Survey Report[15] that these publishers have "published over 50% of the total number of publications, costing more than 75% of the total spend by subscription-fee paying research institutions and libraries on academic journals in Europe".[16] For the Jisc consortium in the UK, the 'Big Five' publishers are usually viewed as Elsevier, Springer Nature, Wiley, Taylor & Francis and SAGE. However, at our institution, the ACS is the fifth largest agreement (SAGE is ninth).

Several of these publishers have UK TAs ending in 2024-2025 – will academic libraries, and the institutions they support, be able to continue to justify paying the increasing costs publishers levy against OA?

Looking to the future

Jisc review of transitional OA agreements

Earlier this year, Jisc published a review on whether the TAs have achieved what they set out to do – to help drive the transition to full OA and provide greater transparency and reduce costs for UK institutions. The review notes that the UK TAs have helped to achieve high levels of funder compliance and to reduce and constrain costs to some extent.[17] However, it observes that the rate of OA transition is slower than expected, with the increase in the number of TAs having little impact on overall levels of OA. There also remains a lack of transparency about how OA charges are costed and about publisher strategies for transitioning to OA. The review also notes that some researchers and organisations are excluded from the TAs and there are significant concerns about the long-term sustainability of article-based business models. Critics of the TAs have argued that they further embed the article as the unit of value, incentivising publishers to increase article volume in order to increase their profits.[18] It is also argued that TAs mean the cost of OA is often "pushed back to libraries"[19] and their dwindling budgets.

In their guidance on the implementation of Plan S with regards to TAs, cOAlition S advises that, "recognising that a fundamental principle of these transformative arrangements is that they are temporary and transitional, where cOAlition S members provide funding to support publication fees of journals covered by such arrangements, this funding will cease on the 31 December 2024".[20] The rules on how funder OA block grants can be used are also expected to change in 2025. What will this mean for researchers wanting to publish OA and meet funder requirements? Researchers may need to rely on institutional RRPs and SAPs if they want to continue to publish in hybrid journals and meet funder requirements.

Next generation OA agreements and supporting researchers

Jisc are proposing a fundamental change to the way that OA agreements are evaluated in the future. They are proposing an increased emphasis on equity and inclusion in research (with the adoption of 'equity indicators') and new requirements for publishers to reduce complexity and demonstrate a commitment to the OA transition. They are also recommending that institutions financially divest from underperforming TAs and use the funds to support equitable and sustainable OA agreements instead. Jisc is now consulting the UK HE sector on this proposed new approach. Its success will depend on gaining support from stakeholders across the UK HE sector. This approach will require a change in author publishing behaviour – but are researchers ready for this? Librarians must educate researchers on the costs involved in the TAs and highlight the alternative publishing options available to them.[21]

It is also recognised by those advocating for OA, that research assessment reform is needed in order to support open research. As of November 2024, 3353 organisations (which includes universities) worldwide have signed the Declaration on Research Assessment (DORA). Currently, research assessment places too much significance on publishing in traditional high-impact journals, and this can dissuade researchers, particularly early career researchers, from seeking to publish in newer OA journals.[22] This is echoed in cOAlition S's feedback on the draft OA policy for REF 2029, which calls on UK funders to adopt a stronger OA policy.[23] As the OA landscape continues to evolve, librarians will need to further engage with the academic community, guiding and supporting researchers on how to navigate the changes ahead.

References

(1) Castle, C. *How Open are Chemists? An Academic Librarian's Perspective*. Open Chemical Science, UK, 9-13 November 2020, RSC Open Access Publishing for Chemistry. *Apollo,* November 10, **2021**. https://doi.org/10.17863/CAM.69417 (accessed 2024-11-29)

(2) Eglen, S.J. *UK Institutional Rights Retention Policies*. https://sje30.github.io/rrs/rrs.html (accessed 2024-11-30)

(3) Jisc. *Working with transitional agreements*. **2020** (updated 2022) https://www.jisc.ac.uk/guides/working-with-transitional-agreements (accessed 2024-11-30)

(4) *DOAJ*. https://doaj.org/ (accessed 2024-11-23)

(5) *SCImago Journal & Country Rank Help Page*. https://www.scimagojr.com/help.php?q=FAQ (accessed 2024-11-29)

(6) SCImago, (n.d.). *SJR — SCImago Journal & Country Rank* [Portal]. Retrieved November 23, 2024 https://www.scimagojr.com/journalrank.php?area=1600&type=j (accessed 2024-11-23)

(7) Brayman, K.; Devenney, A.; Dobson, H.; Marques, M.; Vernon, A. Figure 13 in *A review of transitional agreements in the UK*. **2024**. https://doi.org/10.5281/zenodo.10787392 (accessed 2024-11-29)

(8) Novara, F.R.A. Golden Ten: A Decade of Open Access Society Publishing. *ChemistryOpen.* **2022**, *11*, e202100270. DOI: 10.1002/open.202100270

(9) Coudert, F.X. The rise of preprints in chemistry. *Nat. Chem*. **2020**, *12*, 499-502. DOI: 10.1038/s41557-020-0477-5

(10) Mudrak, B., Bosshart S.; Koch W.; Leung A.; Minton D.; Sawamoto M.; Tegan, S. Five years of ChemRxiv: where we are and where we go from here. *Chem. Sci.* **2022**, *13* (48), 14210–14212. DOI: 10.1039/D2SC90224A

(11) Ma, L. Open Access. In *The Scholarly Communication Handbook: From Research Dissemination to Societal Impact*. Facet, **2023**, p. 34.

(12) *ACS Publications Open Access Pricing Page.* https://acsopenscience.org/researchers/oa-pricing/#adc (accessed 2024-11-29)

(13) Rumsey, S. American Chemical Society (ACS) and author's rights retention. *sOApbox*, October 17, **2023**. https://www.coalition-s.org/blog/american-chemical-society-acs-and-authors-rights-retention/ (accessed 2024-11-28)

(14) Ivy Plus Libraries Confederation. *IPLC response to the article development charge proposed by the American Chemical Society*, November 9, **2023**. https://ivpluslibraries.org/2023/11/iplc-response-to-the-article-development-charge-proposed-by-the-american-chemical-society/ (accessed 2024-11-28)

(15) Morais, R.; Stoy, L.; Borrell-Damián, L. *2019 Big Deals Survey Report: An updated mapping of major scholarly publishing contracts in Europe.* European University Association, **2019**. https://www.eua.eu/downloads/publications/2019%20big%20deals%20report%20v2.pdf (accessed 2024-11-27)

(16) Ma, L. Critical issues and the future of scholarly communication. In *The Scholarly Communication Handbook: From Research Dissemination to Societal Impact*; Facet, **2023,** p. 104.

(17) Brayman, K.; Devenney, A.; Dobson, H.; Marques, M.; Vernon, A. *A review of transitional agreements in the UK*. **2024**. https://doi.org/10.5281/zenodo.10787392 (accessed 2024-11-29)

(18) Mudditt, A. Transitional Agreements Aren't Working: What Comes Next? *The Scholarly Kitchen*. 4 April **2024**. https://scholarlykitchen.sspnet.org/2024/04/04/transitional-agreements-arent-working-what-comes-next/ (accessed 2024-11-29)

(19) Butler, L-A.; Matthias, L.; Simard, M-A.; Mongeon, P.; Haustein, S. The oligopoly's shift to open access: How the big five academic publishers profit from article processing charges. *Quantitative Science Studies*. **2023**, *4* (4), 778–799. DOI: 10.1162/qss_a_00272

(20) cOAlition S. *Guidance on the Implementation of Plan S.* https://www.coalition-s.org/guidance-on-the-implementation-of-plan-s/ (accessed 2024-11-28)

(21) Ma, L. The Platformisation of Scholarly Information and how to Fight It. *LIBER Quarterly: The Journal of the Association of European Research Libraries.* **2023**, *33* (1), 1–20. DOI: 10.53377/lq.13561

(22) Ma, L. Critical issues and the future of scholarly communication. In *The Scholarly Communication Handbook: From Research Dissemination to Societal Impact*; Facet, **2023**; p. 108.

(23) cOAlition S. *cOAlition S urges the four UK higher education funding bodies to adopt a stronger Open Access policy for the next Research Excellence Framework*, June 18, **2024** https://www.coalition-s.org/coalition-s-urges-the-four-uk-higher-education-funding-bodies-to-adopt-a-stronger-open-access-policy-for-the-next-research-excellence-framework/ (accessed 2024-11-30)

The Open Reaction Database

Contribution from Benjamin J. Deadman, Connor Coley, Steven Kearnes, email: ben@open-reaction-database.org

The Open Reaction Database (ORD) is an open-access schema and infrastructure for structuring and sharing organic reaction data, including a centralised data repository. The ORD schema supports conventional and emerging technologies, from benchtop reactions to automated high-throughput experiments and flow chemistry. Our vision is that a consistent data representation and infrastructure to support data sharing will enable downstream applications that will greatly improve the state of the art with respect to computer-aided synthesis planning, reaction prediction, and other predictive chemistry tasks (Figure 1).

Since our initial meeting in October 2019, the database has grown to include 2M reactions (including a large dataset of reactions extracted from USPTO sources) and received contributions from academic and industrial users, both from published and unpublished work. To keep updated with the latest ORD news please follow on LinkedIn, and subscribe to our New ORD Datasets Newsletter.

The database can be quickly explored using the online graphical browse and search interface, which includes chemical structure searching functionality. Datasets and search lists can be downloaded as an ORD protocol buffer file from the interface and the open source ord-schema package allows users to work with these files in Python. The full ORD database is also available on GitHub as the ord-data repository as the source of truth, and users are encouraged to clone the repo for more complex querying and applications. The schema also includes an Object Relational Mapper (ORM) which allows the serialised protocol buffer data files to be converted into a PostgreSQL relational database.

The ORD is currently supported by philanthropic funding under the recommendation of Schmidt Sciences, formerly Schmidt Futures. Under this support, the ORD is in an exciting phase of growth and adoption including the following activities.

Improving user experience

Active development has started on the new interface for contributing reaction data to the ORD. This will be a complete replacement for the first generation form that is currently in use, and is designed with synthesis chemists in mind, to capture reaction data into a structured format as painlessly as possible.

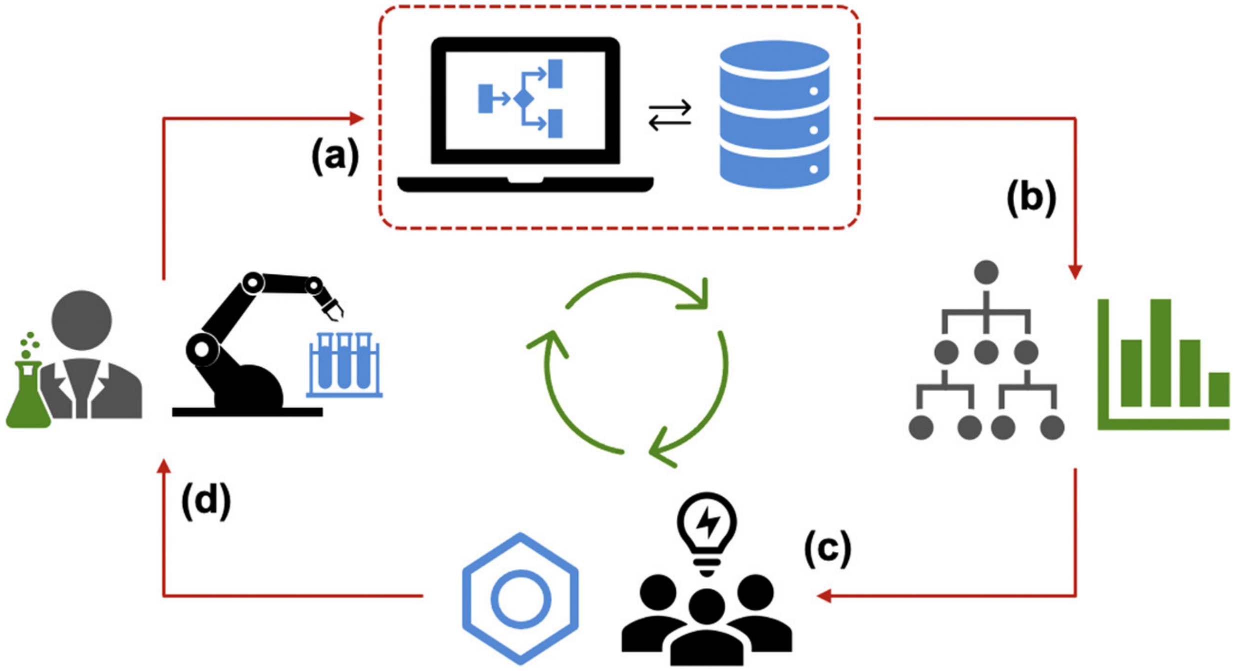


Figure 1. Computer-aided chemical discovery cycle: (a) the Open Reaction Database; (b) machine learning and cheminformatics; (c) human or automated interpretation and material design; (d) manual or robotic chemical synthesis. Image reproduced with permission from American Chemical Society (CC-BY-NC-ND 4.0).[1]

We are also experimenting with dashboard style visualisations in the ORD data browser. These updates are currently in testing but hopefully the first visualisations will be live by the time this newsletter is published.

Front line support is available to help you get started with the ORD, both preparing datasets for submission, and also using the ORD in downstream applications. To request support please submit an issue on GitHub, or contact us by email.

Funding for dataset generation

Over the summer of 2024 we ran the first call for proposals for academic high-throughput experimentation labs to collect new, large reaction datasets in funded projects ($110,000 each). The first round has now completed and we expect to announce the chosen labs in early 2025, with the new datasets expected later in the year.

We hope to run a second call for proposals in 2025, and interested applicants should look for the announcement on LinkedIn and/or email Benjamin Deadman (the ORD Program Manager and Data Evangelist) to find out more.

Training

In July we also announced the ORD Trainee Program, a funded opportunity for students and early-career research chemists in the UK and USA. The program sees trainees working with the ORD Program Manager to prepare 1-2 datasets in the ORD format, from previously published research. It is expected that applicants spend 10 to 20 hours on the ORD training over a three month period and there is some flexibility to adapt this to the applicant's situation.

Participants gain valuable knowledge and experience of the ORD toolset to define reaction data in a structured format. All dataset submissions to the ORD receive formal recognition and a persistent identifier which can be included in CVs, professional web pages and professional development records. Time spent on pre-approved

ORD activities is also reimbursed at a rate of up to 30 USD per hour, with a maximum limit of paid hours per dataset.

If you are interested in joining the program then please submit an Expression of Interest through our Google Form. Trainees are inducted in batches, after selection from our waiting list.

How you can get involved

The Open Reaction Database is a community effort and we welcome you to get involved. Some areas where we are seeking the community's support are:

- Please use the ORD data to do great things, and if possible tell us about it. The data is all published under a CC-BY-SA license to permit both academic and commercial use.
- Do you have some reaction data that could be made open access? It could be some historical data, or a new paper that you are publishing soon. Please get in touch so we can support you in preparing your first ORD dataset.
- Do you want more or different reaction data than what we have in the ORD already? Please work with us as we connect experimental data generators with the machine learning community. We think there is huge potential in our community and we are working together on collaborative funding bids for further dataset creation.
- We can help software vendors to implement ord-schema input and output tools in their software. While we don't have the resources to develop translators for every ELN or reaction data management software, we can usually provide advice on how the ord-schema relates to your data structures.
- Finally, show your support for the ORD with one of our laptop stickers. Look for Ben at conferences, or get in touch to request some by post.

Upcoming events

The ORD will be presented at the following events:

- 1st ICIQ High-Throughput Experimentation Symposium. Accelerating synthetic development: HTE, AI/ML and the Future of Autonomous Labs, 11-12 February 2025, ICIQ, Catalonia Spain.
- London Lab Live, 14-15 May, London, United Kingdom
- More to be announced.

Reference

(1) Kearnes, S.M. et al. The Open Reaction Database. *J. Am. Chem. Soc.,* **2021,** *143* (45), 18820-18826. https://doi.org/10.1021/jacs.1c09820

- -

Meeting Report – Molecular Simulations for Chemistry

Contribution from Hannah Bruce Macdonald, Computational Chemist, Charm Therapeutics, email: *hannah@charmtx.com*

Hannah Bruce Macdonald at the meeting. (Image credit: Julien Michel, University of Edinburgh, CCPBioSim.)

The first Molecular Simulations for Chemistry Meeting was held on 14 June 2024 at the RSC's Burlington House in London to discuss methodologies and applications of simulations. 70 delegates attended the in-person meeting for talks covering a broad range of applications in fields such as pharmaceuticals, materials science, and energy – from both academic and industrial viewpoints.

The one-day meeting consisted of eight talks by renowned speakers. The morning session started with a biological focus, with talks from Victor Guallar (Barcelona Supercomputing Center) who discussed Monte Carlo methods for protein-protein docking (PELE) followed by a view of how they are trying to streamline preclinical drug development using molecular dynamics and free-energy perturbation methods from Silvia Lovera (UCB). Fernanda Duarte (University of Oxford) discussed how her group is developing tools for automating reaction modelling using their own potentials for organic chemistry. The focus then shifted to materials-based methods, with Phillip Camp (University of Edinburgh) talking about his group's research in soft-matter simulation which involves out-of-equilibrium simulations and the large space of complex mixtures and Micaela Matta (King's College London) discussing the use of DFT and MD for polymers, with a focus on melanin and the complexities of the biomaterial.

Graeme Day (University of Southampton) outlined both forwards and backward predictions for crystal structure prediction: both trying to predict the structure of an input molecule, or conversely trying to predict a molecule that can provide a material of given properties, before Daniel Cole (University of Newcastle) discussed the novel methods of forcefield development which underpin molecular simulation (DE-FF and MACE-OFF). The conference was wrapped up by Livia Bartók-Pártay (University of Warwick) with a talk on developing new sampling methods, including nested-sampling for materials.

In addition to the talks, the poster session was well-attended, featuring around 20 posters that were discussed during coffee breaks and lunchtime, with most of the presenters being junior researchers from academia. Friday evening began with a networking session over some post-conference drinks in the Council Room and was a great opportunity to catch up with acquaintances and make some new connections while discussing the day. Discussions are ongoing about a repeat event for 2025, so please watch out for communications from the RSC-CICAG group, or reach out directly if you are interested in being involved.

The conference was organised by Samantha Hughes (AstraZeneca, RSC-CICAG committee), Julien Michel (University of Edinburgh, CCPBioSim), and myself (Charm Therapeutics, RSC-CICAG committee). Special thanks go to CICAG committee members Chris Swain (Cambridge MedChem Consulting) and Neil Berry (University of Liverpool) for their contributions. The conference was sponsored by CCPBioSim, EPCC, and AstraZeneca.

Update from the AIchemy Hub

Contribution from Dr Chris Mellor, email: c.mellor@imperial.ac.uk, and
Dr Ben Alston email: ben.alston@liverpool.ac.uk, AIchemy Hub Managers

Introduction to AIchemy

AIchemy is one of nine AI Hubs funded by the EPSRC, focused on pioneering AI-driven chemistry. We are a UK hub with an interdisciplinary leadership team from Imperial College London, University of Liverpool, University of Cambridge and University of Southampton. Our main aim is to transform the chemistry-AI interdisciplinary research landscape by conducting world-leading research in open AI and catalysing its wider adoption in experimental and computational chemistry communities. Driven by those already working at the AI-chemistry interface, AIchemy will also be inclusive to those new to this field, and has been designed to be broad, spanning molecules through materials to devices, reflecting the broad cross-sector need for AI in UK chemical and chemical-related industries. More information can be found on our website.

Our activities are split across three themes: events, training and community building, and funding. Read about some of our recent and upcoming activities below.

Events

Our event schedule kicked off with our digital launch on 5 September 2024, and in October we hosted the inaugural webinar in our monthly webinar series. Plans are also underway for a variety of other events across 2025, with our inaugural annual conference taking place at Imperial College on Monday 17 March. Further details are available on our website.

Digital launch

At our digital launch, the AIchemy Hub unveiled our vision for integrating AI into chemistry, attracting nearly 150 attendees.

The hub has kicked off with five forerunner projects:

1. **Human-in-the-Loop Platform** – a real-time platform where researchers inject insights as robotic searches evolve, scaling to global crowdsourced experiments.
2. **Large-scale Crystal Structure Prediction (CSP)** – creating large-scale data, to describe relationships between crystal structure, molecular structure, and functional properties.
3. **Data-Driven Materials Discovery** – developing a ‘design-to-device’ pipeline that uses AI, machine learning, and high-throughput testing for materials discovery for chemical applications.
4. **Generative AI for Materials** – scalable generative models that harvest chemical knowledge from large-scale multi-modal chemistry data.
5. **Multi-Fidelity Reaction Optimisation** – an automated platform using AI to identify and optimise reaction mechanisms.

These projects set the stage for collaborative, data-driven chemistry innovation. Postdoctoral researchers are being recruited to these projects, and further research projects will be aligned to the Hub through Flexible Funding calls and industry engagement.

AIchemy encourages collaboration across sectors, offering engagement opportunities for co-funded projects and secondments. These partnerships will play a key role in advancing innovation within the AI and chemistry research, bridging the gap between industry and academia.

Our Digital Launch also outlined the variety of activities the Hub will undertake and included an interactive session where we asked the audience for their views on topics such as what they wanted from the Hub and the timing of activities. The response from attendees during our interactive session highlighted a real need for community-building around AI in chemistry and will guide our future events and initiatives.

Webinars

In October, we hosted the inaugural webinar in our monthly webinar series, organised by the AIchemy ECR Committee. The webinar series allows for a short talk by an early career speaker, followed by a more established speaker. Our aim is to cover a wide range of topics in relation to AI for chemistry, covering how AI is being used within both experimental and computational chemistry.

Our first webinar featured Annabel Basford, a PhD student at Imperial College London, who shared her work on automating the synthesis of porous organic cages, showcasing how robots and high-throughput systems can speed up material discovery and improve efficiency. Then, Professor Anna Slater from the University of Liverpool talked about the wonders of flow chemistry, which allows for precise control over complex reactions, leading to better yields and quicker processes.

In our second webinar on 20 November, our speakers were Abdoulatif Cisse, a PhD student at the University of Liverpool, who explored how expert human hypotheses can be integrated with Bayesian optimisation to navigate unexplored scientific search spaces, followed by Dr Adam Clayton from the University of Leeds, who discussed how machine learning and adaptive algorithms could be used within self-optimising approaches for flow synthesis.

These webinars will take place every month with details of upcoming webinars available on the events page of our website. If you would like to nominate a speaker (including ECR speakers – late stage PhD students or postdoctoral researchers) you can do so via this online form. If you missed a webinar, they are all available via our YouTube channel.

Annual conference

Our first annual conference will take place at Imperial College White City Campus on Monday 17 March 2025. This conference will be focussed on the challenges and opportunities at the AI-chemistry interface. Details of speakers and how to register are available on our website.

Training and community building

Machine learning training school

The first activity within our training and community building theme is a Chemical and Materials Machine Learning School (CAMMLS) taking place at Daresbury Laboratory from 31 March to 4 April 2025. This training school is being run in collaboration with Physical Sciences Data Infrastructure (PSDI) initiative and with support from STFC-SCD, PSDS, CCP5 and CCP9, and is a follow up to the 2023 Machine Learning for Atomistic Modelling Autumn School run by PSDI. The training is targeted towards PhD students, in particular those 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 for the atomistic simulation of materials.

Applications closed on 1 December 2024, with more than double the number than the spaces available. Applicants will be notified of the outcome of their application by late January 2025.

Undergraduate internships

Summer 2024 saw eight undergraduate students take part in eight-week research projects across four of our partner institutions. Projects spanned various applications of AI and data approaches to discover new materials, predict and compare crystal structures, predict chemical properties, embed robotics into chemistry labs, optimise experimental measurements, map chemical space and predict materials properties.

Applications for the 2025 Internship scheme will open in January 2025, with project supervisors and students able to jointly apply for an internship taking place between June and September 2025.

Funding

The AIchemy Hub has a flexible fund available to support projects with the wider community. The first of our funding calls has recently closed (on 6 December 2024). The AIchemy Pump Priming Fund offered up to £25,000 per project (award made at 80% full economic cost) to support short-term research at the AI-Chemistry interface. This funding aimed to help researchers develop and test innovative ideas in areas such as AI applications in chemistry, software development, and data improvements. Applications were encouraged from ECRs, with Postdoctoral researchers also eligible to apply, with PI support.

The review process is now in full swing, with a distributed peer review process being used whereby all applicants take part in the review process, with a small review panel made up of Leadership Committee members making the final funding decisions.

Our next funding call with be for larger grants – up to two years of PDRA time – and is due to open in late Q2 2025, with a closing date in September 2025. A sandpit will also be organised around the call opening for researchers and industry partners to come together and discuss project ideas and collaborations. Details will be on our website in due course.

Stay in touch

Website: https://www.aichemy.ac.uk
Twitter (X): @aichemyhub
LinkedIn: @aichemy-ukhub
YouTube: @AIchemyhubUK
Bluesky: @aichemyhub.bsky.social
Contact us at: info@aichemy.ac.uk
Mailing List: https://www.jiscmail.ac.uk/AICHEMYHUB

- -

ChEMBL@15: Fifteen Years of Drug Discovery Data in the Open

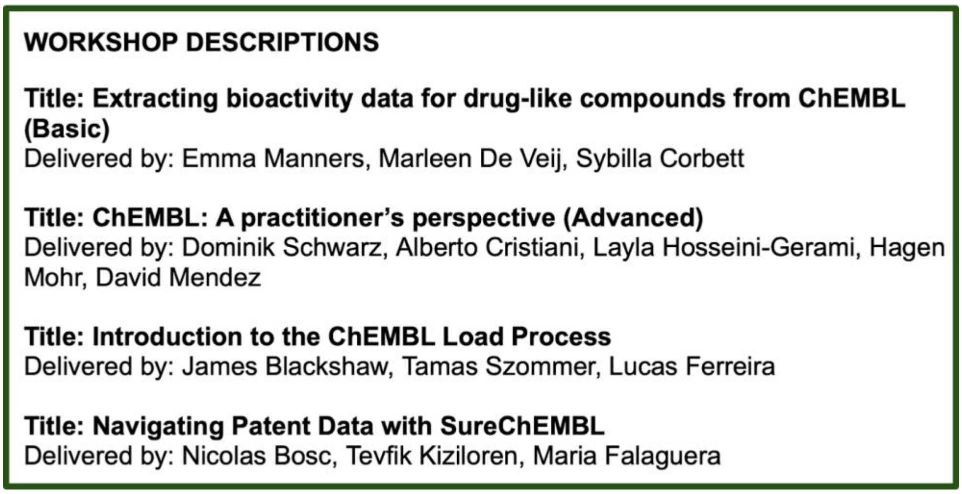
Contribution from Emma Manners, Barbara Zdrazil & the ChEMBL team, EMBL-EBI (Hinxton, UK),
email: chembl-help@ebi.ac.uk

In October 2024, ChEMBL celebrated 15 years as a public data resource for drug discovery. To mark this milestone, the ChEMBL team, our EUbOPEN collaborators, and others within the cheminformatics community, gathered for a two-day symposium. The event was held at EMBL-EBI in Hinxton (UK) and supported by the Royal Society of Chemistry's Chemical Information and Computer Applications Group (RSC CICAG).

The ChEMBL@15 cake and attendees at the ChEMBL@15 symposium. (Credit: C. Palferman, image reproduced with permission from C. Palferman, EMBL-EBI.)

The first day of the event was dedicated to in-person workshops; these focused on best practice for deposition of data to ChEMBL, beginner and advanced ChEMBL tutorials, and exploration of the SureChEMBL patent resource.

Day 1: workshops



WORKSHOP DESCRIPTIONS

Title: Extracting bioactivity data for drug-like compounds from ChEMBL (Basic)
Delivered by: Emma Manners, Marleen De Veij, Sybilla Corbett

Title: ChEMBL: A practitioner's perspective (Advanced)
Delivered by: Dominik Schwarz, Alberto Cristiani, Layla Hosseini-Gerami, Hagen Mohr, David Mendez

Title: Introduction to the ChEMBL Load Process
Delivered by: James Blackshaw, Tamas Szommer, Lucas Ferreira

Title: Navigating Patent Data with SureChEMBL
Delivered by: Nicolas Bosc, Tevfik Kiziloren, Maria Falaguera

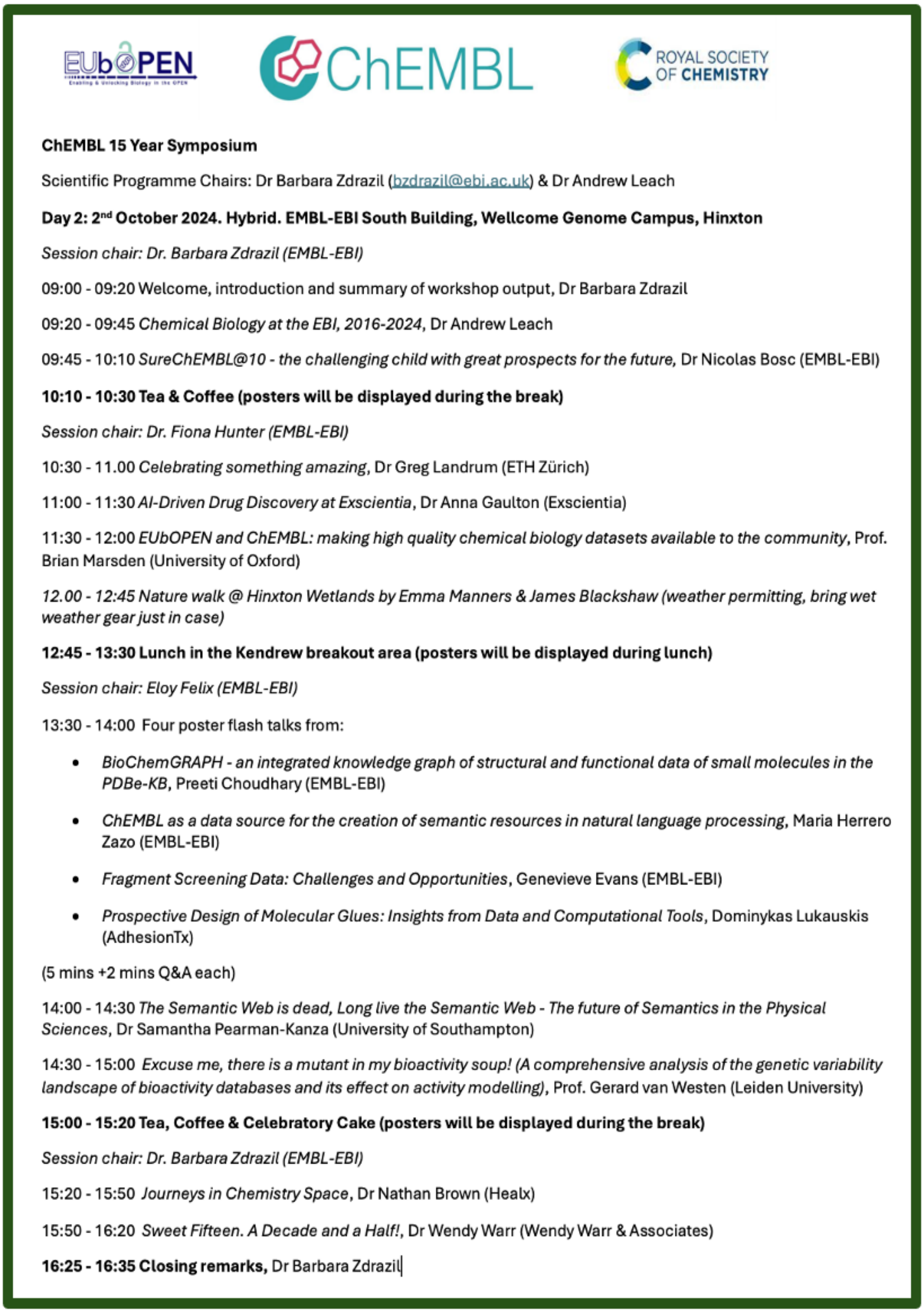
Four in-person workshops ran on Day 1 of the ChEMBL@15 symposium. (Image edited and reproduced with permission from B. Zdrazil, EMBL-EBI.)

Workshops 1 and 2:
(1) Extracting bioactivity data for drug-like compounds from ChEMBL (basic) and (2) ChEMBL: A practitioner's perspective (advanced)
In the morning session, two complementary ChEMBL workshops were run to demonstrate workflows to access ChEMBL's core data, either through the interface (beginner's workshop) or programmatically (advanced workshop). Within the beginner's session, the importance of thorough data checks, including clean-up and tailored curation, were highlighted as critical for high quality analyses and explainability.

Workshop 3:
Introduction to the ChEMBL load process (data deposition)
The core medicinal chemistry literature continues to provide a large source of new data. However, direct depositions constitute an increasing proportion of ChEMBL's bioactivity collection. ChEMBL balances a flexible format to handle varied data with guidance on best practice to ensure consistency across the platform. In recent years, data provenance, structure, and FAIRness has moved to the forefront of bio- and cheminformatics. Therefore, guidelines for deposition to ChEMBL are key. This workshop, jointly led by EMBL-EBI and EUbOPEN colleagues, addressed the deposition process, best practice, and common pitfalls, from the perspective of both the depositor and repository.

***Workshop 4*:**
Navigating Patent Data with SureChEMBL
The SureChEMBL team provided an overview of their patent database which has recently undergone an interface upgrade and implementation of a suite of new biomedical annotations. Focus was on different methods for searching relevant patent data. SureChEMBL has recently reached its 10-year milestone and shared in our celebrations.

Day 2: Presentations and posters
On the second day, we heard excellent presentations from speakers with a wealth of experience using ChEMBL, plus poster sessions, networking opportunities and (of course) the ChEMBL cake! The event was an opportunity to reflect on the contribution of ChEMBL to the field, reconnect with past and present colleagues and collaborators, and to discuss recent advances and future goals of open chemistry data in general. Presentations provided the highlights of ChEMBL's progress and impact. Andrew Leach spoke about major achievements of his team towards delivering ChEMBL, SureChEMBL, and ChEBI during the past eight years, and Nicolas Bosc gave a timely update on progress made for the SureChEMBL database. Speakers included both academic and industrial representatives. From the academic perspective, Greg Landrum's talk delved deeper into the challenges inherent in open-access data and the importance of carefully defined datasets and appropriate analyses.[1] Gerard van Westen's analysis of compound-target interactions, within the context of protein variants, offered another insight into the significance of (often) overlooked annotations and the identification of trends in variant-selective compounds.[2] On the other hand, Anna Gaulton delivered some interesting insights into AI-driven drug discovery at Exscientia and Nathan Brown spoke about his "journeys in chemistry space". Brian Marsden presented outcomes from the EUbOPEN project that enable, via a close collaboration of the Structural Genomics Consortium (SGC) with the ChEMBL team and the University of Oxford, a streamlined deposition of chemical probe data into ChEMBL. In her talk, Samantha Pearman-Kanza invited us to join a short journey into the Physical Sciences and the Semantic Web. Last but not least, Wendy Warr gave a brilliant final presentation on "A Decade and a Half" of ChEMBL.


EUbOPEN
Enabling & Unlocking Biology in the OPEN
ChEMBL
ROYAL SOCIETY OF CHEMISTRY

ChEMBL 15 Year Symposium

Scientific Programme Chairs: Dr Barbara Zdrazil (bzdrazil@ebi.ac.uk) & Dr Andrew Leach

Day 2: 2nd October 2024. Hybrid. EMBL-EBI South Building, Wellcome Genome Campus, Hinxton

Session chair: Dr. Barbara Zdrazil (EMBL-EBI)

09:00 - 09:20 Welcome, introduction and summary of workshop output, Dr Barbara Zdrazil

09:20 - 09:45 *Chemical Biology at the EBI, 2016-2024*, Dr Andrew Leach

09:45 - 10:10 *SureChEMBL@10 - the challenging child with great prospects for the future*, Dr Nicolas Bosc (EMBL-EBI)

10:10 - 10:30 Tea & Coffee (posters will be displayed during the break)

Session chair: Dr. Fiona Hunter (EMBL-EBI)

10:30 - 11.00 *Celebrating something amazing*, Dr Greg Landrum (ETH Zürich)

11:00 - 11:30 *AI-Driven Drug Discovery at Exscientia*, Dr Anna Gaulton (Exscientia)

11:30 - 12:00 *EUbOPEN and ChEMBL: making high quality chemical biology datasets available to the community*, Prof. Brian Marsden (University of Oxford)

12.00 - 12:45 Nature walk @ Hinxton Wetlands by Emma Manners & James Blackshaw (weather permitting, bring wet weather gear just in case)

12:45 - 13:30 Lunch in the Kendrew breakout area (posters will be displayed during lunch)

Session chair: Eloy Felix (EMBL-EBI)

13:30 - 14:00 Four poster flash talks from:

- *BioChemGRAPH - an integrated knowledge graph of structural and functional data of small molecules in the PDBe-KB*, Preeti Choudhary (EMBL-EBI)
- *ChEMBL as a data source for the creation of semantic resources in natural language processing*, Maria Herrero Zazo (EMBL-EBI)
- *Fragment Screening Data: Challenges and Opportunities*, Genevieve Evans (EMBL-EBI)
- *Prospective Design of Molecular Glues: Insights from Data and Computational Tools*, Dominykas Lukauskis (AdhesionTx)

(5 mins +2 mins Q&A each)

14:00 - 14:30 *The Semantic Web is dead, Long live the Semantic Web - The future of Semantics in the Physical Sciences*, Dr Samantha Pearman-Kanza (University of Southampton)

14:30 - 15:00 *Excuse me, there is a mutant in my bioactivity soup! (A comprehensive analysis of the genetic variability landscape of bioactivity databases and its effect on activity modelling)*, Prof. Gerard van Westen (Leiden University)

15:00 - 15:20 Tea, Coffee & Celebratory Cake (posters will be displayed during the break)

Session chair: Dr. Barbara Zdrazil (EMBL-EBI)

15:20 - 15:50 *Journeys in Chemistry Space*, Dr Nathan Brown (Healx)

15:50 - 16:20 *Sweet Fifteen. A Decade and a Half!*, Dr Wendy Warr (Wendy Warr & Associates)

16:25 - 16:35 Closing remarks, Dr Barbara Zdrazil

Presentations on Day 2 of the ChEMBL@15 symposium. (Image edited and reproduced with permission from B. Zdrazil, EMBL-EBI.)

ChEMBL: fifteen years of growth in drug discovery data

The ChEMBL database is focused on drug discovery and was accredited as a Global Core Biodata Resource in 2022. It contains curated data for approved drugs, clinical candidates and other drug-like preclinical compounds, alongside their physicochemical and biorelevant properties. Compounds are mapped to biological targets and their activity (both positive and negative) is captured to provide a profile of their biological impact. Comprehensive annotation of targets with their properties, structural features and active small molecules offers a target-centric view of the data. Since its acquisition by EMBL-EBI in 2009, there have been 34 releases of ChEMBL. Changes to the size, content, and scope of ChEMBL have been substantial. From version 1 to 34, compound numbers have quintupled from 440k to 2.4M, the database has increased from 15 to 78 tables reflecting the increasingly comprehensive data coverage, and content has expanded to include agrochemicals, pesticides, veterinary medicines and drug metabolism data providing an insight into the chemical space of bioactive small molecules. ChEMBL captures the medicinal chemistry literature and therefore reflects trends in drug discovery. Notable examples include the increase in bifunctional small molecules such as those eliciting targeted protein degradation (i.e. PROTACs), and biological drugs and gene therapies of increasing complexity. Biological targets may be nucleic acids or specific protein variants, illustrating the influence of genomics and personalised medicine in drug discovery, or previously considered 'undruggable' proteins or protein-protein interactions that new chemical modalities may address.

Engagement with the chemistry community

ChEMBL's primary goal is the provision of chemistry data to the industrial and academic drug discovery community; engagement is central to this aim. Our on-demand and live training materials, as well as a dedicated ChEMBL Helpdesk, offer general and individual-level support enabling access to our database. Attendance at conferences and workshops, coupled with regular communications through our blog, social media and mailing list, maintains contact with the community leading to productive collaborations and ideas that have shaped ChEMBL. Regular publication of updates and enhancements allows users to follow changes [for recent data highlights see our recent update paper in NAR (3)]. Engagement with the public is also core to our values and the team is committed to regular involvement in public engagement initiatives including our own activities such as the on-site 'nature and chemistry' trail.

*The Chemistry and Nature Trail.
In 2024, we also said farewell to Andrew Leach and the ChEMBL@15 symposium offered a chance to further thank him for his huge contribution to the Chemical Biology Services at EMBL-EBI.
(Credit P.Mynott. Image reproduced with permission from EMBL-EBI.)*

References

(1) Landrum, G. A.; Riniker, S. Combining IC_{50} or K_i Values from Different Sources Is a Source of Significant Noise. *J. Chem. Inf. Model*. **2024,** *64* (5), 1560-1567. DOI: 10.1021/acs.jcim.4c00049

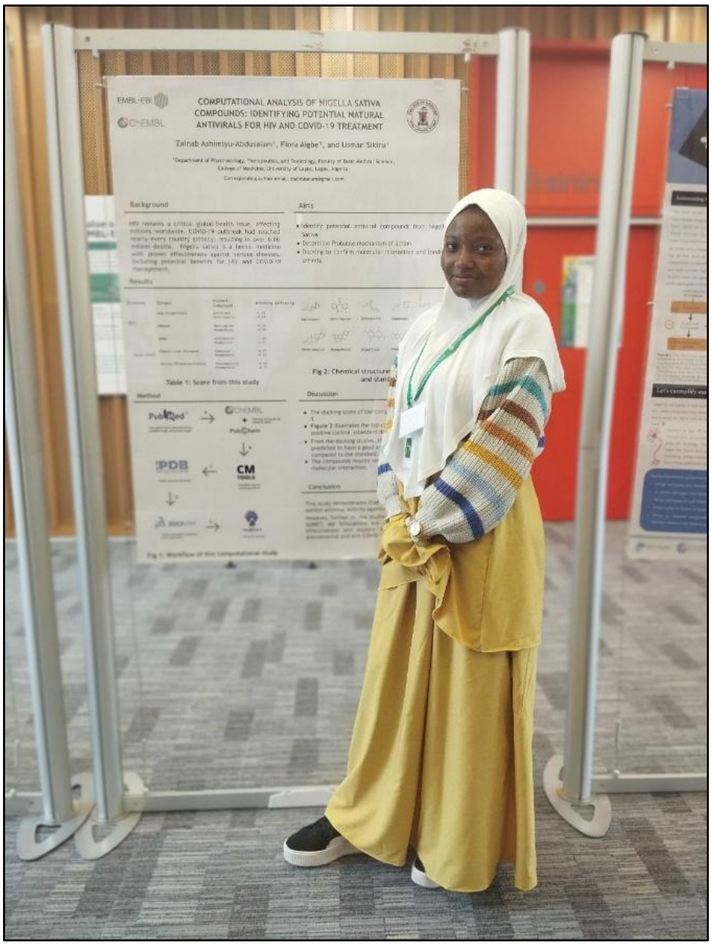
(2) Gorostiola González, M. et al. Excuse me, there is a mutant in my bioactivity soup! https://chemrxiv.org/engage/chemrxiv/article-details/66729e49c9c6a5c07ad1b0a8

(3) Zdrazil, B. et al. The ChEMBL Database in 2023: a Drug Discovery Platform Spanning Multiple Bioactivity Data Types and Time Periods. *Nucleic Acids Res.* **2024**, *52*, D1180– D1192. DOI: 10.1093/nar/gkad1004

- -

Meeting Report: ChEMBL@15 Symposium

Contribution from Zainab Ashimiyu-Abdusalam, Trainee at ChEMBL, EMBL-EBI, email: ziabdusalam@gmail.com, zainab@ebi.ac.uk


COMPUTATIONAL ANALYSIS OF NIGELLA SATIVA COMPOUNDS: IDENTIFYING POTENTIAL NATURAL ANTIVIRALS FOR HIV AND COVID-19 TREATMENT
Zainab Ashimiyu-Abdusalam¹, Flora Aigbe¹, and Usman Sikiru¹
Background
Aims
Results
Table 1: Score from this study
Method
Discussion
Conclusion

In October 2024, I had the opportunity to attend the ChEMBL 15 Year Symposium which took place at the EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridge. It was a great symposium, and I am grateful to have received a travel bursary from the Royal Society of Chemistry's Chemical Information and Computer Applications Group. This support enabled me to attend the meeting, where I learnt a great deal about the effort it takes to create datasets that go into a lot of drug discovery projects today, present my research findings, and also enjoy inspiring talks from distinguished scientists in the field, some of whom I've followed online to learn. It was a great experience to learn about their journey, and their achievements.

My name is Zainab Ashimiyu-Abdusalam. I completed my bachelor's degree at the University of Lagos, Nigeria, where my thesis explored the potential of natural compounds in *Nigella sativa* for managing SARS-CoV-2 and HIV using molecular docking studies. At the symposium I presented a poster outlining my findings, focusing on identifying active phytochemicals, their mechanisms of action, binding modes, and interaction methods. The study suggested potential therapeutic applications of these compounds, particularly for HIV and COVID-19, while emphasising the need for further *in vivo* validation. Currently, I am involved in annotating action-types for active compound-target pairs in ChEMBL as part of a short internship programme. This experience has deepened my understanding of the intricacies of data curation, a theme that resonated throughout the symposium.

The talks were both enlightening and inspiring. Dr Greg Landrum highlighted the importance of data quality and the responsibility of users in ensuring accurate results. Dr Andrew Leach and Dr Nicolas Bosc shared the evolution of ChEMBL and SureChEMBL, while Dr Nathan Brown detailed his extensive journey in the chemical sciences. Dr Samantha Pearman-Kanza's session on the semantic web and the challenges of integrating multiple ontologies struck a chord, particularly as I've encountered similar issues in database mapping. The symposium concluded with a heart-warming tribute to ChEMBL by Dr Wendy Warr, celebrating its contributions and looking forward to its future.

As my first in-person scientific symposium, this experience was incredibly impactful. I am deeply thankful to the RSC CICAG for their generous support and to the ChEMBL committee, especially Dr Barbara Zdrazil and

Ms Clare Kim Palferman, for their guidance during the abstract submission and grant application process. Here's to many more impactful years for ChEMBL!

- -

Open-Source Cheminformatics Toolkits

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

The increased interest in machine learning and artificial intelligence (ML/AI) has highlighted one of the key challenges for data analysis involving molecules, how to encode molecules in the computer in a chemically intelligent manner. Fortunately, there are several open-source toolkits that can be used and there is no need to create another.

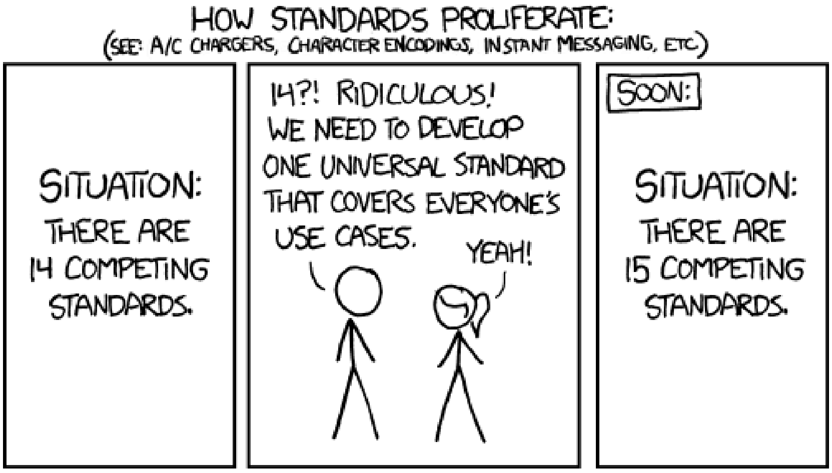

HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC)
SITUATION: THERE ARE 14 COMPETING STANDARDS.
14?! RIDICULOUS! WE NEED TO DEVELOP ONE UNIVERSAL STANDARD THAT COVERS EVERYONE'S USE CASES.
YEAH!
SOON:
SITUATION: THERE ARE 15 COMPETING STANDARDS.

Image credit XKCD; creative commons licence CC BY-NC 2.5.

OpenBabel

An open chemical toolbox, Open Babel presents a solution to the proliferation of multiple chemical file formats. In addition, it provides a variety of useful utilities from conformer searching and 2D depiction, to filtering, batch conversion, and substructure and similarity searching. For developers, it can be used as a programming library to handle chemical data in areas such as organic chemistry, drug design, materials science, and computational chemistry. There are also cheminformatics nodes for KNIME.

- Extensively used in nearly 50 projects. Installs available for Linux, MacOSX and Windows
- OpenBabel is written in C++ and source code is available. Bindings are also available to allow scripting access using Java, .NET, Perl, Python or Ruby. Licence GNU GPL
- Keep up to date by subscribing to the openbabel-discuss list
- Reference: O'Boyle, N.M. et al. Open Babel: An open chemical toolbox. *J Cheminform.* **2011**, *3*, 33. https://doi.org/10.1186/1758-2946-3-33

RDKit

The RDKit is an open-source toolkit for cheminformatics, 2D and 3D molecular operations, descriptor generation for machine learning, etc. There's also a molecular database cartridge for PostgreSQL and cheminformatics nodes for KNIME (distributed from the KNIME community site).

- Installs available for Linux, MacOSX and Windows. The RDKit core algorithms and data structures are written in C++. Wrappers are provided to use the toolkit from either Python (2.x and 3.x), Java, or C#.
- Licence: BSD
- Source code

CDK

The Chemistry Development Kit (CDK) is a collection of modular Java libraries for processing chemical information (cheminformatics). The modules are free and open-source and are easy to integrate with other open-source or in-house projects. Also cheminformatics nodes for KNIME.

- The latest release JAR with all dependencies included from GitHub
- CDK is written in Java
- Licence: GNU Lesser General Public License, version 2.1 (or later)
- Source code
- Reference: Willighagen, E.L. et al. The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *J. Cheminform.* **2017**, *9*(3), https://doi.org/10.1186/s13321-017-0220-4

Indigo

Indigo is a universal molecular toolkit that can be used for molecular fingerprinting, substructure search, and molecular visualisation. Also capable of performing a molecular similarity search, it is 100% open source and provides enhanced stereochemistry support for end users, as well as a documented API for developers. Also cheminformatics nodes for KNIME.

- Installs available for Linux, MacOSX and Windows
- Indigo is written in C++ and source code is available; bindings are also available to allow scripting access using Java, .NET, Python
- Licence: GNU General Public Licence

OpenChemLib

Open source Java-based chemistry library, also openchemlib-js JavaScript interface with the openchemlib java library.

- Licence
- Source code

ChemDoodle Web Components

The ChemDoodle Web Components library is a pure Javascript chemical graphics and cheminformatics library derived from the ChemDoodle® application and produced by iChemLabs. ChemDoodle Web Components allow the developer to present publication quality 2D and 3D graphics and animations for chemical structures, reactions and spectra.

- Licence: GNU Public License (v3.0)
- Source code

Kekule.js

Kekule.js is an open source JavaScript library for cheminformatics released under MIT licence. Currently, it is molecule-centric, focusing on providing the ability to represent, draw, edit, compare and search molecule structures on web browsers.

- Licence: MIT
- Source code

WebMolKit

Cheminformatics toolkit built with TypeScript. Can be used to carry out some fairly sophisticated cheminformatics tasks on a contemporary web browser, such as rendering molecules for display, format conversions, calculations, interactive sketching, among other things. The library can be used within any JavaScript engine, including web browsers, NodeJS and Electron.

- Demo of molecular sketcher
- Written in TypeScript. Requires the TypeScript compiler (tsc) to cross-compile into JavaScript
- Apache 2.0 licence
- Source code

Chempy

A Python package useful for chemistry (mainly physical/inorganic/analytical chemistry).

- Numerical integration routines for chemical kinetics (ODE solver front-end)
- Integrated rate expressions (and convenience fitting routines)
- Solver for equilibria (including multiphase systems)
- Expressions in physical chemistry
- Author: Bjoern I. Dahlgren
- Licence: BSD
- Source code

ChemmineR

A cheminformatics package for analysing drug-like small molecule data in R. Its latest version contains functions for efficient processing of large numbers of small molecules, physicochemical/structural property predictions, structural similarity searching, classification and clustering of compound libraries with a wide spectrum of algorithms.

- Authors Kevin Horan, Yiqun Cao, Tyler Backman, Thomas Girke
- Licence: Artistic-2.0
- Source code

MolecularGraph.jl

A graph-based molecule modelling and cheminformatics analysis toolkit fully implemented in Julia.

- Author: Seiji Matsuoka
- Licence: MIT
- Source code

LillyMol

A C++ library for cheminformatics. This repo also contains a variety of useful command line tools, built with LillyMol. LillyMol does only a subset of cheminformatics tasks, but tries to do those tasks efficiently and correctly. LillyMol has some novel approaches to substructure searching, reaction enumeration and chemical similarity. These have been developed over many years, driven by the needs of computational and medicinal chemists at Lilly and elsewhere. LillyMol is fast and scalable, with modest memory requirements. This release includes a number of C++ unit tests. All tests can be run with address sanitiser, with no problems reported. The file Molecule_Tools/introduction.cc provides an introduction to LillyMol for anyone wishing to develop with C++.

- Authors: Xuyan Ru, Ian Watson, G-Huang
- Licence: Apache 2.0
- Source code

Chython

Library for processing molecules and reactions in python way. Chython is fork of CGRtools for which there has been no development for several years, but could be reinitiated. Features: Read/write/convert formats: MDL .RDF (.RXN) and .SDF (.MOL), .MRV, SMILES, InChI (InChI Trust library), .XYZ, .PDB.Standardize molecules and reactions and valid structures checker; supported python-magic; tetrahedron, allene and cistrans stereo supported; perform subgraph search; build/edit molecules and reactions with Python API; produce template based reactions and molecules atom-to-atom mapping, checking and rule-based fixing; perform MCS search 2D coordinates generation (based on SmilesDrawer); 2D/3D depiction with Jupyter support; SMARTS parser with restrictions; protective groups remover; common reaction templates collection.

- Authors Ramil Nugmanov
- Licence: GNU Lesser General Public License version 3
- Source code

CDPkit

CDPKit (short for Chemical Data Processing Toolkit) is an open-source cheminformatics toolkit implemented in C++. CDPKit comprises a suite of software tools and a programming library called the Chemical Data Processing Library (CDPL) which provides a high-quality and well-tested modular implementation of basic functionality typically required by any higher-level software application in the field of cheminformatics. In addition to the CDPL C++ API, an equivalent Python-interfacing layer is provided that allows to harness all of CDPL's functionality easily from Python code.

- Author Thomas Seidel
- Licence: GNU Lesser General Public License version 2 or later
- Source code

- -

Advancing Transparency in Chemical Sciences: an Update to the *Digital Discovery*'s Data and Code Policy

Contribution from Anna Rulka, Executive Editor, Royal Society of Chemistry, email rulkaa@rsc.org, DigitalDiscovery-rsc@rsc.org

GOLD OPEN ACCESS
Digital Discovery
A journal for new thinking on machine learning, robotics and AI

In the frequently changing world of scientific research, there is a clear need for transparency and reproducibility. In response to these changes, *Digital Discovery*, the Royal Society of Chemistry's (RSC) journal on AI and automation in scientific discovery, has taken a step forward with a significant update to its data and code availability policy. The new policy sets clearer rules for how the data and code are shared and aims to address the need for accessible and reproducible research. This step reflects the RSC's ongoing commitment to advancing chemical sciences while upholding the highest standards of research integrity and transparency in the scientific process.

At its core, our policy reflects our commitment to open science and the FAIR (Findable, Accessible, Interoperable, and Reusable) principles for data management. This commitment is especially important in the context of high-throughput computational methodologies where complex datasets are often used to train the models. Through increasing the ease of access to data and code we support researchers in building on the previously published work and advancing scientific discovery even faster.

Prior to the update in our data policy, researchers submitting to *Digital Discovery* were encouraged, but not required, to submit their code and data to publicly recognised repositories. This approach lacked the consistency needed to ensure that every manuscript was accompanied by the representative data and code necessary for verification and replication.

In the current policy, authors are required to submit their code and datasets to persistent repositories like Zenodo, Code Ocean, or Mendeley as part of the manuscript acceptance process and ensure that their data and code have been assigned Digital Object Identifiers (DOIs) before publication. This makes the data and code associated with a manuscript citable and easy to locate, even years after the publication.

At the same time, similarly to our previous policy, *Digital Discovery* referees must be provided access to the code and data during the peer-review process. This ensures that reviewers can properly assess the validity and reproducibility of the findings, further increasing the credibility of the research published in the journal.

While openness is a foundation of our updated policy, we are conscious of the complexities of industry-funded research, which often faces unique challenges, particularly when it comes to sharing sensitive code or datasets that may be protected by intellectual property rights. These concerns are addressed in the new policy and thus in cases where full code cannot be made publicly available, authors are allowed to provide pseudocode instead. Where appropriate, a binary version of the software is still required to be shared confidentially with editors and reviewers for testing. This ensures that even when full access to the code isn't possible, reviewers can still assess the functionality of the software and allows us to hold industry-related research to the same standards of transparency and reproducibility.

One of the other significant changes in the updated policy is the requirement for authors to provide complete datasets alongside their code while, in the past, authors were strongly encouraged to publicly share as much data as possible. In cases where sharing the full dataset is not possible due to privacy or other restrictions, authors are now required to provide a representative sample that allows others to reach similar conclusions. This ensures that the findings of the study can be verified and that other researchers can still build on the work.

By mandating data and code deposition, we emphasise that reproducibility is fundamental to scientific progress. It is no longer enough to publish findings without the necessary resources to verify them. The new policy requires data to be available in a form that can be used to reproduce the results.

In the updated policy we also acknowledge the rise of artificial intelligence (AI) and large language models (LLMs) in scientific research by introducing new guidelines for researchers working with these technologies. To ensure the reproducibility of research involving LLMs, we now require authors to provide log files detailing the inputs and outputs used in their studies, enhancing the transparency of the study. When using commercial LLMs, authors must also specify which model was used and its generation date, as these models may change over time. We are establishing clear guidelines for using commercial LLMs to promote responsible and reproducible research and ensure that studies can be validated and built upon.

Researchers will continue to leverage digital methods and artificial intelligence, and it is our responsibility to not only improve research integrity but also encourage a more collaborative scientific community. We believe this can be achieved by implementing updated data and code deposition policies.

- -

Catalyst Science and Discovery Centre: Celebrations, Revelations and Donations

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

In the Summer 2024 *Newsletter* I wrote about the Synergy Project in which Catalyst has been involved for the last 18 months. I said that we had submitted a Delivery Phase bid to the National Heritage Lottery Fund for just over £1 million on 24 May 2024 and were waiting to hear whether we had been successful. We waited all summer for news which we hoped to have in September 2024. I am pleased to say that we were successful but the news had to be kept under wraps until the National Lottery agreed that it could be made public. That happened on 9 October 2024. Thanks to all of you who congratulated us on our success. NLHF were pleased that we had managed to raise 10% of the total as grant funding from several organisations which is a mandatory requirement.

The real hard work begins in early January 2025 as we start to put our plans in to action which will take over two years. Catalyst won't close as the heritage work is just part of what we do and our STEM education delivery will continue throughout. A strategy group has been created under the leadership of one of the Trustees to work out the logistics and timing of everything. New staff, such as a Volunteer Coordinator and a Heritage Officer, will need to be recruited and jobs will be advertised soon.

We had decided in the summer that we needed to have an event at Catalyst for our Patrons and supporters and at that, if we were successful, we could celebrate the success of our bid and explain to everyone what we would be doing. The date chosen back in June 2024 was 15 November but we couldn't issue invitations until we could make the news about the bid public.

From mid October 2024 I personally sent out over 130 invitations apologising about the short notice and hoping invitees would come. Thankfully we had over 50 people there from industry, public offices, Patrons, etc. and I am pleased to say that Helen Cooke was able to join us representing CICAG. Neil Berry was invited but was at a meeting in London. The Vice-Chancellor of the University of Liverpool, Professor Tim Jones, couldn't attend but Professor Karl Coleman, Dean of Science and Engineering did. We also had to raise money as our costs are escalating and will increase as they are for all museums, given the rise in minimum wage and the fact that employers have to bear the increases in National Insurance (NI). We decided we would concentrate on trying to persuade attendees to sponsor a chemical element on our Interactive Periodic Table and one of our Patrons gave a speech explaining why he and his family supported Catalyst and tried to encourage others to do the same. It was successful and we had the sponsoring of several elements by those present including Helen and the Lord Lieutenant of Cheshire, Lady Redmond and her husband Sir Phil Redmond (do you remember Grange Hill and Brookside?).

One special guest was Professor Gill Reid FRSC, FRSE, immediate Past President of the Royal Society of Chemistry who came up specially from Southampton where she is a Professor in the Chemistry Department (one of Jeremy Frey's colleagues) to see Catalyst. She had never been there before. I am delighted to say that she was very pleased with what she saw and heard and has agreed to become a Patron of Catalyst.

Professor Gill Reid, Past President of the Royal Society of Chemistry, with our Chair of Trustees Hugh Dowding and Emeritus Professor David Hornby (Sheffield).

We had been involved for several months in the Halton 50 celebrations as it was 50 years since the Borough of Halton was created in 1974. A small grant had enabled us to have a mural created by two local artists which depicted events surrounding Catalyst during that time. Fortuitously it was installed on the wall in the café on 13 November and the Mayor of Halton, Councillor Kevan Wainwright, agreed to symbolically unveil it, which he did. Kevan is one of our Trustees but on sabbatical while he is the Mayor. He is also a shift supervisor at Castner Kellner Works of INEOS/Inovyn where they still make chlorine as they did back in 1897 by the electrolysis of brine pumped from central Cheshire. Kevan was photographed with three attendees all of whose great grandfathers had been mayors of Widnes (1892, 1911 and 1917-1918) and one being our Patron, Peter Gossage was allowed to wear the chain his great grandfather had been the first to wear in 1892.

Mayor of Halton and the artist in front of the Halton 50 Mural.

Visit to the Catalyst Archives by attendees during the celebratory event.

Sir Hugo Brunner, another of our Patrons, brought more archives and artefacts from his home in Oxfordshire to add to our heritage collections. All were related to the chemical industry and Cheshire. So 15 November was a very successful celebratory day.

Our Heritage and Collections Manager/Archivist Judith Wilde continues to receive new material for our archives virtually every day. Recently she visited part of Astra Zeneca at Macclesfield in Cheshire and met with the archivist there. She was delighted to return to Catalyst with a set of publications called *Fulshaw Times*. These were a publication of Central Toxicology Laboratory (CTL) which was based at Fulshaw Hall near Alderley Park and were an ICI publication we did not have. The volunteers who she organises are already digitising these publications.

Our next 'Chemistry at Work Week' for secondary school pupils is due to take place in the last week of January 2025. The Institute of Chemical Engineers (IChemE) and specialty chemicals company Lanxess are providing the financial support on this occasion. It is already fully booked by schools and industrial firms.

Some attendees at a workshop in Catalyst's Laboratory with Lucinda, our Education Manager.

Catalyst is determined to become more inclusive in its provision and to bring under-represented groups in to attend its workshops and shows at weekends and during the holidays. I am delighted to report that my application for a Large Grant to the RSC Outreach Fund for money to support a project entitled 'IDHHP: Interpreters for Deaf and Hard of Hearing People' was successful. We have been working with the local Deafness Resource Centre to encourage and enable deaf people to come to our workshops and shows but with a BSL Interpreter present. So far we have had four workshops and there are two more to come in January and February 2025. A lot has been learnt by participating in these and I have attended them all as they are being delivered by our Education Team members including what works and what doesn't work. I have to do the evaluation of this project and the use of the grant in late March 2025.

On 18 January 2025 it will be 200 years since a renowned Victorian chemist, Sir Edward Frankland FRS, KCB was born at the little village of Caterall near Garstang in Lancashire. Not only is he known as the 'Father of Organometallic Chemistry', the namer of the element Helium, and the person who defined the theory of valence, but is also renowned for his work on water analysis and water purification from the 1870s onwards. He is credited with saving the lives of thousands of people through his work. He was also the first professor of chemistry at Owens College in Manchester when it opened in 1851. I am putting together a small exhibition about him which will run for the month of January 2025 at Catalyst, so more in the next issue of the *CICAG Distillate* about this event.

So lots of activities happening at Catalyst. Do come and visit us if you are in the area. You can find out more about Catalyst by visiting our website.

- -

Certara Completes Acquisition of Chemaxon

Press release

Radnor, PA, 2 October 2024: Certara, Inc. (Nasdaq: CERT), a global leader in model-informed drug development, today announced it has completed the acquisition of Chemaxon.

The combined organisation offers life sciences companies predictive biosimulation and scientific informatics capabilities, improving certainty in decision-making from discovery through commercialisation.

Chemaxon develops leading scientific informatics software products used by the life sciences industry for *in silico* research. Certara develops advanced modelling and biosimulation solutions used to predict the pharmacokinetic and pharmacodynamic properties of large and small molecules. “Combining Chemaxon’s expertise with Certara’s biosimulation capabilities provides life sciences companies with unique solutions to enhance productivity and increase their scientific innovation success rates,” said William Feehery, Certara’s CEO. “Together, we offer scientists more precise insights throughout drug discovery and development.”

Near-term priorities for the combined organisation include incorporating precision chemistry structures, calculators, and predictors into the Certara D360 scientific informatics applications and Simcyp™ PBPK Simulator for improved prediction accuracy. Longer-term plans include leveraging Certara.AI’s life sciences specialised GPT capabilities and bringing knowledge of pharmacokinetics and pharmacodynamics more broadly into the drug discovery process and Chemaxon products including Design Hub and JChem Engines.

“Our teams are excited to have an even greater impact on drug discovery and development practices,” said Richard Jones, Chemaxon CEO. “As pipelines shift to precision medicine therapies, accurate scientific predictions and biosimulation are more crucial to success than ever.”

In 2024, Chemaxon is expected to generate software revenue greater than $20 million. Certara will update its 2024 guidance to include the contribution from Chemaxon when the Company reports third-quarter earnings in November.

A frequently asked questions document regarding the transaction is available on the Company’s investor relations website.

About Certara
Certara accelerates medicines using biosimulation software, technology, and services to transform traditional drug discovery and development. Its clients include more than 2,400 biopharmaceutical companies, academic institutions, and regulatory agencies across 66 countries. Learn more at certara.com.

About Chemaxon
Chemaxon is a leading cheminformatics company that provides platforms, applications, and solutions to handle chemical entities in life sciences, biotechnology, agrochemicals, new materials, education, and other research industries. Its products and services help the capture and processing of chemical information that increases its value and results in more efficient decision-making for life sciences and other R&D environments. Learn more at Chemaxon.

- -

They Cracked the Code for Proteins' Amazing Structures

Press release: The Nobel Prize in Chemistry 2024

The Royal Swedish Academy of Sciences has decided to award the Nobel Prize in Chemistry 2024 with one half "for computational protein design" to:

David Baker
University of Washington, Seattle, WA, USA
Howard Hughes Medical Institute, USA

and the other half jointly "for protein structure prediction" to:

Demis Hassabis
Google DeepMind, London, UK

and

John M. Jumper
Google DeepMind, London, UK

The Nobel Prize in Chemistry 2024 is about proteins, life's ingenious chemical tools. David Baker has succeeded with the almost impossible feat of building entirely new kinds of proteins. Demis Hassabis and John Jumper have developed an AI model to solve a 50-year-old problem: predicting proteins' complex structures. These discoveries hold enormous potential.

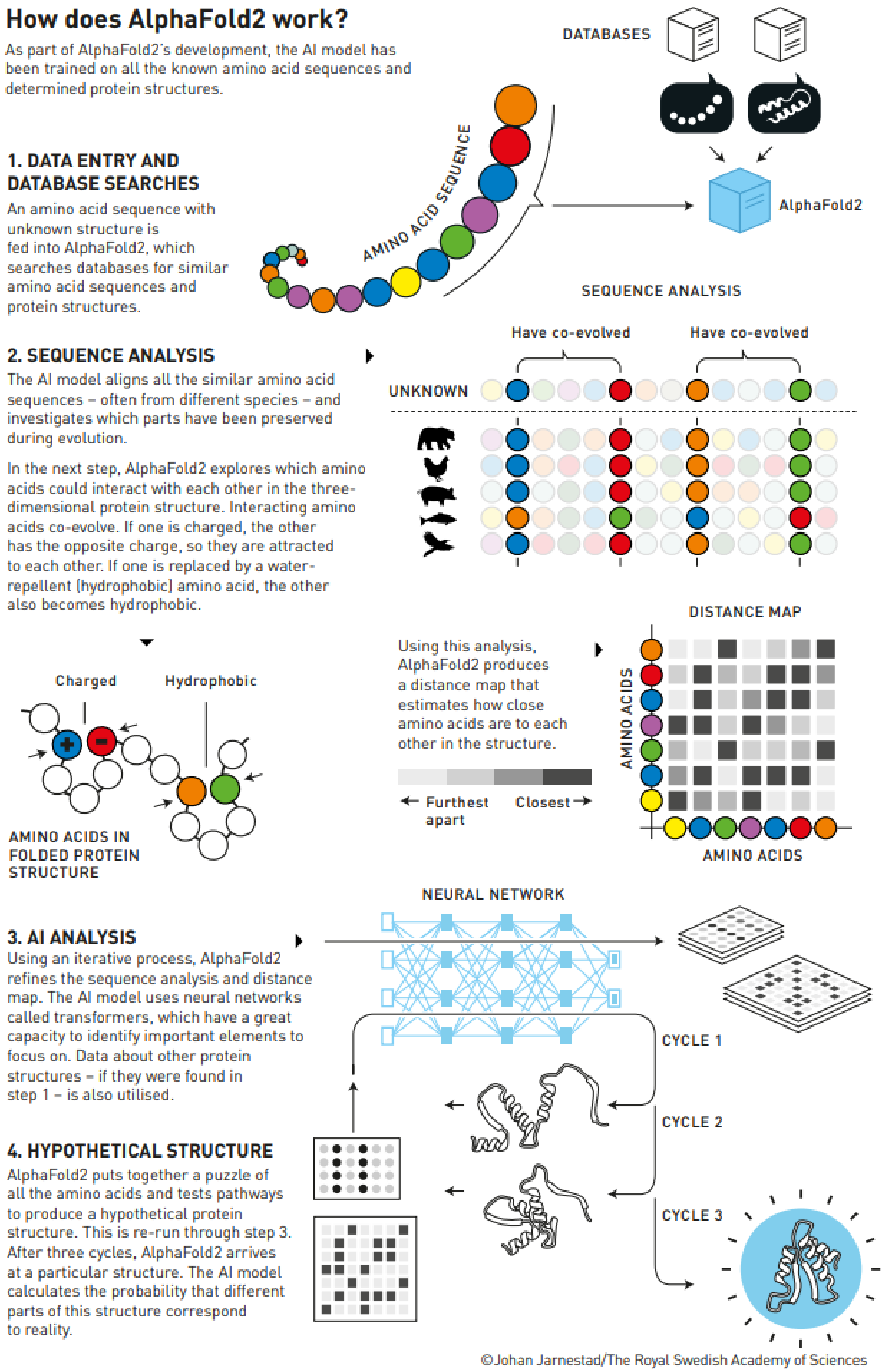
The diversity of life testifies to proteins' amazing capacity as chemical tools. They control and drive all the chemical reactions that together are the basis of life. Proteins also function as hormones, signal substances, antibodies and the building blocks of different tissues.

"One of the discoveries being recognised this year concerns the construction of spectacular proteins. The other is about fulfilling a 50-year-old dream: predicting protein structures from their amino acid sequences. Both of these discoveries open up vast possibilities," says Heiner Linke, Chair of the Nobel Committee for Chemistry.
Proteins generally consist of 20 different amino acids, which can be described as life's building blocks. In 2003, David Baker succeeded in using these blocks to design a new protein that was unlike any other protein. Since then, his research group has produced one imaginative protein creation after another, including proteins that can be used as pharmaceuticals, vaccines, nanomaterials and tiny sensors.

The second discovery concerns the prediction of protein structures. In proteins, amino acids are linked together in long strings that fold up to make a three-dimensional structure, which is decisive for the protein's function. Since the 1970s, researchers had tried to predict protein structures from amino acid sequences, but this was notoriously difficult. However, four years ago, there was a stunning breakthrough.

In 2020, Demis Hassabis and John Jumper presented an AI model called AlphaFold2. With its help, they have been able to predict the structure of virtually all the 200 million proteins that researchers have identified. Since their breakthrough, AlphaFold2 has been used by more than two million people from 190 countries. Among a myriad of scientific applications, researchers can now better understand antibiotic resistance and create images of enzymes that can decompose plastic.

Life could not exist without proteins. That we can now predict protein structures and design our own proteins confers the greatest benefit to humankind.



How does AlphaFold2 work?

As part of AlphaFold2's development, the AI model has been trained on all the known amino acid sequences and determined protein structures.

1. DATA ENTRY AND DATABASE SEARCHES

An amino acid sequence with unknown structure is fed into AlphaFold2, which searches databases for similar amino acid sequences and protein structures.

2. SEQUENCE ANALYSIS

The AI model aligns all the similar amino acid sequences – often from different species – and investigates which parts have been preserved during evolution.

In the next step, AlphaFold2 explores which amino acids could interact with each other in the three-dimensional protein structure. Interacting amino acids co-evolve. If one is charged, the other has the opposite charge, so they are attracted to each other. If one is replaced by a water-repellent (hydrophobic) amino acid, the other also becomes hydrophobic.

Using this analysis, AlphaFold2 produces a distance map that estimates how close amino acids are to each other in the structure.

3. AI ANALYSIS

Using an iterative process, AlphaFold2 refines the sequence analysis and distance map. The AI model uses neural networks called transformers, which have a great capacity to identify important elements to focus on. Data about other protein structures – if they were found in step 1 – is also utilised.

4. HYPOTHETICAL STRUCTURE

AlphaFold2 puts together a puzzle of all the amino acids and tests pathways to produce a hypothetical protein structure. This is re-run through step 3. After three cycles, AlphaFold2 arrives at a particular structure. The AI model calculates the probability that different parts of this structure correspond to reality.

©Johan Jarnestad/The Royal Swedish Academy of Sciences

How does AlphaFold2 work? (pdf)
©Johan Jarnestad/The Royal Swedish Academy of Sciences

Notes from Chris Swain, CICAG Committee Chair
These awards highlight the critical contributions of machine learning/artificial intelligence (ML/AI) in all areas of science.

In June 2018 RSC-CICAG and RSC-BMCS organised the first 'Artificial Intelligence in Chemistry' meeting at the RSC headquarters Burlington House, attended by fewer than 100 people. In the years that have followed, this annual meeting has expanded to become a three-day event with 250 participants and including two days of presentations, posters and a workshop. The continued expansion of the meeting recognises the influence of ML/AI in all areas of chemistry.

It was perhaps fitting that the 7th RSC-CICAG/RSC-BMCS AI in Chemistry meeting in September 2024 was opened by John Jumper giving a keynote address.

The advances in protein structure prediction have been a highlight of the impact ML/AI in science and underline how these technologies can be used to address the most challenging of problems. Whilst the first iterations of AlphaFold AI-assisted protein prediction tools sought to predict the structure of single uncomplexed proteins, subsequent iterations included multimeric proteins, ligand bound and protein-protein interactions. These tools have now been used to produce the AlphaFold Protein Structure Database of over 200 million entries, covering the human proteome and for the proteomes of 47 other organisms important in research and global health. However, John Jumper was very clear that these advances rely of the curation efforts of the Protein Databank.

Access to high quality data is essential to facilitate the adoption of ML/MI in all areas of chemistry, such as molecular force fields, drug discovery, protein design, environmental issues, and energy production.

See also: How protein structure prediction and design won the Nobel prize. *Chemistry World*, **2024**, *21* (11), 24-29.

- -

Professor Matthias Rarey wins the Herman Skolnik Award for 2025

Press release, September, 2024

The American Chemical Society Division of Chemical Information is pleased to announce that Professor Matthias Rarey has been selected to receive the 2025 Herman Skolnik Award for his contributions to the development of foundational algorithms in cheminformatics, education and training in the field of cheminformatics and his activities bridging academia and industry.

Professor Rarey's research has concentrated primarily on methods for ligand-based and structure-based molecular design. His early work focused on the development of novel techniques for flexible ligand-protein docking and his postdoctoral research at SmithKlineBeecham resulted in the Feature Tree concept, an approach to the representation and searching of small molecules by reduced-graph structures that has found particular application for exploring combinatorial chemistry and fragment spaces, and for scaffold-hopping searches in chemical databases. More recent studies have led to novel approaches for the drawing and visualisation of molecular structures, chemical patterns, and biological macromolecules, for *de novo* design, for the analysis of protein binding sites and torsion-angle distributions, and for 3D shape similarity.

While much of his published work is accompanied by freely accessible software, he has exemplified translational research by commercialisation of his research via BioSolveIT GmbH, a scientific software company for virtual screening and lead discovery of which Professor Rarey was one of the co-founders in 2002. Examples of systems derived from his research include the FlexX protein-ligand docking program, PoseView for generating 2D diagrams of complexes with known 3D structures, FTrees for similarity searching in chemical spaces, HYDE for scoring hydrogen bond and dehydration energies in protein-ligand complexes, and ReCore for scaffold-hopping in lead-discovery programs. Recently, SpaceLight and SpaceMACS were added enabling topological similarity searching in combinatorial make-on-demand catalogues like Enamine REAL for the first time.

Professor Rarey obtained MSc and PhD degrees in computer science from Paderborn University and University of Bonn in 1992 and 1996 respectively. Following positions at GMD and the Fraunhofer Institute for Algorithms and Scientific Computing with outward stays at SmithKlineBeecham (Philadelphia) and Roche (Palo Alto), he became the founding director of the Center for Bioinformatics at the University of Hamburg in 2002. As Director, he established new teaching programs in bioinformatics (including cheminformatics) and in computing in science Many students from these programs have gone on to undertake PhD research and then to work in industry in the life sciences.

Within the Center, he leads the Research Group for Computational Molecular Design, focusing on the development of new algorithms for addressing problems in molecular design, in visualisation and in cheminformatics more generally. His commitment to education is evidenced by many of his students emerging as leaders in the field as a result of his mentorship. He is also one of the spokespersons of the DASHH Helmholtz Graduate School for the Structure of Matter jointly established with the Deutsches Elektronen-Synchroton DESY and the Technical University of Hamburg. From 2014-2023, Professor Rarey was an Associate Editor of the ACS *Journal of Chemical Information and Modeling*. He has published extensively, accruing more than 18,000 citations, and has received multiple awards, most notably the 2005 Corwin Hansch Award of the QSAR and Modelling Society and the Emerging Technologies Award of the ACS Division of Computers in Chemistry in 2011.

The prize consists of a $3,000 honorarium and a plaque. Professor Rarey will also be invited to organise an award symposium at the Fall 2025 ACS National Meeting to be held in Washington, D.C.

Robert D. Clark
Chair, ACS CINF Awards Committee

- -

UKeIG: 2024 Tony Kent Strix and Jason Farradane Award Winners

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

Tony Kent Strix Award

The UK electronic information Group (UKeiG) is pleased to announce that the winner of the prestigious international Tony Kent Strix Memorial Award for 2024 is one of the world's leading scientists in information retrieval, Nicola Ferro, Professor in Computer Science at the Department of Information Engineering of University of Padua. The award celebrates his distinguished research record and sustained, cumulative contribution to the field of information retrieval (IR).

The Strix award honours outstanding contributions to the field in memory of Tony Kent, a remarkable IR pioneer. It was inaugurated in 1998 by the Institute of Information Scientists and is now presented by UKeiG in partnership with the International Society for Knowledge Organisation UK (ISKO UK), the Royal Society of Chemistry Chemical Information and Computer Applications Group (RSC CICAG) and the British Computer Society Information Retrieval Specialist Group (BCS IRSG). Whereas at the time of Kent's ground-breaking achievements IR was a fascinating sideline for relatively few specialists, nowadays the Search function underpins everyday living at home, at work and across the globe. This Award is more widely relevant than ever before.

Nicola's IR research focuses on evaluation, from both a theoretical and an experimental perspective. He chairs the Steering Committee of CLEF (Conference and Labs of the Evaluation Forum), the internationally renowned European initiative for the evaluation of multilingual and multimodal information access systems, and through his efforts has made it central to the IR evaluation community. He was inducted into the Special Interest Group in Information Retrieval (SIGIR) Academy in 2023 in recognition of his significant research, innovation and service delivery. The Strix judging panel noted that his contributions to IR are extensive, evidenced by an outstanding and impactful publications record. It would like to congratulate him on his prolific and significant leadership and contribution to the profession on multiple fronts.

Nicola was delighted to receive the news and will celebrate his award in a special free Zoom lecture on the afternoon of Thursday 9 January 2025, where he will reflect on the challenges and opportunities of generative AI on the IR community.

> *"I'm honoured to receive this prestigious award and thank the Strix judging panel for such important recognition. This achievement would have not been possible without the mentors I've encountered throughout my career and without the collaboration with and inspiration from numerous colleagues, in Padua and around the world. I consider this award an appreciation of not only my work but also of the contributions of the international CLEF community."*

Professor Norbert Fuhr, Information Engineering, Department of Computer Science and Applied Cognitive Science at the University of Duisburg-Essen in Germany, submitted the nomination. It included five letters of support from well-known scientists, one of whom was 2017 Strix award winner Professor Maarten de Rijke, University of Amsterdam Informatics.

> *"Early in his career, Nicola Ferro made ground-breaking contributions on reference models for digital libraries, on annotations of digital content, and on system design for digital libraries. His work in this area has been very influential. He is best known for his impressive contributions to the evaluation of information retrieval systems. He is among the world's top experts in the field. Thanks to Nicola's tireless work, his people skills, and his scientific vision, CLEF unites a broad set of tasks, from traditional document retrieval to environmental image classification to medical data integration. CLEF has been, and continues to be, instrumental in supporting and advancing the research agendas of countless researchers."*

The Strix award judging panel would like to thank all colleagues who submitted nominations, and we look forward to your submissions in 2025. The excellence and quality of the entries is proof positive that the information retrieval community is not just thriving, but expanding.

Links
https://www.dei.unipd.it/~ferro/
https://scholar.google.com/citations?user=MP_m6wgAAAAJ
https://dblp.uni-trier.de/pid/f/NicolaFerro.html

Jason Farradane Award
The UK electronic information Group (UKeiG) is pleased to announce that the winner of its prestigious international Jason Farradane Award for 2024 is Karen Blakeman.

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

Over decades Karen's consultancy and professional development role has impacted many sectors and subjects, including business, marketing, company and health information. She has a degree in biological sciences from the University of Birmingham and worked as a microbiologist before joining Wellcome as an information scientist. She then spent ten years in the pharmaceutical and healthcare industry before moving to the international management consultancy group Strategic Planning Associates. In 1989, she set up RBA (Rhodes Blakeman Associates) Information Services and went on to become a leading figure in the use of the web and social media as research tools.

Karen has made a substantial and generous contribution to search tool awareness by openly sharing her knowledge of current search technologies with the community, including search strategy formulation and awareness of the deep web and dark web. Her work on information quality and commitment to the absolute importance of information quality, currency, integrity and provenance is of note.

Karen has also helped raise the profile of the information community through active involvement with professional organisations, including CILIP – the library and information association – both as a Councillor and an Honorary Fellow. Her contribution to the UKeiG management committee in a range of roles from Chair to Honorary Treasurer, has been extensive and helped shape the professional development portfolio the Group offers today.

Chris Armstrong, retired information consultant and Hazel Hall, Emeritus Professor, Edinburgh Napier University submitted the nomination.

"Karen is a worthy winner. Her influence through consultancy and professional development has been invaluable in the rapidly changing world of information retrieval and research skills. She is indisputably a distinguished, influential and inspirational figure, facilitating the effective use of information resources in a digital world."

The UKeiG judging panel would like to congratulate Karen on her significant contribution to the profession. It is a timely award when the critical appraisal of retrieved search outputs, the need to identify robust, transparent and trustworthy information, is increasingly important.

She was delighted to receive the news.

"I am honoured to receive this award. It has been a privilege over the years to meet through my work so many fantastic professionals who adhere to Jason Farradane's scientific approach to information. Now, more than ever, we need the skills to help us identify AI-generated "hallucinations" and disinformation in our professional and personal lives. With appreciation, many thanks."

Links

[Karen Blakeman, RBA Information Services]

[Facebook]

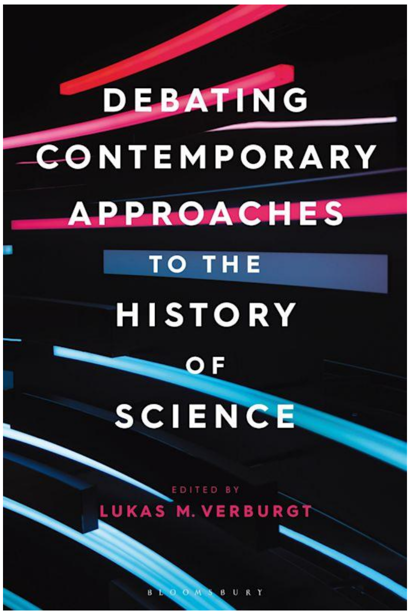
[Bluesky]

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

- -

Book Review – Debating Contemporary Approaches to the History of Science

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



Debating Contemporary Approaches to the History of Science
Lucas M. Verburgt (ed.). Bloomsbury Academic, 2024, London, New York. 396 pp + vii, ISBN 9781350326231, £24.99 (pbk), £22.49 (eBook)

So what's a review of a book like this doing in the *CICAG Distillate*? The history of science is fundamental to all aspects of science including chemistry and chemical information (since it is indeed science information). Granted, the science emphasised is physics, and there are no index references to chemistry, but this reviewer thinks the book is still relevant to CICAG and the chemical information community. Besides being of international interest (with authors from five countries), the book is especially relevant to British audiences since four of the thirteen authors work at British universities, several at Cambridge.

The editor, besides writing one of the chapters, outlines the book and its approach in the Introduction. The book is a series of debates in print (a rare species?) and, along with some additional chapters, is an outgrowth of a symposium on the subject. Each chapter is followed by a comment by another historian and a response by the chapter author. There are notes at the end of each chapter. The authors are international with the majority British or American.

In the Introduction, the editor describes the topics covered and the methods used. He points out that prior to the middle of the last century, science was understood as a historical activity, admittedly with different timescales. 'Circulation' of methods and ideas are described, which many would describe as the flow and growth of information and knowledge.

Several topics are discussed in the 13 chapters, including global history, gender history (a current hot topic), computational history, the history of knowledge, and environmental and material aspects. The last two chapters are on the history of ignorance, or 'agnotology', a novel topic, emerging in the last two decades. It can be defined as the study of history and philosophy of ignorance (time to end our ignorance of ignorance).

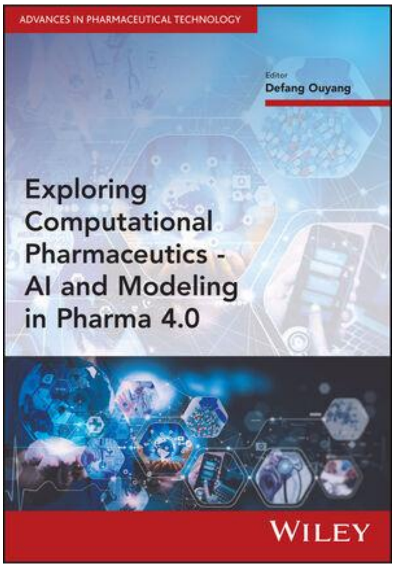
The book concludes with lists of further reading for each chapter, and a book index. Recommended for science libraries and individuals with an interest in history and philosophy as well as science. (A review of this book will also appear in the Winter edition of the *Chemical Information Bulletin* of CINF/ACS.)

- -

Cheminformatics and Chemical Information Books

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

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



Exploring Computational Pharmaceutics: AI and Modeling in Pharma 4.0

This book introduces a variety of current and emerging computational techniques for pharmaceutical research. Bringing together experts from academia, industry, and regulatory agencies, this edited volume also explores the current state, key challenges, and future outlook of computational pharmaceutics while encouraging development across all sectors of the field. Throughout the text, the authors discuss a wide range of essential topics, from molecular modelling and process simulation to intelligent manufacturing and quantitative pharmacology.

Building upon *Computational Pharmaceutics: Application of Molecular Modeling in Drug Delivery,* this new edition provides a multi-scale perspective that reveals the physical, chemical, mathematical, and data-driven details of pre-formulation, formulation, process, and clinical studies, in addition to *in vivo* prediction in the human body and precision medicine in clinical settings. Detailed chapters address both conventional dosage forms and the application of computational technologies in advanced pharmaceutical research, such as dendrimer-based delivery systems, liposome and lipid membrane research, and inorganic nanoparticles.

Covering introductory, advanced, and specialist topics, *Computational Pharmaceutics: From Multi-Scale Modeling to Artificial Intelligence* is an invaluable resource for computational chemists, computational analysts, pharmaceutical chemists, process engineers, process managers, and pharmacologists, as well as computer scientists, medicinal chemists, clinical pharmacists, material scientists, and nanotechnology specialists working in the field.

Edited by Defang Ouyang. Wiley & Sons Ltd, October 2024. Hardcover ISBN: 978111998713-0.

Advanced Modelling and Simulation in the Chemical and Biochemical Process Industry

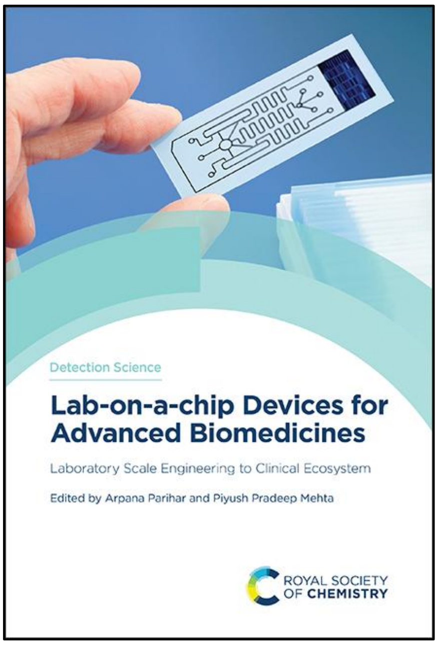
This book explores modelling and simulation of chemical and biochemical processes at the industrial scale using a variety of approaches. Particular attention is devoted to simulations in different scales, which help achieve a wide spectrum and more efficient analysis of several problems, ranging from the design of novel materials to the optimisation of industrial processes as a function of the operating conditions. This book not only covers optimisation with experimental data but also offers readers a thorough understanding and analysis of different parameters of a whole process stream.

- Covers a wide range of advanced modelling and simulation of chemical technologies: ab initio, atomistic molecular dynamics (MD), Lattice-Boltzmann (LB), dissipative particle dynamics (DPD), computational fluid dynamics (CFD), and finite element (FEM)

- Addresses issues associated with process control in different phases of the chemical industry
- Features modelling approaches that allow the design of novel processes/materials in a faster and more reliable way

This book will be of interest to researchers and advanced readers in chemical, biochemical, environmental, and materials engineering and industrial chemistry.

Edited By Sudip Chakraborty, Stefano Curcio. CRC Press, October 2024. ISBN 9781032563695


Detection Science
Lab-on-a-chip Devices for Advanced Biomedicines
Laboratory Scale Engineering to Clinical Ecosystem
Edited by Arpana Parihar and Piyush Pradeep Mehta
ROYAL SOCIETY OF CHEMISTRY

Lab-on-a-chip Devices for Advanced Biomedicines: Laboratory Scale Engineering to Clinical Ecosystem

The global miniature devices market is poised to surpass a valuation of $12–$15 billion USD by the year 2030. Lab-on-a-chip (LOC) devices are a vital component of this market.

Comprising a network of microchannels, electrical circuits, sensors, and electrodes, LOC is a miniaturised integrated device platform used to streamline day-to-day laboratory functions, run cost-effective clinical analyses and curb the need for centralised instrumentation facilities in remote areas. Compact design, portability, ease of operation, low sample volume, short reaction time, and parallel investigation stand as the pivotal factors driving the widespread acceptance of LOC within the biomedical community.

In this book, the editors meticulously explore LOC through three key 'Ts':

- Theories (microfluidics, microarrays, instrumentation, software)
- Technologies (additive manufacturing, artificial intelligence, computational thinking, smart consumables, scale-up tactics, and biofouling)
- Trends (biomedical analysis, point-of-care diagnostics, personalised healthcare, bioactive synthesis, disease diagnosis, and space applications)

This comprehensive text not only provides readers with a thorough understanding of the current advancements in the LOC domain but also offers valuable insights to support the utilisation of miniaturised devices for enhanced healthcare practices. Aimed at career researchers looking for instruction in the topic and newcomers to the area, the book is also useful for undergraduate and postgraduate students embarking on new studies or for those interested in reading about the LOC platform.

Edited by Arpana Parihar, Piyush Pradeep Mehta. Royal Society of Chemistry, August 2024. Hardback ISBN: 9781837672370; PDF ISBN: 9781837673476; EPUB ISBN: 9781837673483.

- -

News from CAS

Contribution from Dr Anne Jones, Senior Customer Success Specialist, email: *ajones2@acs-i.org*

CAS SciFinder Discovery Platform™

The CAS SciFinder Discovery Platform consists of multiple solutions that connect you with the world's scientific knowledge to find the answers you need to advance your research. These solutions are enhanced regularly, providing easier-to-use tools and more information to improve your work. Here are several notable updates for the platform's solutions:

- Stereoselective labelling was introduced in the retrosynthesis tool in CAS SciFinder®. This precise labelling of stereoselective steps is a unique capability, vital for researchers to design molecules precisely.
- Natural language query improvements in CAS SciFinder now recognise searches containing transformation names, substance classes, functional groups, and reaction participant roles.
- Substance text search support in CAS SciFinder was expanded to new functional group and substance class searching capabilities, allowing substance discovery by describing moieties contained or substances belonging to a class.
- Plan pricing estimates in CAS SciFinder retrosynthesis have been AI-optimised for commercially available compounds with limited or incomplete pricing information.
- The Projects feature in CAS SciFinder was expanded to include downloading the information saved in the Project for easier collaboration and the creation of reports or citation lists.
- Reference results exports have been updated to include reactions information with links when present, making it easy to view a reaction set or individual reactions from the download file.
- User interface improvements were made to the supplier results and details pages in CAS SciFinder for easier viewing.
- The Compare workflow in CAS Formulus® was enhanced to allow an easier comparison of selected formulations.

More details on the latest enhancements can be found by viewing the 'What's New?' section found in CAS SciFinder and CAS Formulus. You can also contact us for more information when using the CAS SciFinder Discovery Platform.

STN IP Protection Suite™

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

- The Role Indicator (RL) search field, available in CAplus, has been enhanced to help searchers pinpoint patents in which a substance of interest is tagged in the claims as part of the CAS PatentPak® workflow.
- The FI thesaurus, a unique patent classification system based on IPC codes and produced by the Japanese Patent Office, has been updated for Derwent's WPINDEX/WPIDS/WPIX databases in CAS STNext, and newly added to INPADOC and INPAFAM. The FI thesaurus supports efficient patent document searching and can be used to expand retrieval.
- Ultimate Owner data has been added to eight databases in CAS STNext to provide insight into current ownership of IP rights for a patent of interest.

- CAS PatentPak links are now accessible in MARPAT records as available. Clicking any of these links will take users directly to the full text of the original patent.
- Structure search highlighting in databases DCR, DWPIM, and REAXYSFILESUB has been enhanced and is now streamlined with other structure files on CAS STNext such as CAS REGISTRY® and MARPAT. Structure search results now include an indicator that clearly identifies the part of the structure matching the query.

More details on the latest enhancements to CAS STNext can be found on our product help site. You can also reach out to us directly with any questions or training needs.

CAS Insights™
Actionable insights from emerging research trends across industries
CAS Insights is an open resource for actionable perspectives on the latest developments across science, technology, and innovation powered by CAS human-curated data collection and the expertise of our science team.

Stay ahead of the latest trends in emerging areas:

- Inverse vaccines: New research on how to teach the immune system to determine friend-from-foe.
- Drug repurposing: Can we use existing drugs to cure Parkinson's disease?
- Nanotechnology: Using the CAS TrendScape of nanomaterials to get a bird's eye view of the landscape.

Want to make sure you don't miss anything? Subscribe to CAS Insights and get new insights delivered straight to your inbox.

- -

RSC Chemistry Databases Update

Contributions from Tamara Hughes, email: hughest@rsc.org; Sarah Rogers, email: rogerss@rsc.org; Susan Richardson, email: richardsons@rsc.org

Celebrate a decade of MarinLit with the Royal Society of Chemistry
MarinLit is the go-to platform for marine natural products research, proudly hosted by the RSC since 2014. What makes MarinLit indispensable?

- Comprehensive coverage of marine compounds, synthesis pathways, ecological insights, and biological activities.
- Innovative tools like advanced search features and dereplication workflows that save you time and streamline discovery.

Originally created in the 1970s by Professors John Blunt and Murray Munro at the University of Canterbury, New Zealand, MarinLit continues to evolve, fuelling breakthroughs in marine science and supporting the annual Marine Natural Product Review in *Natural Product Reports*.

🎥 Want to know more? See our YouTube video.
🚀 Ready to elevate your research? Email us and discover what MarinLit can do for you!

Milestones for The Merck Index* Online

The Merck Index Online has reached new heights, continuing to provide valuable resources for the scientific community:

- A record number of **expert-curated monographs** have been added to the database.
- A **new video** offers an in-depth look at the platform's features and capabilities.

🎥 Learn more by viewing our YouTube video

**The name The Merck Index is owned by Merck & Co., Inc., Rahway, NJ, USA, and its affiliates, and is licensed to the Royal Society of Chemistry for use in the USA and Canada.*

ChemSpider: 17 years of supporting chemical discovery

Powered by the Royal Society of Chemistry, ChemSpider has been a trusted resource for chemical data since 2007, offering researchers free and easy access to millions of chemical compounds.
What makes ChemSpider unique?

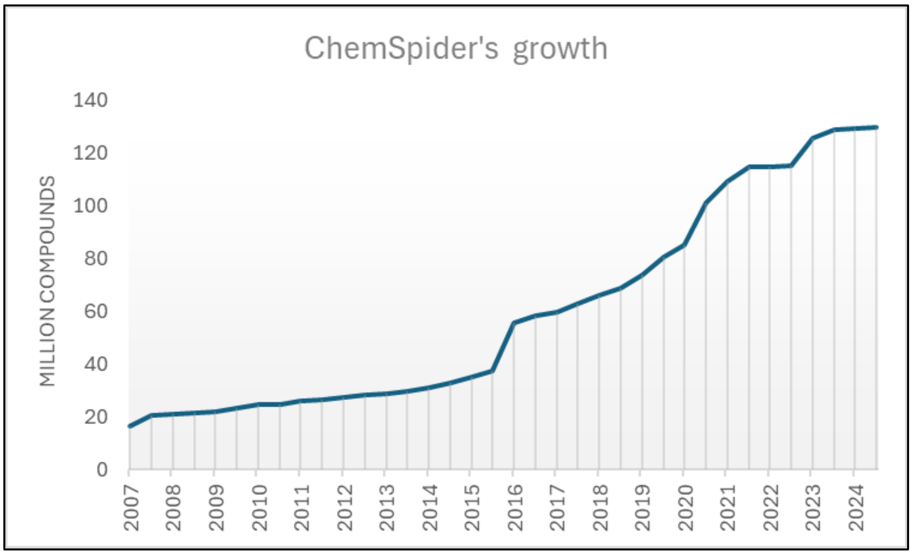
- Included data: 2D and 3D structures, synonyms, identifiers, properties, spectra, and links to specialised resources.
- Broad scope: from elements and laboratory chemicals to pharmaceuticals, food additives, dyes, small peptides, glycans, and nucleic acids.
- Accessible to all: designed for everyone, from students to seasoned researchers.

From its beginnings with 11 million records, ChemSpider has grown to include nearly 130 million distinct chemicals today, evolving to meet the needs of the chemical science community.

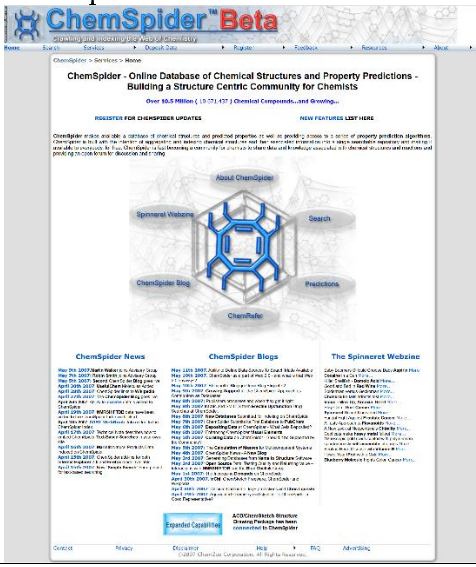
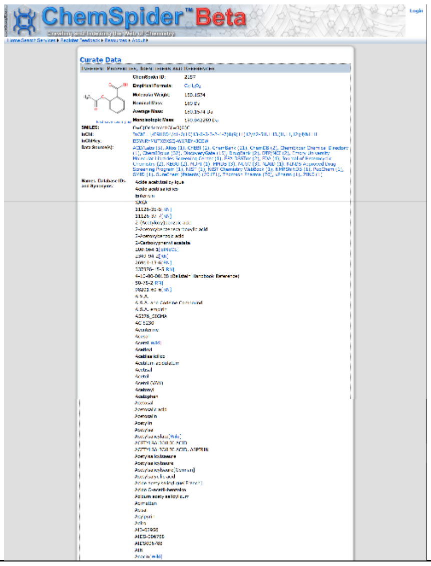
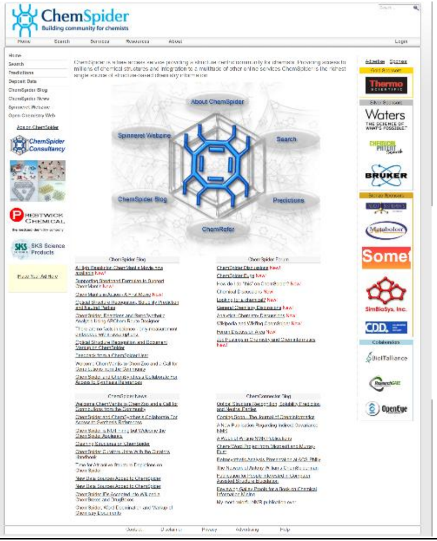
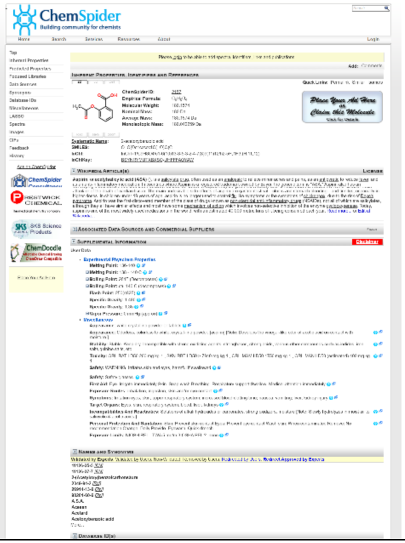
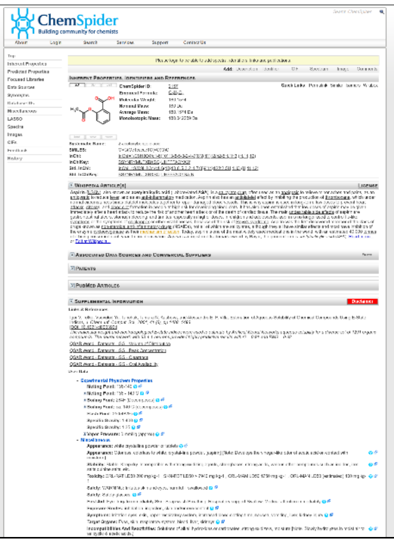
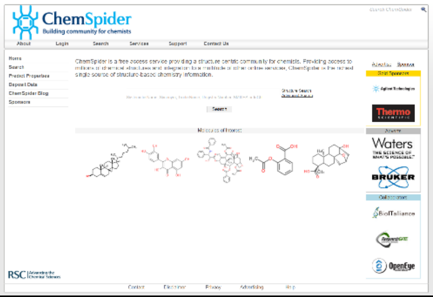
Discover how ChemSpider can support your research and education – wherever you are in your journey.

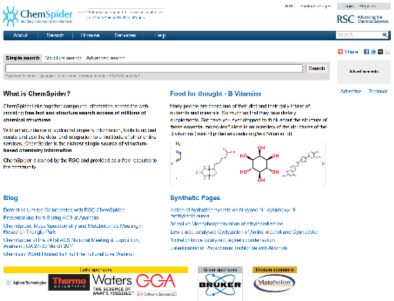
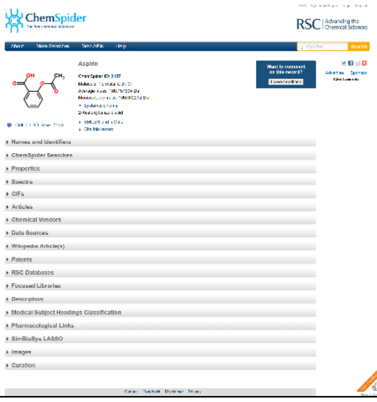
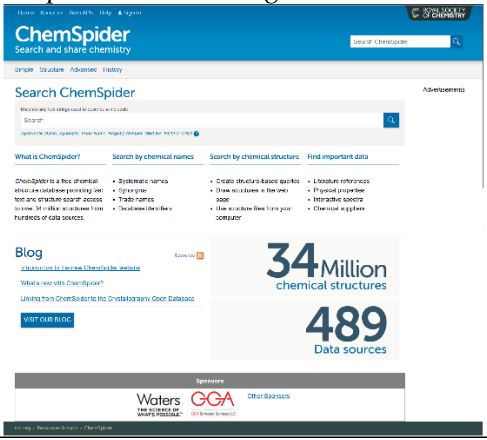
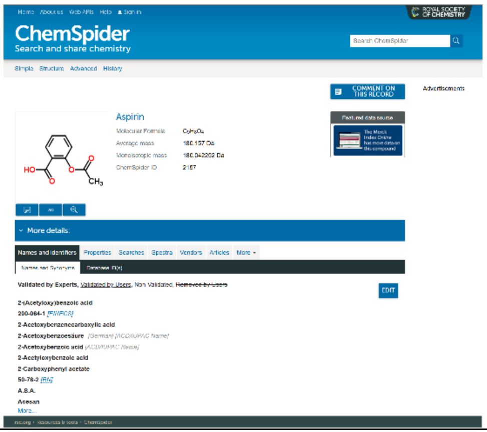
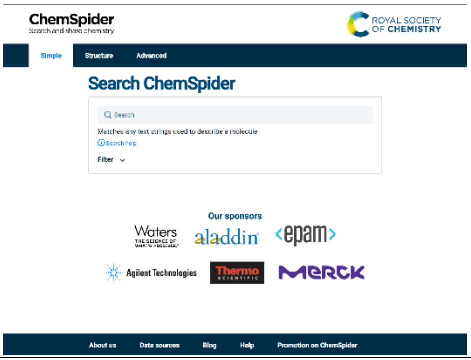
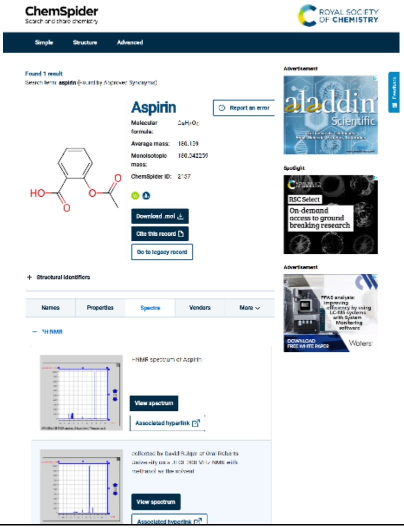
ChemSpider over time

ChemSpider has been offering chemical scientists easy and free access to data about millions of chemical compounds for 17 years. During that time, the site has grown, expanded, and evolved. When we started in 2007, we had fewer than 11 million compound records. Now, we cover just under 130 million distinct chemicals.



A brief timeline of ChemSpider development

2007	ChemSpider is Beta released.  ChemSpider™ Beta ChemSpider - Online Database of Chemical Structures and Property Predictions - Building a Structure Centric Community for Chemists  ChemSpider™ Beta Curate Data
2008	ChemSpider is released.  ChemSpider Building community for chemists  ChemSpider Building community for chemists
2009	The RSC acquires ChemSpider, an updated site is launched.  ChemSpider Building community for chemists  ChemSpider Building community for chemists

2011	Site revamp goes live.  
2015	Responsive site redesign is introduced.  ChemSpider Search and share chemistry Search ChemSpider Blog 34 Million chemical structures 489 Data sources Sponsors  ChemSpider Search and share chemistry Aspirin More details:
2016	A new data pre-processing workflow is introduced to remove common structural errors from files to be uploaded to the site.
2017	Synonym filtering is integrated into the pre-processing workflow.
2018	APIs are rebuilt and relaunched with a modern and robust architecture. A large purge of out-dated data sources was performed.
2021	Back-end updates, and data cleanup work commenced.
2024	Launch of a new ChemSpider site, with an entirely rebuilt codebase and a more robust and modern nosql database.  ChemSpider Search and share chemistry ROYAL SOCIETY OF CHEMISTRY Simple Structure Advanced Search ChemSpider Search Matches any text strings used to describe a molecule Filter Our sponsors About us Data sources Blog Help Promotion on ChemSpider  ChemSpider Search and share chemistry ROYAL SOCIETY OF CHEMISTRY Aspirin Report an error

- -

Cambridge Structural Database (CSD) Updates

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

CSD-materials in action: an integrated approach for solid form derisking

Identifying the best solid form for a drug candidate is not straightforward. A crystalline solid can present itself as a salt, a hydrate, a cocrystal, or a solvate. Different packing arrangements of the same molecular entities can also arise – forming the so-called polymorphs. Learn how CCDC and Pfizer scientists combined approaches to de-risk solid form selection, in our case study and paper.

How to create attention grabbing graphics of crystal structures

Visually showing complex concepts helps understanding, an aspect that is particularly relevant for crystallography and chemistry-related topics: a picture is indeed worth a thousand words.

The CCDC's software Mercury offers a selection of tools that can be helpful when preparing graphics of a crystal structure. Learn how to create eye-catching pictures for your paper, proposal, or presentation with both the free and full versions of Mercury.

2024.3 CSD software update: enhanced disorder handling and visualisation, and data for semiconductor research

This release includes enhanced disorder handling and visualisation, which allows disordered structures to be analysed with a variety of Mercury tools. It introduces new semiconductor data fields that are accessible via the CSD Python API. There have also been updates to the van der Waals radii for most atoms, resulting in changes to some calculations across the CSD software portfolio. Additionally, the CSD-Discovery has received visualisation improvements, including ChemDraw compatibility via Hermes.

Latest CSD data update

We are also pleased to present the latest data update to the Cambridge Structural Database (CSD). The CSD now contains 1,308,545 unique structures and 1,341,400 entries of small molecule organic and metal-organic experimental crystal structures.

As ever, every structure originates from experimental work, reported in the literature, PhD theses, patents, or deposited directly into the database as a CSD Communication. Every structure is validated and curated by manual and automatic methods, to make it as accessible and usable as possible.

Learn more about our latest software and data updates.

2025 events – save the date

- CCDC webinar: 23 January
- User webinar: 13th February
- Cambridge Cheminformatics Network meeting – hybrid: 19 February
- Virtual workshops: 11 and 25 March, and 8 April
- CCDC and ICDD/powder in-person training event: 6-9 May

Visit our website events page to learn more and register.

- -

Physical Sciences Data Infrastructure (PSDI) News

Contribution from Dr Samantha Pearman-Kanza, Dr Nicola Knight, Dr Aileen Day, Professor Jeremy Frey and Professor Simon Coles, University of Southampton, email: psdi@soton.ac.uk

As ever we have been very busy at PSDI over the last few months with active development and testing of our technologies and services as well as a lot of community engagement activities. We ran a Townhall to engage with the community at large about our plans for PSDI version 1, in addition to running some community workshops in Southampton and Edinburgh. Members of the PSDI team have also been presenting our work at a range of national and international conferences. Read about a lot of our latest activities in this instalment! If you want to be kept up to date more regularly don't forget to sign up to our mailing list or follow us on social media.

PSDI Details

🌐 www.psdi.ac.uk
🐦 @PSDI_UK
in linkedin.com/company/psdiuk
🦋 @psdi-uk.bsky.social
m @PSDI@mstdn.science
Mailing List: https://www.jiscmail.ac.uk/PSDI

PSDI Townhall

The first Physical Sciences Data Infrastructure (PSDI) Townhall, held on 20 June 2024, at the Institute of Materials, Minerals and Mining (IOM3) in London, brought together 58 attendees from across the physical sciences community. This event aimed to foster engagement, showcase ongoing projects, and gather feedback on the future direction of PSDI.

The agenda included presentations from UKRI and EPSRC representatives, an overview of PSDI by the principal investigators, and demonstrator presentations showcasing current developments. Key highlights included:

- **Biomolecular Simulation Workflows**: Tools for capturing simulation steps to enhance reproducibility.
- **Data to Knowledge (to Data)**: Infrastructure for machine learning models in materials modelling.
- **Format Conversion**: A web-based tool for easily converting scientific data formats without any software installation.
- **PSDI Gateway**: A prototype website for accessing PSDI resources, guidance and community engagement.

An exhibition session allowed attendees to interact with demonstrators and provide feedback. The event also featured lightning talks on various data challenges and initiatives, and an interactive feedback session to discuss future steps.

PSDI webinars

Since our last update, we have held two more webinars on the NOMAD ecosystem and Knowledger project, and are adding to the recordings available on our YouTube channel. We are busy planning many more webinars for the rest of the year, if you are interested in presenting your work at a PSDI webinar do get in contact via our website. You can find out more about our events here.

PSDI internships

This summer PSDI ran a cohort of nine internships with funding from PSDI, Southampton University's research intern scheme, the AIHub, AIChemy and sepNET schemes. The interns worked with colleagues from across PSDI and Chemistry on projects ranging from Data Conversion to Scientific Communication and Accessibility. As part of the internships, we also ran training events through the 'Skills 4 Scientists' programme including research skills such as literature searching, poster presentations and other technical and soft skills. Keep an eye out for this training programme which we intend to run again next year. There were also two workshops which were as part of the research concordat provision. The first workshop was on the Art of Storytelling and discussed how cartoons can be used to demonstrate complex scientific ideas. At the end of the event participants had the chance to draw their own cartoons which produced some interesting results! The second workshop focused on Ethics and Responsible Research Innovation (RRI).

The internships culminated in a final presentation and report, during which everyone did a fantastic job on reporting on their work. One of the interns, Ashley Doel, was also awarded a runner-up prize in the Undergraduate Research Internship Showcase for their poster on "Automated Metadata Generation and Semantic Tagging Investigation". Congratulations to Ashley!

PSDI received some fantastic feedback on the programme, including from one intern – "Wholeheartedly recommend PSDI to any applicants. They have a great objective, with a brilliant set of skills courses that benefited everyone. They're a wonderful group of people from a plethora of backgrounds and did an amazing job at providing a warm and welcoming environment to all interns."

Well done to all of the PSDI interns on a successful internship and keep an eye out for opportunities for internships and our training programme in 2025.

Upcoming events featuring PSDI

The PSDI team will be featuring at some national and international conferences over the next few months.

Future Labs, Automation & Technology Europe

On 3 December 2024 Dr Philip Leadbitter attended this event on behalf of PSDI. As a pre-cursor to this conference, some of the previous speakers including Dr Samantha Pearman-Kanza were invited to contribute to an article about "How will Automation and Digital Technology Shape the Lab of the Future" written by Blake Forman.

IDCC Workshop

PSDI will be running a workshop at the International Digital Curation Conference 2025 in February 2025. This workshop "Creating communities around best practices and common challenges in data" will provide opportunities for data stewards to come together to share their challenges, best practices and solutions, and to discuss how PSDI might be able to help provide skills and guidance for data stewards working with scientists in the scientific community. You can find out more and sign up if you are going to be at IDCC via Eventbrite.

Research Data in the Physical Sciences: A Forum for Librarians and Research Support Professionals

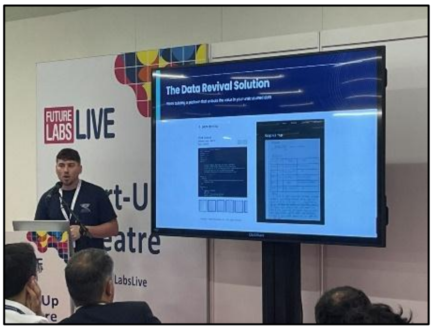
In March 2025 in collaboration with the Digital Curation Centre we will be running a community forum and knowledge exchange event for data librarians and research support professionals. More information and registration will be available on our Events page.

Recent events and presentations

The PSDI team has been very busy over the last couple of months attending a variety of conferences, workshops and events. We are updating our zenodo community with many of the presentations and publications made by the team. Highlights from the last few months include:

Future Labs Live Basel

In late June, Dr Samantha Pearman-Kanza, Professor Jeremy Frey and Mr Samuel Munday attended [Future Labs Live]() in sunny Basel. Samantha chaired two panels, one on training and educational needs for the community and one on the controversial topic of should we ever delete data! She also took a group down the rabbit hole to present on [Mirror mirror on the wall, how do I make the FAIRest metadata of them all?](). Jeremy took us on a fascinating journey with lots of wonderful stories about his work with Smart Laboratories, and Samuel presented on Data Revival as part of the startup pitch competition and made it to the final! It was a fantastic event with lots of rich discussion, and a key takehome for us was how important data stewardship is and will continue to be if we are to ever truly succeed on our FAIR journey.



FUTURE LABS LIVE
SHOULD WE EVER DELETE DATA?
Data Management, Standards & Analytics

Helmholtz metadata conference

In November, Dr Aileen Day attended the [Helmholtz Metadata Collaboration Conference](), a virtual conference focused on metadata. It was not only directly useful to our PSDI metadata development in terms of seeing many metadata examples, best practices and tools, but also for wider inspiration for PSDI and community engagement. One of the most notable features of the conference was that it was entirely virtual, so that on login one can customise an avatar and then walk it around the conference rooms: into rooms with presentations of interest, up to a speaker point to ask a question; over to the poster session and into a poster booth to talk to someone about their poster; and over to workshops with whiteboards to work on together. It was a great way to attend a conference and network while minimising academic travel carbon footprint and maximising inclusivity for those for whom travel is difficult. In terms of content, [Datathons: fostering equitability in data reuse in ecology]() described a 'Datathon' to encourage reuse of a dataset in a different field to PSDI but this could be an interesting type of community activity to consider. A keynote presentation [Keynote: A Glimpse into the Future of Metadata - Practical Challenges in Development and Application of a Metadata Standard]() mentioned the concept of awarding [badges]() to encourage full and correct metadata population. [FAIRly Intelligent – What LLMs Bring to the Research Data Management Table]() was a keynote presentation which described experimenting with LLMs to perform various data management and curation tasks of the kind that PSDI will need to consider (spoiler – they're not perfect!). The workshop [2. Comparative Analysis of Automated FAIR Assessment Tools' results: F-UJI, FAIR Enough, and FAIR Checker]() dissected the results of three automated fair assessment tools for some example datasets and is something that we will need to revisit to investigate their assessments of data made available via PSDI. Presentations from the conference are available at [Zenodo collection for Helmholtz Metadata Collaboration | Conference 2024.]()

EUChemS mini symposium on FAIR data for Chemistry

In July 2024 Dr Samantha Pearman-Kanza and Dr Nicola Knight collaborated with colleagues (John Jolliffe and Jochen Ortmeyer) from NFDI4Chem to run a symposium at the 9th EUChemS conference in Dublin. This session focused on National infrastructures, standards, tools and resources and it discussed some of the drivers behind the work being done in PSDI and NFDI4Chem as well as the solutions that are in development. This session and the discussions afterwards highlighted that data, tooling and skills are key topics that should be discussed more at major conferences like this.

Semantics Conference

In September 2024, Dr Samantha Pearman-Kanza attended the 20th International Conference on Semantic Systems in Amsterdam. This was a very informative conference, with many talks that are highly relevant to our community, especially the ones on materials ontologies, ontologies for unit measurements, and using SHACL to assess the fairness of software repositories. The conference also hosted a fascinating debate about the future of AI and our ethical responsibilities within the realms of the Semantic Web. Samantha highly recommends this conference to any of our semantic colleagues, and warns against getting lost in the pretty Amsterdam forests on morning runs.

ChEMBL 15 Year Symposium

At the beginning of October 2024, Dr Samantha Pearman-Kanza was invited to speak at the ChEMBL 15 Year Symposium in Cambridge. This event was organised to celebrate the 15th anniversary of the first public release of the ChEMBL database and the 10th anniversary of SureChEMBL. Day 1 was a set of workshops geared towards using ChEMBL and SureChEMBL, and day 2 featured invited talks from a range of speakers. Samantha delivered the somewhat provocative presentation The Semantic Web is Dead – Long Live the Semantic Web, which gave a holistic overview of the Semantic Web, discussing common misconceptions, barriers and challenges, mitigations and suggestions for best practice, and a commentary on emerging use cases for semantics for the physical sciences. The talks were excellent, and it was wonderful to connect with other stringent advocates of the Semantic Web Community. The ChEMBL team also commissioned a wonderful and very delicious cake to commemorate the occasion.

Lab Innovations

Once again, just before All Hallows' Eve, Dr Samantha Pearman-Kanza and Mr Samuel Munday braved the spooky and the supernatural to attend Lab Innovations. Dr Samantha Pearman-Kanza presented on The Bare Necessities: How to Implement Electronic Lab Notebooks properly. Implementing ELNs is no trivial matter and goes far beyond just selecting the correct software (although this in itself remains a major hurdle). This presentation provided an overview of ELNs, discussing the barriers, challenges and offerings available on the market, and details the wide range of considerations that need to be taken into account for a successful implementation. However, whilst implementing ELNs is a step to improving data capture and research data management in the future, they do not fix past mistakes and aren't designed to deal with the plethora of paper-based data that came before. Thankfully, one of PSDI's collaborators Data Revival is working in that exact area, and Samuel Munday presented on how Data Revival can be used to scan in your paper-based notebooks, turn them into machine-readable content and extract the meaningful chemical information from them.

Science and Engineering South - Research Data Café

Team members at Southampton were honoured to host their colleagues from SES South in November for a hybrid meeting of the research data café. These meetings bring together researchers and university professionals who have an interest in the management of research data and allow knowledge exchange and sharing of best practices across a range of topics. This particular meeting included presentations on professionalisation of data stewards, research data catalogues, implementation of electronic lab notebooks and data management plans alongside lots of active discussion.

Other useful links / articles

At PSDI, we like to disseminate our learnings across the communities in a range of different formats, geared towards different audiences. So you can also find articles from our team in other places, including:

Articles in Lab Horizons

Dr Samantha Pearman-Kanza contributes regular articles under the CompSci Cat Column. She has written about the following topics:

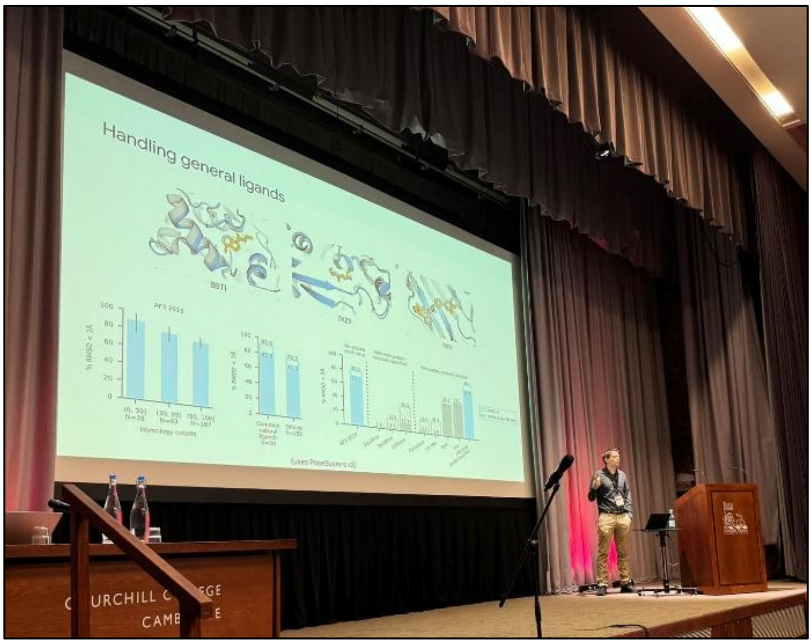
- Issue 1 – Failed it to nailed it: How to avoid mishaps in your data and code recipes (p.8-11)
- Issue 2 - Electronic Lab Notebooks: One notebook to rule them all, a dream or a curse? (p.8-11) And Completing the Quest for an Electronic Lab Notebook (p.22-26)
- Issue 3 – The Evolution of The Scientists Notebook: One step forward or two steps back? (p.10-13)
- Issue 4 – Thats One Small Step For Research, One Giant Leap for Data (p.8-11)
- Issue 5 – Step Aside Gen-Z, Gen-A

- -

Meeting Report: RSC-CICAG and RSC-BMCS 7th Artificial Intelligence in Chemistry Symposium

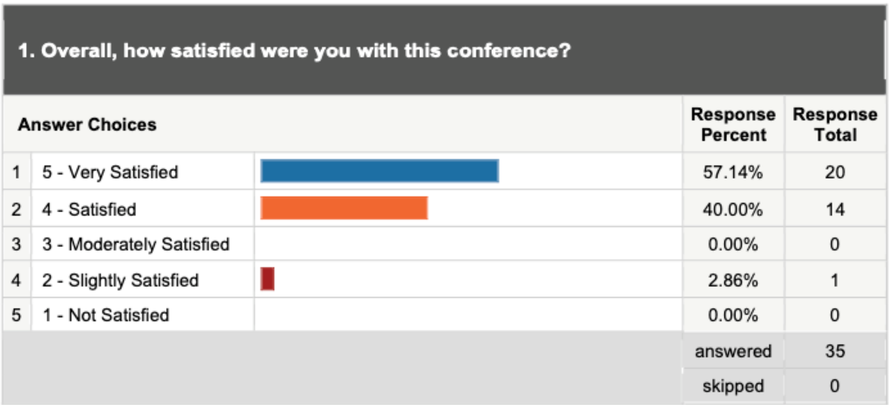
16-18 September 2024

Contributor for correspondence Dr Chris Swain, email: swain@mac.com



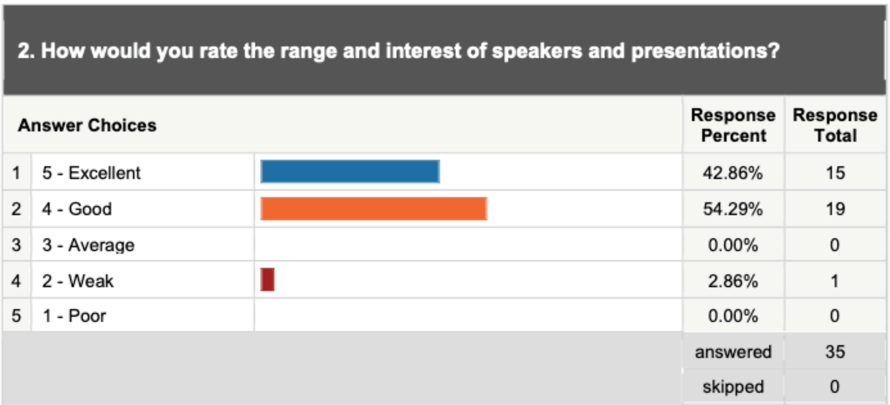
John Jumper giving a keynote at the AI in Chemistry meeting this year.

The *7th Artificial Intelligence in Chemistry* meeting took place on 16-18 September 2024 at Churchill College, Cambridge. The first day of the meeting consisted of a workshop intended to provide an introduction to artificial intelligence/machine learning (AI/ML) with worked examples using Google Colab. Hopefully this will be a resource that attendees can refer back to. The other two days provided a mixture of keynote talks, oral presentations, flash presentations, posters and opportunities for open debate, networking and discussion.



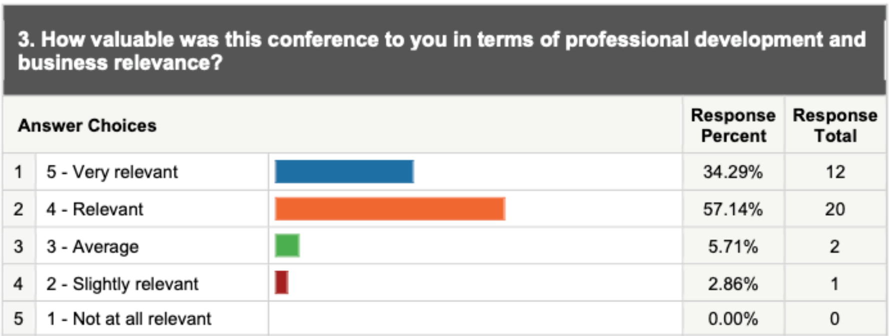
1. Overall, how satisfied were you with this conference?

	Answer Choices	Response Percent	Response Total
1	5 - Very Satisfied	57.14%	20
2	4 - Satisfied	40.00%	14
3	3 - Moderately Satisfied	0.00%	0
4	2 - Slightly Satisfied	2.86%	1
5	1 - Not Satisfied	0.00%	0
		answered	35
		skipped	0



2. How would you rate the range and interest of speakers and presentations?

	Answer Choices	Response Percent	Response Total
1	5 - Excellent	42.86%	15
2	4 - Good	54.29%	19
3	3 - Average	0.00%	0
4	2 - Weak	2.86%	1
5	1 - Poor	0.00%	0
		answered	35
		skipped	0



3. How valuable was this conference to you in terms of professional development and business relevance?

	Answer Choices	Response Percent	Response Total
1	5 - Very relevant	34.29%	12
2	4 - Relevant	57.14%	20
3	3 - Average	5.71%	2
4	2 - Slightly relevant	2.86%	1
5	1 - Not at all relevant	0.00%	0

The meeting was attended by 241 delegates with 199 attending in person. The delegates came from 27 different countries and whilst the UK and US provided the bulk of the attendees it was great to see sizable contingents from Germany, Switzerland and Singapore. In addition there were attendees who had travelled from Zambia, China and Australia. 81 of the attendees were from universities, and the remainder from industry or non-profit organisations. It was notable that the meeting attracted a significant number of PhD students, with 11 bursaries awarded. The majority of total attendees (172) were not RSC members, but it was encouraging that half of the student attendees were RSC members. Including the workshop there were 20 presenters in total 11 female and 9 male, the audience was 65 female and 165 male with 11 other/prefer not to say.

The post event feedback was extremely positive.

Monday 16 September

Workshop

Training a Recurrent Neural Network (RNN) to Generate SMILES
Prior
Embedding
RNN
Output
etc.
etc.
etc.

The workshop was a new addition to the meeting and was presented by Pat Walters, Andrea Volkamer, Michael Backenköhler and Raquel López-Ríos de Castro; it was limited to 100 attendees and was oversubscribed. The workshop was divided into three sessions:

Session 1

- An introduction to AI and ML
- Molecular representations
- AI architectures

Session 2

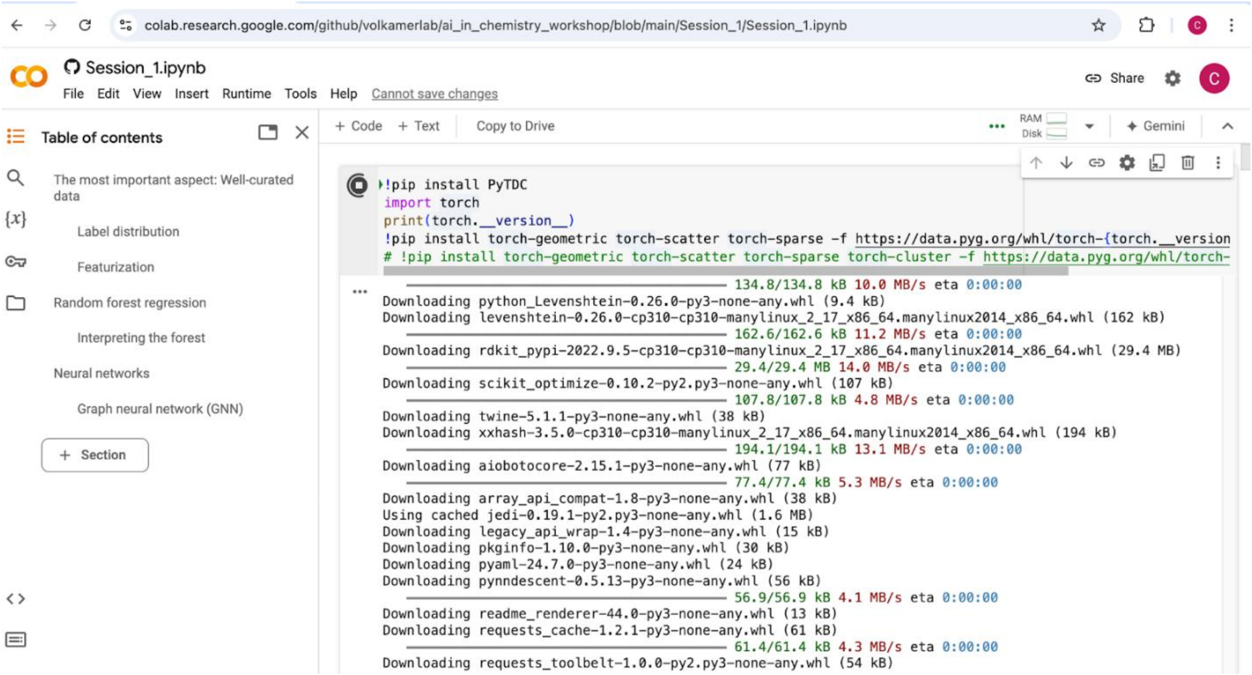
- The importance of data quality for AI/ML
- Exploratory data analysis
- Data preprocessing
- Applicability domains

Session 3

- AI in practice
- Molecule generation
- Active learning

Each session started with a lecture detailing the topics followed by a hands-on session using Google Colab.

All slides and notebooks are freely available online.


colab.research.google.com/github/volkamerlab/ai_in_chemistry_workshop/blob/main/Session_1/Session_1.ipynb
Session_1.ipynb
File Edit View Insert Runtime Tools Help Cannot save changes
Share
Table of contents
The most important aspect: Well-curated data
Label distribution
Featurization
Random forest regression
Interpreting the forest
Neural networks
Graph neural network (GNN)
+ Section
+ Code + Text Copy to Drive
RAM
Disk
Gemini

```
!pip install PyTDC
import torch
print(torch.__version__)
!pip install torch-geometric torch-scatter torch-sparse -f https://data.pyg.org/whl/torch-{torch.__version
# !pip install torch-geometric torch-scatter torch-sparse torch-cluster -f https://data.pyg.org/whl/torch-
```

```
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 134.8/134.8 kB 10.0 MB/s eta 0:00:00
Downloading python_Levenshtein-0.26.0-py3-none-any.whl (9.4 kB)
Downloading levenshtein-0.26.0-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (162 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 162.6/162.6 kB 11.2 MB/s eta 0:00:00
Downloading rdkit_pypi-2022.9.5-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (29.4 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 29.4/29.4 MB 14.0 MB/s eta 0:00:00
Downloading scikit_optimize-0.10.2-py2.py3-none-any.whl (107 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 107.8/107.8 kB 4.8 MB/s eta 0:00:00
Downloading twine-5.1.1-py3-none-any.whl (38 kB)
Downloading xxhash-3.5.0-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (194 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 194.1/194.1 kB 13.1 MB/s eta 0:00:00
Downloading aiobotocore-2.15.1-py3-none-any.whl (77 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 77.4/77.4 kB 5.3 MB/s eta 0:00:00
Downloading array_api_compat-1.8-py3-none-any.whl (38 kB)
Using cached jedi-0.19.1-py2.py3-none-any.whl (1.6 MB)
Downloading legacy_api_wrap-1.4-py3-none-any.whl (15 kB)
Downloading pkginfo-1.10.0-py3-none-any.whl (30 kB)
Downloading pyaml-24.7.0-py3-none-any.whl (24 kB)
Downloading pynndescent-0.5.13-py3-none-any.whl (56 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 56.9/56.9 kB 4.1 MB/s eta 0:00:00
Downloading readme_renderer-44.0-py3-none-any.whl (13 kB)
Downloading requests_cache-1.2.1-py3-none-any.whl (61 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 61.4/61.4 kB 4.3 MB/s eta 0:00:00
Downloading requests_toolbelt-1.0.0-py2.py3-none-any.whl (54 kB)
```

Tuesday 17 September

Keynote

Predicting general biomolecular interactions with AlphaFold 3
John Jumper, *Google DeepMind, UK*
John Jumper gave an insightful talk into the architectural decisions behind the latest incarnation of AlphaFold 3 (AF3), describing how they rethought their approach after developing AF2. John described how they worked to build our understanding of physics and evolution into the neural network architecture itself, and not rely solely on the training data. While many were 'distracted' by geometric deep learning in AF2, John said that the Evoformer was where the work was done. He pointed out that the switch to focus on residue pairs in AF3 was inspired by geometric interactions. John also described how they interrogated the network at each block then predicted the structure, to learn what happens in the network. They learned that quite rapidly, the structure is 'solved' and "the thing that really mattered" was the refinement engine in the middle.

For AF3, the whole Protein Data Bank was their goal, but unlike proteins, there is less data for RNA and DNA; similarly less for small molecules than for amino acids. He also asserted that their predictions for antibody:antigen complexes were worse than other kinds of protein:protein complexes because they lacked the coevolutionary signal of antigens. This elicited one of the most memorable quotes of John's talk: they "just deep learned harder to get better" at predicting antibody:antigen complex structures. John also reported that the PoseBusters benchmark and tool proved invaluable while developing the protein-ligand prediction capability of AF3. John went on to describe the role of diffusion in AF3, pointing out that it forces the model to "think a lot about local geometry", but the "dark side of diffusion is hallucination".

It was also fascinating to hear about how they controlled the rate of introduction of different types of training data at different stages during the training. John concluded by discussing some of the limitations of AF3, including new modalities, hallucination, allostery remains inconsistent, and stereochemistry can be problematic.

The keynote prompted several questions, including whether AF3 had learned the underlying physics or was memorising its training data.

Keynote

Leveraging community knowledge in transition metal complex and metal organic framework discovery
Heather Kulik, Massachusetts Institute of Technology, USA
Heather Kulik gave broad ranging and engaging talk on the application of machine learning to problems in material science and catalysis. She began by describing tuning the properties of ferrocene mechanophores such as hemilability – relevant for the strength of polymers incorporating such mechanophores. Building models that predicting with confidence whether a ligand is hemilabile or not, she showed that models can go beyond the conventional design rules used by humans. The next problem described was predicting the ligand coordination in catalysts using CSD data and incorporating QM descriptors. Lastly she described the discovery of new metal-organic frameworks (MOFs) for applications in gas storage and catalysis. Typically the discovery of new MOFs is very slow; Heather showed that using machine learning to evaluate hypothetical MOFs to identify those with suitable properties such as pore size and high thermal stability can accelerate that process.

To address dataset challenges such as lack of consistent naming and reporting which makes it challenging to leverage knowledge, Heather described examples of natural language processing and text mining methods for the curation of (more) specific datasets for different applications, such as the classification of CSD structures as catalysts, and the expansion of a MOF water stability dataset. These expanded datasets will enhance the design of new, more stable MOF catalysts.

Oral presentations

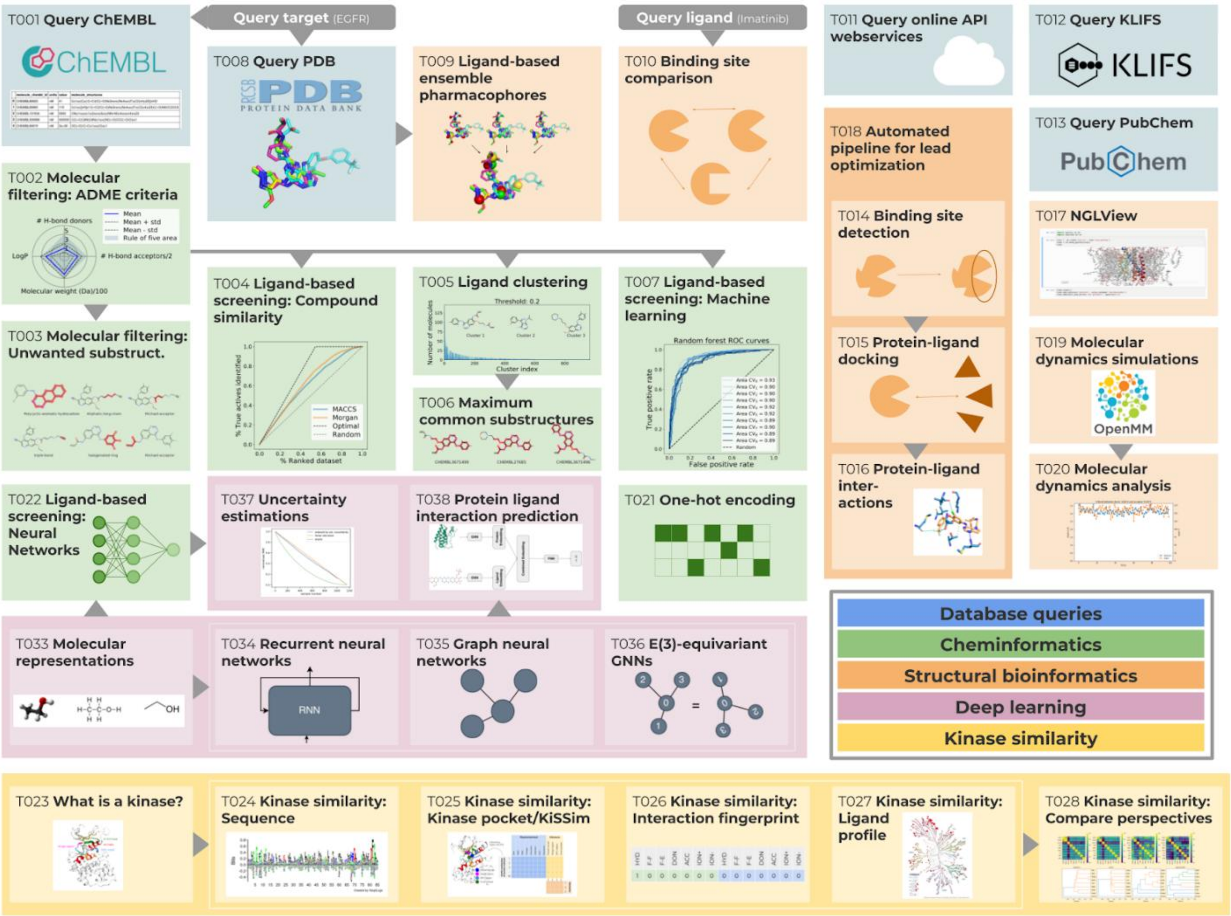
Directional multiobjective optimisation of metal complexes in vast chemical spaces
Hannes Kneiding, University of Oslo, Norway
Hannes Kneiding described the preparation of the tmQMg-L dataset, which includes ligands from 30K unique transition metal complexes (TMC) from the Cambridge Structural Database and then discussed how this was leveraged to construct a 1BN chemical space of putative palladium complexes by restricting the charge and geometries to those ligands observed in the tmQMg-L set. A genetic algorithm (GA) with multiobjective selection was used to generate a set of TMCs with both high polarizability and a high HOMO-LUMO gap while enforcing diversity. Hannes then went on to describe how the ligands in the tmQMg-L dataset were used to train a generative neural network model (junction-tree VAE) to generate both monodentate and bidentate ligands which can then be used in a multiobjective GA. In this case synthetic accessibility also needs to be considered in the selection of ligands.

CHILI: Chemically-Informed Large-scale Inorganic nanomaterials dataset for advancing graph machine learning
Ulrik Friis-Jensen, Department of Chemistry, University of Copenhagen, Denmark
Ulrik Friis-Jensen described the generation of open-source nanomaterial datasets with high structural and elemental diversity: CHILI3K is a medium scale data set containing >6M nodes and >49M edges of mono-metallic oxide nanomaterials while CHILI-100K is a large scale open graph benchmark with >100 Bn nodes and >1Bn edges. These were generated from the structural data in the Crystallographic Open Database by expanding the unit cell, cutting and padding to make sure that molecules were fully coordinated at boundaries. They simulated the scattering data for these molecules, to define the ground truth for the prediction tasks which included predicting the crystal system (a relatively easy task) and space group (a relatively hard task). He benchmarked the performance of a set of different models for these tasks but in most cases the models predicted

the most frequent class. Ulrik also attempted structure generation tasks but concluded that this looked to be too challenging for such nanomaterials.

Molecular set representation learning
Maria Boulougouri, EPFL, Switzerland
Maria Boulougouri gave an interesting presentation on set representation of chemical structures, highlighting the limitations of graph-based methods in ionic or metallic structures. MSR1 and MSR2 encode atoms and bonds respectively, and perform similarly to benchmarks in MoleculeNet compared to D-MPNN but offer no significant advantage. However, there is concern about the suitability of the MoleculeNet dataset. Looking at newer more challenging datasets [OCELOT chromophore for quantum-chemical property prediction, Bhat, V. *et al*. Electronic, redox, and optical property prediction of organic π-conjugated molecules through a hierarchy of machine learning approaches. *Chem. Sci.* 14, 203–213 (2022)], highlighted the limitations of MSR1 and MSR2. In contrast, set representation - enhanced GNN (SR-GINE) performed better in these newer more challenging datasets. All code is available on GitHub.

Hybrid AI and open-source for molecular design
Andrea Volkamer, Saarland University, Germany
Andrea Volkamer was one of the organisers of the AI for Chemists pre-conference course and she again highlighted the fantastic resources that her group have made available. The TeachOpenCADD platform provides a modular approach to learning computer-aided drug design, covering topics such as database access, filtering based on molecular properties, clustering, ligand-based machine learning. Each module is accompanied by an interactive Jupyter notebook. KLIFS is a kinase database that catalogues the way that kinase


T001 Query ChEMBL
ChEMBL
Query target (EGFR)
Query ligand (Imatinib)
T011 Query online API webservices
T012 Query KLIFS
KLIFS
T008 Query PDB
PDB PROTEIN DATA BANK
T009 Ligand-based ensemble pharmacophores
T010 Binding site comparison
T018 Automated pipeline for lead optimization
T013 Query PubChem
PubChem
T002 Molecular filtering: ADME criteria
T014 Binding site detection
T017 NGLView
T004 Ligand-based screening: Compound similarity
T005 Ligand clustering
T007 Ligand-based screening: Machine learning
T003 Molecular filtering: Unwanted substruct.
T006 Maximum common substructures
T015 Protein-ligand docking
T019 Molecular dynamics simulations
OpenMM
T016 Protein-ligand inter-actions
T020 Molecular dynamics analysis
T022 Ligand-based screening: Neural Networks
T037 Uncertainty estimations
T038 Protein ligand interaction prediction
T021 One-hot encoding
T033 Molecular representations
T034 Recurrent neural networks
RNN
T035 Graph neural networks
T036 E(3)-equivariant GNNs
Database queries
Cheminformatics
Structural bioinformatics
Deep learning
Kinase similarity
T023 What is a kinase?
T024 Kinase similarity: Sequence
T025 Kinase similarity: Kinase pocket/KiSSim
T026 Kinase similarity: Interaction fingerprint
T027 Kinase similarity: Ligand profile
T028 Kinase similarity: Compare perspectives

catalytic domains interact with ligands. However, this dataset is insufficient for training machine-learning models. By combining structural kinase data from the KLIFS database and a curated set of ligand–kinase activity measurements derived from ChEMBL a guided docking approach generated around 120,000 in silico kinase-ligand complexes. These complexes were then used to generate a machine learning model to predict binding affinities. The results show that this approach has good predictive power.

Perspective

AI in drug discovery – where are we making an impact?
Patrick Walters, Relay Therapeutics, USA
Pat Walters gave an entertaining pre-conference dinner talk, taking the opportunity to highlight common issues, lack of quality of often used data sets, failure to adequately curate data prior to model building. Lack of statistical rigour in describing results, the dreaded bolded results table where the new method always seems to offer a numerical improvement but no details to demonstrate statistically significant improvement.

MoleculeNet was highlighted as problematic, some of the datasets have dynamic range of 10 log units, whereas usual project datasets are only 3-4 log units. Models trained on data with appropriate ranges may be more applicable to discovery projects.

The structures of molecules in many datasets are sometimes incorrect (there are molecules with uncharged tetravalent nitrogens), and have inappropriate or inconsistent tautomerisation or ionisation. It is essential to standardise the molecular structures prior to model building. Stereochemistry may be very important, especially for binding affinity and needs to be consistently encoded within the chemical structure. Pat has written a blog post highlighting the issues.

Wednesday 18 September

Oral presentations

Interpreting neural network models for toxicity prediction
Val Gillet, University of Sheffield, UK
Val Gillet's work has made contributions to the field of interpreting neural network models for toxicity prediction. Her talk focussed on developing and applying cheminformatics techniques for designing bioactive compounds, with a particular emphasis on toxicity prediction. This included comparison to existing methods through attribution methods such as SHAP and validation using structural alerts for mutagenicity from the Derek Nexus expert system. In addition, she discussed improving the interpretability and reliability of neural network models in toxicity prediction, potentially enhancing their utility in drug discovery and chemical safety assessment. Please refer to Val's work for more information.

Training instruction-tuned and byte-level language models for organic reaction prediction
Jiayun Pang, University of Greenwich, UK
Jiayun Pang's talk was based on training instruction-tuned and byte-level language models for organic reaction prediction. This work evaluates the effectiveness of FlanT5 and By T5 models, which were originally pretrained on language data, for predicting organic chemical reactions through task-specific fine-tuning. FlanT5 and ByT5 models can be effectively specialised for reaction prediction without extensive pre-training on chemical datasets, which achieve comparable Top-1 and Top-5 accuracy in predicting organic reactions. In addition, tokenization and vocabulary trimming can speed up training and inference with only slight effects on

performance. This would suggest that GPU-intensive pretraining on large datasets of unlabelled molecules may not be essential for leveraging language models in chemistry tasks. The talk provided an insight into more effective use of state-of-the-art language models for chemistry-related applications, potentially streamlining the process of organic react ion prediction. Please refer to Jiayun's work for more information.

Practical machine learning for organic small molecule modelling

Emma King-Smith, University of Cambridge, UK

Emma King-Smith started her excellent talk with a description of running computational modelling on mechanisms during the synthesis of the polycyclic caged molecules, picrotoxins and related natural products. Amongst one common synthetic step, it had been unclear why some substrates underwent a hydrogen atom transfer (HAT) process, and others performed β-scission. Computational chemistry was able to reveal an answer that had not been obvious to the chemists, which would allow them to more reliably synthesise targets in this family. Emma pointed out that the paradigm where computational chemistry is run in advance of synthetic chemistry to inform future work is unfortunately not so common in academia as using computational chemistry to explain work that has already happened.

With this in mind, the remainder of her talk focussed on using transfer learning to gather new insights. First, her team worked with a dataset of 8.5K Suzuki cross-couplings, and separately, 4.6K Buchwald-Hartwig cross-couplings. They trained a graph neural network (GNN) on crystal structures from the CSD, and used this to predict the yields of the reactions. Emma explained that transfer learning is ideal for the "small" datasets we typically see in chemistry – these cross-coupling datasets being very large compared to most real-world chemistry datasets – and indeed the predictor did show a reasonably low level of error (the work is now published in Chemical Science). An improved process on a yet-smaller, 2.6-K dataset of Minisci reactions was able to predict yields effectively by using a GNN trained initially on ^{13}C NMR shifts, which relate directly to the electronics of the substrates. Lastly, the approach of a ^{13}C shift-trained model could also be applied to biocatalysis using cytochrome P_{450} enzymes.

Incorporating synthetic accessibility in drug design: predicting reaction yields of Suzuki cross-couplings by leveraging AbbVie's 15-year parallel library dataset

Priyanka Raghavan, Massachusetts Institute of Technology, USA

Priyanka Raghavan's talk features her work on synthetic accessibility determination in drug design. Using a large (24-K; 3.5K unique substrates) Suzuki reaction dataset from 384- and 1536-well high-throughput plates from AbbVie, she was able to run a number of classification and regression models to predict isolated yield outcomes, learning that a random forest model was initially the most effective tested – though still struggled to obtain higher than 72–77%. Although the underlying dataset was very large for a synthesis problem, it does illustrate human bias effectively, as only 201 unique condition sets were used across these many plates. Additionally, by the nature of design and synthesis, t-SNE analysis of the scaffolds and building blocks showed strong clusters of preference for certain types of boronate substrate.

The modelling work was made more useful with a clever turnaround, dubbed "monomer rescue". For substrates ("monomers" in this work) that do not produce viable yields of desired product, suggestions for their replacement were built into the chemists' workflow in a user-friendly way which offered an informed choice to the chemists. This could be used both predictively (if the substrate was expected by the model to fail) or retrospectively (if the attempted synthesis failed experimentally), and exploited 3D shape and electrostatic potential scores for molecular similarity. Please refer to Raghavan's work for more information.

Learning the language of crystal chemistry: using concepts from natural language to model solid state chemistry
Keith Butler, UCL, UK
Keith Butler gave an interesting perspective on the use of machine learning and representations within materials chemistry, particularly focusing upon inorganic crystalline materials. He spoke at length about the work of his colleague Luis Antunes, who used autoregressive large language modelling for the generation of crystal structures. Their model, CrystaLLM was trained on millions of crystal structures of inorganic solids from the Cambridge Structure Database to then reproduce Crystallographic Information File (CIF) formats. This CrystaLLM model can be accessed online. Considering the relative simplicity of the approach, the model produces plausible crystal structures highly reliably, especially when combined with predictors of formation energy.

Enhancing drug discovery through representation learning in sequence and structure spaces
Alexis Molina, *Nostrum Biodiscovery, Barcelona, Spain*
Alexis Molina spoke about how biomolecular language models have driven significant breakthroughs in biological research. However, most models have been fine-tuned rather than used as foundational tools, missing out on enhanced representations due to high computational costs. To tackle this, he introduced, MatMulFree, a protein language model that reduces computational expenses by 30-40%, allowing training on larger protein datasets and pushing benchmark boundaries. Additionally, he presented ChaRnaBert, an RNA model with improved tokenization and an extended context window of 8096 base pairs, offering new insights into RNA biology. The models support various applications, from RNA engineering to drug-target interaction predictions, making foundational training on large datasets more accessible and impactful.

MolSnapper: conditioning diffusion for structure based drug design
Yael Ziv, *University of Oxford, UK*
Yael Ziv explained that generative models are promising for molecular design, but they still face challenges in creating molecules that bind effectively to targets and are synthetically and physically realistic. She highlighted the need for controlling the design process and incorporating prior knowledge to better tailor molecules to specific binding sites. Yael introduced MolSnapper, a new tool that conditions diffusion models for structure-based drug design by integrating expert knowledge through 3D pharmacophores. Through extensive testing on CrossDocked and Binding MOAD datasets, MolSnapper was shown to generate molecules that fit binding sites better, with high structural and chemical similarity to the originals. Additionally, it produces about twice as many valid molecules compared to other methods.

***Continuous monitoring of molecular data and model drift to improve reliability and support of QSAR model**s*
David Marcus, GSK, UK
David Marcus (GSK, UK) presented his innovative work on maintaining the reliability of predictive models in drug discovery, focusing on quantitative structure-activity relationship (QSAR) models. As drug discovery projects navigate new chemical spaces, models risk degradation, reducing their accuracy and utility. To address this, Marcus introduced QSARstudio, an in-house machine learning platform developed at GSK that supports global and local QSAR models while proactively monitoring their performance.

A key feature of QSARstudio is its ability to track model drift and usage patterns. By identifying shifts in chemical space or declining performance early, the platform allows for timely recalibration, localised updates, or experimental data augmentation. This ensures models remain fit-for-purpose, minimising disruptions to drug discovery workflows.

Marcus emphasised the significance of continuous and automated learning systems in cheminformatics. Unlike traditional static models, QSARstudio integrates real-time data and self-monitoring algorithms, enabling adaptive evolution alongside project needs. This approach reduces resource wastage and enhances decision-making. The work can be seen as similar to previous work, such as DiscoveryBus and AutoQSAR.

Building on his expertise in cheminformatics and predictive sciences, Marcus demonstrated how adaptive platforms like QSARstudio are transforming drug discovery pipelines. By ensuring model reliability and preemptively addressing performance gaps, such systems play a vital role in advancing pharmaceutical research and development.

Machine learning and AI for targeted protein degradation

Eva Nittinger, AstraZeneca, Sweden

Eva Nittinger of AstraZeneca showcased how machine learning (ML) and artificial intelligence (AI) can advance the development of PROTACs (proteolytic targeting chimeras), a groundbreaking therapeutic approach that induces protein degradation through the cell's ubiquitin-proteasome system. Unlike traditional inhibitors, PROTACs target proteins for removal, enabling intervention in previously "undruggable" targets.

Nittinger addressed the challenge of PROTACs' complex structure, comprising a protein-of-interest (POI) binder, an E3 ligase ligand, and a linker. Her team developed a "splitter" tool to systematically segment these components, allowing for detailed analysis of how each contributes to degradation efficiency. This facilitates data generation for training ML models and supports matched-pair analyses to refine PROTAC design.

By leveraging public datasets such as PROTAC-DB, and PROTACpedia her team created predictive models that assess degradation outcomes while accounting for the individual roles of PROTAC components. This approach moves beyond traditional modelling techniques, providing insights that streamline optimisation.

Nittinger's work highlights how AI can tackle complex drug discovery challenges, offering tools to design more effective and efficient PROTACs. Her research represents a significant step forward in targeted protein degradation, providing a roadmap for developing therapies with profound potential in oncology and other areas.

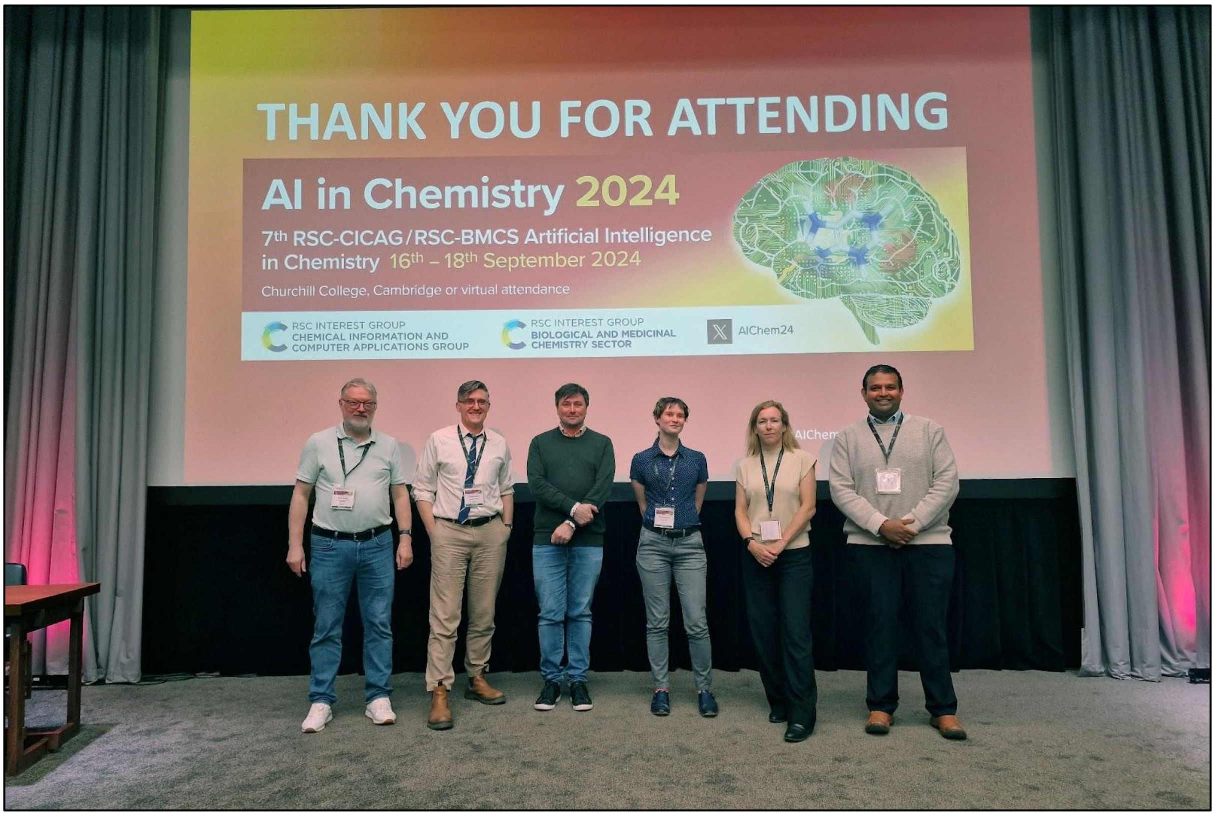
Poster prizes

The three poster prizes were kindly sponsored by RSC publishing and the winners were:

Industry: "SAFEPATH: Using AI to understand the molecular mechanisms causing safety failures, enabling drug optimisation and turnaround", Layla Hosseini-Gerami, Ignota Labs, UK.

Academia: "Predicting the biochemical activities of unidentified chemicals from MS2 spectra to pinpoint potential toxic agents", Ida Rahu, Stockholm University, Sweden.

People's Prize: "How to make machine learning scoring functions competitive with FEP", Isak Valsson, University of Oxford, UK.


THANK YOU FOR ATTENDING
AI in Chemistry 2024
7th RSC-CICAG/RSC-BMCS Artificial Intelligence in Chemistry 16th – 18th September 2024
Churchill College, Cambridge or virtual attendance
RSC INTEREST GROUP CHEMICAL INFORMATION AND COMPUTER APPLICATIONS GROUP
RSC INTEREST GROUP BIOLOGICAL AND MEDICINAL CHEMISTRY SECTOR
AIChem24

- -

CICAG's Social Media

Contribution from Dr Samantha Kanza, email: S.Kanza@soton.ac.uk

Gone are the days when tweeting was the only way of engaging with like-minded professionals with similar research interests. Here at CICAG we have been broadening our social media presence so that you can follow us and engage with us and our community on a range of platforms.

RSC-CICAG Website

rsc-cicag page, rsc-cicag group

@RSC_CICAG

@rsc-cicag.bsky.social

@rsccicag

@RSCCICAG

Why should you engage with us?

1. **Stay updated**: Through its many channels, RSC-CICAG shares the latest developments in chemical information, data management, and computer applications, keeping you informed about cutting-edge research and technological advancements.
2. **Community engagement & networking opportunities**: Whatever your preferred flavour of social media is, use it to connect with professionals, researchers, and academics in the CICAG community outside of our face-to-face meetings. This can help you connect with like-minded researchers and open doors to collaborations, job opportunities, and professional growth.
3. **Events**: CICAG organises a wide range of conferences, webinars, and training workshops. These events cover a wide range of topics, from cheminformatics to machine learning applications in chemistry. Engage with us to be the first to hear about these events, how to apply/attend, and to find out more about the bursary support that CICAG makes available.
4. **Recognition and awards**: Learn about and celebrate the achievements within the community, such as the Inspirational Committee Award for the development of the Open Chemistry series.

How can you engage with us?

Website

Our website provides details of all our events, in addition to giving details on how to formally join our interest group. You can use our website to join our group, sign up to new events, and read our bi-annual newsletters.

LinkedIn

CICAG now has both a LinkedIn Interest Group and a LinkedIn Page. If you wish to start discussions with our community, you can put them in the group or tag our page in your posts. Follow our page for all our latest updates, and please do tag us in all your posts about attending RSC/RSC-CICAG events, or in other relevant computational chemistry content.

X, BlueSky, Mastadon

CICAG has a presence on all of these, so whichever you are on you can follow us for latest updates.

YouTube

Our YouTube channel contains very informative training videos and recorded talks from our events. Follow us to see the latest videos.

- -

Finding Chemical Information on Bluesky

Contribution from Dr Chris Swain, email: swain@mac.com and Srijit Seal, Postdoctoral Research Associate in Cheminformatics, Broad Institute of MIT and Harvard, Cambridge MA, USA email: seal@broadinstitute.org

Over the years, CICAG has relied on social media to share updates, advertise events, and connect with the computational chemistry, cheminformatics and chemical information communities. LinkedIn has become a growing hub for these interactions, but X (formerly Twitter) has declined in popularity.[1] Platforms like Mastodon initially showed promise but didn't immensely gain traction due to their federated server model, which made finding content tricky.

Enter Bluesky, a fresh social media platform that feels like the early days of Twitter. Bluesky is quickly becoming a favourite among scientists looking for a place to connect with colleagues and rebuild lost communities. One of its standout features is curated Starter Packs that help you dive straight into the academic and research communities that matter most to you.

What are Starter Packs?

Starter packs are curated lists of accounts tailored to specific fields or interests – perfect for those who want to skip the hassle of finding people to follow from scratch. Whether you're into cheminformatics, AI in drug discovery or bioinformatics, these packs immediately connect you with experts, organisations, and vibrant discussions. They're like ready-made networks designed to save time and help you feel part of the community from day one.

Dr Ellis Crawford, @elliscrawford.bsky.social, Scientific Editor, at the Royal Society of Chemistry, has assembled an excellent list of Bluesky Chemistry (and related) starter packs, which is regularly updated.

To search for Starter Packs visit Starter Pack - Bluesky Directory. Here are some specifically relevant to CICAG. Click and sign up using the links below, and you will start with a group of people so you never start from zero.

- Cheminformatics Starter Pack: A curated list of social media accounts that delve into the world of chemical informatics, molecular modelling, and data analysis techniques. It's designed to help users quickly connect with relevant communities in this field
- Computational and Theoretical Chemists Starter Pack: Quantum and computational side of chemistry
- AI in Drug Discovery Starter Pack: Researchers use AI to innovate in drug discovery, from chemical structures to omics datasets
- Medicinal Chemistry Starter Pack: Developments in medicinal chemistry for researchers in the field.
- Bioinformatics and Machine Learning (BioML), Bioinformatics and AI researchers working at the cutting edge of biology and machine learning: BioML Starter Pack #1; BioML Starter Pack #2
- Drug Developers Starter Pack: Researchers involved in the drug development pipeline, from discovery to clinical and regulatory work
- AI Science: Emerging trends and thought leaders in AI across scientific disciplines; AI Science Connect Pack #1; AI Science Connect Pack #2
- Chemistry Creators and Communicators Starter Pack: A collection of accounts creating chemistry-related content, posting about chemistry for a broad audience, or writing popular chemistry books
- STEM Editorial and Publishing Starter Pack: Various people working in STEM as editors, publishers, and the like. Priority is people full-time in the industry.
- RSC Journal Editors Starter Pack: Here's a group of journal editors currently on Bluesky from the Royal Society of Chemistry
- University Presses to Follow Starter Pack: Association of University Presses members to follow
- Open Research – Library Teams and Librarians Starter Pack: Librarians and library teams interested in open research (including open access, FAIR research data management, open licensing, open educational resources and more)

Note: Some of these packs have been designed and curated by individuals on BlueSky.

Reference

(1) Like 'old Twitter': The scientific community finds a new home on Bluesky

- -

Other Chemical Information News

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

A map of every conceivable molecule could be possible with AI

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

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

Source: New Scientist

Enabling AI to explain its predictions in plain language

Using LLMs to convert machine-learning explanations into readable narratives could help users make better decisions about when to trust a model.

https://www.eurekalert.org/news-releases/1067785

Source: AAAS EurekAlert!

The changing face of Internet Search

Phil Bradley considers whether generative AI is a true "Google killer" and how newer AI technologies redefine search as conversations that transform information seeking behaviour.

https://www.infotoday.eu/Articles/Editorial/Featured-Articles/The-changing-face-of-internet-search-164285.aspx

Source: Information Today

96% of Researchers say AI will be used for Misinformation

A survey of 300 corporate researchers in industries including pharmaceuticals, life sciences and chemicals has been published by Elsevier. The 2024 Elsevier Attitudes on AI Report explores how corporate researchers feel about the use of AI and Generative AI; finding that more than a third (38%) have already used AI for work purposes and three quarters (76%) expect to use AI within the next two to five years. The survey is part of a larger Elsevier study of the attitudes towards AI of 3,000 researchers and clinicians across 123 countries.

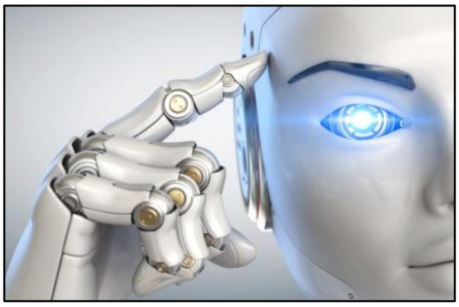


Image courtesy of Research Information

https://www.researchinformation.info/news/96-researchers-say-ai-will-be-used-misinformation/

Source: Research Information

Can Google Scholar survive the AI revolution?

The largest scholarly search engine is celebrating its 20th birthday, but AI-driven competitors offer advantages.

https://www.nature.com/articles/d41586-024-03746-y

Source: Nature

Scientists use Maths to predict Crystal Structures in hours instead of months

Researchers have devised a mathematical approach to predict the structures of crystals - a critical step in developing many medicines and electronic devices - in a matter of hours using only a laptop, a process that previously took a supercomputer weeks or months.

https://www.sciencedaily.com/releases/2024/11/241114161259.htm

Source: Science Daily

ChatGPT turns two: how the AI chatbot has changed scientists' lives

How many researchers are using the AI tool? Nature gathers data and talks to members of the academic community.

https://www.nature.com/articles/d41586-024-03940-y

Source: Nature

Clarivate launches Generative AI-Powered Web of Science Research Assistant

Web of Science™ Research Assistant is an AI-powered tool which helps researchers find key papers faster, handle complex research tasks and visualise connections. The chat interface combined with the Web of Science knowledge graph allows researchers to get more out of their interactions with 120 years of trusted publication and citation data in the Web of Science Core Collection™.

https://clarivate.com/news/clarivate-launches-generative-ai-powered-web-of-science-research-assistant-4/

Source: Clarivate

A 'chemical ChatGPT' for new Medications

Researchers have trained an AI process to predict potential active ingredients with special properties. Therefore, they derived a chemical language model -- a kind of ChatGPT for molecules. Following a training phase, the AI was able to exactly reproduce the chemical structures of compounds with known dual-target activity that may be particularly effective medications.

https://www.sciencedaily.com/releases/2024/10/241023130911.htm

Source: Science Daily

Five key trends in the shifting world of search

Expert Mary Ellen Bates breaks down the seismic shift impacting the information landscape, including the search trends everyone must pay attention to.

https://www.copyright.com/wp-content/uploads/2024/04/Five-Key-Trends-in-the-Shifting-World-of-Search.pdf

Source: CCC

The chemical enterprise braces for a second Trump presidency

Experts anticipate big changes in the life sciences and environmental policy.

https://cen.acs.org/policy/regulation/The-chemical-enterprise-braces-for-second-Trump-presidency/102/i35

Source: C&EN News

Using the world's fastest exascale computer, ACM Gordon Bell Prize-winning team presents record-breaking algorithm to advance understanding of Chemistry and Biology

High-performance computing innovation breaks million-electron and 1 EFLOP/s barrier.

https://www.eurekalert.org/news-releases/1065783

Source: AAAS EurekAlert!

Five ways Cambridge embraced AI in 2024

As new technologies such as generative AI emerge and mature, Cambridge has been harnessing opportunities and exploring ways to respond to the changing needs of learners, teachers and researchers.

https://www.cambridge.org/news-and-insights/five-ways-cambridge-embraced-AI-2024

Source: CUP

How to fix computing's AI energy problem: run everything backwards

Artificial intelligence wastes an extraordinary amount of energy - but running every computer calculation twice, first forwards and then backwards, could drastically curb that problem.

https://www.newscientist.com/article/mg26435231-300-how-to-fix-computings-ai-energy-problem-run-everything-backwards/

Source: New Scientist

Clarivate Launches Generative AI-Powered Primo Research Assistant

Primo™ Research Assistant, developed in collaboration with partners from the library community, this new generative AI-powered library discovery solution offers a seamless experience for students and researchers. It provides immediate answers to natural language queries and offers expansive visibility into sources and references.

https://clarivate.com/news/clarivate-launches-generative-ai-powered-primo-research-assistant/

Source: Clarivate

A quick journey through the expanding world of AI and Copyright Litigation

Just three short years ago, copyright litigation discussions centered around whether it is fair use to copy declaring code or make unlicensed use of Lynn Goldsmith's photographs of Prince. How AI technologies intersect with copyright was but a twinkle in most judicial systems' eyes. But in the brief time that followed, generative AI systems exploded into the public consciousness, and their interaction with copyrighted works likewise dominated copyright litigation.

https://www.copyright.com/blog/a-quick-journey-through-the-expanding-world-of-ai-and-copyright-litigation/

Source: CCC

CAS introduces world's first Stereoselective Labelling capability in CAS SciFinder® Retrosynthesis Tool

CAS has unveiled a groundbreaking feature in its SciFinder® retrosynthesis tool: stereoselective labelling. This first-of-its-kind capability in scientific technology solutions allows researchers to precisely label stereoselective steps within predictive retrosynthesis, marking a new milestone for drug development and chemical synthesis.

https://www.knowledgespeak.com/news/cas-introduces-worlds-first-stereoselective-labeling-capability-in-cas-scifinder-retrosynthesis-tool/

Source: Knowledgespeak

2025 will see AI PCs become the new normal, but ARM-based PCs will not grow out of its minority segment

In its new whitepaper, 101 Technology Trends That Will—and Won't—Shape 2025, analysts from global technology intelligence firm ABI Research. ABI Research analysts identify 54 trends that will shape the technology market and 47 others that, although attracting vast amounts of speculation and commentary, are less likely to move the needle over the next twelve months.

https://www.prnewswire.co.uk/news-releases/2025-will-see-ai-pcs-become-the-new-normal-but-arm-based-pcs-will-not-grow-out-of-its-minority-segment-302340351.html

Source: PR Newswire

An AI tool saves time in improving Protein drugs

Researchers skip high-throughput screen while optimising antibody affinity.

https://cen.acs.org/pharmaceuticals/drug-development/AI-tool-saves-time-improving/102/web/2024/07

Source: C&EN News

Enamine partners with Elsevier

Pharmaceutical chemical provider Enamine has uploaded 43 million of its make-on-demand compounds to Elsevier's Reaxys database. The compounds are synthetically feasible molecules derived from Enamine's library of building blocks and screening compounds that the company says it can synthesise in a few weeks using well-validated chemistry.

https://cen.acs.org/pharmaceuticals/drug-discovery/Enamine-partners-Elsevier/102/i38

Source: C&EN News

Paris Declaration calls for data-driven forensics to spearhead the fight against fake science

Research integrity champions say Forensic Scientometrics (FoSci) will decontaminate "polluted" science and scholarly literature.

https://www.digital-science.com/news/paris-declaration-calls-for-data-driven-forensics-to-fight-fake-science/

Source: Digital Science

Crystal-hunting DeepMind AI could help discover new wonder materials

We know of around 48,000 inorganic crystal structures, which provide materials with a range of properties. Now, an AI created by Google DeepMind has predicted over 2 million more possibilities.

https://www.newscientist.com/article/2404929-crystal-hunting-deepmind-ai-could-help-discover-new-wonder-materials/

Source: New Scientist

Understanding Compound Selectivity with Data-Driven Drug Design

Selectivity is a crucial property in the development of new active pharmaceutical ingredients (APIs). Binding site comparisons within a protein family are key to modulating the selectivity profile of a potential new API, which includes understanding both on- and off-target effects. This white paper outlines the importance of data-driven-drug design to maximise compound selectivity.

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

Source: Scientific Computing News

Clarivate Launches AI-Powered Patent Search Solution in Derwent WPI

Clarivate has announced the launch of AI Search in Derwent™. Combining AI with Derwent World Patents Index (DWPI)™, the solution will enable IP professionals to make reliable innovation decisions by finding more relevant patents in less time.

https://clarivate.com/news/clarivate-launches-ai-powered-patent-search-solution-in-derwent/

Source: Clarivate

New molecule-creation method a 'powerful tool' to accelerate Drug Synthesis and Discovery

Rice researchers develop novel two-step approach to functionalise complex molecules.

https://www.sciencedaily.com/releases/2024/12/241219190304.htm

Source: Science Daily

Wiley expands its chemical compound coverage with new SmartSpectra databases

Wiley has announced the release of two new Wiley SmartSpectra Database Collections generated using the most current machine-learning techniques to significantly expand the number of spectral data available for analysis.

https://newsroom.wiley.com/press-releases/press-release-details/2024/Wiley-expands-its-chemical-compound-coverage-with-new-SmartSpectra-databases/default.aspx

Source: Wiley

Elsevier supports Pistoia Alliance in accelerating safe and responsible AI adoption in drug discovery

Elsevier addresses AI-related challenges, including the need for trusted data, and AI transparency, at life sciences industry workshops, webinars, and conference.

https://www.elsevier.com/about/press-releases/elsevier-supports-pistoia-alliance-in-accelerating-safe-and-responsible-ai

Source: Elsevier

Open Access Partnerships Key to Increasing the Global Impact of African Research

Collaboration between libraries and publishers accelerating access to the work of African scholars and researchers in Africa is experiencing increased international reach, supported by new open OA partnerships between research libraries and publishers.

https://newsroom.taylorandfrancisgroup.com/open-access-partnerships-key-to-increasing-the-global-impact-of-african-research/

Source: Taylor & Francis

AI protein-prediction tool AlphaFold3 is now more open

The code underlying the Nobel-prize-winning tool for modelling protein structures can now be downloaded by academics.

https://www.nature.com/articles/d41586-024-03708-4

Source: Nature

AI is coming. Is the chemistry world ready?

At the 2024 BIO International Convention, AI was deemed central to chemistry's future.

https://cen.acs.org/business/informatics/AI-coming-chemistry-world-ready/102/web/2024/06

Source: C&EN News

AI-Driven Mobile Robots Team Up to Tackle Chemical Synthesis

Researchers have developed AI-driven mobile robots that can carry out chemical synthesis research with extraordinary efficiency.

https://www.sciencedaily.com/releases/2024/11/241106132220.htm

Source: Science Daily

How Is Generative AI Transforming Clinical Trial Work?

Generative AI could enhance and accelerate the way people work on clinical trials. In this Q&A, a management consultant shares his insights on benefits, risks and more.

https://www.biospace.com/career-advice/how-is-generative-ai-transforming-clinical-trial-work

Source: BioSpace

New method for producing innovative 3D Molecules

Chemists have synthesised so-called heteroatom-substituted cage-like 3D molecules. The innovative structures are created by precisely inserting a triatomic unit into the strained ring of a reaction partner. They could help address key challenges in drug design by serving as more stable alternatives to traditional, flat, aromatic rings.

https://www.sciencedaily.com/releases/2024/10/241023131343.htm

Source: Science Daily

Clarivate reveals highly cited Researchers 2024 List

Clarivate has revealed its 2024 list of Highly Cited Researchers™ – influential researchers at universities, research institutes and commercial organisations around the world who have demonstrated significant and broad influence in their field(s) of research.

https://clarivate.com/news/clarivate-reveals-highly-cited-researchers-2024-list/

Source: Clarivate

Cloud computing captures chemistry code

An innovative all-of-computing approach offers the potential for sustainable cloud computing applications to address urgent energy needs.

https://www.sciencedaily.com/releases/2024/10/241021123021.htm

Source: Science Daily

Springer Nature launches AskAdis: An AI-powered conversational interface for Pharma sector

AskAdis will provide more immediate and relevant answers to drug development and research questions as draws on exclusive validated information. AskAdis is a new cutting-edge conversational chat interface developed by Springer Nature for the pharmaceutical drug development market using its AdisInsight , a market-leading drug development intelligence database with over half a million users annually.

https://www.stm-publishing.com/springer-nature-launches-askadis-an-ai-powered-conversational-interface-for-pharma-sector/

Source: STM Publishing News

AI and quantum mechanics team up to accelerate Drug Discovery

SMU have created SmartCADD. This open-source virtual tool combines artificial intelligence, quantum mechanics and Computer Assisted Drug Design (CADD) techniques to speed up the screening of chemical compounds, significantly reducing drug discovery timelines.

https://www.sciencedaily.com/releases/2024/10/241007134022.htm

Source: Science Daily

Open-Source AI Language Model with a Distinctly European Perspective

OpenGPT-X research project releases open-source AI model Teuken-7B suporting all 24 official EU languages, with data remaining securely with its owners.

https://www.chemistryviews.org/open-source-ai-language-model-with-a-distinctly-european-perspective/

Source: Chemistry Views

Wiley Expands Advanced Journal Portfolio into Life and Health Sciences, Deepens Physical Science Offering

Led by the flagship OA journal, Advanced Science, and the world-renowned Advanced Materials , the Advanced Portfolio currently encompasses 22 high-impact titles built and driven by a team of full-time, professional editors.

https://newsroom.wiley.com/press-releases/press-release-details/2024/Wiley-Expands-Advanced-Journal-Portfolio-into-Life-and-Health-Sciences-Deepens-Physical-Science-Offering/default.aspx

Source: Wiley

40 years of scientific triumph from net zero to cancer research

https://www.ukri.org/news/40-years-of-scientific-triumph-from-net-zero-to-cancer-research/

Source: UKRI

Dimensions launches AI-based natural language feature
Digital Science's Dimensions product is launching a beta "to explore the responsible use of a new AI-based Natural Language to Query technology".
https://www.researchinformation.info/news/dimensions-launches-ai-based-natural-language-feature/
Source: Research Information

'We are in the century of the protein'
How Nobel-winning algorithms are fuelling biotech today.
https://cen.acs.org/people/nobel-prize/century-protein/102/web/2024/10
Source: C&EN News

Clarivate launches new Sustainability Research Solution
ProQuest™ One Sustainability is an expansive, curated, multi-format content collection designed to meet the growing demand for sustainability curricula across research, teaching and learning.
https://clarivate.com/news/clarivate-launches-new-sustainability-research-solution/
Source: Clarivate

From LIS to KM: How Library and Information Professionals are playing their part in the AI Revolution
In an era where generative AI is transforming how organisations manage and leverage their knowledge assets, library and information professionals are emerging as even more crucial players in the knowledge management landscape.
https://www.infotoday.eu/Articles/Editorial/Featured-Articles/From-LIS-to-KM-How-Library-and-Information-Professionals-are-playing-their-part-in-the-AI-Revolution-167047.aspx
Source: Information Today

Can AI-generated podcasts boost Science Engagement?
Researchers are using AI to keep up with the literature and spread the word about their work.
https://www.nature.com/articles/d41586-024-03960-8
Source: Nature

New Report explores the impact of GenAI on Scholarly Publishing
In a landmark analysis of the future of scholarly communication, Ithaka S+R has released a new report, A Third Transformation? Generative AI and Scholarly Publishing, investigating the profound effects of generative AI (GenAI) on academic publishing. The report, supported by STM Solutions and leading publishers such as the ACS, IEEE, Elsevier, Springer Nature, Taylor & Francis, and Wiley, highlights the rapid evolution of GenAI and its potential to reshape research workflows, operational strategies, and communication processes.
https://www.knowledgespeak.com/news/new-report-explores-the-impact-of-genai-on-scholarly-publishing/
Source: Knowledgespeak

How a Biochemistry Department used redacted job applications to achieve gender parity
Anonymising job applications before shortlisting helped to boost the number of women appointed.
https://www.nature.com/articles/d41586-024-03944-8
Source: Nature

Cyberthreats are growing – so are Patents for technology to combat them
Patent data analysis highlights the leading companies in cybersecurity innovations.
https://www.digital-science.com/news/cybersecurity-patents-growing/
Source: Digital Science

Taylor & Francis Announces Subscribe to Open Journals Pilot

Taylor & Francis has announced its first Subscribe to Open (S2O) pilot, one of several innovative options it is trialing to accelerate OA publishing. S2O enables a journal's subscribers to support its conversion to OA, making new articles available to readers everywhere.

https://newsroom.taylorandfrancisgroup.com/taylor-and-francis-announces-subscribe-to-open-journals-pilot/

Source: Taylor & Francis

First draft of the General-Purpose AI Code of Practice published by Independent Experts

The first draft of the General-Purpose AI Code of Practice has been released, marking a significant step in ensuring safe and trustworthy development of AI models. The draft, prepared by independent experts and facilitated by the European AI Office, will be further discussed with around 1,000 stakeholders. This milestone represents the completion of the initial phase of a four-stage drafting process set to conclude by April 2025.

https://www.knowledgespeak.com/news/first-draft-of-the-general-purpose-ai-code-of-practice-published-by-independent-experts/

Source: Knowledgespeak

Addressing common assumptions about Copyright & AI

Copyrighted materials are the fuel for artificial intelligence (AI) systems, but misunderstandings persist about how copyright applies to the use of content as training material for AI models.

https://www.copyright.com/resource-library/insights/addressing-common-assumptions-about-copyright-ai/

Source: CCC

AI-generated images threaten science — here's how researchers hope to spot them

Generative-AI technologies can create convincing scientific data with ease — publishers and integrity specialists fear a torrent of faked science.

https://www.nature.com/articles/d41586-024-03542-8

Source: Nature

AI Is accelerating Biopharma innovation but not erasing a human's touch

Artificial intelligence won't replace people in biopharma, but it is infiltrating every step of drug development, including in some ways that aren't so obvious.

https://www.biospace.com/business/ai-is-accelerating-biopharma-innovation-but-not-erasing-a-humans-touch

Source: BioSpace

5 AI-Related Topics every Information Professional should think about in 2024

Learn how information professionals can approach the changing environment caused by the rapid-fire advancements in AI technology to raise the profile of the information center and provide value.

https://www.copyright.com/wp-content/uploads/2024/01/5-AI-Related-Topics-in-2024-Tip-Sheet.pdf

Source: CCC

AI will surpass human brains once we crack the 'Neural Code'

Understanding how 'visual thinking' works is key to building human-level AI, says AI expert It may be possible for computers to emulate a type of consciousness, he suggests Expert also warns that society must control AI technology and have 'sole control of the off switch'.

https://newsroom.taylorandfrancisgroup.com/ai-will-surpass-human-brains-once-we-crack-the-neural-code/

Source: Taylor & Francis

ReadCube expands Literature Management with new AI Assistant and Comprehensive Search
Applying AI to literature workflows with ReadCube gives researchers more time to focus on life-changing discoveries.
https://www.digital-science.com/news/readcube-expands-literature-management-ai-assistant-comprehensive-search/
Source: Digital Science

From Research to Manufacturing: Overcoming Data challenges in the Drug Development Lifecycle
Data decisions made at each step of the drug development process – from target identification and discovery right through to final recipes in manufacturing – have an impact on time-to-market. In pharmaceuticals, as in most industries, time is money, so any aspect that can be modified to shorten the time taken at any step will bring about cost efficiencies.
https://www.scientific-computing.com/white-paper/research-manufacturing-overcoming-data-challenges-drug-development-lifecycle?
Source: Scientific Computing News

The AI revolution is running out of data. What can Researchers do?
AI developers are rapidly picking the Internet clean to train large language models such as those behind ChatGPT. Here's how they are trying to get around the problem.
https://www.nature.com/articles/d41586-024-03990-2
Source: Nature

Wiley expands KnowItAll Suite with advanced LC-MS tools, enhanced Data Management, and Spectral Innovations
Wiley has announced the release of KnowItAll 2025, the latest version of its powerful software suite for spectral analysis and analytical data management. This new version introduces significant advancements, including a revolutionary tool for automating LC-MS analysis, streamlined enterprise data management, and expanded spectral capabilities, making it a comprehensive solution for researchers and analysts.
https://www.knowledgespeak.com/news/wiley-expands-knowitall-suite-with-advanced-lc-ms-tools-enhanced-data-management-and-spectral-innovations/
Source: Knowledgespeak

Sage acquires the Scientific and Medical Publisher Mary Ann Liebert
Sage, a leading independent academic publisher, has acquired Mary Ann Liebert, Inc., a renowned global media company publishing more than 100 peer-reviewed journals in biotechnology and the life sciences, specialised clinical medicine, public health and policy, and technology and engineering, as well as the leading B2B media brands *Genetic Engineering & Biotechnology News* (GEN) and *Inside Precision Medicine*.
https://www.eurekalert.org/news-releases/1068221
Source: AAAS EurekAlert!

cOAlition S releases independent study assessing the impact of Plan S on Scholarly Communication
cOAlition S has announced the release of an independent study evaluating the impact of Plan S on the academic publishing landscape, five years after its launch. Conducted by scidecode science consulting, the study, titled "Galvanising the Open Access Community: A Study on the Impact of Plan S", offers an in-depth assessment of the policy's influence on the push for full and immediate Open Access.
https://www.knowledgespeak.com/news/coalition-s-releases-independent-study-assessing-the-impact-of-plan-s-on-scholarly-communication/
Source: Knowledgespeak

Digital Science unveils Papers Pro: Revolutionising Scholarly Research with advanced AI-powered Features
New AI features enable users to easily discover and engage with research articles.
https://www.digital-science.com/news/papers-pro-scholarly-research-ai-powered-features/
Source: Digital Science

Elsevier partners with Pistoia Alliance to drive safe AI adoption in Drug Discovery
In a strategic move to accelerate the responsible use of AI in drug discovery, Elsevier has announced a commitment to support the Pistoia Alliance, a global non-profit organisation promoting collaboration in life sciences. This partnership seeks to address core challenges in AI adoption within the pharmaceutical and research sectors, including the need for trusted data, transparency in AI applications, and effective regulatory alignment.
https://www.knowledgespeak.com/news/elsevier-partners-with-pistoia-alliance-to-drive-safe-ai-adoption-in-drug-discovery/
Source: Knowledgespeak

World-leading researchers named as Turing AI fellows
Three leading AI researchers have been appointed to lead bold new work through Turing AI World-Leading Researcher Fellowships.
https://www.ukri.org/news/world-leading-researchers-named-as-turing-ai-fellows/
Source: UKRI

New report suggests open data on edge of becoming a recognised global standard for scholarly output
Latest report in the State of Open Data series, released by partners Digital Science, Figshare and Springer Nature, provides quantitative analysis on growth of open data sharing globally.
https://group.springernature.com/gp/group/media/press-releases/state-of-open-data-report-2024/27724212
Source: Springer Nature

The great Pharmaceutical-Academic Merger
As drug companies fret over their finances, they are increasingly partnering with universities to help with early-stage research.
https://cen.acs.org/pharmaceuticals/drug-discovery/great-pharmaceutical-academic-merger/102/i31
Source: C&EN News

Publishers are selling papers to train AIs — and making millions of dollars
Generative-AI models require massive amounts of data — scholarly publishers are licensing their content to train them.
https://www.nature.com/articles/d41586-024-04018-5
Source: Nature

Modern AI systems have achieved Turing's vision, but not exactly how he hoped
A recent perspective published Nov. 13 in Intelligent Computing, a Science Partner Journal, asserts that today's artificial intelligence systems have finally realised Alan Turing's vision from over 70 years ago: machines that can genuinely learn from experience and engage in human-like conversation.
https://www.eurekalert.org/news-releases/1068915
Source: AAAS EurekAlert!

CAS and PetroChina Shanghai Advanced Materials Research Institute announce a collaboration to accelerate new materials discovery and innovation

CAS and PetroChina Shanghai Advanced Materials Research Institute Co., Ltd are collaborating for use of the CAS SciFinder Discovery Platform™ to accelerate research and discovery of new chemical materials.
https://www.prnewswire.co.uk/news-releases/cas-and-petrochina-shanghai-advanced-materials-research-institute-announce-a-collaboration-to-accelerate-new-materials-discovery-and-innovation-302338527.html
Source: PR Newswire

2024 Heroes of Chemistry celebrated at ceremony in Washington, DC

Scientific teams in industry are honored for their contributions to benefit humankind.
https://cen.acs.org/acs-news/2024-Heroes-Chemistry-celebrated-ceremony/102/web/2024/10
Source: C&EN News

Real-world chemists are more diverse than generative AI images suggest

Asking children 'What does a scientist look like?' now results in more illustrations of women and people of color than decades ago. But do generative artificial intelligence (AI) tools also depict the diversity among scientists? Researchers prompted AI image generators for portraits of chemists. They found that none of the collections accurately represents the gender, racial or disability diversity among real chemists today.
https://www.sciencedaily.com/releases/2024/11/241120121558.htm
Source: Science Daily

Nvidia is going for quality not quantity with AI chip Patents

Patent analysis shows how world leader in microchips has risen to the top in a constrained global market.
https://www.digital-science.com/news/nvidia-is-going-for-quality-not-quantity-with-ai-chip-patents/
Source: Digital Science