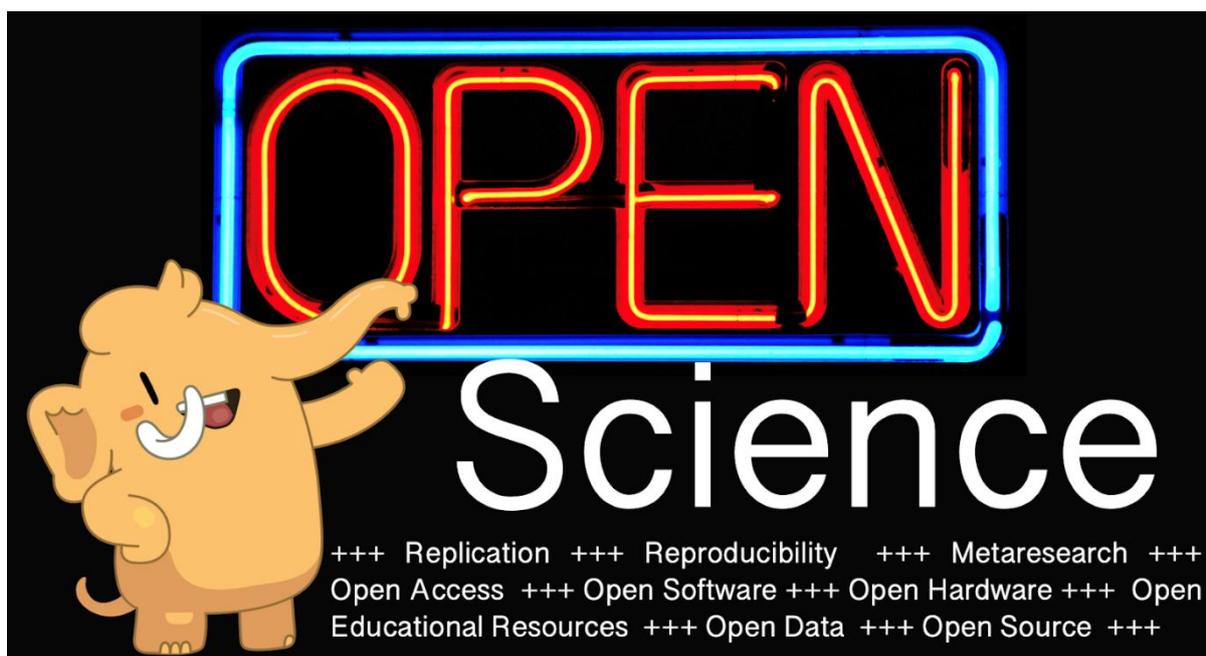




RSC INTEREST GROUP CHEMICAL INFORMATION AND COMPUTER APPLICATIONS GROUP

NEWSLETTER Winter 2022-23

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



Landing page of the [Online Mastodon directory](#).

CICAG Websites and Social Media

The screenshot shows the RSC website header with the logo and navigation menu. The main content area features a large banner for the Chemical Information and Computer Applications Group with the text "Connect with colleagues in your field and become involved in your scientific community". There are two prominent buttons: "Forthcoming events" and "How to join".

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

The screenshot shows the RSC Interest Group website for the Chemical Information and Computer Applications Group. It includes a "WELCOME" message, a navigation menu, and a "Welcome" section with a description of the group. There are also sections for "YouTube Channel" and "Newsletters". A "Tweets by @RSC_CICAG" section is visible on the right side.

<http://www.rscicag.org>



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



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



@RSC_CICAG

https://twitter.com/RSC_CICAG

Contents

Chemical Information and Computer Applications Group Chair's Report	4
CICAG Planned and Proposed Future Meetings	5
Social Media Migration – Opening up Mastodon as a Tool for Scholarly Communication.....	5
Cheminformatics: A Digital History – Part 2. A Personal Perspective of the Role of the Web During the Period 1993-1996.....	9
InChI Technical Developments.....	13
Update from the Royal Society of Chemistry Library	14
The Open Free Energy Project.....	16
Meeting Report: SCI-RSC Workshop on Computational Tools for Drug Discovery 2022	19
A Crystallography Papermill: The CSD Response.....	20
Meeting Report: Ultra-Large Chemical Libraries	22
Open Science in the Royal Society of Chemistry	25
The Davy Notebooks Project.....	27
This JACS Does Not Exist: Generating Chemistry Abstracts with Machine Learning	29
Meeting Report: RSC-CICAG and RSC-BMCS 5th Artificial Intelligence in Chemistry Conference.....	31
EU-OPENSREEN ERIC: an Open-access Research Infrastructure for Chemical Biology and Early Drug Discovery	47
DECIMER – An Open Toolkit for Optical Chemical Structure Recognition and Document Analysis.....	52
Cryo-EM for Industrial-Scale Structure-Based Drug Design.....	55
Cryo-EM & Drug Discovery	58
2022 CSD Updates	62
News from ACS CINF.....	63
News from CAS	64
RSC Databases Update.....	66
UKeiG: Winners of the Prestigious Tony Kent Strix Award 2022.....	68
AI4SD News	69
Book Review: Digital Transformation: New Tools and Methods for Mining Technological Intelligence	74
Cheminformatics and Chemical Information Books.....	74
2022 Reflections on Life at the Catalyst Science and Discovery Centre and Museum in Widnes	77
Other Chemical Information News	79

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

Chemical Information and Computer Applications Group Chair's Report

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

I am very pleased to report that Sam Whitmarsh has been co-opted to the CICAG Committee. Sam's PhD, awarded by the University of Bristol, was on the synthesis of novel steroid macrocycles and measurement of their complexes by high resolution mass spectrometry. In 2007 he joined AstraZeneca as an Analytical Scientist, leading analytical components of drug development across pre-clinical to Phase III drug candidate projects. In 2010, Sam moved to BP and led the formation of the high-resolution mass spectrometry petroleomics facility in the UK. In 2020 Sam joined CatSci Ltd as Head of Scientific Operations becoming Director of Digital Transformation in 2022.

As events slowly return to normality after lockdown there have been several in-person CICAG events. The first was on [Ultra Large Chemical Libraries](#) (10 August) at Burlington House followed by the 5th RSC-BMCS / RSC-CICAG [Artificial Intelligence in Chemistry](#) (1-2 September) at Churchill College Cambridge. CICAG also helped run the [SCI-RSC Workshop on Computational Tools for Drug Discovery 2022](#) in collaboration with the SCI. Reports from all meetings are in this Newsletter.



Chris and others visit the Observatory at Catalyst.

I was delighted to be invited to be the CICAG representative at the [Catalyst Science Discovery Centre and Museum](#) in Widnes at the opening of the Wellcome Trust – Inspiring Science Fund Project. We were also given a tour of the centre and I can recommend a visit if you are in the area.

Social media became an increasingly important way for communicating with members (and non-members) during lockdown and the trend continues, with [Twitter](#) now having 1506 followers, and [LinkedIn](#) with 607 members, we will keep an eye on Twitter and may create a Mastodon account. The CICAG [website](#) is often updated and we would be very interested to hear suggestions for additional content for all channels. CICAG's [YouTube](#) channel now has 922 subscribers and contains the 13 video presentations from AI4Proteins meetings in addition to all 20 of the [Open-Source Tools for chemistry](#) workshops. These workshop videos have proved to be very popular and have been watched a nearly 27,000 times.

In addition to the meeting reports, this newsletter also includes a fascinating personal historical perspective on the role of the web on cheminformatics from Henry Rzepa, an update on the Open Free Energy project, a piece by Anna Rulka on Open Science at the RSC and a description of DECIMER an Open Toolkit for Optical Chemical Structure Recognition and Document Analysis. There are also a couple of articles highlighting the increasing use of cryo-EM in drug discovery.

CICAG came into existence in 2007 with the merger of the Chemical Information and Computer Applications groups. Currently membership stands at nearly 700 members and is increasing steadily. Whilst RSC members can join interest groups for free, in practice many members do not take up this opportunity. You can make a request to join a group via email (membership@rsc.org), telephone (01223 432141) or via the [website](#). Once again, I'd like to invite contributions to the CICAG Newsletter that would be of interest to the CICAG community. Please contact the Newsletter Editor, [Helen Cooke](#), or me to discuss your ideas.

CICAG Planned and Proposed Future Meetings

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

Meeting	Date	Location	Further Information
Solutions in Science	4-6 July 2023	Cardiff	Organised by RSC Separation Science group, supported by CICAG. See SINS website .
6th Artificial Intelligence in Chemistry Meeting	4-5 Sept 2023	Churchill College, Cambridge	Joint event from RSC-CICAG and RSC-BMCS division. See RSC Events page.
Centenary of Markush Structures	Q3/4 2023	Burlington House	Details to follow.
Python for Chemists	TBD	TBD	Details to follow.
Molecular Simulation and Free Energy Methods	TBD	TBD	Details to follow.
7th Artificial Intelligence in Chemistry Meeting	July 2024	TBD	Organising committee decided to bring the date forward to avoid a clash with other events, school holidays, and the start of school term.

Social Media Migration – Opening up Mastodon as a Tool for Scholarly Communication

Contribution from Susann Auer, Maximilian Frank, Rima-Maria Rahal and Guido Scherp. German Reproducibility Network, email: info@reproducibilitynetwork.de

Researchers on social media – a love story

For many researchers, using social media has become part of their academic life, for a variety of reasons. Displaying their own achievements and insights, staying up-to-date with developments in the field, and informal exchange with colleagues are central motivators for individual researchers to create an online presence on platforms like Twitter, Facebook, LinkedIn or Mastodon.^{1,2} Social media offers researchers a global meeting point where research structures would make it difficult to reach colleagues beyond localised communities.

In addition to professional exchange, many scientists seek to contribute to the solution of societal challenges in their activities on social media. Whether in the scope of discussing research findings regarding COVID and public health, climate action, employment conditions in academia or research practices, sharing facts, opinions and arguments on social media is part of researchers' engagement for what they consider the right course of action. The Open Science movement in particular has built a strong community on social media, where researchers support each other and share best practices in creating robust and transparent research.

This discourse is not limited to researchers' debates among peers. Science communication, i.e. the communication about research findings with the public, further draws researchers towards tools with which they can reach out to non-scientific audiences, beyond the ivory tower. As the importance of science communication increased, so did the importance of Twitter as a tool to reach a wider audience.³

Musk takes over Twitter – a wake-up call

Where researchers' discourse takes place (or should take place) has been the object of much debate in the past, with some propagating the benefits of mailing lists while others advocate the advantages of social media to reach an interested public. In the Open Science movement, however, using for-profit platforms built on non-open source solutions has often been regarded as a problematic issue. A peer-led platform that meets the needs of the community would have been the "right" way to situate the discourse from the start, but practicalities have often hindered the adoption of such platforms.

Where social media has been used, however, this practice has always been bound to the policies implemented by the platform in question. Platform policy problems – or changes in the policies – have often sparked further disagreement on whether researchers should subject their discourse to these conditions.

Recently, the take-over of Twitter by Elon Musk has given rise to a renewed discussion about what a good forum for researchers' exchange would be.⁴ With the potential for hate speech and fake news sky rocketing due to layoffs and neglected content moderation, revised account verification procedures and reinstating previously blocked accounts – the famous one being the former president Donald J. Trump whose account was under a "lifelong" ban from the platform – Twitter seems in bad shape. Further, the legal usage of the platform in the EU stands on feet of clay: due to massive layoffs in the privacy department, it is unclear whether Twitter can fulfil the demands of the EU's General Data Protection Regulation (GDPR). Investigations by the Irish Data Protection Commission, where Twitter is registered for offering services in the EU, are currently ongoing.⁵

All of these points combined have left many users, not only scientists, questioning whether Twitter is still a viable platform. Through this adversity, a window of opportunity has opened for relocating the scholarly discourse to a different platform.

Social media migration – a search for alternatives

Looking for alternatives to Twitter, the attention is drawn to [Mastodon](#). Mastodon is a decentralised microblogging service launched in 2016 by the German Software Developer Eugen Rochko and runs on open source software. Due to the distributed approach, there is no longer only one central instance, but rather several instances are connected to a (global) network. After creating an account on an instance, users can also switch between them, taking their profile and followers with them. Furthermore, Mastodon uses open communication protocols that allow exchange with other platforms, which makes Mastodon part of the so-called [Fediverse](#). In other words, an overarching, federated network of different platforms. One key feature of these open communication protocols is a federated identity. With Mastodon, you only need one account on one instance to interact with other instances as well. And since the software is open source and thus freely available, in principle anyone can run an own instance that becomes part of the Mastodon network.

Thus, many researchers have taken the opportunity to explore Mastodon, in the hope of finding a platform that fits their needs better – one that works against fake news and hate speech, with codes of conduct enforced by community moderators, and is run by the community itself (and not by commercial players) on the basis of open source software. In the end, one may be faced with the question of whether and when it is worth running an own Mastodon instance. This sounds quite tempting, especially if there is no suitable instance for your own "community". But operating such a (community-led) instance, there are a number of things to consider in order to create trust and a reliably available platform. In addition to data protection (data policy, GDPR compliance) and data security (backup, updates), this includes transparent community rules and how and by whom compliance is ensured.⁶ Here, the scientific community should cooperate more closely and try to agree on common standards.

While motivations for moving to a new platform can certainly be individual, there is a clear correlation with the particular decisions made by Twitter's new management. When a problematic policy change was announced, the number of new users on Mastodon reached new heights (see Figure 1). The service now has more than eight million users across all instances (compared to 238 million daily active users on Twitter worldwide, eight million of whom are in Germany).⁷ Especially for the academic community, there seems to be momentum for a substantial and sustainable shift to Mastodon. For individuals, the barriers to jump on Mastodon are quite low, while for institutions, the switch is a bit more complex. Usually, they have already invested resources in building a social media presence on Twitter and established appropriate processes for publishing their content. In addition, a Twitter account of a university is not primarily aimed at scientists but rather targeted at the interested public. There is a risk that these followers will not leave Twitter to the same extent as the scientific community, and therefore an institutional account will lose a larger proportion of its followers when switching to Mastodon.⁸

New mastodon accounts (daily)

Data from @mastodonusercount@bitcoinhackers.org

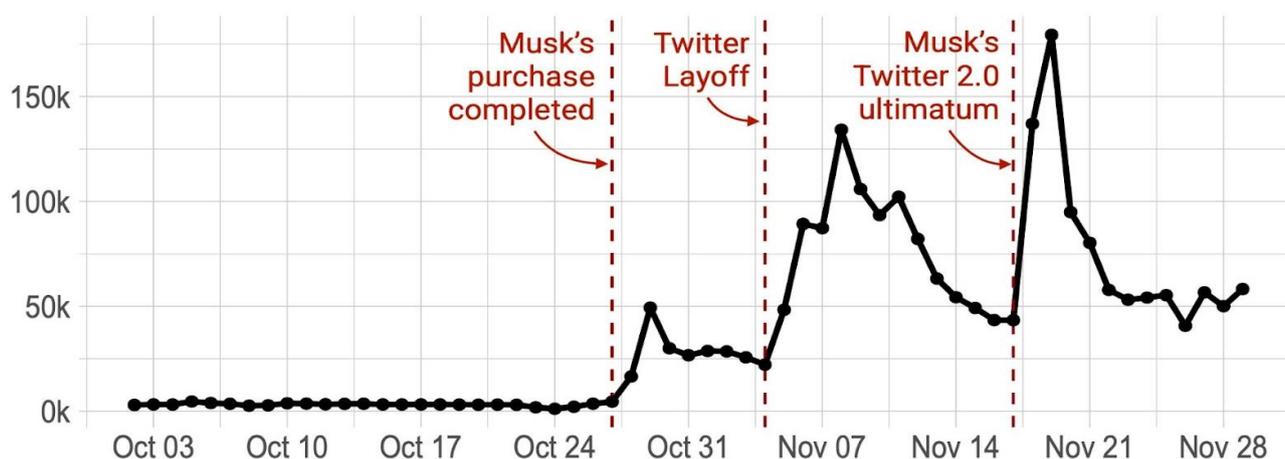


Figure 1: New Mastodon accounts matched with Twitter policy changes.⁹

The way people interact on Mastodon differs from Twitter, in some areas quite considerably. While on Twitter the engagement is more performative and metric-driven due to the number of likes and retweets, on Mastodon it feels more like a natural conversation with individuals.¹⁰ Mastodon offers tools such as content warnings that hide triggering pictures and content until a user actively clicks to open it which might help people feel safer on this platform.¹¹ Moderators and users themselves can block whole anti-social instances and thus [limit exposure to harassment and negativity](#) from individuals, trolls and bot armies.

Mastodon also offers tools that researchers in particular may find helpful. For instance, the instance [mastodon.xyz](#) renders LaTeX input. For users in need of communicating precisely with formulas and equations, such an instance offers advantages that one-size-fits-all platforms cannot. Users can – and do – continue to develop the software behind Mastodon, one strength of a community driven platform based on openly available code.

The open science community on Mastodon – finding and being found

In addition to the technical differences compared to Twitter, which are certainly unfamiliar at first, but which one can get used to, there are also social hurdles for a change. Every social media platform lives from the

interaction of people and the content they share there. Twitter wouldn't be a place of interest without the Twitter bubble. Thus, when switching to a new platform, everyone faces the potential risk of losing their social ecosystem, especially if other people don't change as well.

Many users have started to add their Mastodon handle to their Twitter account name or biography to express their move to the new platform and to help their followers find them on Mastodon. Another route to staying connected is offered by [phone-book-like compendia of academics](#) on Mastodon that have recently begun to surface. To aid the Open Science community in the transition towards Mastodon, the [German Reproducibility Network](#) (GRN)¹² also provides such an online directory (Figure 2). It allows researchers who self-identify as members of the Open Science community to add their Mastodon handle and to bulk-follow the already listed accounts thus practically solving the aforementioned problem of losing the social bubble while leaving Twitter. Currently comprising 200 entries and awaiting more community members joining the list, this tool makes it easy to locate and follow accounts dedicated to the many facets of the Open Science movement. Researchers' keywords indicate their work and interests range from open data to open source software, open educational resources and practices, to meta-research and big-team science.

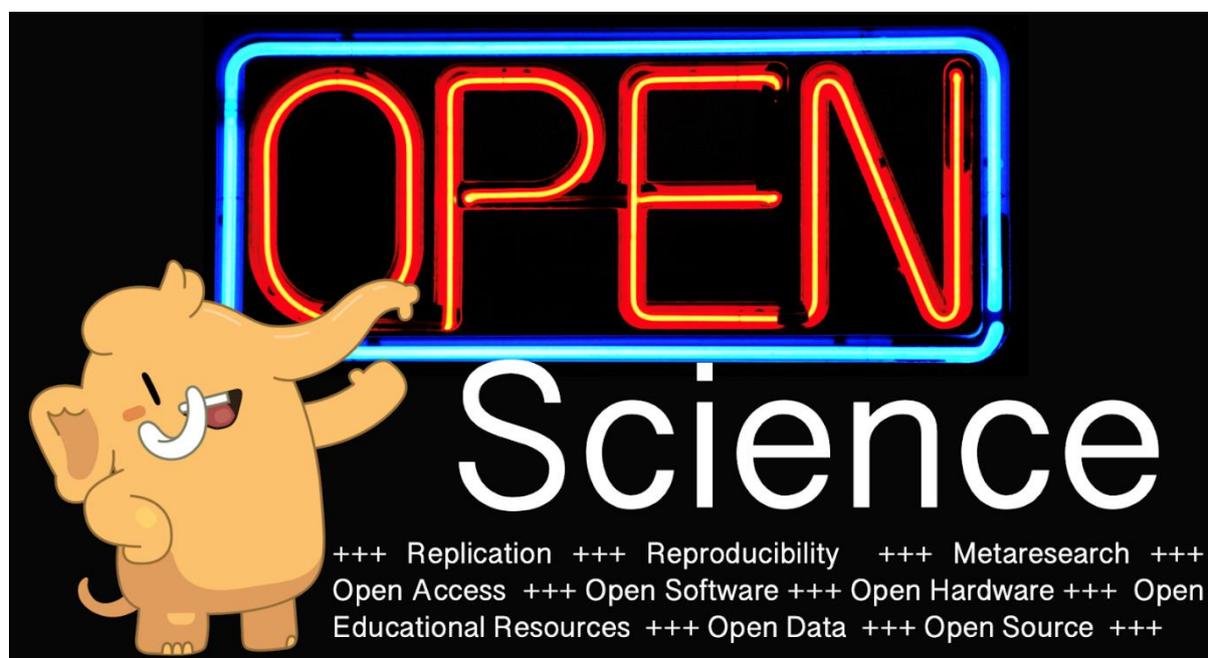


Figure 2: Landing page of the Online Mastodon directory.¹³

While we wish the Mastodon community continued growth, it remains unclear in which direction researchers' use of these platforms will develop. It is conceivable that more and more scholars would gradually turn their backs on Twitter and set up their new home there. It would also be possible that technical hurdles or a lack of interesting content cue a return to Twitter. In the transition phase, many users are working with cross-posting on both platforms to make their content accessible. Although practical, this comes at the risk of creating less incentive for others to switch platforms, since the content can still be found in the familiar environment of Twitter. Either way, the disruptive development of Twitter has brought movement to the social media landscape and it will be exciting to see how it and its scientific users evolve in the coming year. Tools such as the GRN Mastodon Open Science directory can help lower the barriers for switching to Mastodon and to explore the opportunities of a community-operated platform.

Notes and references

- (1) Small, G. Time to tweet. *Nature*. 2011, 479(141). <https://doi.org/10.1038/nj7371-141a>
- (2) Smith, A. "Wow, I didn't know that before; thank you": How scientists use Twitter for public engagement. *Journal of Promotional Communications*. 2015, 3(3). <http://www.promotionalcommunications.org/index.php/pc/article/view/61>
- (3) Côté, I.M.; Darling, E.S. Scientists on Twitter: Preaching to the choir or singing from the rooftops? *FACETS*. 2018, 3(1), 682-694. <https://doi.org/10.1139/facets-2018-0002>
- (4) Campbell, M. Should #SciTwitter migrate elsewhere? *Technology Networks*. 07.11.2022. <https://www.technologynetworks.com/tn/articles/should-scitwitter-migrate-elsewhere-367346>
- (5) Lomas, N. Is Elon Musk's Twitter about to fall out of the GDPR's one-stop shop? *TechCrunch*. 14.11.2022. <https://techcrunch.com/2022/11/14/is-elon-musks-twitter-about-to-fall-out-of-the-gdprs-one-stop-shop/>
- (6) ZBW MediaTalk-Team. Self-organised network: does Mastodon have what it takes to become the "scholarly-owned social network"? *ZBW MediaTalk*. 30.11.2022. <https://www.zbw-mediatalk.eu/2022/11/self-organised-network-does-mastodon-have-what-it-takes-to-become-the-scholarly-owned-social-network/>
- (7) Another estimate of 486 million users refers to the number of accounts in total and not the number of daily active users. Twitter 22.07.2022. Number of Twitter's monetizable daily active users (mDAU) worldwide from Q4 2020 to Q2 2022. In *Statista*. Zugriff. 08.12.2022. <https://de.statista.com/statistik/daten/studie/1032299/umfrage/monetarisierbare-taeglich-aktive-nutzer-von-twitter-weltweit/>
- (8) Tattersall, A. Academics can easily leave Twitter's town square, but it will be much harder for their institutions. *LSE Blogs*. 10.11.2022. <https://blogs.lse.ac.uk/impactofsocialsciences/2022/11/10/academics-can-easily-leave-twitters-town-square-but-it-will-be-much-harder-for-their-institutions/>
- (9) Moro, E. 30.11.2022. <https://datasci.social/@estebanmoro/109429814361434072>
- (10) Stokel-Walker, C. Should I join Mastodon? A scientists' guide to Twitter's rival. *Nature*. 10.11.2022. <https://www.nature.com/articles/d41586-022-03668-7>
- (11) Navarro, D. Everything I know about Mastodon. A hastily written guide for data science folks trying to navigate the fediverse. 03.11.2022. https://blog.djnavarro.net/posts/2022-11-03_what-i-know-about-mastodon/
- (12) The German Reproducibility Network (GRN) was founded in 2019 and aims to increase trustworthiness and transparency of scientific research. GRN's activities span multiple levels, including researchers, institutions and other stakeholders (e.g., funders, publishers, and Academic Societies). For more information visit: www.reproducibilitynetwork.de
- (13) German Reproducibility Network. 2022. Open science on Mastodon. <https://germanrepro.github.io/Mastodon-OpenScience/>

Cheminformatics: A Digital History – Part 2. A Personal Perspective of the Role of the Web During the Period 1993-1996

Contribution from Henry S. Rzepa, Department of Chemistry, Imperial College London, email: h.rzepa@imperial.ac.uk

Cheminformatics was a relatively mature subject by 1993, but the explosion of the world-wide web onto the scene around that period induced a seminal change in how the subject was exposed to and exploited by most chemists. Here I recount some aspects of how this came about, mostly from the personal perspective of "I was there when it was happening".¹

I should start the story a few years earlier around 1988, and the scene is an editorial advisory board meeting at the Royal Society of Chemistry, specifically relating to the journal *Perkin Transactions II*. This was the primary vehicle for publishing in the expanding area of computational organic chemistry, accompanied by the realisation by authors and editors alike that the traditional mechanisms for making available the data underpinning the modelling were very much not



fit for purpose. They consisted of printing the information onto paper and then sending it to the British library to be deposited in a box in their cellars at Boston Spa, with a label indicating the journal article it related to. I once remember phoning them up and asking about this “supporting information” for a particular article. The curator told me that very few of the boxes containing these papers were ever asked for by people such as myself; most would probably decay into illegibility without ever being inspected again. Against this background, I was tasked at the RSC editorial meeting with investigating more modern electronic methods for depositing supporting information; thus ESI or electronic supporting information was born. Over the next few years various Internet protocols such as FTP (file transfer protocol) and Gopher appeared, but these did not seem sufficiently seamless to incorporate into the procedures associated with the then still firmly paper-bound journal publishing and cheminformatics procedures.

In May of 1993, I wandered into the office of a colleague I knew, in the then named computer centre. He was someone who had almost a hobby of keeping abreast of recent developments in how documents were starting to be handled on the Internet. On this occasion he showed me what appeared to be yet another protocol, called the World-Wide-Web or WWW for short. It was about four years old then and starting to make waves amongst computer informaticians, having emerged from the high energy particle physics laboratories at CERN in 1989. I made two requests. Show me some chemistry. After a few minutes, we had not found any – remember, this was in the days before Google. Not entirely convinced, I asked my second question: what language would I need to learn to create some Web resources? He showed me some HTML source code and I went away only mildly intrigued. About a month later the realisation struck me that if I learnt HTML, I could indeed start to populate the Web with chemistry. Of course, I was not in fact the first to have this thought. I give the credit at least in part to Ben Whitaker at Leeds and Mark Winter at Sheffield, who had started a year or so earlier. At any rate, by July 1993 I had found someone in the computer centre who was prepared to install a Web server for me and so the domain www.ch.ic.ac.uk was born. It has been serving continuously since 1993. Well, 29 years old may not seem that old, but I think it must rank amongst the most veritable.

My first lecture course went on-line in November 1993² and soon more content followed. After 29 years, this web server is replete with what might now be considered historically significant documents constituting diverse experiments in the medium. Soon after, I discovered a world index of Web servers (it had about 50 entries then, it has long since stopped being updated) which in turn led me to Ben and Mark. By May 1994, between the three of us we had accumulated enough examples to convince ourselves that wider dissemination of this potentially wonderful cheminformatics tool should be attempted. We decided to write an article for *Chemical Communications*, an RSC journal. I should imagine the editor had an interesting challenge finding referees, who themselves must have been intrigued by this rather unusual article, reporting no new molecules or chemical research, but it was indeed published.³ It covered a wealth of applications, amongst them suggestions on how electronic journals of the future could exploit the mechanisms. About that time, a funding call from the UK JISC funding board went out for projects and a group of us submitted one based on our vision of how an electronic journal might offer up chemical information and data wrapped in highly accessible visual presentations.⁴ That RSC editorial board meeting in 1988 had finally borne fruit! Nowadays, one of the prime aspects of this project,⁵ which *inter alia* involved embedding rotatable molecule models within journal articles, chemical information resources such as the CCDC⁶ and databases and repositories, is now routine (Figure 1).⁷

In May of 1994, I also went with Peter Murray-Rust to the inaugural World-Wide-Web conference at CERN⁸ to get a perspective on the future of this phenomenon. I was running the chemistry workshop there and had been given an SGI computer on which to display examples. Unfortunately it had next to no software installed and I still remember the excitement when Peter and I had to compile RasMol⁵ from scratch about five minutes before the workshop started. We succeeded and the demonstrations went ahead on time with 10 seconds to spare!⁸ I recollect another meeting there; Tim Berners-Lee was on hand to talk about the future of HTML, and I

approached him to enquire about an equivalent that might handle chemistry! He immediately suggested we adapt MathML, which was (then a proposed) display markup language for mathematical expressions. I think after about 30 minutes of discussing how benzene might be represented by MathML, we both realised a new approach would be needed. That new approach is described below.

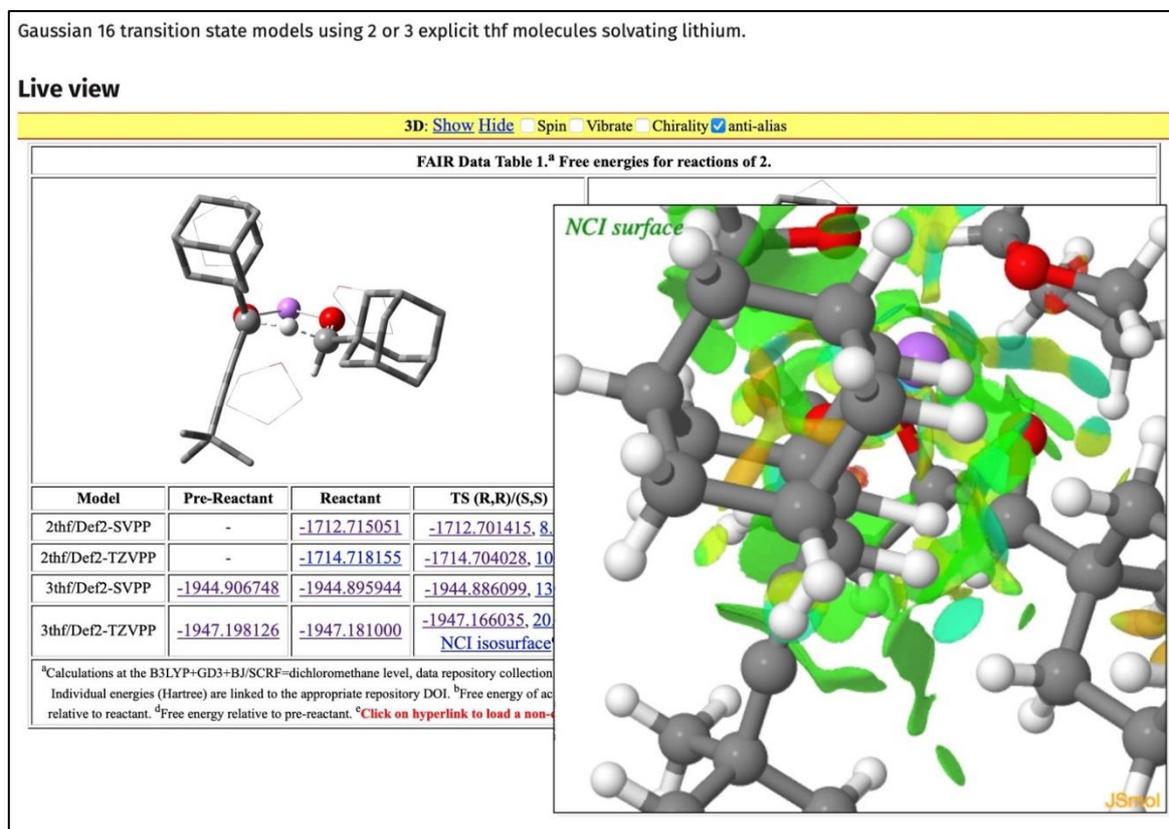


Figure 1. An interactive 3D model showing an isosurface for interpreting molecular non-covalent-interactions as part of a FAIR data table in a journal article.⁷

In parallel with these Web activities, in late 1993 Peter Murray-Rust, Ben Whitaker and I had been pondering the chemical application of another Internet standard then known as MIME but now called Media Types.⁹ Since chemical information is often carried in well-defined file types (for example the well-known MDL molfile) we thought it appropriate to propose a more general **chemical** media type to enable chemical information transfer using the Internet. This led to all sorts of fascinating interactions with the IETF (Internet Engineering Task Force) and a discussion of the proposal in Stockholm in 1994. Our draft proposal¹⁰ turned out to be widely adopted by the chemical and life sciences communities, but it never progressed to IETF formally approved status at that stage. As it happens, in 2022, there are moves originating from the W3C (the organisational body set up as a result of that inaugural WWW conference in 1994 which now oversees Web standards) to consider whether to make the proposal more formal. Also from 2022 is an example of the application of this proposal to searching for chemical data in repositories.¹¹

In 1994, we also saw the first online conferences in chemistry start to emerge, using the Web as a medium. Armed with my now good working knowledge of HTML, a call for papers (a misnomer if ever there was one) went out and ECTOC (Electronic Conferences on Trends in Organic Chemistry) was born.¹² Participants uploaded photos of themselves, posted comments on articles and generally avoided long distance travel – at a time when climate change and carbon dioxide emissions from aircraft was not yet topical! Four conferences

were organised in the end,¹³ before the realisation that in-person interactions were not going to be quickly replaced by this virtual medium. Who knows, with powerful new VR/AR/MR equipment on the horizon and the metaverse currently topical, they may yet become the norm. Meanwhile, who would have thought that for the last two years and running, students and researchers both would be attending virtual lectures as that norm.²

I will end with the idea seeded above by the discussion with Tim Berners-Lee and elaborated further by Peter Murray-Rust and myself during 1995, the aim being to produce the chemical equivalent of the HTML web language to be called CML (Chemical Markup Language). HTML was by then maturing rapidly and in 1996 a more general framework called XML (extensible markup language) had been proposed by computer scientists, with the intention that HTML itself be recast as the first such application of XML; hence XHTML. Peter and I were fortunate to be part of these early XML discussions in the Rembrandt Hotel, London and I emerged somewhat bemused to find ourselves the new moderators of the about to be set up XML-DEV discussion list. Over the next two years or so, some 20,000 posts to the lists by information professionals in all disciplines charted that early development of XML. CML itself took several more years to mature; it was very much informed by the frenetic development of XML during that period and emerged as one of the earliest XML languages.¹⁴

Over a short three years or so, many of the fundamentals of what we now consider the online information age, and its application to molecular sciences such as chemistry, were born. In the current age of the COVID pandemic, are we not lucky that those foundations laid not even 30 years ago yet have transformed how cheminformatics is practised today. What of the future? Well, my attention has been in the area of what is now called FAIR data.¹¹ But that will be another story.

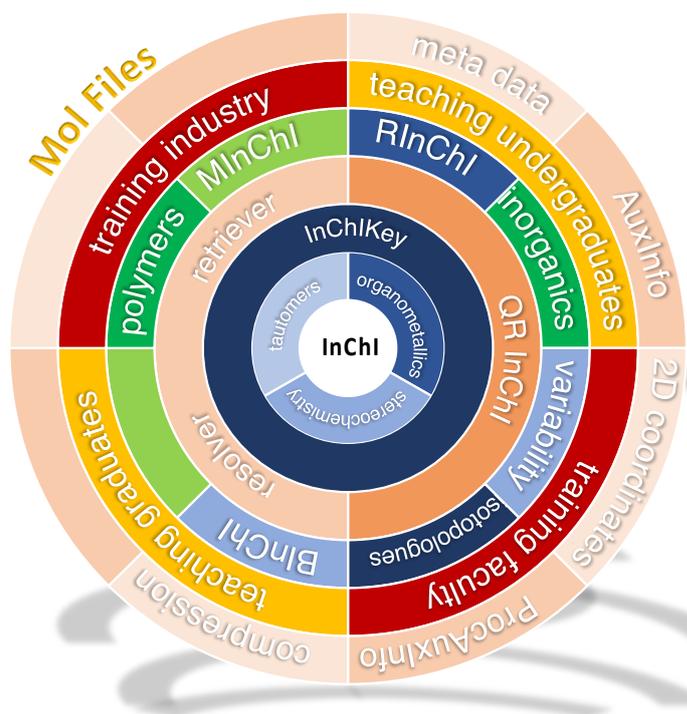
Notes and references

- (1) Some of this is recounted in a Rosarium; Rzepa, H. S. The long and winding road towards FAIR data as an integral component of the computational modelling and dissemination of Chemistry. *Israel J. Chem.* **2021**. <https://doi.org/10.1002/ijch.202100034>
- (2) Rzepa, H. S. NMR Spectroscopy. Principles and application. Six second year lectures given at Imperial College by Henry Rzepa. <https://doi.org/10.14469/hpc/10722>
- (3) Rzepa, H. S. et al. Chemical applications of the World-Wide-Web. *J. Chem. Soc., Chem. Commun.* **1994**, 1907-1910. <https://doi.org/10.1039/C39940001907>
- (4) James, D. et al. The case for content integrity in electronic chemistry journals: The CLIC Project. *New. Rev. Information Networking.* **1995**, 1(1), 61-69. <https://doi.org/10.1080/13614579509516846>
- (5) Casher, O. et al. Hyperactive molecules and the world-wide-web information system. *J. Chem. Soc., Perkin Trans. 2.* **1995**, 7-11. <https://doi.org/10.1039/P29950000007>
- (6) See <https://www.ccdc.cam.ac.uk/structures/>
- (7) Originating from Braddock, D.C. et al. A stereoselective hydride transfer reaction with contributions from attractive dispersion force control. *Chem. Comm.* **2022**, 58, 4981-4984. <https://doi.org/10.1039/D2CC01136K> and the associated data repository collection at <https://doi.org/10.14469/hpc/9237>
- (8) Rzepa, H. S. The first international conference on world-wide-web, May **1994**. <https://doi.org/10.14469/hpc/10724>
- (9) Rzepa, H. S. et al. The application of chemical multipurpose internet mail extensions (Chemical MIME) internet standards to electronic mail and world-wide web information exchange. *J. Chem. Inf. Comp. Sci.* **1998**, 38, 976-982. <https://doi.org/10.1021/ci9803233>
- (10) See The Chemical MIME home page, <https://doi.org/10.14469/hpc/10725>
- (11) Rzepa, H. S.; Davies, A. N. Open publishing FAIR spectra for and by students. *Spectros. Eu.* **2022**. <https://doi.org/10.1255/sew.2022.a10>
- (12) Proceedings of the first electronic conference on trends in organic chemistry (ECTOC-1). Rzepa, H. S. et al. (eds), ISBN 0 85404 899 5, CD ROM version. The Royal Society of Chemistry, **1996**.
- (13) Rzepa, H. S. (ed.). Electronic conferences on trends in organic chemistry. <https://doi.org/10.14469/hpc/10726>
- (14) Murray-Rust, P.; Rzepa, H. S. Chemical markup language and XML part I. Basic principles. *J. Chem. Inf. Comp. Sci.* **1999**, 39, 928. <https://doi.org/10.1021/ci990052b>

InChI Technical Developments

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

The International Chemical Identifier, InChI, is a structure-based molecular identifier, strictly unique, non-proprietary, open source and freely accessible. It is a canonical molecular identifier: one InChI for every molecule and one molecule for every InChI. This makes it ideal for checking for duplicates in molecular databases, and for merging molecular databases. The InChI is widely used to represent molecules as text strings, and is particularly useful for cataloguing organic molecules. The length of the string depends on the size of the molecule, which is essential to provide a complete account of large molecules. The latest version of the InChI has been tested on the hundred million molecules in [PubChem](https://pubchem.ncbi.nlm.nih.gov/) and found to be highly reliable.¹



InChI version 1.06 is extremely useful, but not quite perfect. The [InChI Trust](https://www.inchi-trust.org/) coordinates [working groups](#) which are developing ways of expanding the use and applicability of the InChI. The working group projects are at different stages: some are exploratory and others are well developed. Some expand the core InChI functionality and others build on the InChI to use it for molecule-related subjects such as reactions and mixtures.

The Reaction InChI working group has released software for an InChI-based description of reactions.² This has been tested on a database of more than a billion reactions.³ The Mixture InChI working group has published an InChI-based machine-readable format for mixtures.⁴

InChI version 1.06 is extremely good, but does not describe all molecules perfectly. One issue is tautomerism: for many molecules, such as nucleotide bases, there may be several reasonable locations for a hydrogen atom. It may be convenient for them all to have the same InChI. What is the right choice? A detailed study of all the many possibilities has been published by the tautomerism working group.⁵

Another issue is a small set of stereochemical issues. The InChI describes most stereochemistry very effectively. Work is continuing to improve the InChI's description of partially defined stereochemistry, atropisomers and some aspects of non-tetrahedral stereochemistry.

InChI version 1.06 generates valid identifiers for organometallic and inorganic molecules. These are useful, but insufficient to identify uniquely all of organometallic and inorganic chemistry. The organometallic working

group is currently addressing these issues so that a future version of the InChI will be able to distinguish all such molecules uniquely and canonically.

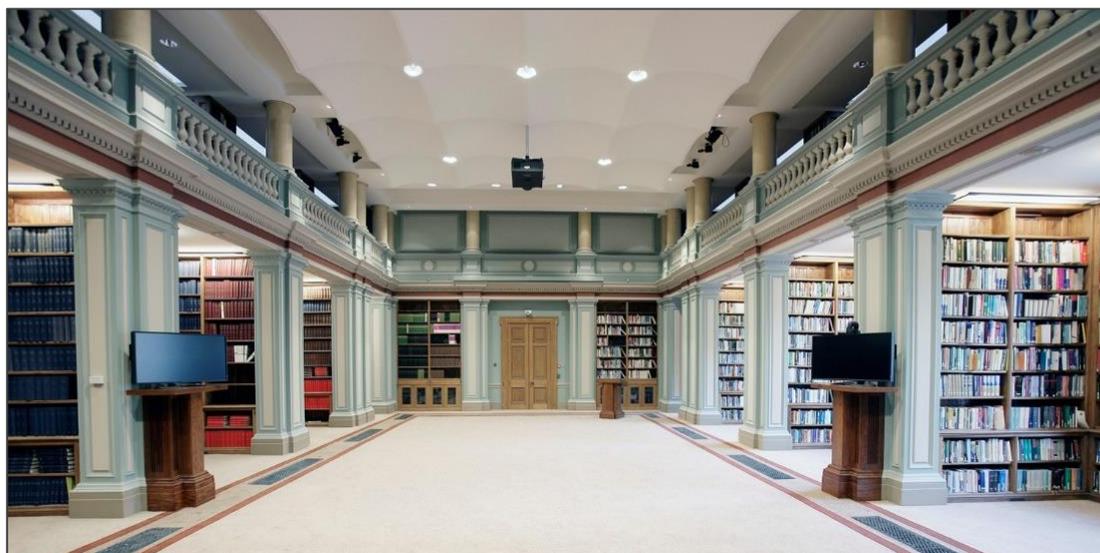
The current version of the InChI describes almost all of chemistry. The InChI working groups are filling the gaps and building new InChI-applications to make the International Chemical Identifier even more useful.

References

- (1) Goodman, J.M. et al. InChI version 1.06: now more than 99.99% reliable. *J. Cheminformatics*. **2021**, 13(40). <https://doi.org/10.1186/s13321-021-00517-z>
- (2) Grethe, G. et al. International chemical identifier for reactions (RInChI). *J. Cheminformatics*. **2018**, 10(22). <https://doi.org/10.1186/s13321-018-0277-8>
- (3) Goodman, J.M. et al. Analysing a billion reactions with the RInChI. *Pure and Applied Chemistry*. **2022**, 94, 643-655. <https://doi.org/10.1515/pac-2021-2008>
- (4) Clark, A.M. et al. Capturing mixture composition: an open machine-readable format for representing mixed substances. *J. Cheminformatics*. **2019**, 11(33). <https://doi.org/10.1186/s13321-019-0357-4>
- (5) Dhaked, D.K. et al. Toward a comprehensive treatment of tautomerism in chemoinformatics including in InChI V2. *J. Chem. Inf. Model.* **2020**, 60, 1253-1275. <https://doi.org/10.1021/acs.jcim.9b01080>

Update from the Royal Society of Chemistry Library

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



Over the past few years, the Library at the Royal Society of Chemistry has continued to operate as the custodian of the Royal Society of Chemistry's archive and to provide access to that information to its members. The following is an overview/reminder of what we offer, along with some news on recent developments.

Visits to the Library at Burlington House had been by appointment only since September 2021 but since September 2022, we have reopened to walk-in visitors. Members should check the [Library calendar](#) before planning their visit to see whether the library space is open, but even when not, we still provide a space to work and have tea and coffee. As there are limited facilities, booking for the Computer Room has become a permanent feature and so members must book a place via the website to avoid disappointment.

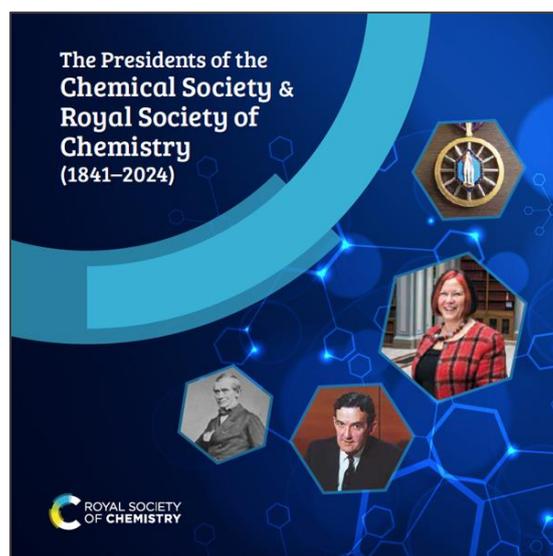
Services that we continue to offer include:

- The enquiry service. We can help with trouble-shooting remote access issues; fulfil copy requests and reference enquiries (for items held physically on-site); undertake historical/biographical enquiries. If we're unable to deal with an enquiry directly, we're usually able to identify another department within the RSC, or another organisation, that can.
- Book loans. The [Library catalogue](#) can be searched online. We offer members loans of up to ten items for a month, with five renewals, effectively giving a six-month total loan period. Loans can be requested by e-mail with the RSC paying for the postage out; the member will then pay for their return, if not doing so in person.
- The [Virtual Library](#). Providing access to third-party content including EBSCO eBooks, Knovel eBooks and databases, 22 titles from the ScienceDirect journal archive, Springer eJournals and eBooks as well as access to Kirk-Othmer & Ullmann's from Wiley. If a member has a suggestion for any relevant additional eBooks we could acquire for the EBSCO eBook collection, we would be happy to consider them (please note that not all eBooks are available on this platform). Please e-mail library@rsc.org with the title, author, and ideally the ISBN.

One of the projects undertaken by the Library during lockdown was to compile a booklet on the past presidents of the Society, going back to 1841. The booklet provides brief biographies of each president from Professor Thomas Graham to Professor Gill Reid and is available for free as a PDF on the library website, or in hardcopy at Burlington House.

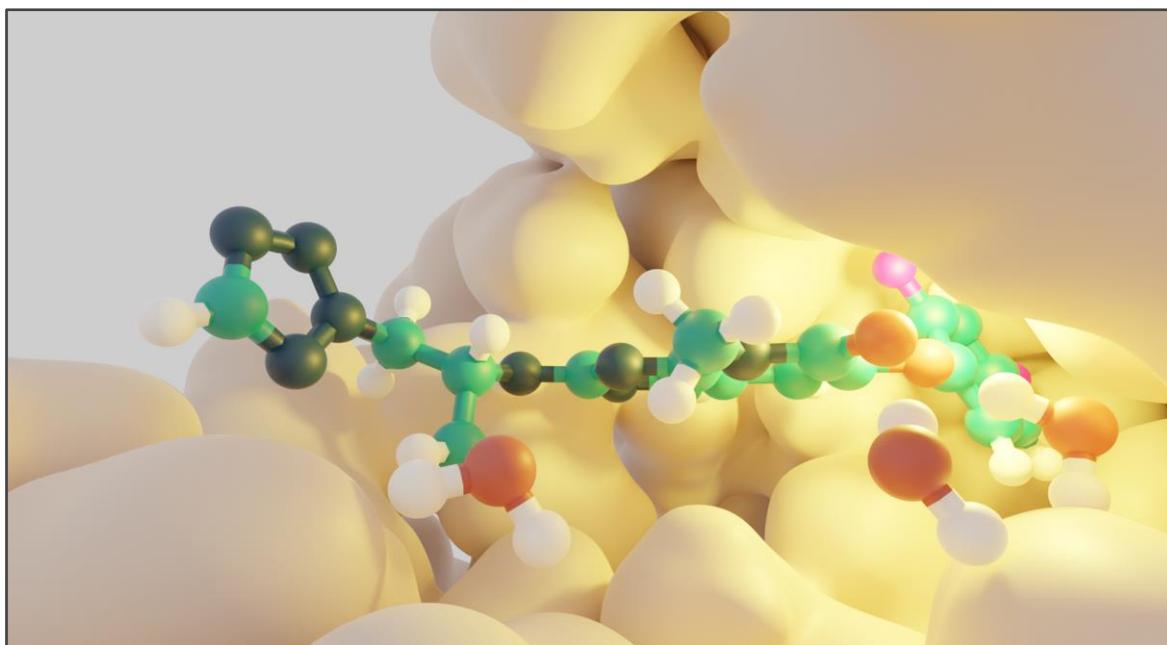
Since its launch in 2015, the [Historical Collection](#) platform has given members access to most of its digitised content, including *Chemistry in Britain*, *Education in Chemistry*, lectures, annual reports, council minutes and much of the older book collection. As the site has remained relatively unchanged since launch, we are now embarking on a major redevelopment which will entail redesigning the user interface, reorganising some of the content and improving the search function. We are also planning to make available some new content, including the Roscoe letters & notes and the Sir Frederick Abel Papers. We don't have a definite timescale for this as yet, but we are looking at a relaunch in late 2023.

We have also recently acquired a sizeable donation of books from Professor William Brock which we are now cataloguing into a dedicated space within the History of Chemistry section of the East Gallery in the main Library.



The Open Free Energy Project

Contribution from Irfan Alibay, Richard J. Gowers, Mike M. Henry, Diego Nolasco, Benjamin Ries and David W.H. Swenson. Email: openfreeenergy@omsf.io



Modelling the binding of a small molecule. (Artwork by Benjamin Ries made using Molecular Nodes and Blender.)

Project introduction

Alchemical free energy methods have become key components of drug discovery pipelines over the last few decades.^{1,2} These methods use classical molecular simulation techniques, with the molecular models then perturbed through non-physical (alchemical) states to allow the sampling of processes that would otherwise not be computationally feasible. Such techniques are typically able to predict free energies of binding for small molecules up to within an accuracy limit of approximately 1 kcal/mol.^{1,3,4} Despite the power of these methods, widespread adoption has been hindered by the fragmented state of the software landscape for these methods, with poor interoperability, reproducibility and accessibility of novel methods.

Addressing these problems is the goal of the Open Free Energy (OpenFE) project, which aims to provide robust, high-quality workflows under a permissive open source licence. The project employs a full-time software development team, currently six members, working in collaboration with various academic groups. The project is funded and directed by a pre-competitive consortium of industrial partners and is hosted by the Open Molecular Software Foundation ([OMSF](https://www.omsf.io)), a 501(c)(3) US based non-profit organisation. OMSF was founded last year to host such open source initiatives; other software projects hosted by the OMSF include Open Forcefield (which featured in the Winter 2019 edition of this newsletter), OpenFold and WESTPA.

Project outputs

The OpenFE package

The OpenFE project has now been operating for a year and, as a primarily open source software project, has been contributed to the development of various existing projects including OpenFF-toolkit, Lomap, RDKit, Protein-Ligand benchmark and OpenMMTools to name a few. Our flagship project, however, is the `openfe` Python package which brings these various elements together into a single cohesive interface for planning and executing campaigns of free energy calculations.

Key to enabling the interoperability and ease of use that this project aims to provide, is the design of good interfaces (APIs) between different software components. For example, for the planning of a perturbation network between many ligand compounds, we have split the task into three roles of Atom Mapper, Scorer and Network Planner (see Figure 1). This allows a modular approach, where, for example, the cheminformatics toolkit providing the mapping could easily be swapped out, and indeed we currently have implementations which use either RDKit or OEChem.

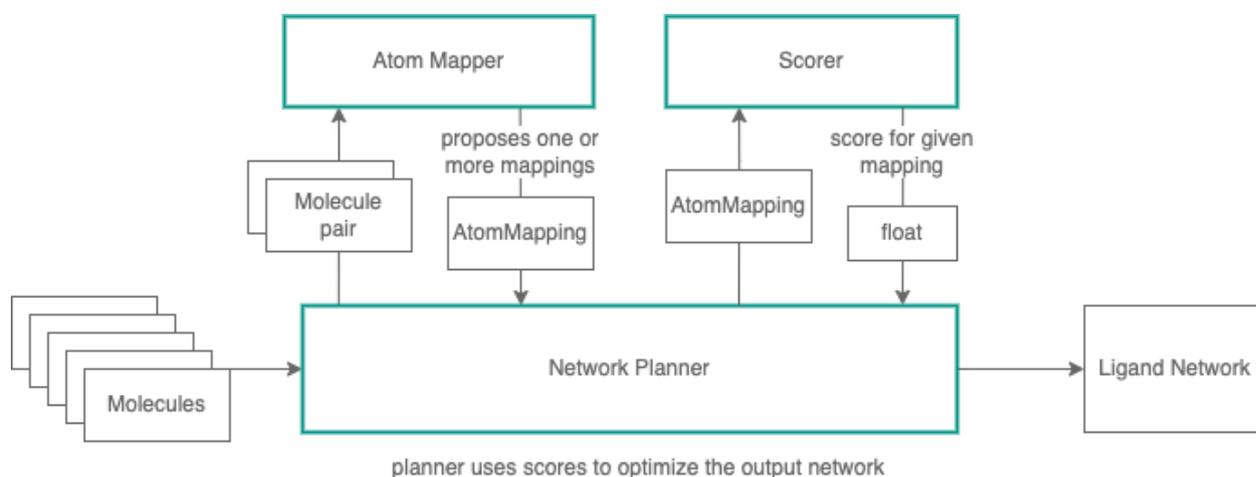


Figure 1. The modular API for planning a ligand perturbation network.

A further example is the Protocol system, which allows authors of scientific methods to define a complete computational workflow for use by others (Figure 2). From a Protocol, a directed acyclic graph (DAG) of individual computational tasks can be generated, then farmed out to a HPC computer resource, the results of which will yield an estimate of the free energy difference. Packaging definitions of workflows in this way aids in reproducibility as the definition of the method (the Protocol) can be stored on disk or shared with others.

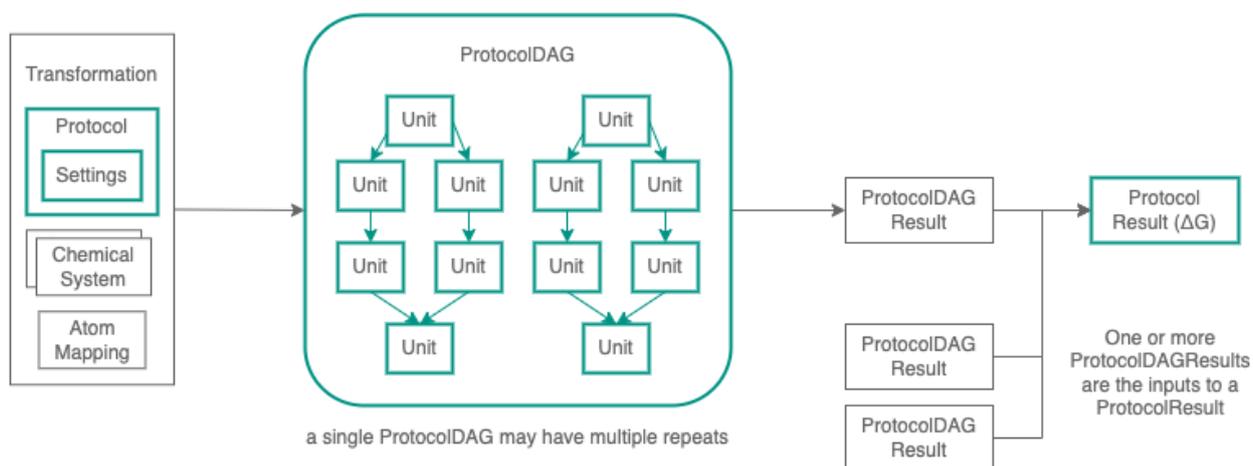


Figure 2. The Protocol workflow system. Shown is the generation of a series of work units, which are executed then gathered together, along with any previous results to form an estimate of free energy difference.

A relative binding free energy Protocol

In collaboration with the Chodera Lab at Memorial Sloan Kettering Cancer Center, we have adapted a code originally in the perses Python package to create the first working implementation of a Protocol. This allows users to execute a relative binding free energy (RBFE) calculation to compare the relative binding affinity of

two ligands. This is implemented using the OpenMM molecular dynamics engine and is using the latest OpenForceField parameter sets. Results from our benchmarking against a well known experimental reference data set are shown in Figure 3.

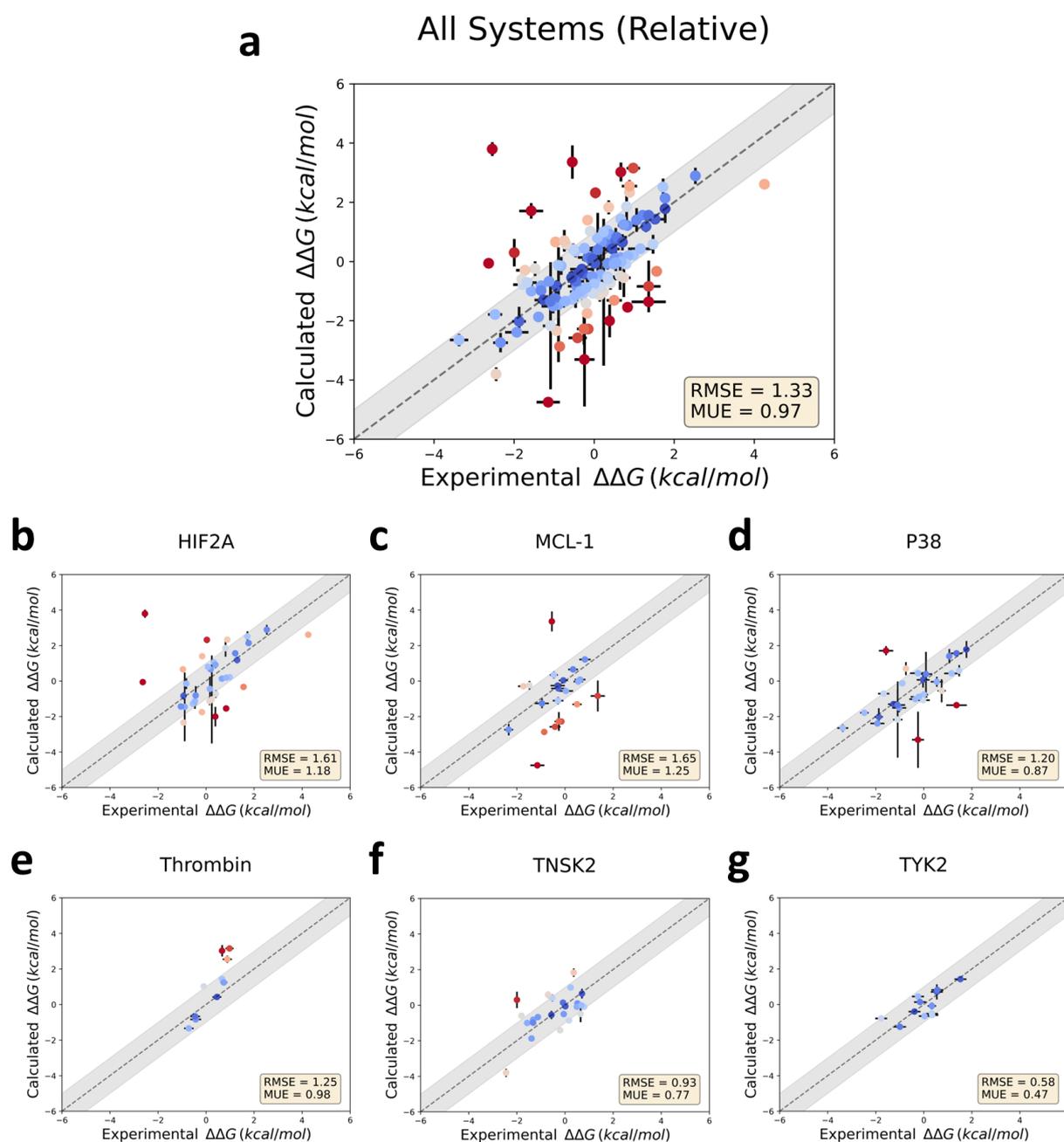


Figure 3. Relative binding free energy results for six protein-ligand benchmark systems, a) all six systems, b) HIF2A, c) MCL-1, d) P38, e) Thrombin, f) TNSK2, g) TYK2. All free energies, including RMSE and MUE measures, in kcal/mol.

Future outlook

Thanks to the backing of our 15-strong industrial partner consortium, we are hopeful for the outlook of the project as we enter our second year. We will be continuing to develop infrastructure and novel methods for molecular simulation; interested parties, both academic and industrial are invited to contact the project at openfreeenergy@omsf.io or try the toolkit out in their [web browser](#).

References

- (1) Schindler, C.E.M. et al. Large-scale assessment of binding free energy calculations in active drug discovery projects *J. Chem. Inf. Model.* **2020**, *60*, 5457–5474.
- (2) Sherborne, B. et al. Collaborating to improve the use of free-energy and other quantitative methods in drug discovery. *J. Comput. Aided Mol. Des.* **2016**, *30*, 1139–1141.
- (3) Ross, G.A. et al. The maximal and current accuracy of rigorous protein-ligand binding free energy calculations. *ChemRxiv*. <https://doi.org/10.26434/chemrxiv-2022-p2vpg>
- (4) Fu, H. et al. Meta-analysis reveals that absolute binding free-energy calculations approach chemical accuracy *J. Med. Chem.* **2022**, *65*, 12970–12978.

Meeting Report: SCI-RSC Workshop on Computational Tools for Drug Discovery 2022

*Contribution from Professor Neil Berry, Department of Chemistry, University of Liverpool, email:
ngberry@liverpool.ac.uk*

This meeting, held on 23 November 2022 at The Studio, Birmingham, was organised by SCI's Fine Chemicals group and RSC's Chemical Information and Computer Applications Group. The organising committee was Neil Berry, Al Dossetter, Daniel Hamza, Caroline Low, Jayshree Mistry, and Chris Swain. The workshop offered delegates a unique opportunity to try out a range of software packages for themselves with expert tuition in different aspects of pre-clinical drug discovery. Attendees were able to choose from sessions covering data processing and visualisation, ligand and structure-based design, or ADMET prediction run by the software providers. All software and training materials required for the workshop was provided for attendees to install and run on their own laptops and use for a limited period afterwards. There were over 40 delegates from across academia (mainly PhD students and post-doctoral research assistants) and industry.

Collaborative drug discovery

The workshop was a live demonstration and hands-on training of the [CDD Vault](#) platform and showed how it can help with organising and analysing scientific data and help progress research faster. CDD Vault is a complete and essential informatics platform used by drug discovery researchers around the globe to manage their data. It helps project teams manage, analyse, and present data for biotech companies, CROs, academic labs, research hospitals, agrochemical and consumer goods companies. This modern web application can be used to register molecules, biological entities, mixtures and all the assay data associated with experiments, allowing scientists to organise their data in a simple to use and secure system.

Schrodinger

Following a brief introduction to [Schrodinger's](#) Design Platform, a number of tools were presented that are aimed at empowering chemists to become more fluent with 3D design. PyMOL is a familiar tool for many non-modellers, and this was demonstrated using its easy-to-use features showing “why we should consider 3D visualisation” to assist our projects. Schrodinger's Ligand Designer in Maestro, which transitions us into a more powerful 3D environment, was presented. Ligand design in the context of the protein was shown using various built-in fragment and enumeration libraries, and important features such as water and “growth space” to assist the decision-making process. Ligand designer has many expert scientific tools under-the-hood, but hides those complexities under a single interface – this makes it ideal for brain-storming ideas with colleagues in a real-time way. The workshop highlighted design candidates, in order to progress them to the next steps such as ranking using free energy calculations.

Optibrium

After a brief introduction to StarDrop, [Optibrium](#)'s complete platform for small molecule design, optimisation, and data analysis, there was a focus on StarDrop's Inspyra module that combines the user's chemistry knowledge with the exploratory power of generative chemistry methods to design optimal compounds faster. Users experienced Inspyra generating new compound ideas in the background while dynamically learning from your interactions using a unique AI "inference engine". User responses to Inspyra's suggestions guided generative chemistry algorithms to explore the most relevant chemistry spaces and suggested optimisation strategies that are most likely to succeed in the user's project.

Chemical computing group

The workshop focussed on drug discovery and design using MOE (Molecular Operating Environment). MOE is one of the world's leading platforms for computational drug discovery and design for both small molecule and macromolecular therapeutics. Produced by [Chemical Computing Group](#), it is very widely used in pharmaceutical and biotech companies, and academic institutions, worldwide. The workshop showed how MOE can be used to analyse and exploit both modelled and known crystal structures to find new molecules that may be superior binders to those already known, and with acceptable properties. Techniques described include pharmacophore derivation and search, molecular editing and transformation in situ, and docking.

Alvascience

The workshop focused on QSAR processes, from data curation to model deployment using [Alvascience](#) software solutions. They started with data curation to show how to handle molecular datasets. This led to demonstrations on how to calculate more than 5000 molecular descriptors and fingerprints using alvaDesc. These features can then be used to generate machine learning models to predict given endpoints. Finally, they showed how to deploy such models so that they can easily be used by clients and fellow colleagues.

Cresset

The workshop covered comprehensive molecule design using [Cresset](#)'s Flare™, a comprehensive drug design platform where ligand and structure-based methods work in synergy to offer more effective molecule design for medicinal and computational chemists. Flare was demonstrated showing how it can design and prioritise new molecules for synthesis. Analysis of protein-ligand interactions, Electrostatic Complementarity™ maps and scores, virtual screening, and calculations of water stability and locations.

A Crystallography Papermill: The CSD Response

Contribution from Suzanna Ward, Head of Data and Community at CCDC, email: ward@ccdc.cam.ac.uk

In April 2022 a [pre-print article](#) was published alleging a papermill in crystallography. It highlighted at least 800 papers of concern due to their use of repetitive phrases and incorrect references. These included 992 crystal structures. We are thankful for the work the author of the pre-print has undertaken, as well as others that have taken efforts to alert us when issues like this arise. We had already become suspicious of some structures during the course of our normal checks, but on learning about the pre-print from Retraction Watch and the community, we launched a larger investigation.

What have CCDC done in response?

We rely on the peer review process to spot issues in publications, and while we work with publishers when we spot issues the goal of the CSD is to reflect the published literature. This means we only retract structures when

the associated publication is retracted. However, we do wish for high quality, accurate data to be available to the community, and so have been investigating and working with the publishers on this matter. Thankfully standard file formats and data sharing through curated databases is standard. Along with community tools for data validation, investigation is more achievable.

We began a full investigation on 28th April after becoming aware of the papermill preprint. We made a [public statement](#) on the same day on Twitter, linking to further details on our [website](#). We continue to share updates on the situation here.

As an initial step, we added a note reading: "This structure is currently under review following a 2022 study of a prolific papermill <https://doi.org/10.21203/rs.3.rs-1537438/v1>" to all 992 structures linked to articles implicated in the papermill pre-print. This is an editorial comment, we use this field to provide additional information about the dataset, flag up potential issues to our users, or in cases where we have needed to contact the authors so our users can decide if they want to investigate further.

Crystal details	
Space group	C 2/c (15)
Unit cell	a 25.932(2)Å b 13.4420(12)Å c 9.9870(17)Å α 90° β 93.2310(10)° γ 90°
Cell volume	3475.71
Reduced cell	a 9.987Å b 13.442Å c 14.604Å α 117.400° β 92.868° γ 90.000°
Z, Z'	4, 0.5
Habit	block
Colour	colorless
Remark	This structure is currently under review following a 2022 study of a prolific papermill https://doi.org/10.21203/rs.3.rs-1537438/v1

An editorial comment is displayed in the Remark field when viewing structures online that are linked to the pre-print.

Have any data been retracted from the CSD?

Currently, we only retract data from the released version of the CSD if the accompanying scientific article has been retracted. Of the 992 structures initially flagged, 14 have been retracted from the CSD because the accompanying papers have been retracted. When data is retracted from the CSD the entry with associated bibliographic details still exists, but the scientific content is removed and the entry is linked to the retraction notice in the literature. We have relationships and workflows with all the major publishers to help keep us informed on retractions and we also monitor resources like Retraction Watch.

There are currently around 400 retracted entries in the CSD, out of 1.2 million. 14 of the 400 were implicated by this papermill pre-print. The other retracted entries are not related to the pre-print, they had been retracted previously due to the associated scientific articles being retracted in the literature.

What is happening to the other structures not retracted?

Although we reflect data in scientific articles, for the remaining 978 structures implicated by the papermill pre-print, we are assessing if any structures require further investigation. We have extended our existing data integrity checks to do more extensive analysis and will be following COPE guidelines and communicating with

publishers and authors when appropriate, depending on the outcomes of this analysis. Our investigations include a more in-depth analysis of the underlying datasets and comparisons between the datasets and the CSD.

Our team of PhD-level editors and deposition coordinators are involved in the investigation, including our dedicated data integrity scientist. We have also reached out to other crystallographic community experts who are supporting us in these efforts. Going forward we are adding more automated checks which will help us to identify and prioritise structures which need further manual examination.

Since publication retraction timescales can take time and our more extensive checks will also take time, we decided to add the additional editorial comment so our users can be informed while the investigation is ongoing. We can also provide a complete refcode list of the structures affected on request, should users wish to exclude these from their work.

What does reviewing the structures entail?

Every structure in the CSD undergoes automated and manual checks. For this investigation, we have added more of both.

Specific checks involve looking for similarities with existing data in the CSD. This includes unit cell checking, assessment of underlying reflection data, finding CIF similarities, using the IUCr's checkCIF service, using Mogul to identify unusual bond lengths and angles, and overlays of exact coordinates.

How is CSD data quality maintained?

As ever, each structure deposited into the CSD undergoes a series of automated and manual checks. We have a dedicated data integrity scientist in the team, and are constantly working to improve the quality and accessibility of the data.

We have an ongoing program of data improvements, enriching and making targeted improvements to existing entries in the CSD to again make the individual datasets and the knowledge from the collection more discoverable to both human and machine interrogation. This year's enhancements included improvements to melting point data, enriching how we capture polymorphic structures and standardisation of structures from new and emerging techniques, such as structures determined by electron diffraction.

Meeting Report: Ultra-Large Chemical Libraries

Contribution from RSC CICAG Chair Chris Swain, email: swain@mac.com and Ilaria Proietti Silvestri, email: ilaria.proietti@liverpoolchirochem.com

This in-person meeting, organised by Ilaria Proietti Silvestri and Chris Swain on behalf of [CICAG](#), took place on 10 August 2022 at Burlington House, London, attracting around 60 people from industry and academia. About half of the presenters were from overseas and the gender balance for oral and poster presenters combined was 1:1.

A decade ago a chemical library of a million compounds was considered large, but over the last few years there has been a period of continuous growth in the size of both physical and virtual chemical libraries. As the libraries have grown, the conventional search technologies have become unsustainable and new technologies

are needed. This meeting looked at the challenges and solutions used to design, create, compare and search these ultra-large chemical libraries.

Recent advances in chemical search of ultra-large databases

The opening presentation was by Roger A. Sayle (NextMove Software, UK), highlighting Enamine's "make-on-demand" database of molecules available for purchase. The database contained 647 million compounds in 2018 – by February 2022 there were over 22 billion. Keeping up with this exponential growth represents an ongoing challenge to the field of cheminformatics, requiring continual innovation and advances in the algorithms and techniques for working with large data sets. One particularly striking illustration of the problem is shown below, with conventional searching techniques failing at around 1 billion molecules due to memory bandwidth limitations.

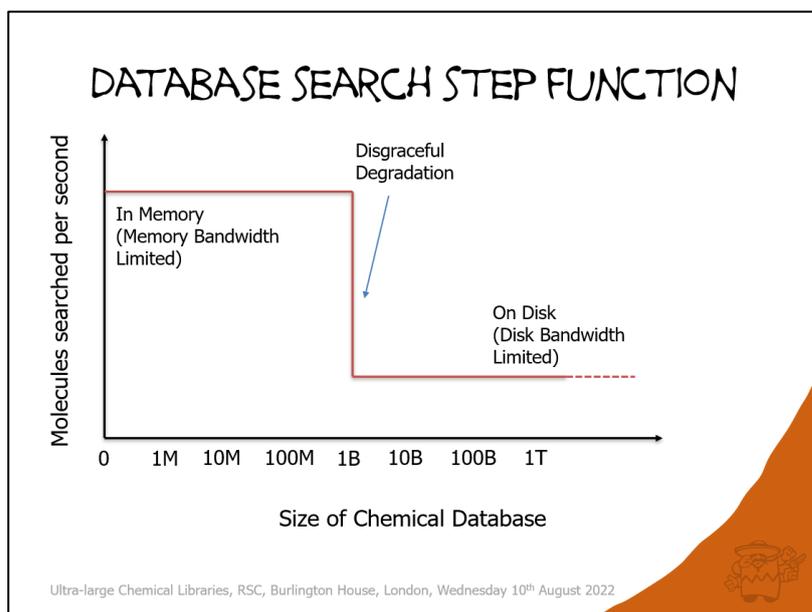


Image: Roger Sayle, NextMove Software, Cambridge, UK.

A variety of possible approaches were discussed. More cores, faster processors/memory, GPUs. More memory, distributed servers. Pre-touch database, dedicated servers. Alter on-disk performance using faster disks, NVME or SSD vs HD, local vs NFS, RAID0.

Software considerations were also highlighted: for similarity searching, one billion molecules (memory requirement depends on the fingerprint length), 256 bit fingerprint (34GB), 512 bit fingerprint (68GB), 1024 bit fingerprint (136GB), 2048 bit fingerprint (252GB). For ultra-large libraries folding fingerprints up to 256 bits seems to be fine. This means that for smaller databases GPU acceleration might be useful, but for larger databases CPU implementation is usually preferred. The NextMove's SmallWorld indexes molecular subgraphs are an efficient molecular representation allowing the searching of ultra-large libraries.

It should be remembered that "search" is just one aspect of ultra-large chemical database management and that indexing, deduplication, hashing and pre-calculation of properties are all useful tools for optimising specific types of search.

References

- Irwin, J. J. et al. ZINC20 – a free ultralarge-scale chemical database for ligand discovery, *J. Chem. Inf. Model.* **2020**, 60(12), 6065-6073.
- Grabowski, S.; Bieniecki, W. Tight and simple web graph compression for forward and reverse neighbor queries. *Discrete Applied Mathematics.* **2014**, 163(4), 298- 306.
- Lemire, D.; Boytsov, L. Decoding billions of integers per second through vectorization. *Software: Practice & Experience.* **2015**, 45(1), 1-29.
- Sayle, R.; Delany, J. SMILES multigram compression. Presented at Daylight User Group Meeting (MUG01), **2001**, Santa Fe, New Mexico.

Baeza-Yates, R. A. A fast set intersection algorithm for sorted sequences. In Proceedings of the 15th Annual Symposium on Combinatorial Pattern Matching (CPM). *Lecture Notes in Computer Science*. 2004, 3109, 400-408.

Processing and searching chemical data sets at GSK

This presentation, by Peter Pogany (GlaxoSmithKline, UK), described pre-processing to ensure efficient searching of chemical datasets. At GSK several tools which facilitate pre-processing of datasets are used: ChemAxon's MadFast and NextMove's SmallWorld and Arthor. These tools are asked either via a user interface or through a variety of APIs. Many of the datasets, which can contain billions of molecules, are provided by third parties and are filtered and standardised in house.

Unmasking false theoretical friends in the COVID-19 era

The first presentation after lunch was given by José Pedro (Cerón-Carrasco Centro Universitario de la Defensa, Universidad Politécnica de Cartagena, Spain). One of the responses to the COVID-19 pandemic was an international effort to identify novel therapeutic agents. Much of this effort centred around virtual screening/docking of huge chemical libraries. These efforts served to underline that whilst pose prediction is encouraging, the subsequent prediction of binding affinity was very unreliable. In an effort to refine two massive screening efforts a rescoring protocol was implemented. However, it was clear from the discussion the generated by the presentation that this remains an unsolved problem.

Synthon-based docking of ultra-large virtual libraries

The next talk was given by Oliver Vipond (Evariste Technologies, UK), continuing the theme that the expansion of synthetically accessible chemical space, brought about by the ever-increasing size of make-on-demand compound libraries, necessitates more efficient means to undertake virtual screens. In particular, the pairing of machine-learning algorithms with molecular-docking procedures to allow the screening of ultra-large chemical libraries was explored. Evariste use search algorithms to explore reagent space; by doing so they are able to approximate docking scores for entire combinatorial libraries while only explicitly evaluating an extremely small proportion. To achieve this, they reward molecules found in areas of high-scoring chemical space. To date, four virtual-screening campaigns have exhibited an average hit rate of 16%, typically completed in a matter of days using 32 CPUs.

[Synthon-GA: searching make-on-demand libraries using a genetic algorithm](#)

Jan H. Jensen (University of Copenhagen) spoke about a new software tool. Whilst there are a number of algorithms for generating virtual libraries a major issue is they generate molecules that not readily synthesisable. By using commercially available synthons (building blocks/reagents) as genes in a genetic algorithm, using a defined set of reactions it is hoped that the generated virtual library will contain relatively accessible molecules. As an example, this approach was used to search for molecules in the Enamine REAL Space with good docking scores towards the leucine-rich repeat serine/threonine-protein kinase 2 (LRRK2) as part of the CACHE-1 challenge. Since the exact reactions underlying Enamine REAL Space is proprietary the search will not necessarily identify the exact molecules in this space, so the closest analogs in REAL Space must be found and re-docked.

Synthon-GA is not on GitHub yet so for more information please contact Jan Jensen directly, @janhjensen.

Applications of DNA-encoded library technology toward the discovery of potential antibody-recruiting small molecules

Carol Mulrooney (GSK, USA) gave a virtual presentation providing an excellent description of DNA-encoded library (DEL) technology, used to synthesise and screen combinatorial libraries of up to billions of unique chemical structures ligated to unique corresponding DNA sequence tags. DEL screens have been reported to deliver hits for previously "undruggable" targets. Carol also described a screening campaign to discover novel antibody-recruiting small molecules (ARMs) potentially effective against lectin-type oxidised LDL receptor 1 (LOX-1), which has been implicated in atherosclerosis and myocardial ischemia.

Chemical Space Docking: novel ROCK1 kinase inhibitors found by large-scale structure-based virtual screening

The final talk was given by Franca-Maria Klingler (MSD, UK). Chemical Space Docking is a novel virtual screening method that initially docks building blocks; these are then grown using a collection of known chemical reactions and rescored. Thus it is possible to “explore” a vast virtual library without the need to enumerate every member of the library. The process is analogous to fragment-based drug design. Chemical Space Docking was used to identify inhibitors of ROCK1 kinase from almost one billion commercially available synthesis-on-demand compounds. From 69 synthesised molecules, 39% had K_i values below 10 μM . Two leads were crystallised with the ROCK1 protein and the structures showed excellent agreement with the docking poses. This approach scales roughly with the number of building blocks that span a chemical space and is therefore multiple orders of magnitude faster than traditional docking of fully enumerated libraries.

Posters

Presenter	Title
Esperanza Pearl	Connecting virtual libraries and synthesis: predictable, reliable, and robust synthesis of virtual hits using Synple automation
Debora Zian	Axxam compound collection next-generation small-molecule collections for hit-finding by HTS
Olga O. Tarkhanova	Approaches to build and explore ultra-large chemical spaces
Javier Vázquez	Huge chemical space exploration using 3D-hydrophobic profiles
Samantha Kanza	The AI 4 Scientific Discovery Network+
Samantha Kanza	The Physical Sciences Data Infrastructure
Stephanie Wills	The Fragment Network as a novel source of fragment merges
Y. Zabolotna	ChemSpace Atlas: multiscale chemography of ultra-large libraries for drug discovery
Y. Zabolotna	Exploration of the chemical space of DNA-encoded libraries

Open Science in the Royal Society of Chemistry

Contribution from Dr Anna Rulka, Executive Editor, Open Access, Royal Society of Chemistry, email: rulkaa@rsc.org

Open science is a set of methods and practices aimed at making the research process and its results transparent, accessible and reusable. It encompasses open data, software and hardware, and aspects of scholarly communication such as open peer review, open science policies, open licensing and of course open access publishing, which is a substantial movement in its own right. Open science leverages modern technology, services and platforms to put research outputs in the hands of as many as possible. Through this it facilitates more accurate verification of scientific results, reduces duplication of effort and increases innovation and productivity.

The RSC considers the aims of open science as fully aligned with our mission to help chemists make the world a better place. We believe that open science and open access are crucial to driving scientific discovery, fostering collaboration and increasing the interdisciplinarity of research. The need for research to be made immediately and freely available to everyone ensures the broadest dissemination of knowledge and drives our commitment to open science and open access publishing.



Graph showing different aspects relating to the term open science. Reproduced from: [SciencesPo](#).

As part of this, the RSC recently announced the significant decision to [transition all of its fully-owned journals to an open access](#) model within the next five years, becoming the first major chemistry publisher to make such a commitment. In order to achieve full open access whilst maintaining our commitment to inclusion and diversity, we will be developing and implementing a new institutional-level model and moving away from a model in which the author pays article processing charges.

Likewise, the RSC has adopted the FAIR Data Principles which are a set of guiding principles for data management proposed by a consortium of scientists and organisations to support the reusability of digital assets. These principles underline the need to make data machine-readable in addition to openly available under CC-BY licence, mainly because we are experiencing an exponential increase in the volume and complexity of available data. Therefore, data must be easy to find and machine-readable (**F**indable), openly available (**A**ccessible), linked to other relevant data (**I**nteroperable), and needs to be possible to use for other purposes than originally intended (**R**eusable).

One RSC journal which demonstrates our desire to put open science principles into action is [Digital Discovery](#). *Digital Discovery* covers the emerging field of accelerated discovery (use of AI and automation for discovery in materials, chemical and biomedical research), publishing both experimental and computational work on robotics, screening, databases and advanced data analytics.

Not only is *Digital Discovery* an open access journal, but it also contains several other features in support of open science. It is of vital importance that all data supporting the scientific evidence are readily available for both reviewers and readers, thus *Digital Discovery's* authors are required to deposit any code and/or data that is central to the main findings of the study in a public repository. A statement that the code is available upon reasonable request is not accepted and any restrictions on code availability must be shared with the editors at

submission. Exceptions require a valid reason to be supplied for not being able to publicly share the code (e.g. when it is developed for and owned by a private company).

Digital Discovery reviews supporting data and code, utilising an additional dedicated reviewer for the purpose because the journal's editorial mission is to maintain high standards of reliability and reproducibility of all published content. 'Data reviewer' checks if the data and/or code are appropriately presented and documented and verify that the code is functional and reproduces the reported findings. In addition, the journal offers transparent peer review, such that authors have the option to publish the peer review history alongside their article.

Digital Discovery is one of many of the open access journals that the RSC launched in the last couple of years to provide authors with a platform to publish their work open access. We recognise the [benefits of open access](#) to researchers, funders and society at large. Open access speeds up the progress of research and leads to new collaborations that inspires innovation and results in the release of new products or services. Greater openness helps avoid the production of redundant research and leads to increased citations to work published under an open access model.

It is worth noting that our flagship journal, *Chemical Science*, has since January 2015 been published under a 'diamond' open access model, whereby all articles are published fully open access without any author charge. The substantial costs of running the journal continue to be fully subsidised by the RSC. We are also in the process of extending the afore-mentioned transparent peer review to all of our journals.

In addition to the activities within our journal programme, we are also a founding partner of [ChemRxiv](#), an open access preprint archive for chemistry research. We make all of our journals open to direct submission from ChemRxiv and we actively encourage researchers to deposit their manuscripts in the repository prior to submission. We also note that since acquiring [ChemSpider](#) in May 2009 we have also continued to maintain this free chemical structure database providing access to over 100 million structures from hundreds of data sources.

The RSC is fully committed to an open and equitable future for scientific research and will continue to work with and for the scientific community towards that goal.

References

<https://www.nature.com/articles/s41557-022-00910-7>

<https://www.rsc.org/journals-books-databases/author-and-reviewer-hub/authors-information/prepare-and-format/data-sharing/>

<https://www.oecd-ilibrary.org/sites/90ebc73d-en/index.html?itemId=/content/component/90ebc73d-en>

<https://www.silverchair.com/news/2020-publisher-working-groups-text-and-data-mining/>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6245499/>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4837983/>

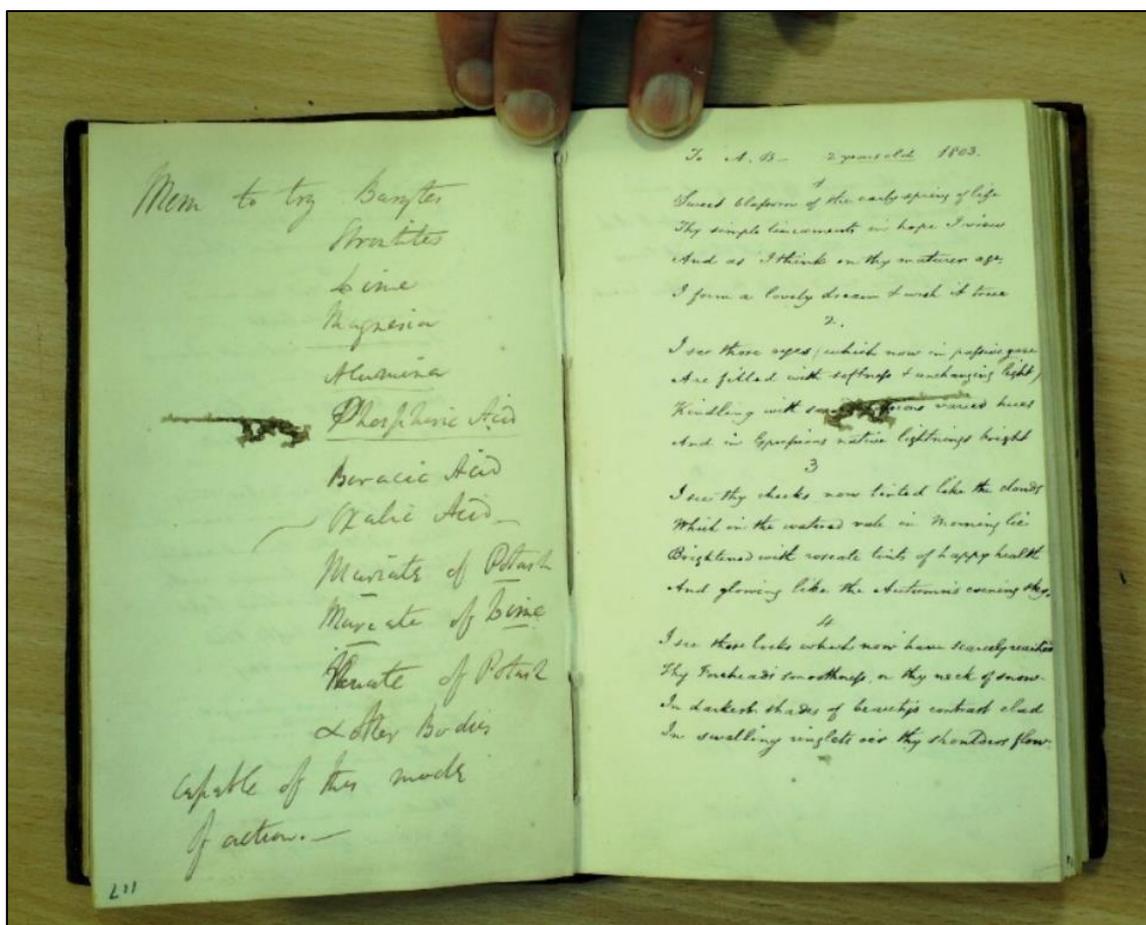
The Davy Notebooks Project

Contribution from Frank James, University College London, email: frank.james@ucl.ac.uk

Humphry Davy (1778-1829) moved from provincial obscurity as an apprentice surgeon and apothecary in Penzance, via the Medical Pneumatic Institution in Bristol to national and international fame as Professor of Chemistry at the Royal Institution and President of the Royal Society of London. In this social climb, Davy, amongst much else, discovered the extraordinary physiological effects of nitrous oxide (which he administered

to his friends including the poets Robert Southey and Samuel Taylor Coleridge), developed a fairly coherent theory of electro-chemistry, a term he coined (isolating and naming sodium and potassium among other chemical elements as a result) and invented the miners' gauze safety lamp. In addition, he wrote poetry (which Southey and Coleridge admired); at the Royal Institution he became the most popular scientific lecturer in wartime London and advised and lectured to the Board of Agriculture.

Much, but by no means all, of this is recorded in some form or other in the 75 or so of his notebooks that have survived held by the Royal Institution in London and Kresen Kernow in Redruth. Anyone who has used them will know that in the main they are not the neat well laid out manuscripts we all remember from O-level chemistry or those of his successor at the Royal Institution Michael Faraday. Instead, many of them have no organisation whatsoever, beginning at both ends, containing scraps of poetry, philosophical and religious musings, chemical notes, and occasionally enlivened with small drawings.



An example of a Davy notebook showing a list of chemicals on one side and a poem (by Davy though in the hand of an amanuensis) to the infant daughter of his former employer in Bristol, Thomas Beddoes. RI MS HD/13/G, pp.118-17.

As Davy is such a significant figure in many early 19th-century contexts, a group led by Professor Sharon Ruston at the University of Lancaster successfully applied for a significant grant from the Arts and Humanities Research Council to undertake their digitisation, transcription and to place them, freely available, on the web as part of Lancaster Digital Collection with a light critical apparatus. In addition to Lancaster University, Kresen Kernow and the Royal Institution, the other institutions involved are Zooniverse based at the Adler Planetarium in Chicago, University College London and the University of Manchester.

Over nearly two years, images of the notebooks have been placed in tranches that are generally up for around three months on the [Zooniverse crowd transcription website](#). There a team of around 2000 volunteers transcribe each line of the notebooks three times and we are always pleased to welcome new volunteers. The vast amount of this work is done remotely, but from time to time transcribeathons have been held either remotely or, more recently, in person. At the time of writing (November 2022) around 70% of the approximately 10,000 pages involved have been transcribed. Once a notebook has been completed, the three transcriptions are reconciled automatically and those lines or pages where there are significant differences (mostly arising from Davy's handwriting or where he has overwritten words) are flagged up for checking by the three research associates on the project and the Principle and Co-Is.

Davy's notebooks have long been studied by researchers coming from history and literature who either had to come to London or Redruth (previously Truro) to read through them. And of course with such a large collection one could not ever be certain that all the relevant passages on a specific topic had been identified. Once placed on the Lancaster Digital Collection website, it will be possible to search through all the notebooks.

While we are perfectly well aware that the technology used in this project was developed to aid scientists in their research, it has allowed the creation of digital editions of many key historical figures. For example, the letters of Davy's friend, the Irish writer Maria Edgeworth are currently being digitised along the same lines. All this effort will help scholars better understand the role of science, literature and much else besides in eighteenth- and nineteenth-century society and culture.

Further details of the project can be found [here](#).

Frank James is Professor of History of Science at University College London (Department of Science and Technology Studies) and a Co-I on the Davy Notebook Project. He is currently writing a biographical study of Davy.

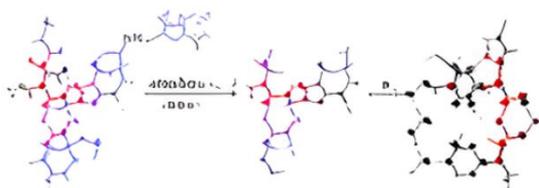
This JACS Does Not Exist: Generating Chemistry Abstracts with Machine Learning

Contribution from Yue Wu, SLAC, email: yuewu@slac.stanford.edu, @_yue_wu

Machine learning models have reached a point where their ability to interpret and generate text often appears almost human-like. You may have wondered how these models work, and whether you need a building full of computers and a power station just to train one. In fact, many pre-trained models are freely available. Here, I'll discuss how I recently built a machine learning project using a range of popular models (all publicly available) to generate fake chemistry paper abstracts.

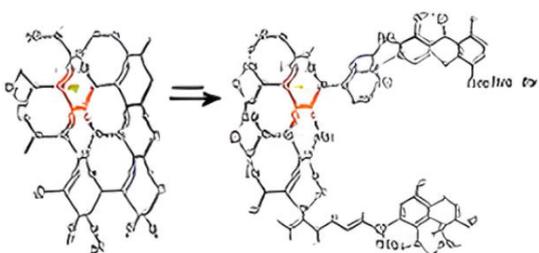
The inspiration for this project was a site that shows ["photos" of people that, in fact, have never existed](#). It is based on a machine learning model known as a generative adversarial network (GAN). GAN models work by training "generator" and "discriminator" networks (generally, deep neural networks of some type). To take the example of generating people's faces, for each training example, the generator creates an image from a random seed. The discriminator then compares it to a real picture of a face drawn from the training examples and tries to choose which is real and which is fake. The two networks are then updated to make them each better at their jobs. After millions of iterations of this arms race, the generator becomes very good at generating images that are almost indistinguishable from real faces. (This process can be applied not just to images but to any suitable representation – for example, [molecules](#).)

Synthesis, Structural, and Electron Paramagnetic Resonance Characterization of N-Heterocyclic Carbene Boron-Stabilized 2,2,6-Diborenes



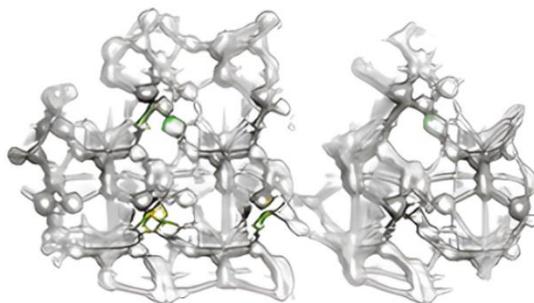
Several N-heterocyclic carbene (NHC)-stabilized 2,2,6-diborenes, $[(\text{NHCH})_2\text{B}_2]$ and π -unsaturated diborenes, have been synthesized and structurally characterized. The boron-containing diborene derivatives are a new class of naphthalenyl-substituted cyclophane-type molecules. X-ray crystallographic analysis of the corresponding dihydro-2,2,6,6 borane compounds reveals that the dihedral angle between the BB bonds is 2.873° .

Characterization of a Paramagnetic, Mononuclear Pt(III)-Alkyl Complex Intermediate in Carbon Monoxide Dehydrogenase



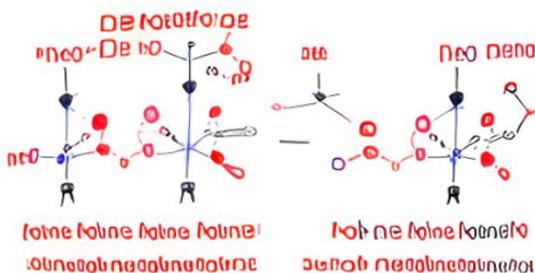
Using a combination of X-ray absorption spectroscopy and density functional theory calculations, we have identified Pt(III)-alkyl intermediates in the active site of CO dehydrogenase from the cyanobacterium *Psococcus* sp. Y218. The ligands are derived from pyridine and thioether groups, and the Pt-C bond distances and bond angles are calculated to be 1.919(9) and 1.16(10) nm, respectively. These results are in good agreement with the experimentally determined values.

Conformational Preferences of Nucleic Acid Bases in the Gas Phase and in Water: The Effect of Alkylation on Base Excision Repair



Molecular dynamics simulations of the base excision repair reaction of adenine in the gas phase and in water have been performed. The results show that the enthalpy of base dissociation is 1 kcal/mol lower than that of nucleotide dissociation. This is consistent with the experimental finding that base cleavage is not favored by the alkylation of an oligodeoxynucleic acid base.

Proton-Coupled Electron Transfer in the Reduction of O₂ to H₂O with [Fe₄S₄] Clusters



$[\text{Fe}_4\text{S}_4]$ clusters are shown to catalyze the reduction of O_2 to H_2O in the presence of a proton source, i.e., H_2 . The reaction is characterized by kinetic isotope effects (KIEs) and spectroscopic characterization. The KIE is found to be 1 compared to that of the corresponding $[2\text{Fe}]$ -cluster. This is consistent with redox potentials that are correlated with the protonic oxidation state of $[\text{4Fe4Cl}_4]$, which is the most favorable pathway for the reaction.

Four examples of artificial chemistry abstracts generated using machine learning models at www.thisjacsdoesnotexist.com.

In this project, I used the latest version of the most well-known GAN model, [StyleGAN 3](#), and scraped around 60,000 table-of-contents (ToC) images from the *Journal of the American Society (JACS)* to train it. This was the most computationally intensive part of the project, but I was able to train the model to a reasonable level running on a consumer-grade graphics card on my home computer for a week or two.

Having obtained a model that could generate somewhat convincing ToC images, I wondered if I could caption them with appropriate titles – given that each training sample includes both an image and its associated title, a machine learning model must be able to learn the link between the title and the ToC image! This turned out to be possible, but a more involved process than training the GAN.

For many text-based models, the [transformer](#) neural network architecture remains the underlying technology behind some of the most powerful models around today, the most well-known of which might be OpenAI's

GPT-3. A key breakthrough of the transformer architecture is that it allows the efficient learning of longer-range connections between elements in training sequences (“attention”) – this meant that language generation models could move from generating sequences that only made sense locally to much longer sequences of structured and meaningful text. It is still extremely expensive to train models such as GPT-3 on huge amounts of text, making it out-of-reach for most individuals. However, once trained, the parameters (or “weights”) of these models can then be used to perform inference (running tasks by passing values through the trained model) very cheaply.

Another important aspect of the transformer architecture is that it was designed for sequence-to-sequence translation. It does this by encoding a sequence into a lower-dimensional space using one transformer network (the “encoder”), then decoding that sequence using a second transformer network (the “decoder”), both of which are trained together.) Interestingly, the individual parts were found to be very powerful in their own right – the BERT model that powers Google’s search results is based on an encoder-only model, while GPT-3 is based on a decoder-only model.

For our purposes, we will treat generating titles for ToC images as translation, but from a picture to a title, rather than, say, English to German. (Many chemical problems can also be formulated as sequence-to-sequence translation using transformers – for example, [retrosynthetic analysis](#).)

To perform this translation, we need a full encoder-decoder model. While training such a model is beyond an individual’s reach, we can take advantage of the millions of dollars large tech companies have already spent to train publicly available large models, then “fine tune” on our data. For this project, I used a software framework from Hugging Face (a machine learning company that develops the most popular open-source implementation of transformer models) to construct an encoder-decoder model consisting of a “vision transformer” encoder (an adaptation of the transformer architecture for images rather than text, specifically Microsoft’s pretrained [BEiT](#)), and a GPT-2 decoder (the publicly available predecessor to GPT-3). Training this “warm-started” model on the 60,000 JACS ToC-title pairs gave a model that could generate a title given a ToC-like input image.

The final part of a paper abstract is the main abstract text. For this part, I fine-tuned a different sequence-to-sequence model pretrained on a large text corpus (T5, trained by Google) using title-abstract pairs. To deploy the final site, I generated 50,000 ToC images using the GAN model; from these, I generated titles, and then from those, I generated abstract text. These are all hosted on a web server, and every time a user accesses [www.thisjacsdoesnotexist.com](#), a random matched ToC-title-abstract set is served. If you would like to test them out, the [ToC-to-title](#) and [title-to-abstract](#) models are also hosted as web apps.

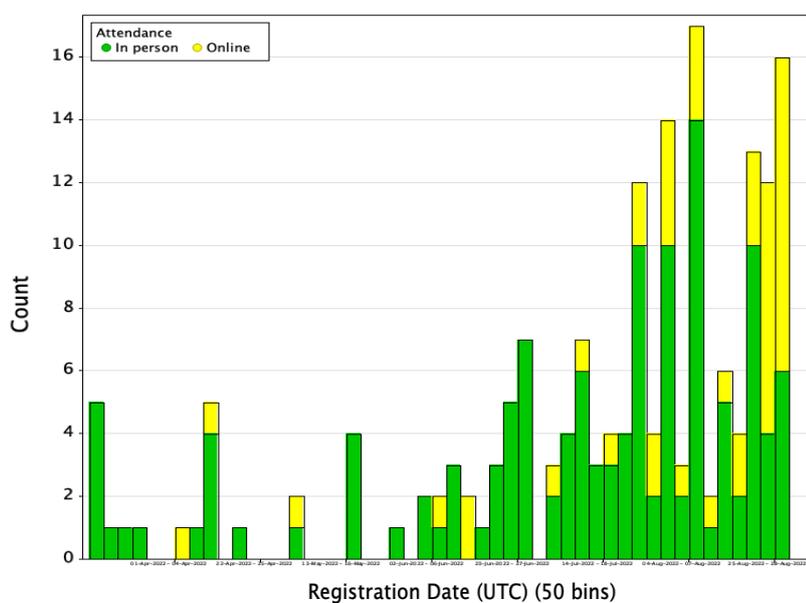
Meeting Report: RSC-CICAG and RSC-BMCS 5th Artificial Intelligence in Chemistry Conference

#AIChem22

Contribution from Samantha Hughes, AstraZeneca, email: Samantha.Hughes@astrazeneca.com and Chris Swain, Cambridge MedChem Consulting, email: swain@mac.com

Meeting write-up contributors: Samantha Hughes (AstraZeneca), Chris Swain (Cambridge MedChem), Noel O’Boyle (Sosei Heptares), Nathan Brown (Healx), Hannes Whittingham (AstraZeneca), Ruben Sanchez Garcia (University of Oxford), Samuel Boobier (University of Nottingham), Garrett M. Morris (University of Oxford) and Morgan Thomas (University of Cambridge).

The meeting held on 1-2 September 2022 at Churchill College, Cambridge, UK, was the 5th meeting organised by the RSC's CICAG and BMCS in a series that started in 2018 as a small one-day meeting at Burlington House in London, which over the years has expanded to a two-day meeting with an increasing number of delegates. The previous two meetings in 2020 and 2021 were held virtually due to the COVID-19 pandemic lockdowns, and this meeting marked the return to an in-person event with a broadcast virtual attendance option. The organising committee comprised of Dr Nathan Brown (Healx), Dr Samantha Hughes (Co-Chair, AstraZeneca), Prof. Garrett Morris (Co-Chair, University of Oxford) and Dr Chris Swain (Cambridge MedChem Consulting).

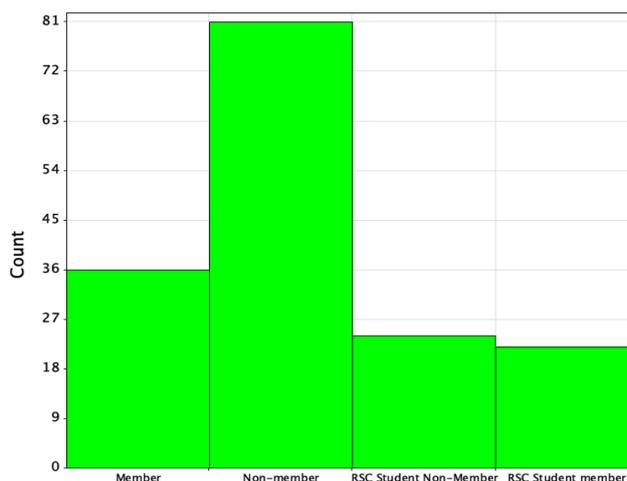


Around 180 people registered for the event including delegates, speakers, committee members and exhibitors, with 130 registering for in-person attendance, and around half of the people who registered for virtual attendance doing so in the week before the event. During the meeting, questions were taken from the audience in the auditorium and online via Zoom. There were delegates from 22 different countries, with the UK, USA and Switzerland being the top three. The male to female ratio was ~60:40, with around 5% "other", or who preferred not to say. Nearly 25%

had dietary requirements and this is becoming an increasingly important part of the conference planning.

Despite the majority of attendees being from the UK, most were not RSC members; however, among students there were an equal number of RSC members and non-members.

The 14 speakers were drawn equally from industry and academia, half were from outside of the UK, and there was an overall gender ratio of six female to eight male speakers. At least seven of the speakers were early career scientists. The 17 flash poster presentations were delivered by five female and 12 male speakers. The session chairs consisted of three female (one early career) and three male chairs. In total, 44 posters were submitted; the Academic and Industry Poster Prizes were judged by a panel of four judges, with even representation across industry and academia. The third Poster Prize, the People's Prize, was voted for by in-person attendees.



There was a panel discussion session at the end of day 1. The panel comprised of keynote and invited speakers, (Prof. Conor Coley, MIT; Prof. Charlotte M. Deane, MBE, University of Oxford; Prof. Kim Jelfs, Imperial; and Dr Teodoro Laino, IBM) as well as a committee member (Dr Nathan Brown), and the panel was chaired by Prof. Garrett M. Morris, University of Oxford. The discussion covered a wide range of topics, including future

directions of AI, how to approach the problem of bias in the data and widely used benchmarks, and incorporating inductive bias into AI models when we know something about the underlying physics.

The conference schedule, abstracts and some speakers' slides are available [online](#).

Teodoro Laino from IBM Zurich, Switzerland, gave the first keynote lecture *From a Combination of Chemical Synthesis and Automation to Enzymatic Design: the Many Opportunities of Language Models in Chemistry*. Summary by Chris Swain, Cambridge MedChem Consulting, UK.

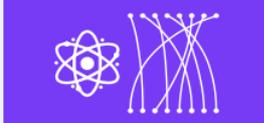
Whilst graph-based methods for encoding chemical structures have proved popular in machine learning, natural language processing (NLP) in organic chemistry has been a very effective way for capturing chemical knowledge and building models of chemical processes. NLP has proved very successful in tasks such as language translation and the approach was to use a similar process, mapping a text sequence that represents the reactants to a text sequence representing the product. In this case the text sequence representing the chemistry used the SMILES notation. This approach is in contrast to other approaches that use hand-coded rules or templates to extract data from published literature.

Most of the training data had been previously extracted from patent data¹ and [Pistachio](#). The limited amount of publicly accessible reaction data was also emphasised by Connor Coley in a later talk, in particular the lack of negative data.

What would you like to do today?

<p>Predict forward reaction</p> <p>Predict the product of a reaction from its precursors</p> 	<p>Predict retrosynthetic routes</p> <p>Predict possible retrosynthetic routes given a target molecule</p> 	<p>Plan a synthesis</p> <p>Plan and execute a synthesis starting from a target molecule, a retrosynthetic route, or an experimental procedure in text format</p> 
---	---	---

There are many things you can do on IBM RXN

 <p>RoboRXN</p> <p>The first remotely accessible, autonomous chemical laboratory.</p> <p>more about RoboRXN</p>	 <p>Text to procedure</p> <p>Translate your chemical experiments from text to exact steps to follow.</p> <p>more about Text to procedure</p>	 <p>Atom mapping</p> <p>It's a chemically agnostic attention-guided reaction mapper with great accuracy and speed.</p> <p>more about Atom mapping</p>	 <p>Model tuner</p> <p>Train your AI model to make it more reliable and effective</p> <p>more about Model tuner</p>
---	--	--	---

They then used a deep-learning model based on the transformer architecture² to translate the largely unstructured experimental procedures into synthesis actions.^{3,4} The molecular transformer was shown to outperform all other published algorithms in reaction prediction.

In the case of retrosynthesis the proposed route to the target is described as a retrosynthetic tree in which the target molecule is successively fragmented to get to simple, easily accessible small molecules. Whilst this simple scheme is useful it would be better if it was annotated with the complete experimental protocol. The synthesis actions provide a simplified description of the experimental protocol. This structured format is particularly automation-friendly and has been used to develop the [RoboRXN](#), an autonomous chemical laboratory.

These tools are all available [online](#) and the code is available on [GitHub](#).

References

- (1) Lowe, D.M. PhD thesis, University of Cambridge, 2012.
- (2) Vaswani, A. et al. Attention is all you need. *Advances in Neural Information Processing Systems*. 2017, 30.
- (3) Schwaller, P. et al. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Central Science*. 2019, 5, 1572–1583.
- (4) Vaucher, A.C. et al. Automated extraction of chemical synthesis actions from experimental procedures. *Nature Communications*. 2020, 11, 3601.

Hélène Gaspar from Benevolent AI, UK, spoke about *Global and Local Experts for Molecular Activity Prediction*.

Summary by Noel O'Boyle, Sosei Heptares, UK.

When developing Quantitative Structure-Activity Relationship (QSAR) models, there is a choice between developing a global model or a local model. Global models including all of the available data, or perhaps a diverse selection thereof, are intended to be applicable to any compound and are especially useful when we don't know the series in which we are interested. Local models are built on a more localised set relevant to a particular series and can give better performance if within their domain of applicability. Typically you would have several local models to cover the series present in a project. For Hit ID, a global model is more useful; later on as a project progresses, local models become of more interest. However, a problem with local models is to know which local model to apply for a particular compound.

In the QSAR literature in the early 2000s, there were several papers comparing global and local models. There was the rise of the AutoQSAR approach, building all possible models and choosing amongst them in an automated pipeline. An alternative is to use ensemble learning, which might be a consensus (average or majority voting), or by combining expert predictions either from pre-defined experts or by weighting predictions. But how to learn the weights?

In the approach described by Gaspar et al.¹ there are several experts. The 'gate' takes the molecule as an input and assigns weights to the experts and we have a weighted average of predictions. The gate learns the parameters. This ultimately gives rise to implicitly localised experts in contrast to explicitly localised experts, where the problem space is partitioned *a priori*. An example of this is the paper by Dörgó et al.² Gaspar et al.'s approach was different – it can also be extended for the multi-label case, where activity is predicted against several proteins and each expert focuses on one protein.

The chemical series or scaffold is described by a SMARTS pattern. These can be hand-defined or can use automatic series detection algorithms (AutoSMARTS in Gaspar et al.'s paper). This is similar to the approach

described by Kruger et al.,³ a tree-based method. Also implemented was a kNN-based method similar to that of Hattori, et al.;⁴ however they don't use a similarity measure but rather discard it if it matches. The two algorithms provide similar results, and they usually use the tree-based one – both are very fast.

Using a dataset from ChEMBL, the tree-based method was applied to find a top chemical series of interest. Time-splitting was used. The training set contained a mixture of molecules that matched or didn't match the pattern, while the validation and test sets just contained molecules that matched the pattern. Several experiments were carried out. One of the main advantages of this method is that you can introspect the data; for example, for the one molecule shown, the highest weights were for the first two local experts; for another it was a local expert and the global model. When you do this for multiple proteins, the model learns to use the right protein spaces.

References

- (1) Gaspar, H.A.; Seddon, M.P. Glolloc: Mixture of global and local experts for molecular activity prediction. <https://openreview.net/forum?id=Mdj229oYWa3>
- (2) Dörgö, G. et al. Mixtures of QSAR models: learning application domains of pKa predictors. *Journal of Chemometrics*. <https://doi.org/10.1002/cem.3223>
- (3) Kruger, F. et al. Automated identification of chemical series: classifying like a medicinal chemist. *J. Chem. Inf. Model.* **2020**, *60*(6), 2888-2902. <https://doi.org/10.1021/acs.jcim.0c00204>
- (4) Hattori, K. et al. Predicting key example compounds in competitors' patent applications using structural information alone. *J. Chem. Inf. Model.* **2008**, *48*, 135-142. <https://doi.org/10.1021/ci7002686>

Kim Jelfs from Imperial College, UK, spoke about *Remembering the Lab in Computational Molecular Material Discovery*.

Summary by Noel O'Boyle, Sosei Heptares, UK.

The Jelfs group work on predicting new materials such as supramolecular materials, host-guest materials and materials built from organic building blocks. One area of interest is porous molecular materials. These don't have extended molecular bonding in 3D dimensions (like MOFs), they just pack together in the solid state. A distinction exists between extrinsic porosity – classical example is dianin – and intrinsic porosity, where there is a cavity. Metal-organic polyhedra, calixarenes, imine cages, boronate cages are examples. These tend to have a few 100 m² per g of surface area – but this is not record breaking. These are useful for sensors or molecular separation, and are typically soluble in common solvents.

Jelfs et al. were trying to assist their experimental collaborations with predictions in a few areas. Given two structures in 2D, they wanted to predict the topology. This can be quite hard. An example was shown where completely different topologies occur despite very similar amines (five-membered vs six-membered rings). Another question is whether the structure is 'shape persistent', i.e., does it require the solvent to maintain porosity? The majority of systems (95%) are not shape persistent and it would be useful to be able to predict this. Prediction of solid-state packing is also of interest. They succeeded in bringing these parts together as part of a study in 2019 – precursors to molecule prediction to crystal packing prediction.

Beyond the structure, Jelfs described how they also want to predict the properties of the system. These are inherently modular systems, and properties of the molecule can be used to predict properties of the solid state. They have developed the open source [pyWindow](#) software, which is used to quantify how many windows are observed during an MD simulation. It is designed to analyse the structural properties of molecular pores (porous organic cages) as well as MOFs and metalorganic cages. Jelfs also described the development of [stk](#) ('Supramolecular toolkit'), software that can automate the assembly and testing of supramolecular systems.

This allows the assembly of cages in different topologies and can also consider chirality. It uses RDKit and is open source on [GitHub](#).

Using a neural network for polymer screening, they were able to identify trends. In predicting shape persistence, they needed to generate their own dataset computationally, consisting of 63K polymers in a variety of different topologies. While they tried neural networks, it turned out that random forest was just as good, overall having 93% accuracy within class. A graph neural network was used to understand collapse, and this improved the prediction. Attribution scores and integrated gradients were used to highlight which parts of the molecules were prone to collapse.

Jelfs et al. wanted to incorporate ease of synthesis, and asked experimental collaborators to score potential precursors, asking, “can you make 1g of this compound in less than five steps?” They built a machine learning model using that data.¹

Jelfs then moved on to the topic of polymer membranes for separation, specifically PIMs: Polymers of Intrinsic Microporosity. PIMs have a *spiro* group, while the polymer chains are rigid and contorted and could be used to separate CO₂ and N₂, for example. Unfortunately, there is not a lot of data, as only 200 or so have been reported. However the Membrane Society of Australasia had gas separation data for a whole variety of polymers. Generally, while wanting high permeability (for fast separation), and high selectivity (for clean separation), this leads to a trade-off because as soon as permeability is increased (pores are made larger), selectivity is lost (which requires smaller pores). As the datasets are sparse, an imputation method was used to fill in the gaps. A prediction method was built with Bayesian linear regression. They made predictions of promising historical (known) polymers, and found an example which was missing from the database but had actually been reported in the literature as useful for separation.

Reference

(1) Bennett, S. et al. Materials precursor score: modeling chemists’ intuition for the synthetic accessibility of porous organic cage precursors. *J. Chem. Inf. Model.* **2021**, 61(9), 4342-4356. <https://doi.org/10.1021/acs.jcim.1c00375>

Julien Michel from the University of Edinburgh, UK, spoke about *Hybrid Alchemical Free Energy / Machine-learning Methodologies for Drug Discovery*.

Summary by Noel O’Boyle, Sosei Heptares, UK

Free Energy Perturbation (FEP) pipelines for the estimation of relative binding free energy are evolving. Such calculations take about 10 CPU hours per compound pair. Given this expense, how can one enhance the data from the simulation after the simulation is over?

With standalone FEP there is an approximate energy function. If the error for this was known, then it could be corrected retrospectively. The Michel group have developed a correction term using machine learning (ML). FEP/ML requires a fraction of the training set ML needs to outperform FEP. A pure ML method would need a sizable dataset. It is much easier to learn the error, rather than the actual value. This was tested on data from SAMPL challenges using a time split. The ML correction term retrospectively boosts FEP performance in the SAMPL4 competition.

The state of the art in the network planning stage in FEP is rule-based edge scoring (e.g. LOMAP). This is a subjective process, and different results will be obtained depending on which edges are included, as some edges are better behaved than others. The issues with rule-based methods are that someone needs to come up with the rules, the rules are not perfect, and they may not extend to new areas of chemical space.

The goal of Relative Binding Free Energy (RFBE) network planning is to pick a graph that maximises accuracy at a minimal computing cost. You need a spanning tree, but a fully connected tree is too expensive. Their hypothesis is that inaccurate edges have large FEP protocol statistical uncertainties. When we get the energy, we also get the mean and uncertainty. However, it would be too expensive to actually measure this for each edge and so they introduce ML to estimate this.

When creating a training set, it's hard to construct a sufficiently large set that would be transferable. They decided to simplify the problem and mined the literature for congeneric series. They threw away the protein information, and grafted the substituent that is changing onto a phenyl ring. Running the simulation of the latter in a water box they obtained a measure of uncertainty in FEP. Does it still relate to the question of interest? They plotted the uncertainty from FEP vs the uncertainty from the water box simulations and showed that changes with large uncertainties in the former also showed large uncertainty in the reduced (water box) system.

They used a Siamese neural network architecture to encode the molecule with/without the R group. They pre-trained the model with a cheap label (delta ESOL), and then used transfer learning to predict the uncertainty of the FEP calculation. Prof. Michel showed an application to a TYK2 dataset. They looked at different edge scoring protocol methods: star shaped networks were as poor as random edge selection, but even a fully connected network does not guarantee best accuracy. The offset measure gives the most accurate results based on 1/6 of the edges. Shallow learning (RF) did not perform too badly but it was worse. In conclusion, RBFENN has comparable performance to LOMAP rule-based scoring, but instead is data-driven.

Miriam Mathea from BASF SE, Germany, spoke on the topic of *Machine Learning for Toxicity Prediction – Applications at BASF*.

Summary by Nathan Brown, Healx, UK.

One of the most challenging aspects of effective molecular design, other than general potency against the biology implicated in disease, is the various challenges of toxicity [prediction], which itself is a convolution of manifold characteristics, and even more challenging when dealing with *in vivo* model systems as opposed to *in vitro* models. Dr Mathea delivered a talk on using machine learning to predict toxicity outcomes from *in vivo* endpoints for targets involved in molecular initiating events of thyroid hormone homeostasis. With over two million animals used per year in Germany alone, it is essential that the community look to improve our models so that we can reduce this number for ethical and economic reasons.

Dr Mathea reported the ChemBioSim approach developed at BASF using conformal modelling approaches using a combination of molecular descriptors and biological fingerprint data derived from bioassays. Since the bioactivity data was sparse, Dr Mathea introduced predicted data points to augment and expand data coverage. The team found that the introduction of the biological descriptors significantly improved the predictive performance of their *in vivo* models. While the team investigated a number of different modelling algorithms, they found that Random Forests appeared to work best for their use-cases.

Daniel Probst from École Polytechnique Fédérale de Lausanne, Switzerland, spoke on the topic of *Socioeconomic, Environmental, and Scientific Considerations for the Recent Technological Shift in Cheminformatics and Computational Chemistry*.

Summary by Hannes Whittingham, AstraZeneca, UK.

Exploring an unusual perspective on the impact of AI in chemistry, Dr Probst examined the external issues associated with its growth, discussing hardware inequalities between researchers, environmental cost, and reproducibility issues. In particular, he highlighted the extreme and exponentially increasing cost of

computational hardware for cutting-edge deep learning research. He presented evidence that the success and volume of such research are therefore strongly predicted by available funding, leading to large inequalities between researchers by nationality, institution and industry. Further to this, he showed that the energy requirement to train a single advanced language model may be as much as the annual electricity consumption of 1,000 European citizens and run to over \$1m in cost, which makes it clear both that the environmental impacts cannot be ignored, and that the sheer expense of such research prohibits reproducibility.

Charlotte Deane, MBE from the University of Oxford, UK, gave the second keynote talk, titled *The Power and Pitfalls of Machine Learning in Early Stage Drug Discovery*.

Summary by Samantha Hughes, AstraZeneca, UK.

Day two of the conference began with a keynote lecture by Prof. Deane, covering the application of machine learning in early stage drug discovery, using structure-aware models. The first part of the talk focussed on the prediction of binding and strength of binding, with Prof. Deane asserting that these aspects are not well predicted. She described DenseFS¹ a method for structure-based virtual screening which uses a convolutional neural network (CNN) model based on the DenseNet architecture. CNNs are suited to this challenge because of their usefulness in image recognition. In DenseFS, determination of binding is reframed as a computer vision problem.

In the first step, the binding site from input 3D protein-ligand complex structures is discretised into a grid format, containing information channels belonging to the protein and the ligand. Receptor channel descriptors included aliphatic carbon, nitrogen acceptors and sulfur, while ligand channel descriptors include aromatic carbons and sulfur acceptors. The binding site is turned into an image via voxels which is used as input to the deep-learning architecture and the CNN is trained to predict whether the ligand is a binder or non-binder. Validation was performed using the DUD-E database. As the true poses were unknown, the models were trained on the top-ranked pose from docking and then tested on the top-ranked docked pose and all generated poses using clustered cross-validation: DenseFS performed best compared to DenseU (a modified DenseFS), a baseline CNN model and AutoDock Vina scores. When applied to a ChEMBL test set, DenseFS dropped in accuracy but was still the best-performing method tested.

The second part of the talk dealt with removing biases in datasets, citing a paper by Chen et al.,² describing deep-learning methods “cheating” at structure-based virtual screening due to hidden bias in the DUD-E dataset. The DenseFS method does not allow the user to control the split therefore if there is bias in the data, then this could be used by the algorithm to do this splitting. This raised the question of does this matter if the model is learning a useful bias? As it is difficult to go back to work through the model to find out what it has learned, they investigated this by deleting the protein from the 3D input, so that it only comprised the small-molecule information. The results showed the same distribution of scores for actives and decoys as the ligand-and-receptor test set, thus they concluded that the model has learned to discriminate *entirely* on the ligands in the dataset due to the biases present in the ligands. To counteract this, they associated three new decoys with each active – the decoys have the same chemical structure but put a random place in the protein structure through application of a random translation, orientation and conformation, to create a new dataset DUD.E-Trans. The new deep learning model TransFS trained on this data set is forced to learn from the protein and the physical interactions between the protein and ligand.

They tested the influence of bias through the process described earlier (providing only the ligand as input) and demonstrated that there was still a signal indicating that they have not completely solved the problem but they have made the model a lot better as TransFS performed as well or better than their original DenseFS on 13/14

unseen targets from the ChEMBL data set. For kinase targets they saw the least difference, as kinases are well represented in the set, while the improvements were the largest for targets underrepresented in DUD-E.

They then demonstrated that TransFS assigns importance to both protein and ligand atoms associated with interactions, by performing a masking process, i.e. deleting an atom and putting the “1-atom-deleted” and the full system through the CNN. The CNN score is calculated in the presence and absence of each atom and the difference in score is assigned as a ‘masking score’ for that atom. This analysis showed that that model recognised as important features from both the protein and ligand that are associated with interactions.

The third part of the talk covered molecular design in the receptor pocket and how to encode the protein pocket in a suitable way. In fragment-based drug discovery (FBDD), fragment hits are used as the basis for designing molecules which bind to the protein with higher affinity. Deep learning based generative models are increasingly proposed as an alternative to human design for FBDD. Generative design tasks can include scaffold hopping, fragment linking, PROTAC design, R-group optimisation and fragment growing as all these approaches change molecules in some way to bind better. They chose a Graph Neural Network (GNN), as nodes and edges naturally correspond to atoms and bonds in molecules and GNNs capture dependencies between nodes and can thus quantify the effect of adding a given functional group to the fragment. Their method was based on a constrained graph variational autoencoder framework which, starting with a single fragment, initialises a pool of fragments that can be added. At each step, the NN decides which atom should be added and with which bond type, until a stop node is sampled. Structural information from the binding site is incorporated through the imposition of pharmacophoric constraints which the functional groups must satisfy. Pharmacophoric restraints are derived from existing active molecules using software from the Deane group, e.g. DEVELOP,³ or extracted directly from the protein target using STRIFE,⁴ or via the Fragment Hotspot API tool.⁵ STRIFE derives a 1D pharmacophoric profile from the 3D structures, these can be coarse-grained (i.e. pharmacophoric points only) or fine-grained (pharmacophore points and path distances). Generation is done in two stages – an exploration stage with the coarse-grained pharmacophore from which a fine-grained pharmacophore profile is generated and used in the subsequent refinement stage. Final elaborations are ranked by ligand efficiency. Five metrics were used to assess the generative designs 1) validity of structures, 2) uniqueness, 3) novelty, 4) recovery of ground truth structures (i.e. the CASF test set), and 5) 2D property filters and attractiveness. Results from the CASF test set showed that STRIFE generates a greater proportion of unique and novel elaborations compared to structure-unaware methods such as Scaffold Decorator and and CREM with slightly lower validity of the structures. A live demo of the software was shown, demonstrating that it is fairly simple to setup the pharmacophore constraints and run the generative design.

To finish, Prof. Deane described some new work from her group to develop ML methods that understand the physics underlying protein-ligand binding, and applied these to identifying binding hotspots. They developed PointVS – a group-invariant GNN method to predict binding affinity and demonstrated good performance vs other scoring functions. The standard in the field is to use the PDBBind General set for training and CASF-2016 as the test set. However an analysis showed that 284 of the proteins in the CASF-2016 set have $\geq 90\%$ similarity to a protein in the general set, and 273 have an identical sequence! They thus set out to debias the training and test sets as far as possible by removing any protein with $>80\%$ sequence similarity in the CASF-2016 data set. There was a drop in performance for all methods on the debiased dataset but greater confidence that the method is not simply learning the dataset.

PointVS outperformed the other methods tested. Feature attributions to edges and nodes highlighted key ligand-protein interactions as edge “hotspots”. To see if the hotspots from the PointVS method are the same as hotspots derived by the Fragment Hotspot API methods, the two methods were applied to the same crystallographic fragment screening set: there is some overlap of hotspots identified by the two methods.

PointVS hotspots were found to generate “better” molecules when incorporated into the STRIFE method (described earlier) applied to various SARS-CoV-2 proteins, this is attributed to PointVS ranking hotspots better than the noisier Hotspot API method.

References

- (1) Imrie, F. et al. Protein family-specific models using deep neural networks and transfer learning improve virtual screening and highlight the need for more data. *J. Chem. Inf. Model.* **2018**, *58*, 2319-2330. <https://doi.org/10.1021/acs.jcim.8b00350>
- (2) Chen, L. et al. Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. *PLoS One.* **2019**, *14*(8), e0220113. <https://doi.org/10.1371/journal.pone.0220113>
- (3) Imrie, F. et al. Deep generative design with 3D pharmacophoric constraints. *Chem. Sci.* **2021**, *12*, 14577-14589. <https://doi.org/10.1039/d1sc02436a>
- (4) Hadfield, T.E. et al. Incorporating target-specific pharmacophoric information into deep generative models for fragment elaboration. *J. Chem. Inf. Model.* **2022**, *62*, 2280-2292. <https://doi.org/10.1021/acs.jcim.1c01311>
- (5) Radoux, C.J. et al. Identifying interactions that determine fragment binding at protein hotspots. *J. Med. Chem.* **2016**, *59*, 4314-4325. <https://doi.org/10.1021/acs.jmedchem.5b01980>

An Goto from the University of Oxford, UK, spoke about *De novo Molecular Design in 3D using Available Reagents, Reactions, and Docking in Deep Reinforcement Learning for SARS-CoV-2 Main Protease*.

Summary by Ruben Sanchez Garcia, University of Oxford, UK.

Compound optimisation, the process in which new compounds are designed to better satisfy a desired set of properties, is one of the most important problems in computational chemistry. Although several approaches have been proposed to tackle it, including reinforcement learning (RL) or generative models, the problem is far from solved. With the aim of addressing the recurrent limitation of low synthetic feasibility designs, Goto et al. proposed RxnDQN, a RL-based method that explicitly incorporates chemical feasibility. RxnDQN is based on the MolDQN framework, in which molecules are generated applying a set of actions on the initial molecule while aiming to maximise the reward that combines the different properties to be optimised such as QED (Quantitative Estimate of Druglikeness) or logP. However, contrary to MolDQN, in which the actions operate at the graph level (atom or bond addition or removal), RxnDQN action space is defined as a set of well-established chemical reactions, thus maximising the chances of suggesting synthetically accessible compounds. In addition, they also proposed a hybrid version, RxnMolDQN, in which the action space of MolDQN and RxnDQN are combined.

Experiments carried out on ten SARS-CoV-2 Mpro fragments showed that both RxnDQN and RxnMolDQN results exhibit far better synthetic accessibility and chemical space coverage than MolDQN designs while achieving comparable if not better values for the property being optimised, namely QED. RxnDQN was shown to be marginally better than RxnMolDQN in terms of synthetic accessibility, but RxnMolDQN showed better chemical space coverage. Finally, a 3D version of RxnDQN and RxnMolDQN, in which docking scores are included into the reward function, was shown to propose sensible designs with improved ligand efficiencies and excellent levels of synthetic accessibility.

Donal O'Shea from the Royal College of Surgeons in Ireland (RCSI), Dublin, Ireland, spoke on *Forecasting Vaping Health Risks Through Neural Network Model Prediction of e-liquid Flavour Pyrolysis Reactions*.

Summary by Nathan Brown, Healx, UK.

The recent rise in use of vaping technologies and range of options in the vaping devices together with the plethora of different flavour options, has led to some concern over longer term use of these medical devices, as licensed by the Federal Drug Administration (FDA). Dr O'Shea presented his talk detailing the size of the

market, 470 brands and 7760 unique flavours, and possible health liabilities, highlighting some recent publications in such journals as the NEJM.

Dr O'Shea was motivated to investigate pyrolysis products using machine learning methods due to the complexity and time involved in conducting the experiments. Typically, the vaping liquids come in two forms, each in with a drug and a carrier/dilutant component: namely THC with Vitamin E acetate and nicotine with propylene glycol. In addition to the chemicals under instigation, Dr O'Shea highlighted that the vaping devices could often be user-controlled to temperatures from 110–1000 deg. C, with temperatures reaching those that enable thermally driven reactions.

Since the real experimental set up would be too time consuming to explore,¹ Dr O'Shea applied published machine-learning models, open sourced by Coley et al.² to predict likely products from pyrolysis, resulting in the identification of a set of 7307 products, which was compared with the NIST EI-MS database of 33,000 compounds. From this, 1166 molecular mass matches were determined. The chemicals identified included 127 with reported acute toxicity, 153 with health biohazard, 225 as irritants, 95 not present in the database, and 566 with no hazard reported. Further investigations are ongoing for a more comprehensive evaluation of chemical hazards that could arise from pyrolysis reactions of these ingredients at high temperatures.

References

- (1) Wu, D.; O'Shea, D.F. Potential for release of pulmonary toxic ketene from vaping pyrolysis of vitamin E acetate. *Proc. Natl. Acad. Sci.* **2020**, *117*, 6349-6355. <https://doi.org/10.1073/pnas.1920925117>
- (2) Coley, C.W. et al. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.* **2019**, *10*, 370-377. <https://doi.org/10.1039/C8SC04228D>

Connor Coley from MIT, USA, spoke on *Models of Chemical Reactivity to Inform Molecular Design*.
Summary by Samuel Boobier, University of Nottingham, UK.

Molecular design concerns generating functional molecules with desirable properties. Screening virtual libraries can be time-consuming and subject to constraints on the search space explored. Alternatively, a range of generative models have been used to explore even larger areas of chemical space.¹ Synthesis planning is the process of designing a route to a desired molecule from available starting materials. Retrosynthesis is the reverse of this process, recursively breaking down a target molecule. There is now the focus in retrosynthetic planning of including details such as additives, solvents, and conditions. This is particularly important for process chemistry and autonomous laboratories, where very precise procedural details are required. In this work, Coley describes the combination of machine-learning-based molecular generation with synthesis planning to produce molecules via reactions which are therefore synthesisable.

Reaction data sources range from general databases such as USPTO (patent data),² Pistachio,³ Reaxys,⁴ and CAS/SciFinder,⁵ on which global models may be built, and more specific high-throughput data such as the Merck CN coupling collection⁶ more suitable for local models. Many datasets are proprietary, and the quality of the data is affected by missing values/details, *e.g.*, *what were the by-products in this reaction?* A cultural shift is required to encourage companies to release data and provide negative examples for machine learning. The Open Reaction database⁷ is one such effort to produce freely available structured reaction data for researchers.

In single step reaction prediction, templates can be inferred from reaction data, where these templates are crudely the “rules of chemistry” to determine the allowed chemical transformations. To combat the rigidity of these rules, there are also template-free methods which directly relate reactants to products, formulated as SMILES-to-SMILES, Graph-to-Graph, and Graph-to-SMILES. The Coley group found particular success with

Graph2SMILES,⁸ combining the power of transformer models for text generation, with the permutation invariance of molecular graphs. The importance of bias in datasets was extensively highlighted in this meeting, and Coley demonstrated that testing via a document split (reactions from a single paper are all placed in the test or training set) exhibited a drop in accuracy compared to a random split. For multistep retrosynthesis, one step strategies can be recursively searched using Monte Carlo Tree Search (MCTS) using reinforcement agent-based learning. Models can create pathways indistinguishable from “real” or human routes in a blind A/B preference survey of chemists.⁹ However, the “over the arrow” problem of predicting conditions is lacking in many of these studies.

ASKCOS¹⁰ is an open-source tool which can predict partial conditions for retrosynthesis, with search time of approximately 15 seconds per target. The conditions are predicted initially as a supervised classification problem. Retrospective analysis of the embeddings revealed clustering of similar reagents together, adding additional verification to the model.

A novel method of molecular design was proposed as a deep generative model, which uses synthesis planning to filter the potential candidates. In this way, molecules are built up by reaction rather than atom so that all additions are actionable. The task is formulated as a single shared task of conditional synthetic pathway generation, with a genetic algorithm.¹¹ With the route to the new molecule automatically generated, suggested molecules are necessarily synthesisable.

Data-driven synthesis planning is becoming more widely used and will continue to do so. Synthetic predictions need to be precise and actionable, especially for the increasing prevalence of robotic systems. Many other challenges remain in the field. Synthesis planning is not yet driving new synthetic innovation, rather proposing existing technologies. Additionally, complex natural products are still beyond the scope of models. In the future, models will need to be more quantitative than qualitative and prospective rather than retrospective.

References

- (1) Coley, C.W. Defining and exploring chemical spaces. *Trends in Chemistry*. **2021**, 3(2), 133-145.
- (2) Lowe, D. https://figshare.com/articles/dataset/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873
- (3) NextMove Software, <https://www.nextmovesoftware.com/pistachio.html>
- (4) Elsevier, <https://www.reaxys.com/>
- (5) CAS, <https://www.cas.org/cas-data/cas-reactions>
- (6) Ahneman, D.T. et al. Predicting reaction performance in C–N cross-coupling using machine learning. *Science*. **2018**, 360(6385), 186-190. <https://doi.org/10.1126/science.aar5169>
- (7) Open Reaction Database, <https://open-reaction-database.org/>
- (8) Tu, Z.; Coley, C.W. *J. Chem. Inf. Model.* **2022**, 62, 3503-3513.
- (9) Segler, M.H.S. et al. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature*. **2018**, 555, 604-610. <https://doi.org/10.1038/nature25978>
- (10) ASKCOS, <https://askcos.mit.edu/>
- (11) Gao, W. et al. Amortized tree generation for bottom-up synthesis planning and synthesizable molecular Design. <https://arxiv.org/abs/2110.06389>

Raquel Rodriguez Perez, NIBR, Basel, spoke on *Machine Learning-based Predictions of ADME Properties in [the] Pharmaceutical Industry*.

Summary by Garrett M. Morris, University of Oxford, UK.

One of the key aspects of developing a successful drug rests on predicting its *in vitro* metabolic clearance. Hepatic cytochromes P450 play a key role in metabolic stability, and can be measured in liver microsomes, including in human, rat, mouse, dog, monkey and mini-pig. For machine learning, this data can be split into more challenging training and test sets by splitting chronologically on registration date.

Low-clearance compounds were defined as having CLint values of <100 μ L/min/mg, high clearance of >300 μ L/min/mg, and medium-clearance in between. This allowed classification models to be constructed; while regression models were also developed to predict the CLint values directly. ML models were explored using either extreme gradient boosting (using the [xgboost](#) library with either Morgan fingerprints or molecular descriptors); or graph neural networks (GNNs) using directed message passing (using the [chemprop](#) library). Results were presented for XGB, GNN, and multi-task GNN classifiers and regression models, and their ability to correctly predict low- and high-clearance molecules. While all models achieved very good accuracies (0.82-0.85) for predicting low-clearance compounds, their performance was worse for high-clearance compounds (0.55-0.65), with the regression models performing slightly better than the classifiers. They also found that ensemble predictions did even better than individual models.

Adrian Roitberg, University of Florida, gave the final keynote talk, speaking on the topic of *Machine Learning for Accurate Energies and Forces in Molecular Systems. Uses in Conformational Searches and Free Energy Calculations*.

Summary by Samantha Hughes, AstraZeneca, UK.

Prof. Roitberg spoke on the topic of Neural Network (NN) potentials and their utility for a variety of computational chemistry applications. Framing the problem as: “we need practical methods for QM, MD etc. and all methods are approximations”. He described the trade-off between accuracy and speed/scaling with forcefield methods computationally scaling as order $O(N^1)$, semi-empirical methods $O(N^2)$, DFT $O(N^3)$ and highly accurate CCSD(T) methods $O(N^7)$; for the last it takes a day for a small molecule single state calculation and it is not possible to sample conformational space. Computational chemists need energies and forces to be as accurate as possible (as accurate as high level quantum calculations) *and* fast (completed in less than ms). Is it possible to train a machine learning algorithm to predict energies and forces within 1kcal/mol given only the structure? Although initially sceptical, he concluded that it does work: they have developed ANI, a neural network potential that is as accurate as QM and can sample a lot of states due to its speed.

The method is called ANAKIN-ME (a reference to Star Wars), and is known as ANI for short. The model was trained on structures and absolute energies from QM and, given a new structure, can infer the energy and forces. He explained how they generated the ANI-I data set in 2016 (the method has moved far beyond this, but it is useful to explain the principles here). To generate the training set, molecules with ≤ 8 heavy atoms were taken from the GDB (*ca* 57K compounds). Conformations were generated for each, resulting in 22M pairs of structure/energy calculated using DFT (total energy) WB97x/6-31g(d). Random molecules and conformations were taken from this to form the ANI-1 training set. The “secret sauce” was Behler and Parinello’s neural-network representation of DFT potential-energy surfaces.¹ In ANI, each NN focusses on an atom, one for each atom type C, O, N, etc. During training, the ANI prediction is compared to the True QM (DFT) Energy, then the network potentials are updated. The molecular representation is an atom plus a sphere of 5 Å radius from the centre (short range interactions), the energy is the sum of the atomic contributions from that Atomic Environment, thus it is possible to differentiate between carbons in the same molecule which have different environments. The current version of ANI handles C, H, N, O, S, F and Cl; in 2021 they introduced charged systems and in 2022 are working on reactions. Ligands or proteins can be input into ANI. ANI is now ported to pytorch (TORCH ANI).

The test set comprised 131 randomly selected molecules with ten heavy atoms, resulting in over 8200 structure/DFT energy pairs with a 300 kcal/mol energy range. Plotting E(ANI) vs E(QM) showed near perfect correlation (RMSE 1.2 kcal/mol). The DFT energies took an average of 1143 s/mol to compute, whereas ANI energies took an average 0.0032 s/mol to compute, representing a 35,700x speed up (the current version is even faster).

They next explored whether it is possible to predict when the model is wrong using ensemble confidence for three NN. Where networks are close in prediction there was less error to the true (DFT) energy while where the networks disagreed the error was found to be larger. An example of ensemble confidence was shown for a molecular dynamics simulation where a collapsed state resulted in the network ensemble average errors increasing dramatically. Such ensemble disagreement can be used drive data generation via active learning: ANI-1X is an automated and self-consistent data generation framework, where ensemble disagreement triggers further QM calculations and retraining of the network models in the ensemble until the networks agree. In this the structure pool can be ligands, proteins, peptides or antibody-drug-conjugates. In testing, ANI-1x predicted harmonic frequencies well for both a small molecule set from Drugbank (with a few outliers) and a tripeptide benchmark set.

CCSD(T) calculations were run for a subset of compounds and used to retrain ANI to create an ANI1-CCX. The performance of these models was assessed on a torsion benchmark set from Genentech. ANI1-CCX was closer to the CCSD (T) energy than the other methods assessed. Other applications to prediction of reaction enthalpies demonstrated that ANI1-CCX could achieve similar accuracies to high level quantum calculations, with a speed of 0.009 s/molecule. Promising applications of ANI to bond breaking and correction of computed (MM) binding free energies were described.

Reference

(1) Behler, J.; Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401. <https://doi.org/10.1103/PhysRevLett.98.146401>

Lucian Chan from Astex spoke on *3D Pride Without 2D Prejudice: Bias Controlled Multi-level Generative Models for Structure-based Ligand Design*.

Summary by Morgan Thomas, University of Cambridge, UK.

This talk expands on some earlier conference discussions on dataset bias, more specifically removing this bias when using generative models. Generative models can be categorised into mostly two types: 1) Those that learn from a latent space represented by SMILES, graphs, or conformers; 2) Those that model input and elaborate on a graph or conformer. Astex have a pipeline of fragment screens that use X-ray crystallography, electron microscopy, thermal shift, and NMR to conduct fragment discovery, therefore, it is the latter type of generative model that is most useful for fragment methods at Astex.

One can form a matched molecular pair database; however, this is not transparent, nor interpretable, and is biased (including explicit reactivity constraints). It contains 1D bias that reproduces common chemistry and 2D bias that has a tendency to reproduce a motif based on synthetic constraints. Moreover, there are distributional shifts between datasets for example, [Enamine](#) and [PDB](#). If you extract chemistry from these databases, you observe different distributions of motif frequencies. Comparing these ratios of frequencies between Enamine Real and PDB highlights that motif frequency can differ by a factor of two or more.

Chan et al. define a multi-level model, named the *PQR* framework. This includes a 1D model based on frequencies (*P*), a 2D model looking at motifs and vectors (*Q*), and a 3D model looking at protein-ligand conformation and growth vectors (*R*). These can be combined to elaborate a 3D molecular fragment that match certain criteria (*PQR*). To achieve this, they use contrastive learning to avoid inheriting the bias through the 1D, 2D, and 3D models. This involves sampling “negative” examples from previous model to debias the data; forcing the model to focus on topology and not frequency for the 2D model, and to focus on growth vector and not topology for the 3D model.

Next they investigated the ability of this PQR framework to reconstruct PDB ligands compared to 0D, 1D and 2D baselines. All models strongly outperformed the 0D uniform sampling baseline, however, there was a performance improvement of the 2D model over the 1D model and more importantly, of the 3D model over the 2D model. The de-coupling of these biases allows this quantification of enrichment obtained by this contrastive learning approach. Next he showed a case where they cut a fragment from a ligand and tried to reproduce it. A dimensionality reduction map shows that the chemical space of likely replacement fragments dramatically decreases through the 1D, 2D and 3D models respectively. However, by using the 2D model, and even more so the 3D model, we increase the chance to reconstruct the true removed fragment. The 3D model is a product of all models and therefore the 1D model still has a strong effect on the 3D model. There exist more examples where higher orders of chemical dimensionality information (the 3D model) increases the correct identification of the true fragments. In summary, this framework decodes the 1D, 2D, and 3D information into separate models without inheriting bias by use of contrastive learning (by sampling negative data). This approach allows the proper evaluation of the bias of each particular model.

Krzysztof Maziarz, Microsoft Research UK, spoke on *MoLeR: Creating a Path to More Efficient Drug Design*.
Summary by Morgan Thomas, University of Cambridge, UK.

Firstly, this work is part of a collaboration between Microsoft and Novartis that allows the prospective validation of MoLeR – a generative model for molecular design. Additionally, Microsoft have a chemistry team in Cambridge that work on molecular generation, molecular properties, protein structures, structure-based drug design and retrosynthesis.

Lead optimisation is typically conducted via the design-make-test-problem and can be augmented by generative models as a smaller *in silico* loop within the design stage. However, as many already exist, what are we looking for from yet another generative model?

1. To be able to constrain it to a scaffold with an arbitrary configuration of attachment points.
2. A model that is not easily exploited during optimisation.
3. Something that is fast in practice.

The first requirement can be satisfied by sequential extension to a graph, so use of a Graph Neural Network (GNN) model is logical. The second requirement can be satisfied by assembling molecules from common fragments and motifs, constraining chemistry to already observed, accessible chemistry. The third requirement is difficult to satisfy with most GNN-based models, as there is a trade-off between control/constraint and simplicity/speed. Overall, it is preferred to do something well which could still be more efficient in total considering the full process (for example, synthesis and testing) but the model should still be engineered well.

MoLeR (the proposed GNN-based generative model) produces *de novo* molecules belonging to the same distribution of a training dataset by training on that dataset of molecules. Molecule generation can be constrained by scaffold, only introduced at inference time. It can conduct multi-objective optimisation by using molecular swarm optimisation.¹ Molecular motifs are extracted from the training dataset which is then used to decompose input molecules into identified motifs. Next, a GNN encoder embeds both atom level and motif level information into a latent vector which is input into a GNN decoder. The decoder learns to construct molecules based the latent embedding and a partial graph by selecting an atom/motif, attachment point or bond type. Due to this separation between encoder and decoder, scaffold constraints can be imposed at inference time only and the decoder can be parallelised to improve computational performance.

How well does MoLeR perform on the GuacaMol benchmark?² It seems the generative model works reasonably well although is not the best performing model, however, when accounting for the quality of molecules it

performs better than most other generative models. When inspecting the scaffold-based benchmarks, some generative models work quite well simply by rewarding the presence of a scaffold. However, MoLeR is more robust as you can always specify the scaffold. Investigating computational efficiency by training and sampling time (mol/sec), it performs better than variational autoencoders but not as fast as recurrent neural networks – in practice, this is fast enough. The code is all open-source and can be found on [GitHub](#).

This approach is undergoing prospective application at Novartis and subsequently being validated on three in-house projects. It should be noted that this is just a pilot, and more projects will be used for validation. Chemists found that MoLeR had a good coverage of chemical space and that proposed molecules looked synthesisable (despite synthesisability not being explicitly encoded). Moreover, the model was useful for generating ideas that the chemists wouldn't have considered or had previously ignored but upon revisiting actually turned out to be reasonable. In summary, good software engineering is important, everything takes longer than expected and connecting different ML algorithms is very complex.

References

- (1) Winter, R. et al. Efficient multi-objective molecular optimization in a continuous latent space. *Chem. Sci.* **2019**, *10*, 8016–8024.
- (2) Brown, N. et al. GuacaMol: Benchmarking models for de novo molecular design. *J. Chem. Inf. Model.* **2019**, *59*, 1096–1108. <https://doi.org/10.1021/acs.jcim.8b00839>

Poster Prizes

At the end of the second day, three poster prizes were awarded:

- Academic Poster Prize Winner: **Leo Klarner** “Bias in the benchmark: Systematic experimental errors in bioactivity databases confound multi-task and meta-learning algorithms”
- Industry Poster Prize Winner: **Lauren Reid** “SARKush: Automated Markush-like structure generation for SAR communication”
- People’s Poster Prize Winner: **William McCorkindale** “Fragment-based hit discovery via unsupervised learning of fragment-protein complexes”

The poster judges were Hannah Bruce Macdonald (MSD), Connor Coley (MIT), Astrid Stroobants (AstraZeneca) and Marton Vass (Schrödinger).

EU-OPENSSCREEN ERIC: an Open-access Research Infrastructure for Chemical Biology and Early Drug Discovery

Contribution from Robert Harmel, EU-OPENSSCREEN ERIC, email: robert.harmel@eu-openscreen.eu, and Tanja Miletić, EU-OPENSSCREEN ERIC, email: tanja.miletic@eu-openscreen.eu



Introduction

The development of high-quality chemical probes is an arduous journey through basic research, discovery, optimisation and biological testing, and the requirements of novel drugs extends this endeavour even further towards clinical trials and regulatory authorities' procedures. The interdisciplinary nature of the research embedded in this typically long and costly process requires an excellent network of scientists with complementary and multidisciplinary expertise. Naturally, it is challenging to find partners that can reliably and efficiently work together on such a project. Therefore, initiatives like the [Structural Genomics Consortium \(SGC\)](#),¹ the [European Lead Factory \(ELF\)](#)² and [EU-OPENSSCREEN](#)³ were established to facilitate interactions between scientists from different countries and different expertise, and to provide resources for supporting these developments. While the ELF offers large-scale proprietary screening campaigns for drug discovery and development within a public-private partnership, the SGC promotes an open collaborative network of scientists that develop chemical probes and new medicine from a structural biology perspective and makes all data available to the public. Compared to the ELF and SGC, the EU-OPENSSCREEN approach is complementary to these initiatives. How does this model work?

Open access

EU-OPENSSCREEN is a not-for-profit European Research Infrastructure Consortium (ERIC) for chemical biology and early drug discovery. The ERIC was established in 2018 by seven founding countries and is in 2022 supported by ten European member countries (CZ, FI, DE, LV, NO, PL, ES, DK, PT, SE, Figure 1). More than 30 academic institutions joined forces to accelerate the development of small-molecule chemical probes and lead compounds by providing access to expert facilities and resources with a special focus on high-throughput screening, small-molecule libraries, medicinal chemistry and open research data. Through this approach, scientists have the opportunity to progress their chemical biology projects from hit identification to hit-to-lead (H2L) and lead optimisation phases, using technology platforms and expertise which are usually available only in big pharma platforms or prestigious institutes.⁴ Any scientist in the world can apply for an EU-OPENSSCREEN collaboration via the central office, based in Berlin, Germany, which acts as a single point of contact for external researchers and coordinates the collaboration between them and the EU-OPENSSCREEN partners. These partner institutions were selected based on to their strong track-record in small molecule screening and medicinal chemistry research projects in diverse disease areas.⁵ They support comprehensively the needs of academic and/or industrial scientists ensuring the quality and efficiency of the research and services provided. The first screening campaigns started in 2019 and were funded by the EC through the H2020 [EU-OPENSSCREEN-DRIVE](#) project.

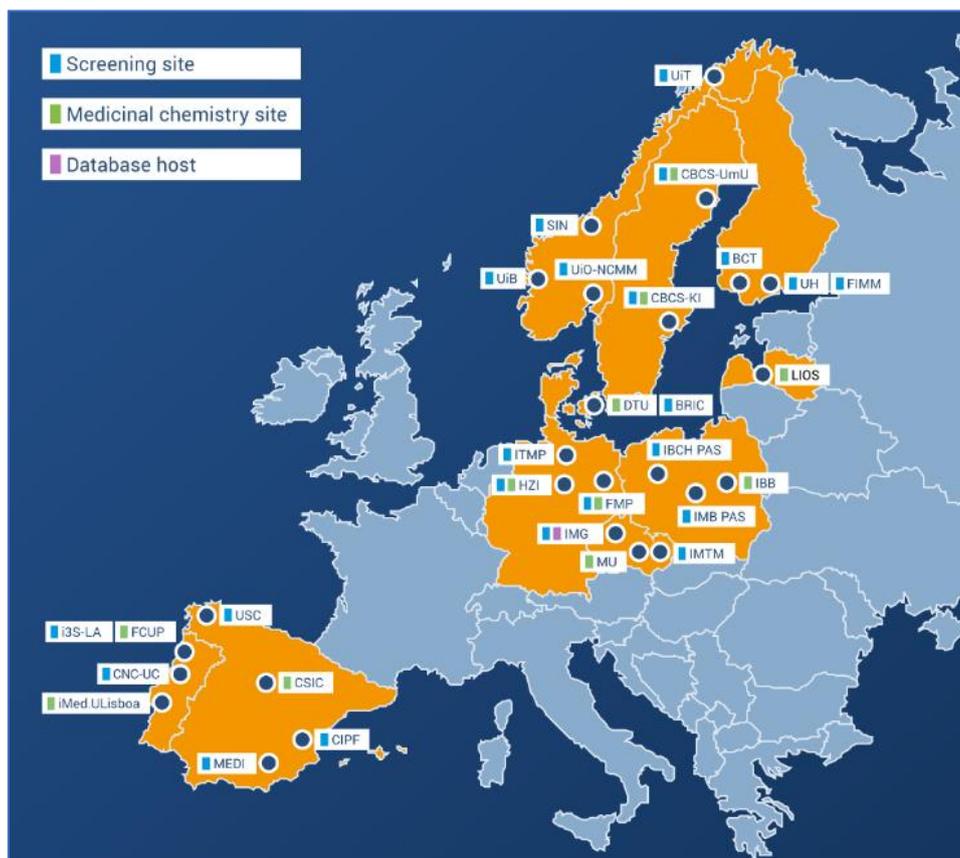


Figure 1. EU-OPENSREEN ERIC consortium for chemical biology and early drug discovery.

Open compound collections

EU-OPENSREEN provides a collection of several small-molecule libraries to be used in high throughput screening projects. The European Chemical Biology Library (ECBL) consists of 100,000 compounds with high chemical diversity. Its main set is composed of approx. 96,000 compounds from commercial vendors, while the pilot collection consists of approx. 5,000 molecules: half of these are bioactives with known biological targets while the other half are representative compounds from the larger ECBL screening set. The ECBL was designed by five academic computational chemistry groups in member countries of EU-OPENSREEN that carefully selected the compounds with an emphasis on chemical stability, absence of reactive compounds, screening-compliant physico-chemical properties, and maximal diversity/coverage of chemical space.⁶ To enlarge our library further and maximise its diversity, EU-OPENSREEN has opened its compound collection to submissions of small molecules from medicinal chemists worldwide to create the European Academic Compound Library (EACL). Collected molecules are registered, quality controlled and reformatted into screening plates at the Central Compound Management Facility in Berlin (Figure 2). The plates are then available for screening campaigns by biologists who access the infrastructure with novel assays and biological targets. Whenever there is a hit identified, chemists are informed and get the opportunity to start a collaboration with the corresponding biologist. In this process EU-OPENSREEN acts only as facilitator and is not taking any IP or fees for this role. In contrast, while biologists have the chance to screen new chemical space, the chemists may find unprecedented activity in their compounds and remain owners of their compounds. Within EU-OPENSREEN, we believe this can lead to exciting new collaborations with high innovative potential.

The ECBL in combination with the capabilities of EU-OPENSREEN partner sites offers scientists the opportunity to transition a project from basic research in chemical biology towards applications in drug discovery or agriculture. EU-OPENSREEN's ambition is to establish the ECBL as a future reference point for small molecule screening campaigns and compile comprehensive datasets of biological and physicochemical data of the compounds. Usually, such screening data sets remain confidential, and other scientist cannot use them to learn or study the results. Therefore, the ECBL is being profiled using diverse assays and biological target classes, and all data are being made available to everyone (with an embargo period of up to three years to allow for patent filing or publications) to increase the re-use of the data. In addition to biological target-specific screening campaigns, in 2021 we started to profile the full ECBL with an activity called "bioprofiling" to increase the knowledge of compound properties through screenings against a panel of predefined assays to



Figure 2. Compound storage at the Central Compound Management Facility at EU-OPENSREEN ERIC in Berlin, Germany.

assess aqueous solubility, interference with common bioluminescent reporter systems, effects on cell viability, antifungal and antibiotic activity, and performance in 'cell-painting' assays.⁷ Altogether, these experiments will generate a comprehensive map of (bio)activities and properties, but also off target effects intrinsic to the ECBL compounds and will be a valuable resource for chemists and biologists who want to develop chemical probes and drugs in the future.

Open and FAIR data

One of the founding principles of EU-OPENSREEN is the commitment to provide open access and FAIR (findable, accessible, interoperable and reproducible) scientific data. To this end, the European Chemical Biology Database ([ECBD](#)) was developed as a readily accessible and user-friendly online portal (Figure 3). The ECBD is the central data sharing environment for all data generated within the EU-OPENSREEN consortium. It was created and is maintained by our partner site Institute of Molecular Genetics (IMG) in the Czech Republic under the leadership of Petr Bartůněk who is also the director of [CZ-OPENSREEN](#). Besides the ECBD, this team also developed and maintains the [Probes & Drugs](#) portal, a powerful tool for the exploration of bio-active compound space.⁸ The team at the IMG works together with CESNET which is the e-infrastructure for the research and development of advanced network technologies and applications. CESNET provides the cloud-based hosting, backup and security for the ECBD.

Scientists can fully access uploaded data without specific software at the ECBD website. Here, primary screening data, biological activities and other properties measured using the EU-OPENSREEN chemical libraries can be found. The ECBD is separated into three distinct but interconnected ways of browsing; each is focused on different aspects of the database: compounds, assays and targets. While the database contains about 100,000 compounds and is divided into several chemical libraries, users can easily browse a subset or individual compounds that contain a certain (sub)structure or properties. From a single compound view, it can be switched to the assay view to find all assays in which a given compound was screened and identified as active or inactive. Therefore, the ECBD does not only show molecules that were identified as hits in a screen but also 'negative' data of inactive compounds.

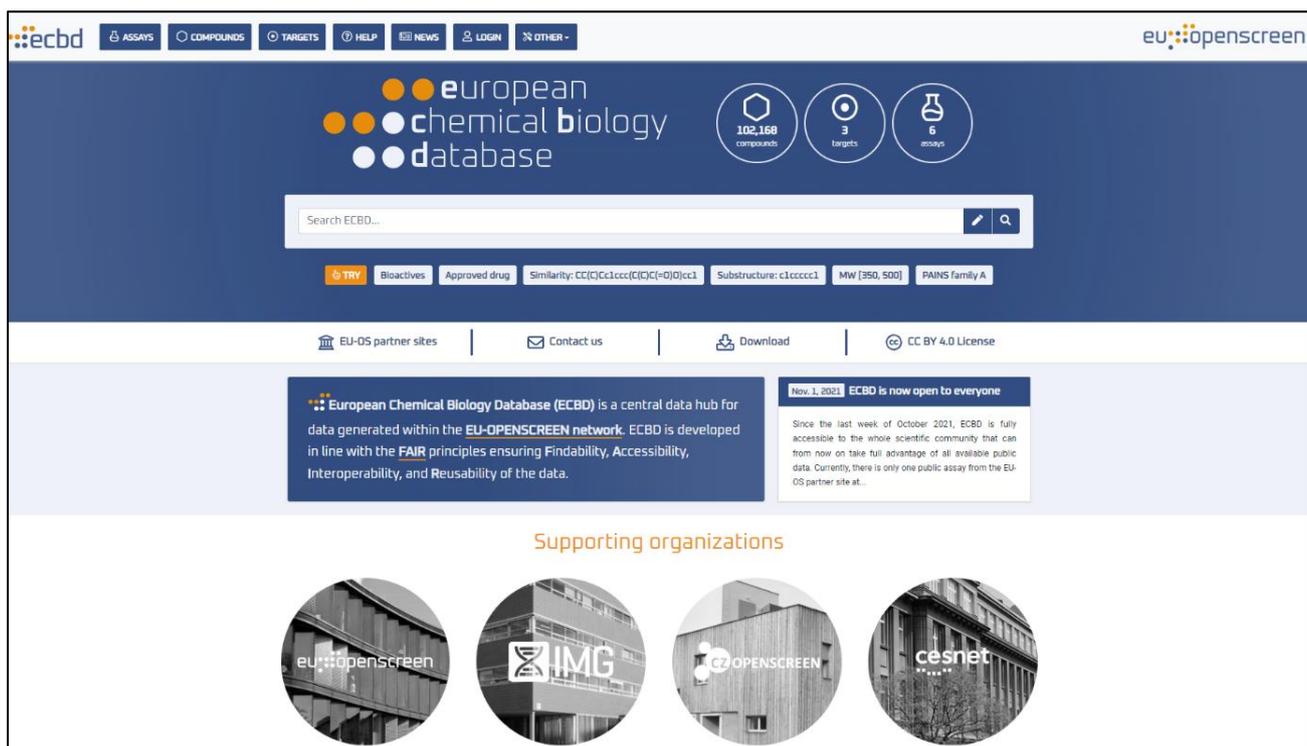


Figure 3. European Chemical Biology Database for chemical biology data.

Interoperability of the ECBD is achieved by annotation and characterisation of all uploaded data by ontology terms, e.g. organisms that are described by the NCBI taxonomy ontology or assays described by the BioAssayOntology (BAO).⁹ To maximise the use of data, users do not need to visit ECBD only via a web browser but alternatively also by an API, prepared for programmatic processing, or data can be downloaded as a database dump to use locally. While EU-OPENSREEN promotes open access and FAIR data policies, EU-OPENSREEN ensures the IP interests of their collaborators by offering an optional embargo period on primary screening data for up to three years, as described above, thereby allowing sufficient time for the characterisation of promising compounds, scientific publications, patent filing and the generation of innovation with commercial partners. After an initial testing period in the beginning of 2021, the ECBD is now fully available to the public, and the first screening and bioprofiling data sets have been uploaded into the portal. The majority of these data are not visible yet and will become available after the expiration of the requested embargo period which will be 2025 for current screening data, and 2023 for bioprofiling data.

Over the last decades the impact of computational approaches in drug discovery, chemical biology and structural biology has tremendously increased. Through the open science and FAIR data policy, EU-OPENSREEN strives to support also the community of computational and data scientists. Researchers will have in fact the option to mine the ECBD data sets for new insights or develop more reliable predictions that have great potential to accelerate research. One example for which we are currently illustrating this possibility is our first EU-OPENSREEN/SLAS (Society for Automation and Screening) Joint Challenge on predicting compound solubility which runs on the [Kaggle](#) platform until the end of 2022 (Figure 4).

In this machine learning competition, we used unpublished aqueous solubility data from the bioprofiling initiative produced at the BioFarma Research group, led by Mabel Loza at the University of Santiago de Compostela, and distributed the ECBL into three categories (highly, medium, low solubility) based on the

measured nephelometry results. 70% of the data were made available to the public while we asked the participants to classify the residual 30% of compounds. The setup and the results of the competition, including discussions on the rational design of the prediction and codes used, will be published in 2023 in the SLAS journal *SLAS Discovery* and presented at the SLAS Europe 2023 in Brussels. Large, high quality and published screening data sets are still rarely accessible and hamper the advancement of computational science. With our data sets we hope to tackle parts of this issue by organising such competitions regularly. This approach may help to benchmark current prediction strategies and encourage the development of new ways to assess small molecules properties or biological activities.



Figure 4. 1st EU-OS/SLAS joint challenge on compound solubility.

Final remarks

EU-OPENSSCREEN ERIC is an open access and open data research infrastructure that offers scientists experimental support for the development of chemical probes and drug candidate molecules by democratising the access to highly equipped academic research facilities. These services and expertise are available to any scientist around the world, from both academia and industry. EU-OPENSSCREEN's compound collections are open to use and open for contributions by chemists. In addition, we offer open data that are freely available to everyone to serve academic and industrial research in biology, chemistry and data science. We have only just started with the provision of screening data sets to the public but in the future, this has great potential to become a unique and valuable asset driving science and innovation. In 2022, there are about 550 chemical probes covering about 450 targets in the public domain, and we are still far from having high-quality tool compounds for the majority of human proteins.¹⁰ Therefore, organisations like EU-OPENSSCREEN or the SGC will further support open science towards the development of chemical probes, eventually leading to a better understanding of biology and to the development of new therapeutic treatments.

Acknowledgments

We acknowledge all consortium partners and our users for their collaboration and feedback. EU-OPENSSCREEN receives funding from the European Union's H2020 research and innovation programme under grant agreement No. 823893 (EU-OPENSSCREEN-DRIVE).

References

- (1) Williamson, A. R. Creating a structural genomics consortium. *Nature Structural Biology*. 2000, 7(11), 953–953. <https://doi.org/10.1038/80726>; Morgan Jones, M.; Chataway, J. The Structural Genomics Consortium: successful organisational technology experiment or new institutional infrastructure for health research? *Technology Analysis & Strategic Management*. 2021, 33(3), 296–306. <https://doi.org/10.1080/09537325.2021.1882673>
- (2) Mullard, A. European lead factory opens for business. *Nature Reviews Drug Discovery*. 2013, 12(3), 173–175. <https://doi.org/10.1038/NRD3956>; Karawajczyk, A. et al. The European Lead Factory: A blueprint for public-private partnerships in early drug discovery. *Frontiers in Medicine*, 2016, 3(Jan.), 75. <https://doi.org/10.3389/fmed.2016.00075>
- (3) Frank, R. EU-OPENSSCREEN - A European infrastructure of open screening platforms for chemical biology. *ACS Chemical Biology*. 2014, 9(4), 853–854. <https://doi.org/10.1021/cb500189k>
- (4) Stechmann, B.; Fecke, W. Accelerating chemical tool discovery by academic collaborative models. *Cheminformatics and Its Applications*. 2020. <https://doi.org/10.5772/INTECHOPEN.91138>

- (5) Brennecke, P. et al. EU-OPENSSCREEN: A novel collaborative approach to facilitate chemical biology. *SLAS DISCOVERY: Advancing Life Sciences R&D*. **2019**, 24(3), 398–413. <https://doi.org/10.1177/2472555218816276>
- (6) Horvath, D. et al. (2014). Design of a general-purpose European compound screening library for EU-OPENSSCREEN. *ChemMedChem*. **2014**, 9(10), 2309–2326. <https://doi.org/10.1002/cmdc.201402126>
- (7) Bray, M. A. et al. (2016). Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature Protocols*. **2016**, 11(9), 1757–1774. <https://doi.org/10.1038/nprot.2016.105>
- (8) Skuta, C. et al. Probes & Drugs portal: an interactive, open data resource for chemical biology. *Nature Methods*. **2017**, 14(8), 759–760. <https://doi.org/10.1038/nmeth.4365>
- (9) Visser, U. et al. BioAssay Ontology (BAO): a semantic description of bioassays and high-throughput screening results. *BMC Bioinformatics*. **2011**, 12, 257. <https://doi.org/10.1186/1471-2105-12-257>
- (10) Škuta, C. et al. Will the chemical probes please stand up? *RSC Medicinal Chemistry*. **2021**, 12(8), 1428–1441. <https://doi.org/10.1039/D1MD00138H>; Antolin, A. A. et al. The Chemical Probes Portal: an expert review-based public resource to empower chemical probe assessment, selection and use. *Nucleic Acids Research*. **2022**, 1. <https://doi.org/10.1093/NAR/GKAC909>

DECIMER – An Open Toolkit for Optical Chemical Structure Recognition and Document Analysis

Contribution from Henning Otto Brinkhaus, email: otto.brinkhaus@uni-jena.de, and Kohulan Rajan, email: kohulan.rajan@uni-jena.de. Institute for Inorganic and Analytical Chemistry, Friedrich-Schiller-University Jena, Lessingstr. 8, 07743 Jena, Germany

Introduction

Information about chemical compounds is typically shared in text and image formats. The textual information can, for example, comprise various types of analytical information, chemical names, biological activities, information about the origin of the substance or synthesis instructions. Chemical structures are normally depicted as 2D structure diagrams (Figure 1). In the era of data-driven applications, it is highly problematic that most chemical information is exclusively published in unstructured, human-readable data formats and not in structured, machine-readable formats.

The field of Optical Chemical Structure Recognition (OCSR) deals with the translation of images of chemical structures to machine-readable representations. This is a key step in the extraction of chemical knowledge from the printed literature. For the public domain, there is an immense interest in this, as publicly funded research results are mostly published in human-readable, unstructured data formats (text and images) that are hidden behind paywalls of scientific publishers. Making chemical knowledge available in open databases in structured data formats makes it accessible and usable for researchers. In the private domain, OCSR methods are used excessively for the automated mining of information from patents.

Since the publication of the first complete OCSR system *Kekulé* in 1992, we have seen three decades of active development. Up until recently, algorithmic rule-based OCSR applications dominated the field. These tools typically apply a vectorisation algorithm to the binarised structure depiction. The nodes and edges of the vectorised images are then used to reassemble the scaffold of the chemical structure while text labels (e.g. 'CH₃') are resolved using optical character recognition (OCR). We would like to emphasise the importance of the open-source projects *OSRA*, *Imago* and *Molvec*. Until *OSRA* was published in 2009, there was no openly available OCSR tool. Unfortunately, the available rule-based systems are not very robust. Recently, Clévert et al. have shown that the introduction of slight shearing and rotation to images from OCSR benchmark datasets leads to a catastrophic failure of the above-mentioned tools.

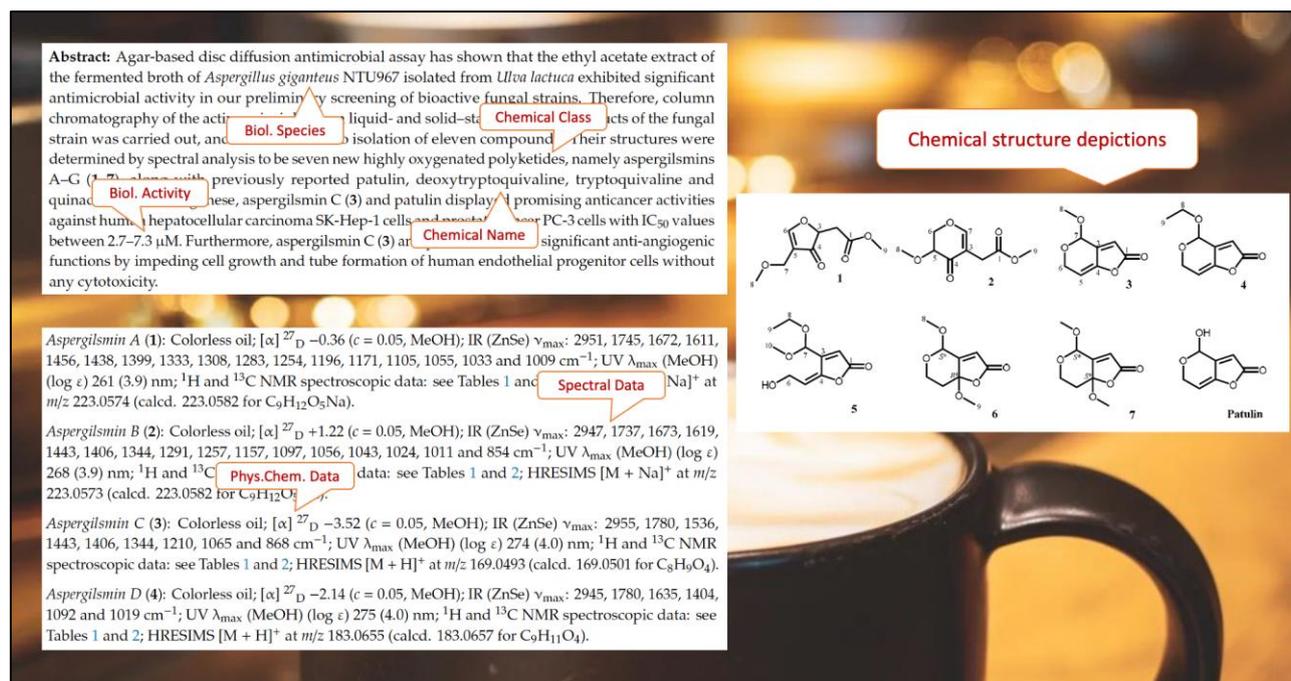


Figure 1. A collection of information about small molecules found in printed literature, including textual information and chemical structure representations. (Background image © Kohulan Rajan, figures and text from Chen et al. 2020.)

In the past years, there has been a shift towards deep-learning-based OCSR systems. These systems use deep-learning architectures to extract features from the given images. Here, there are two categories of tools: The first class of tools segments and identifies elements like bonds and atoms in the image using deep learning and then applies a set of rules to assemble the molecular graph based on them (e.g. *ChemGrapher*). The second class is fully based on deep learning and uses the latent representation of extracted image features to generate sequences of tokens of string representations of the depicted molecules (e.g. *DECIMER Image Transformer*). There are three open-source solutions (that all belong to the second category) – *Img2Mol*, *SwinOCSR* and our tool *DECIMER Image Transformer*.

DECIMER - an open toolkit for optical chemical structure recognition

At Friedrich Schiller University, Jena, Prof. Dr Christoph Steinbeck's lab initiated the Deep Learning for Chemical Image Recognition (DECIMER) project in 2017 in collaboration with Prof. Dr Achim Zielesny of Westphalian University of Applied Sciences to create the first open-source deep learning-based OCSR system.

During the past five years, the DECIMER project has seen significant developments that have led to the development of multiple components (Figure 2). These components are *DECIMER Segmentation*, *DECIMER Image Transformer*, *DECIMER Image Classifier* and the *SMILES-to-IUPAC Translator (STOUT)*. *DECIMER Segmentation* is an application for the segmentation of structure depictions from the scientific literature. Given a PDF document or images of scanned pages, it returns images of segmented structure depictions. *DECIMER Image Transformer* is capable of translating these chemical structure images into SMILES representations of the depicted molecules. Apart from delivering state-of-the-art results on common benchmark sets, the latest version of the *DECIMER Image Transformer* is capable of integrating R-group variables in the SMILES output when a depiction of a Markush structure is processed. *DECIMER Image Classifier* is a classification model that can distinguish between chemical structure depictions and non-chemical content. *STOUT* uses a transformer to translate SMILES strings to IUPAC names and vice versa.



Figure 2: Graphical overview of the DECIMER project.

A web platform incorporating all these components is now openly available under the domain decimer.ai (Figure 3). A user can simply upload a scientific publication or a patent as a PDF document, and chemical structures are automatically segmented and translated into SMILES representations. The recognised molecules are then depicted in an embedded molecular editor window (*Ketcher*). If an image that does not contain a chemical structure is uploaded, our *DECIMER Image Classifier* system is capable of recognising that. If desired, our application SMILES to IUPAC Translator (*STOUT*) can be used to generate the IUPAC name of the recognised structure.

How to use the DECIMER web app?

Just upload a pdf document or one or multiple images that contain chemical structure depictions above. If a pdf document is uploaded, DECIMER Segmentation is used to detect and segment all chemical structure depictions. The detected or uploaded chemical structure depictions are processed using the powerful DCSR engine of DECIMER V2. The chemical structure depictions and the corresponding SMILES representation are presented above. You can edit the structures according to your needs in the [online chemical structure editor](#) windows before downloading the segmented images and the corresponding mol files. Additionally, the IUPAC names of the chemical structures can be realized using STOUT V2.

Cite us

If our toolkit helped your work, please cite our publications.

- DECIMER: towards deep learning for chemical image recognition
Rajan, K., Zalesny, A., Steinhilber, C. *J Cheminform.* **12**, 85 (2020).
- DECIMER-Segmentation: Automated extraction of chemical structure depictions from scientific literature
Rajan, K., Brinkhaus, H.C., Sorokina, M. et al. *J Cheminform.* **13**, 20 (2021).
- DECIMER 1.0: deep learning for chemical image recognition using transformers
Rajan, K., Zalesny, A., Steinhilber, C. *J Cheminform.* **13**, 81 (2021).
- STOUT: SMILES to IUPAC names using neural machine translation
Rajan, K., Zalesny, A., Steinhilber, C. *J Cheminform.* **13**, 34 (2021).

b. The image of a hand-drawn chemical structure is resolved and displayed

c. An image of a depiction of a chemical structure, segmented from printed literature, is resolved and displayed

a. decimer.ai web interface

Figure 3: The web interface of <https://decimer.ai> (a) enables the easy detection and recognition of chemical structure depictions in the literature (b, c).

The source code of all components and the web application itself is openly available and published under permissive licences. DECIMER Segmentation, DECIMER Image Classifier, DECIMER Image Transformer and STOUT are all available on GitHub and can be installed as Python packages. If desired, the user can launch a local copy of the web application using Docker.

The manual extraction of information from printed literature is a time-consuming and error-prone process, but it is possible to automate almost all the components of this task using deep learning. This reduces the amount of manual labour enormously. We initiated the DECIMER project with the intention of creating an open platform for the extraction of chemical information from the literature. All components that are developed as a part of DECIMER are continuously developed, and we hope to contribute to making more chemical knowledge available in structured data formats in the future.

References and additional resources

- Chen, J.-J. et al. Highly oxygenated constituents from a marine alga-derived fungus *aspergillus giganteus* NTU.967. *Mar. Drugs*. **2020**, *18*(6), 303.
- Clevert, D.-A. et al. Img2Mol - accurate SMILES recognition from molecular graphical depictions. *Chem. Sci.* **2021**. <https://doi.org/10.1039/D1SC01839F>
- Filippov, I.V.; Nicklaus, M.C. Optical structure recognition software to recover chemical information: OSRA, an open source solution. *J. Chem. Inf. Model.* **2009**, *49*, 740–743.
- McDaniel, J.R.; Balmuth, J.R. Kekulé: OCR-optical chemical (structure) recognition. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 373–378.
- Oldenhof, M. et al. ChemGrapher: optical graph recognition of chemical compounds by deep learning. *J. Chem. Inf. Model.* **2020**, *60*, 4506–4517.
- Peryea, T et al. (2019) MOLVEC: Open source library for chemical structure recognition. Abstracts of Papers of the American Chemical Society 258; <https://github.com/ncats/molvec>
- Rajan, K. et al. A review of optical chemical structure recognition tools. *J. Cheminformatics.* **2020**, *12*, 60.
- Rajan, K. et al. DECIMER 1.0: deep learning for chemical image recognition using transformers. *J. Cheminform.* **2021**, *13*, 61.
- Rajan K. et al. DECIMER-Segmentation: Automated extraction of chemical structure depictions from scientific literature. *J. Cheminform.* **2021**, *13*, 20.
- Rajan, K. et al. DECIMER: towards deep learning for chemical image recognition. *J. Cheminformatics.* **2020**, *12*, 65.
- Rajan, K. et al. STOUT: SMILES to IUPAC names using neural machine translation. *J. Cheminformatics.* **2021**, *13*, 34.
- Smolov, V. et al. Imago: open-source toolkit for 2D chemical structure image recognition. **2011**. TREC. <https://trec.nist.gov/pubs/trec20/papers/GGA.chemical.pdf>
- Xu, Z. et al. SwinOCSR: end-to-end optical chemical structure recognition using a Swin Transformer. *J. Cheminform.* **2022**, *14*, 41.

Cryo-EM for Industrial-Scale Structure-Based Drug Design

Contribution from Christopher Earl, email: chris.earl@astx.com, and Joanna Brown, email: joanna.brown@astx.com.
Astex Pharmaceuticals, 436 Cambridge Science Park, Cambridge, CB4 0QA, UK

Introduction

Structure-based drug design (SBDD) is the process of progressive rational design and optimisation of a drug compound based on the 3D structure of the compound in complex with its target protein. X-ray crystallography has been the technique of choice for SBDD because of its almost unique ability to reveal the detailed 3D structures of a range of target-ligand complexes at the atomic level. As such, technical development of X-ray crystallography has seen huge investment – resolution, throughput, reproducibility, and automation have all vastly improved over the last 30 years. Powerful though X-ray crystallography is, it relies on crystallisation of a purified protein sample. This limitation means that many important drug targets which are difficult or impossible to crystallise have been structurally inaccessible.

Enter the “Resolution Revolution” of electron cryo-microscopy (cryo-EM).¹ Cryo-EM is the well-established process of freezing a specimen in a thin layer of amorphous ice and imaging in a transmission electron microscope;² no crystal is required (Figure 1). However, it was not until 2012 that transformative developments in microscope hardware, detector technology, and data processing algorithms made near-atomic resolution structure determination by cryo-EM possible. The impact of this on the structural biology field was so profound that the 2017 Nobel prize for chemistry was subsequently awarded to Richard Henderson, Jacques Dubochet, and Joachim Frank for their development of the technique. Now, cryo-EM is being rapidly adopted by many major players in the biopharma space allowing them to access new target classes for a broad range of human diseases.

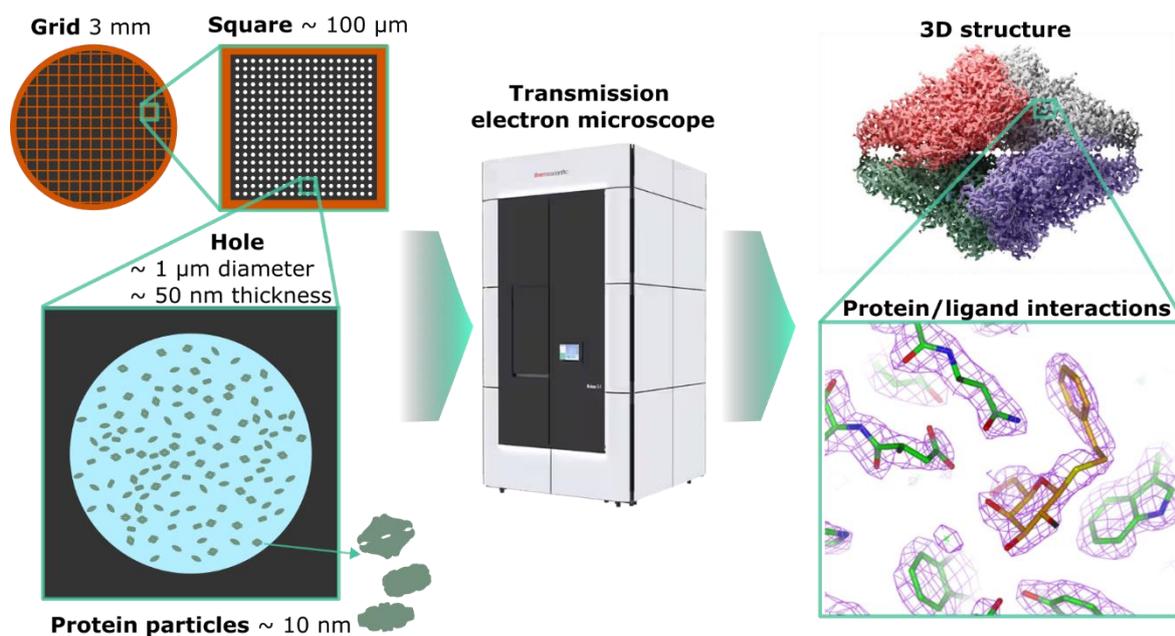


Figure 1. Overview of protein/ligand complex cryo-EM structure determination. The sample support is a metal grid, supporting a thin film of carbon or gold perforated with holes. A thin layer of amorphous ice containing protein/ligand complex particles spans each hole. Images are collected in a transmission electron microscope and processed to compute a 3D map of the protein/ligand complex. This map is used to build an atomic model revealing interactions between protein and ligand [adapted from reference (3)].

Reproducibility and automation

Inside a cavernous, purpose built, environmentally stable room sits a 3m high, multimillion-pound electron microscope surrounded by peripheral equipment and supported by an extensive computing infrastructure. Inside this behemoth sits a 3mm metal grid supporting a foil layer around 50-100nm thick punctured by thousands of holes around 1μm in diameter. Inside each of those holes is a very thin layer of vitreous ice containing the specimen.

Producing this frozen specimen involves several steps where the delicate grid is manually handled with tweezers. To create a thin enough layer of liquid, the sample is manually pipetted onto the grid and the vast majority of it is blotted away by filter paper before the grid is rapidly frozen by plunging into a well of liquid ethane. Though the “blot and plunge” part of the process is now automated, the fundamental technique used to prepare most grids for cryo-EM has not significantly changed since the 1980s. The process introduces a significant amount of grid-to-grid variability, even in the hands of an experienced practitioner. Often, multiple repeats of the same sample must be loaded into the microscope and screened to find a suitable grid for data collection, hampering throughput.

There are several new instruments that employ different methodologies for the application of samples to grids,^{4,5,6} but none of them have yet been shown to out-perform “blot and plunge” technology in the hands of an experienced user. There is light at the end of the tunnel though, as improvements in the grids themselves have the potential to increase sample reproducibility, allow faster image acquisition, and increase the quality of the data obtained.^{7,8}

The variability of sample preparation and the visual nature of cryo-EM data sample assessment mean full imaging automation is not yet possible. Current data collection software packages are capable of queuing multiple data collections⁹ but still require significant user decision making. Data processing, on the other hand, is now highly automatable for images of an optimised sample. We have implemented the open-source software package Relion into our own, bespoke in-house processing pipeline for on-the-fly data processing with minimal user input.^{3,10}

Increasing throughput

Biological specimens are extremely radiation sensitive, so using a low electron dose is essential when imaging them. Consequently, individual cryo-EM images contain very little signal. Very large datasets in the form of thousands of individual images are therefore required; by summing large numbers of noisy, low signal images, we boost the overall signal to produce a clearer 3D map of the protein target. Improvements in detectors and implementation of new data collection strategies mean that in the last few years we have gone from a single data collection taking days or even weeks to now only taking hours.

Modern cryo-EM image processing packages have become increasingly fast and efficient and advances in the algorithms now allow higher resolution information to be reconstructed.^{10,11} All of this means we can now produce high-resolution maps from large amounts of data faster and more efficiently than ever before. We can also leverage these improvements to access ever more challenging targets.

Currently, all electron microscopes capable of protein structure determination at resolutions useful for SBDD cost millions of pounds and have significant setup and maintenance costs. Despite that, the potential of cryo-EM enabling SBDD on important disease targets has led to many pharmaceutical companies investing heavily in in-house cryo-EM facilities and/or access to external facilities.

Cryo-EM has been shown to be useful both in identifying hit matter and determining the high-resolution structures of protein/ligand complexes. For the cancer target PKM2, we screened a sub-set of preliminary fragment library compounds by cryo-EM and identified several compounds bound to the target.³ Although near-atomic resolution structures usually require large datasets from the highest end microscopes, a hybrid screening/structure determination workflow offers a route to reasonable throughput for SBDD. More modest resolution structures can be determined with smaller datasets on slightly cheaper, more accessible screening microscopes. Then, promising samples can be progressed for near-atomic resolution structure determination on the most expensive, large and highest-end microscopes. Such a workflow was recently demonstrated by the Greber lab at the Institute of Cancer Research in London. For the cancer target CDK-activating kinase, structures at moderate resolutions sufficient to determine the presence or absence of a ligand were obtained with short data collections using a screening microscope. Higher-resolution structures were then separately determined to identify the detailed inter-molecular interactions between drug compounds and the protein target.¹²

Future perspectives

The cryo-EM field has enjoyed rapid growth in the last decade. We expect that trend to continue, with increasing throughput and automation of high-resolution structure determination at the forefront. Perhaps most importantly, new grid and grid preparation technologies are expected to improve sample reproducibility.

Machine learning-based approaches for high quality automated data collection are also currently under development. These advances are opening the door to faster and more reproducible data acquisition with minimised need for expert user input. We anticipate that structure-based drug design supported by high throughput cryo-EM structure determination will become routine in the coming years.

References

- (1) Callaway, E. The revolution will not be crystallized: a new method sweeps through structural biology. *Nature*. **2015**, 525, 172–174. <https://doi.org/10.1038/525172a>
- (2) Raimondi, V.; Grinzato, A. A basic introduction to single particles cryo-electron microscopy. *AIMS Biophysics*. **2022**, 9(1), 5-20. <https://doi.org/10.3934/biophy.2022002>
- (3) Saur, M. et al. Fragment-based drug discovery using cryo-EM, *Drug Discovery Today*. **2020**, 25(3), 485-490. <https://doi.org/10.1016/j.drudis.2019.12.006>
- (4) Klebl, D.P. et al. Towards sub-millisecond cryo-EM grid preparation. *Faraday Discuss.* **2022**, 240, 33-43. <https://doi.org/10.1039/d2fd00079B>
- (5) Rima, L. et al. cryoWriter: a blotting free cryo-EM preparation system with a climate jet ad cover-slip injector. *Faraday Discuss.* **2022**, 240, 55-66. <https://doi.org/10.1039/d2fd00066k>
- (6) Al-Otaibi, N. et al. Sample preparation in single particle cryo-EM: general discussion. *Faraday Discuss.* **2022**, 240, 81-100. <https://doi.org/10.1039/d2fd90059A>
- (7) Naydenova, K; Russo, C.J. Integrated wafer-scale manufacturing of electron cryomicroscopy specimen supports. *Ultramicroscopy*. **2022**, 232, article 113396. <https://doi.org/10.1016/j.ultramic.2021.113396>
- (8) Naydenova, K. et al. Cryo-EM with sub-1 Å specimen movement. *Science*. **2020**, 370(6513), 223-226. <https://doi.org/10.1126/science.abb7927>
- (9) EPU Software with Multigrid (Thermo Fisher Scientific)
- (10) Kimanius, D. et al. New tools for automated cryo-EM single-particle analysis in RELION-4.0. *Biochem. J.* **2021**, 478(24), 4169–4185. <https://doi.org/10.1042/BCJ20210708>
- (11) Punjani, A. et al. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nature Methods*. **2017**, 14, 290–296. <https://doi.org/10.1038/nmeth.4169>
- (12) Greber, B.J. et al. 2.5 Å-resolution structure of human CDK-activating kinase bound to the clinical inhibitor ICEC0942. *Biophysical Journal*. **2021**, 120(4), 677-686. <https://doi.org/10.1016/j.bpj.2020.12.030>

Cryo-EM & Drug Discovery

Contribution from Oliver Acton, Fiona Shilliday and Taiana Maia de Oliveira, email: taiana.maiadeoliveira@astrazeneca.com. Mechanistic and Structural Biology, Discovery Sciences, R&D, AstraZeneca, Cambridge, UK

Pharmaceutical companies have been fast establishing cryo-EM capabilities within their R&D. These investments have been driven by the expansion of target space the technique could bring into the Structure Based Drug Design realm. The most transformative developments that propelled the method to generate well resolved structures routinely became widely available about a decade ago, and progress has been steady and fast. In this section, we provide a short review of the impact of cryo-EM in the drug discovery on previously intractable targets including membrane proteins like the TRP channels, G protein-coupled receptors (GPCRs), and multiprotein complexes like the ones mediating protein degradation.

Cryo-EM and membrane protein

Membrane proteins constitute the prime category of drug targets for medicines.¹ They have been particularly challenging for crystallography and are then underrepresented in the pool of x-ray structures. The last years have witnessed a quick expansion in the structural biology of membrane proteins, thanks to the application of cryo-EM.²

TRP channels

Cryo-EM has been a powerful tool in the structural understanding of the TRP ion channel family. Prior to the advances in cryo-EM methodology, only TRPV6 had been successfully crystallised and resolved to a resolution higher than 4Å.³ However, since 2013, several structures of full-length TRP channels have been resolved using single particle cryo-EM exploiting various membrane mimetic technologies to stabilise the transmembrane domains (Figure 1A). Furthermore, as cryo-EM traps samples in near native states, this has also allowed a full range of conformational states to be resolved for these channels revealing insights into their respective mechanisms of action.⁴ As a result, structures of full-length versions in multiple conformations are now available for TRP channels from all seven of the sub-families.

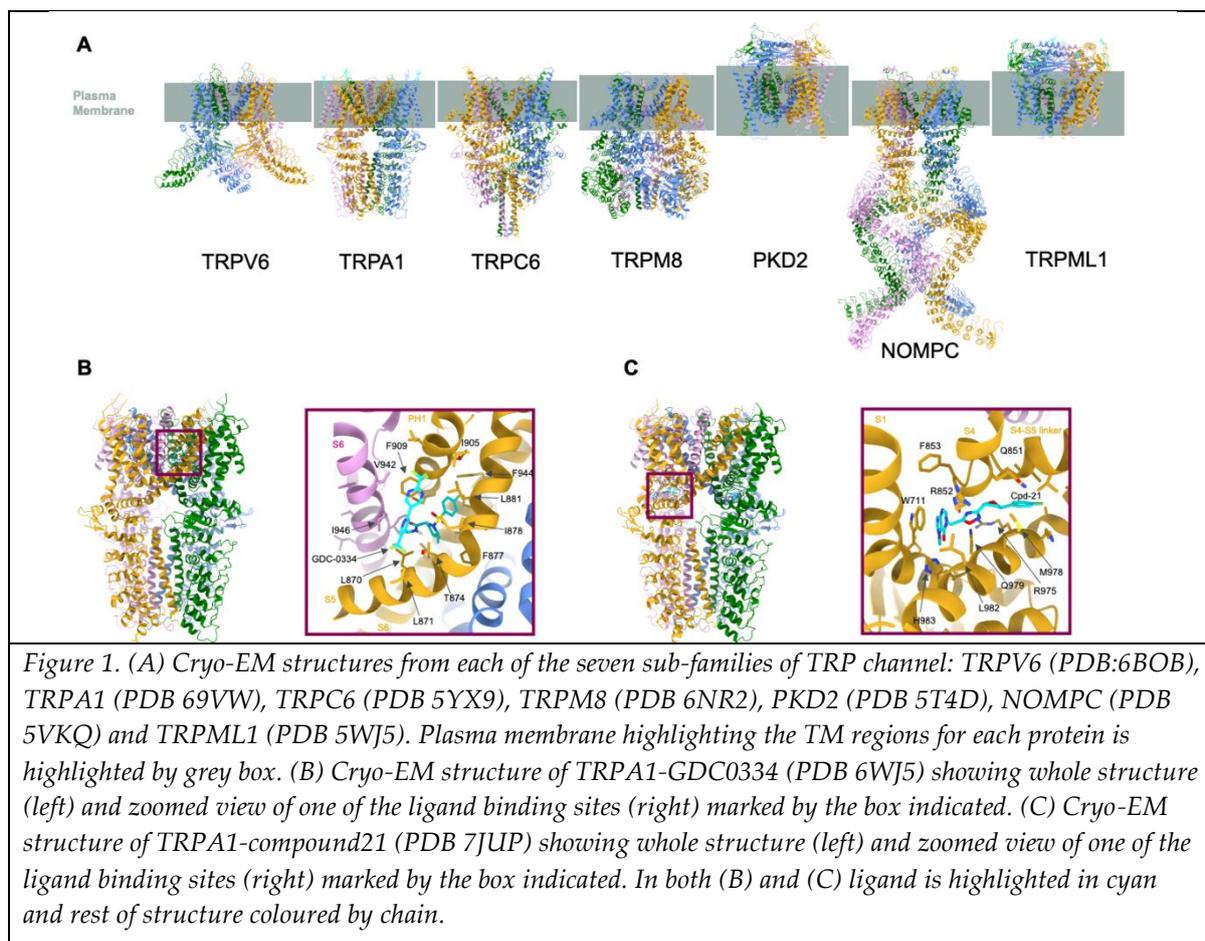
The advancements in structural understanding of TRP channels has now opened the possibility to exploit cryo-EM based drug discovery. Perhaps one of the best examples of this has been the study of small-molecule antagonists against the nonselective cation channel TRPA1. TRPA1 has numerous exogenous and endogenous ligands including several proinflammatory mediators which are often elevated in asthmatics. As a result, antagonists targeting TRPA1 to reduce airway inflammation and chronic cough has proven to be a field of great interest. Cryo-EM structures of TRPA1 bound to the antagonists GDC-0334 (Figure 1B)⁵ and compound-21 (Figure 1C)⁶ proved to be major breakthroughs in the understanding of selective antagonism of TRPA1. The binding site for GDC-0334 demonstrated the strength of cryo-EM in drug discovery as the binding pocket was revealed to be entirely within the transmembrane helices of TRPA1. Specifically, it revealed multiple contact points between the molecule and TRPA1 (including the helices S5-S6, pore helix 1 and linker between helices S4-S5), locking TRPA1 into an inactive conformation characterised by a distinctive kink midway up the S5 helix which closes the transmembrane pore. By comparison, compound-21 was shown to bind a cytosolic, but membrane-proximal, pocket which included contacts to TM and cytosolic residues of TRPA1. Like GDC-0334, compound-21 also binds the S4-S5 linker and appears to use this contact as a mechanism for bending the S5 helix to close TRPA1. In both these cases, the ability to resolve binding membrane associated binding pockets inaccessible to other structural techniques has provided valuable insight into the mechanism of action of antagonists and provided a platform for their further development.

GPCRs

In the case of GPCRs, the method has been transformative especially in the understanding of agonism. Nearly all crystal structures captured nonactive GPCR conformations and the agonistic species are formed by the association of the GPCR with a G protein heterotrimer into a larger complex which are suitable for cryo-EM once stabilised. Several active-state GPCR-G protein structures have been reported.

Oral small molecule GLP1R agonists have been a holy grail within the diabetes area since the discovery of GLP-1. The EM structures of the glucagon-like peptide 1 receptor: G protein transducer (GLP1-G α s, β , γ 2) complex bound to small molecule agonists TT-OAD2² and Pf-06882961⁷ (Figure 2A) were milestones in the use of cryo-EM in drug discovery. These agonists binding sites and poses have few commonalities with the natural agonist GLP but they reorganise the class B GPCR conserved central polar network which is linked to activation in a similar way to the natural agonist. An important GPCR class B drug discovery learning from these structures is that non-peptidic ligands do not need to mimic the extensive contacts established by natural ligands to modulate receptor activity. Still using GLP1-R as an example, another important learning can be gathered: distinct pharmacological profiles sometimes can only be understood through the investigation of conformational landscape. The static consensus structures of GLP1R in complex with peptidic agonists taspoglutide and semaglutide were very similar and could not explain their divergent pharmacological footprints.⁸ Because cryo-EM samples are captured in near native state, a whole range of conformational states is preserved and can be identified within a data set. That allowed for 3D variance analysis which revealed

peptide-dependent dynamics. These may be critical for explaining pharmacological differences between the two peptides.⁸



Protein degradation

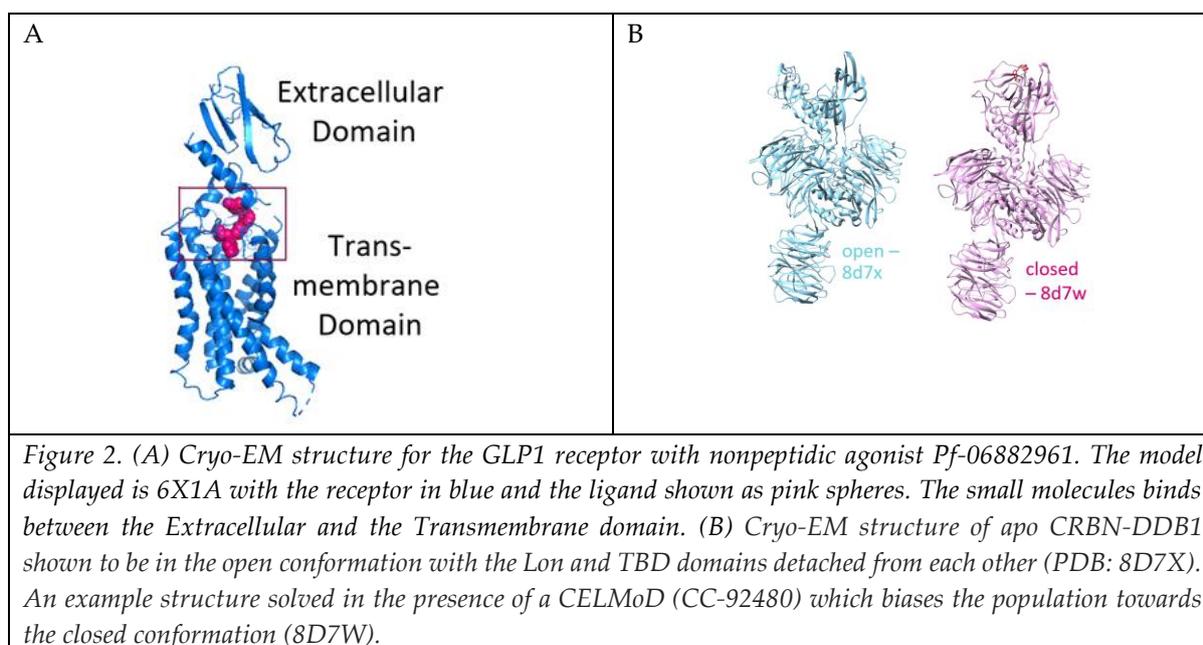
A protein can be targeted throughout its lifetime right from synthesis to degradation. Cryo-EM has been the technique of choice to study various large protein complexes involved in protein synthesis and degradation. Structures for ribosomes and proteasomes in both human and parasite species have been determined and used in development of selective anti-parasite drugs.

Another drug discovery strategy is in targeting specific proteins for degradation. Small molecule compounds can act as molecular glues, binding to substrate receptor proteins such as cereblon (CRBN), part of the CUL4 E3 ligase complex. This creates a novel binding surface shifting the recruitment towards non-native substrates.⁹

An example of this substrate recognition modulation occurs with the drug thalidomide. Although withdrawn from use in its initial indication due to birth defects in children, thalidomide does have a positive effect against leprosy and multiple myeloma. This is because it changes the substrate recognition of CRBN, towards transcription factors Ikaros and Aiolos.⁹ Structures solved by both X-ray crystallography and cryo-EM visually show this mechanism and could guide enhancement of binding affinity or variations that could be made to recruit other substrates.

A recent paper has shown that due to its tolerance of flexibility and heterogeneity, cryo-EM can provide important insights into the mechanism of molecular glues.¹⁰ It confirmed that cereblon exists in both an open

or closed conformation involving movements of the N-terminal Lon- protease-like domain and the thalidomide-binding domain. In its apo form CRBN adopts almost exclusively the open conformation but addition of different CRBN E3 Ligase Modulatory Drugs (CELMoD) bias it to varying extents towards the closed conformation. Addition of first-generation CELMoD pomalidomide shifts 20% of the population to the closed conformation and next-generation CELMoD iberdomide biases further to 50%. Although many CELMoDs were shown to bind both conformations, recruitment of neosubstrates is exclusively to the closed conformation. The enhanced efficacy of next-generation CELMoD mezigdomide likely arises from the fact that it biases the population 100% to the closed conformation, thus enhancing neosubstrate recruitment.



Conclusion

The examples discussed here demonstrate the influence of cryo-EM in discovery portfolios and underline what is achievable. The extent of cryo-EM-facilitated SBDD cannot yet be fully appreciated, though. Publications of drug design stories, irrespective of the method used, are generally slow to emerge. The speed with which insights can be delivered could be exemplified by the rapid determination of the SARS-CoV-2 target:ligand complex structures during the pandemic.

References

- (1) Gong, J. et al. Understanding membrane protein drug targets in computational perspective. *Current Drug Targets*. **2019**, 20(5), 551–64. <https://doi.org/10.2174/1389450120666181204164721>
- (2) Zhao, P. et al. Activation of the GLP-1 receptor by a non-peptidic agonist. *Nature*. **2020**, 577(7790), 432–36. <https://doi.org/10.1038/s41586-019-1902-z>
- (3) Singh, A.K. et al. Structural bases of TRP channel TRPV6 allosteric modulation by 2-APB. *Nature Communications*. **2018**, 9(1): 2465. <https://doi.org/10.1038/s41467-018-04828-y>
- (4) Morano, A. Cannabinoids in the treatment of epilepsy: current status and future prospects. *Neuropsychiatric Disease and Treatment*. **2020**, 16, 381–96. <https://doi.org/10.2147/NDT.S203782>
- (5) Balestrini, A. et al. A TRPA1 Inhibitor suppresses neurogenic inflammation and airway contraction for asthma treatment. *Journal of Experimental Medicine*. **2021**, 218(4), e20201637. <https://doi.org/10.1084/jem.20201637>
- (6) Terrett, J. et al. Tetrahydrofuran-based transient receptor potential ankyrin 1 (TRPA1) antagonists: ligand-based discovery, activity in a rodent asthma model, and mechanism-of-action via cryogenic electron microscopy. *Journal of Medicinal Chemistry*. **2021**, 64(7), 3843–69. <https://doi.org/10.1021/acs.jmedchem.0c02023>
- (7) Zhang, X. et al. Differential GLP-1R binding and activation by peptide and non-peptide agonists. *Molecular Cell*. **2020**, 80(3), 485-500. e7. <https://doi.org/10.1016/j.molcel.2020.09.020>

- (8) Zhang, X. et al. Structure and dynamics of semaglutide- and taspoglutide-bound GLP-1R-Gs complexes. *Cell Reports*. **2021**, 36(2), 109374. <https://doi.org/10.1016/j.celrep.2021.109374>
- (9) Collins, I. et al. Chemical approaches to targeted protein degradation through modulation of the ubiquitin-proteasome pathway. *The Biochemical Journal*. **2017**, 474(7), 1127–47. <https://doi.org/10.1042/BCJ20160762>
- (10) Watson, E. et al. Molecular glue CELMoD compounds are regulators of cereblon conformation. *Science*. **2022**, 378(6619), 549–53. <https://doi.org/10.1126/science.add7574>

2022 CSD Updates

Contribution from Pete Wood, Director of Product Management at CCDC, email:
wood@ccdc.cam.ac.uk



2022 was a busy year for the CSD data and software, with a host of new additions. Here I'll share some of the major updates from the year. Stay up to date with all our latest releases and sign up for email alerts on our website [here](#).

CSD data

In 2022 we added over 57,000 new structures to the CSD. As ever these are from structures associated with articles in the published literature, published directly as CSD Communications by crystallographers, patents, PhD theses, and other sources. Each one undergoes a series of automated and manual checks to make it more findable and accessible.

CSD software and web services

2022 saw three major updates to the web-based services, and three to the desktop CSD software.

In our web-based services:

- The 3D visualiser was updated to improve performance, especially with larger structures, and to allow hit highlighting in substructure searches.
- CSD-Theory Web, our platform to store and manage proprietary crystal structure prediction data, was updated to offer visual overview of structural similarity across structures.
- 3D searching was launched in WebCSD — you can now apply constraints to angle, centroid, plane, vector, or point on line features when sketching a search, to return more exact results.

In the desktop CSD software:

- CSD-Particle was launched, allowing you to assess the mechanical and chemical properties of crystalline particles, to find the root cause of bulk product issues like sticking, tabletability, or wettability.
- CSD Python API now supports Python 3.9.
- SMARTS updates to allow the use of recursive and dot-disconnect SMARTS.
- Usability improvements, performance improvements, and bug fixes across a range of functions.

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

News from ACS CINF

Contribution from Sue Cardinal, 2022 CINF Chair, email: scardinal@library.rochester.edu

ACS CINF members rejoiced with **Wendy Warr** as she received her 2020 **Herman Skolnik Award**, in Chicago, Illinois, USA on Tuesday 23 August 2022 after a long COVID-related delay. Judith Currano wrote a thorough report about the symposium in the Winter issue of the [Chemical Information Bulletin](#). Here is a list of the speakers and their talk titles in the order that they presented.

1. William L. Jorgensen, "Simulation and informatics for drug design"
2. Dusanka Janezic, "Cheminformatics in light of drug development"
3. Paul C. Hawkins, "Macrocycles just like Mother (Nature) makes"
4. Erin Davis, "Digital chemistry at scale: Accelerating drug discovery with democratized computational tools and project learning"
5. Jeremy G. Frey, "Bytes and molecules: Data, data, everywhere and not enough to model"
6. Bonnie Lawlor, "Use of blockchain technology in the scientific research workflow"
7. Carmen Nitsche, "From 'You can't do that' to 'Do it yourself' – a look back at the end user journey"
8. Valentina Eigner-Pitto, "Would you power a rocket with old frying oil? High-quality data and deep chemistry knowledge as key drivers for successful ML/AI projects"
9. Judith Currano, "Artificial intelligence and chemical information retrieval: Who's afraid of the big black box?"
10. Phil McHale, "Chemical structures: The long and winding road or back to the future?"

Most speakers delivered their talks in person to a hybrid audience. Jeremy Frey and Bonnie Lawlor joined the symposium online. The symposium was followed by an awards reception. We toasted Wendy and took selfies of ourselves to share with her for the [Flicker site](#). For the full report on Wendy Warr's symposium, please see the [Chemical Information Bulletin](#) (Winter 2022, Vol. 74, No. 4, pp 4-24).

Carmen Nitsche (CCDC) received the **Val Metanomski Meritorious Service Award** for 2022 at the Awards Reception at the Fall Chicago ACS National Meeting. Staff at the [CCDC interviewed Carmen](#) about the honour of winning the award.

The **ACS Spring 2023** meeting will be in-person in Indianapolis and hybrid on March 26-30, 2023. The CINF symposia schedule, technical program details, and social events will all be posted as we get closer to the meeting, but you can preview the [planned symposia](#).

2023 Skolnik Award – Dr Patrick Walters: The American Chemical Society Division of Chemical Information is pleased to announce that Dr Patrick Walters has been selected to receive the 2023 Herman Skolnik Award for his contributions to the fields of chemical information and cheminformatics applied to computer-aided drug discovery research. [More on his accomplishments](#). The prize consists of a \$3,000 honorarium and a plaque. Dr Walters will also be invited to organise an award symposium at the Fall ACS National Meeting in 2023.

Data Summit for the 2023 Fall Meeting: To align with the theme for San Francisco – *Harnessing the Power of Data*, CINF is planning a Data Summit. Please discuss your ideas with Michelle, Meghan, or Ye Li (yel@mit.edu) as soon as possible. We encourage collaborations with other divisions and external organisations. If additional time is needed to confirm such collaborations, please let us know.

If you have ideas for projects that CICAG and CINF can work on together, please reach out to us. Thank you for reading.

News from CAS

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

ajones2@acs-i.org



Every day CAS scientists collect and analyse published scientific literature from around the globe, building the highest quality and most up-to-date collection of scientific information in the world.

CAS is proud to cover advances in chemistry and related sciences, and at the heart of the CAS Content Collection is human intelligence.

CAS Insights™

A new open-access content hub at the intersection of science, technology, and innovation was launched this year. Offering R&D and business leaders actionable perspectives on the latest developments across science and technology, [CAS Insights](#) draws on the human-curated data collection and deep scientific expertise from CAS to highlight emerging trends, unseen connections, new applications, and future opportunities across disciplines.

CAS Insights features articles, analytical reports, infographics, webinars, videos, and peer-reviewed journal publications on topics including sustainability, biotechnology, drug discovery, materials science, consumer goods, synthetic chemistry, digital R&D, and more. CAS is providing this resource for the scientific community to enable innovation leaders from the boardroom to the bench to gain a clearer view of the landscape and identify opportunities ahead so they can get breakthrough solutions to market faster. Access CAS Insights [here](#).

Webinars and virtual events

CAS continues to offer virtual opportunities for our customers (and prospective customers) to engage, discuss trends, and learn more about growing features and functionality within our solutions. These sessions cover a variety of topics and areas. Our webinars are recorded and are available for users to view after the live event is complete. To see what sessions are coming up soon and to sign up, please visit our [events page](#).

CAS Future Leaders program

CAS Future Leaders, established in 2010, is recognised as the premier science leadership program by delivering a unique, high-quality training experience to elite early-career scientists in chemistry and related sciences. Each year, participants gain the connections, perspectives, and insights they need to successfully navigate career paths that encompass academic, commercial, and government sectors. The 2023 program is currently accepting applications from PhD students and postdoctoral scholars from around the world. Eligible applicants can learn more and submit their [application](#).

CAS Innovation Incubator™

CAS launched the CAS Innovation Incubator to assist early-stage scientific organisations accelerating new ventures, particularly those creating unique applications that could benefit future research and development. CAS will provide early-stage innovators with access to the CAS Content Collection™ and its exclusive technologies that reveal unseen connections among disparate data, as well as CAS expertise in mapping and

managing intellectual property in multiple scientific disciplines, and in some cases, financial support.

CAS SciFinder Discovery Platform

CAS SciFinder Discovery Platform, an enterprise-wide platform solution with workflow tools and capabilities designed to support multiple scientific research requirements was launched in 2021. The CAS SciFinder Discovery Platform includes CAS SciFinderⁿ, CAS Formulus, CAS Analytical Methods, as well as all the new enhancements to Retrosynthetic Planning, and our newest capabilities in Biosequences. Additional enhancements were made throughout 2022. To learn more about the CAS SciFinder Discovery Platform, please click [here](#).

CAS SciFinderⁿ

Expanding our core capabilities to meet the growing needs of the scientific community continued in earnest throughout 2022 within CAS SciFinderⁿ. Our continued focus on improving the research efficiency and effectiveness of our traditional users remains of utmost importance to our team of technology and scientific experts.

Among recent notable enhancements:

- Workflow enhancements to further improve the experience of biologists using our biosequences functionality
- Integration of GetFTR full-text links to give users direct access to the full text of entitled and open access articles
- Inclusion of ORCID iDs as a search option for improved author searching
- The introduction of the CAS Lexicon Query Builder to enable users to quickly build more thorough search queries
- Transferring results from CAS STNext® to CAS SciFinderⁿ via the history tab in REGISTRY, CAplus, or MEDLINE files within CAS STNext. Up to 10,000 answers can be included per request
- Introduction of the Knowledge Graph in CAS SciFinderⁿ allows researchers to view the connectivity and shared indexing of references quickly and interactively
- Multi-term reference text queries conducted within CAS SciFinderⁿ will now take advantage of the new Precision Search functionality. This search advancement helps address the desire of many users to quickly achieve their most relevant search results.

More detailed information on recent enhancements can be found by viewing the monthly “What’s New” release notes withing CAS SciFinderⁿ, or feel free to get in touch and we’ll be happy to provide you with more information about items that are most meaningful to you.

CAS Formulus

2022 saw several significant enhancements within CAS Formulus. The year started by focusing on improving the search experience for our users. This included making enhancements to autosuggest, improving controlled vocabulary handling, and incorporating supplier trade names as searchable ingredient identifiers.

In addition to improving the search experience, CAS invested in expanding our formulations content. CAS Formulus now includes expanded surfactant formulations, covering products such as detergents, hard surface cleaners, and consumer products. In the second half of the year, CAS Formulus also expanded content across formulations from cleaning & personal care, food & related, and inks, paints, & coatings.

CAS Formulus will offer the ability to save items of interest as well as setting alerts on search queries. These key features will offer better workflow support to our formulators. Another aspect of workflow support is expanded exporting capabilities, including the ability to export information such as *Commonly Formulated With...*, *Commonly Used As...*, and formulation-centric regulatory data.

STN IP Protection Suite

The STN IP Protection Suite offers trusted search and monitoring solutions, including CAS STNext®, CAS Scientific Patent Explorer™, and FIZ PatMon, and access to search services and expertise through CAS Search Guard. The Suite also includes extended capabilities to support comprehensive and efficient search, including CAS PatentPak, and access to expanded formulations and sequences content. For more information on the STN IP Protection Suite, click [here](#).

CAS STNext

CAS STNext continues to connect IP searchers to premier content with precision tools and technology built to power a comprehensive and efficient search. Throughout 2022, CAS has continued to enhance this solution to meet the growing search needs of our users.

Enhancements of note include:

- New predictive features to help searchers find additional relevant results and provide key insights into new topic areas
- Availability of links to National Patent Registers in patent databases to provide direct access to information on the status of the patent process and original documents from the patent offices.
- Expanded access to Taiwanese patent information with new full-text database and claims coverage

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

RSC Databases Update

Contribution from Tamara Hughes, email: HughesT@rsc.org, Mark Archibald, and Richard Kidd, Royal Society of Chemistry.



Following the RSC’s recent pledge to support [Open Science](#) and [Open Access](#) commitment make all its fully-RSC owned journals open access by 2028, the chemistry data and databases team is keen to develop and implement strategies supporting the community’s move to more open and FAIR research data practices.

In addition to working with our colleagues in journal publishing to improve their data policies, we hope to work closely with the community to raise awareness, establish needs and develop solutions. If you would like to discuss ideas or help us drive this in the right direction, please get in touch!

To get an idea what we have been up to over the past few months and how our existing data products are developing, please keep reading.

IUPAC FAIRSpec update

RSC, represented by Mark Archibald, is an active member of the IUPAC project [Development of a Standard for FAIR Data Management of Spectroscopic Data](#), abbreviated as FAIRSpec. The aim of the project is to standardise the generation and registration of metadata relating to spectroscopic data in accordance with the FAIR principles (findable, accessible, readable, interoperable).

In practice, it is hoped that:

- In the future more publications are accompanied by raw spectroscopic data (rather than simply descriptions and images of spectra)
- Spectroscopic data deposited anywhere (whether with a publisher or in a specialist or general-purpose repository) can be located by searching on structure, type of experiment, etc.
- The wider availability of spectra accompanied by standardised metadata will enable processes that rely on automated ingestion of large amounts of data, e.g. machine learning models

The project group is beginning to concentrate the past couple of years' work into the final set of recommendations. We hope to share a more detailed update in the next edition of this newsletter.

Database products

MarinLit

[MarinLit](#) is a comprehensive database of the marine natural products literature, covering new and revised structures, synthesis, ecology and biological activities. This subscription-only database is continuously updated with the latest articles and compounds published in the literature, and expert curation ensures comprehensive coverage and dereplication. The Royal Society of Chemistry has been publishing the database since 2014 and in 2022, we have added over 1,500 new compounds to date.

MarinLit was our first data product to undergo modernisation and the new website, launched in 2021, has been receiving very positive feedback from the community. With a simpler and more intuitive design, we are making it easier and quicker for users to find what they are looking for.

The Merck Index Online*

For over a century The Merck Index, chemistry's version of a compound encyclopaedia, has been regarded as the most authoritative and reliable source of key physical, pharmacological and historical information on chemicals, drugs and biologicals. Since 2013, it has been updated by the RSC and was moved online.

The [Merck Index Online](#) contains over 12,000 monographs and is continuously expanded with expert-curated content highlighting only the most relevant literature and patents. This year alone, our team has worked on over 50 new monographs, including compounds such as Copper (⁶⁴Cu) oxodotreotide, a cancer imaging agent, Formetanate, an acaricide and insecticide, and Bedinvetmab, a veterinary analgesic. While the full monographs can only be accessed with a subscription, shortened profiles are openly available to all users.

Earlier this year, we launched a new and improved version of the [website](#), with an enhanced record layout, as well as simpler and faster search functions.

ChemSpider

[ChemSpider](#) is a free database of over 115 million chemical structures, properties, and associated information. It integrates data from hundreds of sources, and links back to these original data sources. With over half a

million users every month, ChemSpider is our most used database and it is ideally placed to support open science initiatives in the future.

Our team has been working hard on improving the quality of the data records and for example an analysis of the synonyms displayed on a compound profile allowed us to delete over one million erroneous entries across the database. Over the coming months, we hope to share more details on how we are curating ChemSpider to ensure users will continue to access the most relevant and reliable information. These efforts also form part of the ongoing work to simplify, improve and modernise the current website. Due to the sheer scale of this database, these product developments are our most ambitious yet and we can't wait to see the results and share them with the community.

Literature updating services

Unfortunately, niche abstracting and indexing alerts without additional data or workflow features have had their day – the alternatives have been reflected in declining usage for a number of years. As we are faced with the need to upgrade and maintain our technology platforms, it's just not realistic to rebuild for a largely vanished audience. This means we have taken some products, including Synthetic Reaction Updates, Methods in Organic Synthesis, and Catalysts & Catalysed Reactions offline as we need to retire the platforms.

*The name THE MERCK INDEX is owned by Merck Sharp & Dohme Corp., a subsidiary of Merck & Co., Inc., Whitehouse Station, N.J., U.S.A., and is licensed to The Royal Society of Chemistry for use in the USA and Canada.

UKeiG: Winners of the Prestigious Tony Kent Strix Award 2022

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

The UK electronic information Group (UKeiG), is pleased to announce that the joint winners of the prestigious international 2022 Tony Kent Strix Memorial Award given in recognition of an outstanding practical innovation or achievement in the field of information retrieval are:

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

The judging panel congratulates them on their prolific and significant contributions to information retrieval research and development on multiple fronts, most notably the search experience. Their high impact publication records and scholarly contributions are peerless and international support for their nominations was overwhelming across the information retrieval community.

Professor Ounis is noted for his sustained contributions to advances in information retrieval, his inspirational leadership, commitment to PhD education and research, and contributions to R&D through open-source software and information retrieval tools. The highly valued Terrier and PyTerrier platforms have been utilised extensively across the information retrieval community and advanced research significantly. He has focused on designing intelligent technology that enables people to access information, developing new models and techniques for search engines. His work is at the intersection of information retrieval, machine learning and big data systems where data-driven models are learned from the users' interactions with the system. His work on many information retrieval tasks including expert search models, search results diversification, search ranking,

recommendation, fake news detection and query performance prediction has furthered the community's understanding of some of the most fundamental information retrieval questions.

"I'm delighted to receive this prestigious award and honoured to join the company of the inspiring past recipients who have influenced my own career in the field. I'm grateful to those colleagues, predecessors and friends who nominated me and/or supported my award application."

Dr White has made important contributions to information retrieval, search interaction models and health informatics, mainly focussed on understanding and enriching user interactions with information retrieval systems. He leads multidisciplinary research teams that have developed new techniques and advanced the state of the art in projects spanning artificial intelligence, human-computer Interaction and systems development. His user- and task-centric collaboration with Microsoft colleagues has pushed the boundaries in web and enterprise search. His research has underpinned the development and enhancement of widely available Microsoft products and services including the Cortana digital assistant, Bing, Xbox, Internet Explorer, Skype, Windows, Office and Azure. He was also the chief scientist at Microsoft Health.

"I am deeply humbled to receive the 2022 Strix Award and join such an illustrious group of fellow awardees. Information retrieval has been a passion of mine for over two decades. Receiving this recognition from the research community is such an incredible honour."

The Strix judging panel would like to thank colleagues who submitted a nomination for the 2022 award and look forward to submissions in 2023. The excellence, quantity and quality of the entries is proof positive that the information retrieval community is thriving.

A Zoom date for your diary

Two online Strix Memorial Lectures will be presented by Professor Ounis and Dr White on the afternoon of Thursday 23 February 2023. Further details and booking information will follow early in the new year. The event will be free of charge.

About the award

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

AI4SD News

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

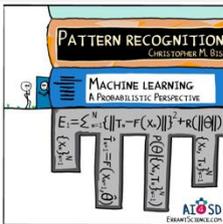
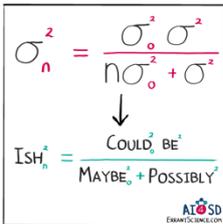
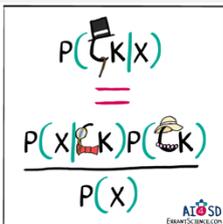
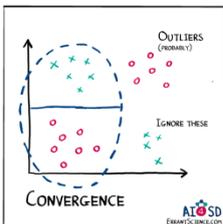
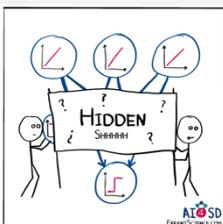


AI4SD conference report

As detailed in the previous CICAG newsletter, on 1-3 March 2022 the AI4SD conference took place at Chilworth Manor in Southampton. We were lucky enough to commission a conference report from the wonderful Wendy Warr who has produced many excellent in-depth reports for CICAG and AI4SD alike. This report is now available in our ePrints repository and can be viewed [here](#). Where speakers gave permission their talks from the AI4SD conference are available on our YouTube Channel in the [AI4SD Conference 2022 Playlist](#).

Machine learning summer school

All [Professor Mahesan Niranjan's](#) talks from the Machine Learning Summer School are now available on our YouTube Channel: [ML Summer School Playlist](#).

Cartoon	Title	Video Link
	ML1: Mathematical Foundations for ML Prof Mahesan Niranjan (University of Southampton)	Video Link
	ML2: Estimation with Machine Learning Prof Mahesan Niranjan (University of Southampton)	Video Link
	ML3: Classification and Clustering Prof Mahesan Niranjan (University of Southampton)	Video Link
	ML4: Linear Regression to Perceptron Convergence Prof Mahesan Niranjan (University of Southampton)	Video Link
	ML5: Radial Basis Functions and Multi-Layer Perceptrons Prof Mahesan Niranjan (University of Southampton)	Video Link

As part of the Summer School the students took part in a hackathon to attempt one of three challenges:

Task 1 – Predict solubility given a large set of calculated features

Task 2 – Event Detection in Nanopore Data

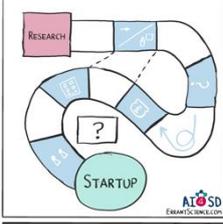
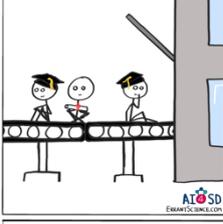
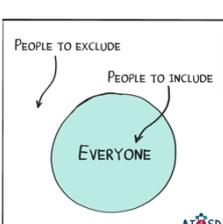
Task 3 – Detect Defects in Electron Microscopy Images

These tasks and the methods our groups used to approach them are detailed in their summer school group reports listed below.

Report Details & Link	Project Team
AI4SD-SummerSchool-Series:Report-1	Jonathan Swain (<i>University of Cambridge</i>) Bradley Patrick (<i>Nottingham Trent University</i>) Andrea Frisco (<i>University College London</i>) Dan Criveanu (<i>University of Nottingham</i>)
AI4SD-SummerSchool-Series:Report-2	Ross J. Urquhart (<i>University of Strathclyde</i>) Chris Woodley (<i>University of Liverpool</i>) Katerina Karoni (<i>University of Edinburgh</i>) Jan Elsner (<i>UCL</i>) Daniel York (<i>Swansea University</i>)
AI4SD-SummerSchool-Series:Report-3 Awarded Runner up for Best Project	Anna Bachs Herrera (<i>Swansea University</i>) Abdoulatif Cisse (<i>The University of Liverpool</i>) Emilio Alexis de la Cruz Nun �ez Andrade (<i>Swansea University</i>) Philipp Deussen (<i>UCL</i>) Ivan Yankov (<i>University of Strathclyde</i>)
AI4SD-SummerSchool-Series:Report-4 Awarded Best Project	Robert Dickson (<i>University of Liverpool</i>) Ben Honore (<i>University of Bristol</i>) Hai Lin (<i>University of Liverpool</i>) Rachael Pirie (<i>Newcastle University</i>)
AI4SD-SummerSchool-Series:Report-5	James Osborne (<i>University of Liverpool</i>) Ellie Nelson (<i>University of York</i>) Edvin Mamo (<i>University College Dublin</i>) Shaoqi Zhan (<i>University of Oxford</i>) Steven Tendyra (<i>University of Manchester</i>)
AI4SD-SummerSchool-Series:Report-6	Wole Ademola Adewole (<i>University of Southampton</i>) Halil Ibrahim Aysel (<i>University of Southampton</i>) Stephen Gow (<i>University of Southampton</i>) Zheng Jiang (<i>University of Southampton</i>) Dimitrios Stamatis (<i>University of Southampton</i>)
AI4SD-SummerSchool-Series:Report-7 Awarded Runner up for Best Project	Peng Bao (<i>University of Liverpool</i>) Jack Macklin (<i>University of Bath</i>) Masood Gheasi (<i>University of Southampton</i>)
AI4SD-SummerSchool-Series:Report-8	Xuerui Guo (<i>University of Southampton</i>) Zien Ma (<i>Cardiff University</i>) Jayanta Kumar Pal (<i>University of Liverpool</i>)

AI4SD early career researcher event

In July we held a two-day hybrid event at Chilworth Manor in Southampton. This event was for Early Career Researchers working across the domains of Computer Science and Chemistry. It was specifically designed to inform, upskill and facilitate networking between Early Career Researchers. The event contained talks on scientific publishing, ED&I, grant and fellowship applications, advice on CVs, networking and much more. There was also a dedicated time for networking.

Cartoon	Title	Speaker	Video Link
	Introduction to EPSRC and Funding Opportunities Dr Liam Boyle (EPSRC)		Video Link
	Opportunities for ECRs in the Royal Society of Chemistry Robert Bowles (RSC)		Video Link
	Research to Startup Mr Samuel Munday (University of Southampton)		Video Link
	Transitioning to Industry: A Long, Short Road Will Bowers (Dotmatics)		Video Link
	Create a Killer CV Robert Bowles (RSC)		Video Link
	Introduction to equality, diversity and inclusion and development of your code of conduct Debra Fearnshaw (University of Nottingham)		Video Link

AI4SD interviews

We now have 39 Humans of AI4SD Interviews published on our [website](#). We are working on creating an interview booklet with these so keep your eyes peeled for that!

Introducing AI4SD Instagram and Sharkcat

Exciting news! We now have an AI4SD Instagram Account, and a new mascot! It gives us great pleasure to introduce SharkCat!



Sharkcat is somewhat of an ML guru and is highly dedicated to accelerating scientific discovery using AI and ML methods. He is a stickler for high quality data, woe betide you if he spots any inconsistent datasets or finds out that you have submitted a research paper without the proper supplementary data! Follow us on instagram at [@aiscinet](https://www.instagram.com/aiscinet) to keep up to date with Sharkcats latest exploits!

What's next for AI4SD?

Rest assured you will still be hearing from us! We are still planning to run some AI4SD events in conjunction with other initiatives such as CICAG and the [Physical Sciences Data-science Service \(PSDS\)](#) and the [Physical Sciences Data Infrastructure \(PSDI\) Initiative](#). We are still working hard to finalise the last set of resources from the most recent events so keep checking out our [YouTube Channel](#). The AI4SD Team are also working on a number of other projects some of which are detailed below.

Open & transparent research practices: case study

Dr Samantha Kanza and Dr Nicola Knight interviewed Professor Jeremy Frey to form a case study on openness and transparency in Chemistry. You can read the full case study [here](#).

Openness and transparency constitute a foundational principle for research integrity, as set out in the UK Concordat to Support Research Integrity. Openness can promote rigour, constructive scrutiny, accountability and can enable others to build on research. However, it can also bring challenges. Critically, what openness and transparency can and should mean varies across disciplines and fields of study. This is one of a series of case studies in a wide range of disciplines that illustrate these differences. The series is intended to enable researchers to see similarities and differences between fields, and to inform those supporting open research through, for example, training, policies or incentives. This Case Study is based on a single interview with Professor Jeremy Frey, and is therefore illustrative rather than representative.

PSDI launch

The Physical Sciences Data Infrastructure (PSDI) initiative is an EPSRC Digital Research Infrastructure project that has been funded to create a more integrated data infrastructure within the physical sciences. Members of AI4SD from the University of Southampton are contributing to the PSDI Team.



Despite the growing use of digital tools in physical sciences, research infrastructure is still fragmented across many systems. The aim of PSDI is to enable researchers in the physical sciences to handle data more easily by connecting the different data infrastructures they use. PSDI

will connect and enhance existing infrastructure in Physical Sciences. An initial pilot phase took place from Nov 2022 – Mar 2023 (Grant [EP/W032252/1](#)) and work is just beginning on Phase 1b (Grants [EP/X032701/1](#) and [EP/X032663/1](#)) where further engagement and design is being undertaken alongside development work on the technology platform and demonstrator pathfinders. You can find out more information about the PSDI initiative on their website: <https://www.psdia.ac.uk/> or on Twitter [@PSDI_UK](#) where updates will be posted as their work begins.

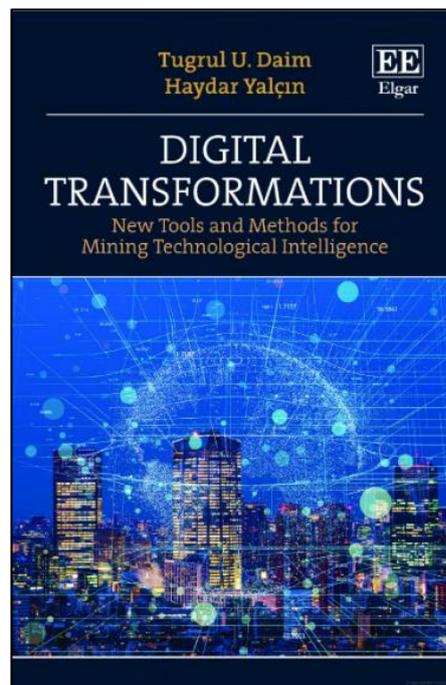
Book Review: Digital Transformation: New Tools and Methods for Mining Technological Intelligence

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

This excellent book covers three methods for technology management, especially for evaluation of emerging technologies: analyses by bibliometric, patent-based, and network-based analyses. Based on the authors' researches, for their own research and consulting, the book is jam-packed with background information and data in the form of graphs, tables, tag clouds, and analysis maps. Topics covered cover the spheres of emerging technologies from engineering to computation and chemistry. Potential audiences include industry and government professionals, researchers, professors, and students (especially graduate students). Highly recommended.

Reviews of this book will appear the December issues of [ALA/CHOICE](#) and the [Chemical Information Bulletin](#) (CIB) of the ACS Division of Chemical Information (CINF).

Daim, Tugrul U., Yalcin, Haydar. *Digital Transformation: New Tools and Methods for Mining Technological Intelligence*. Edward Elgar Publishing, Cheltenham, UK, Northampton, USA, 172p + xi, hardcover ISBN 978-1-789900-862-6, \$110.00. Ebook ISBN 978-1-789908-863-3.



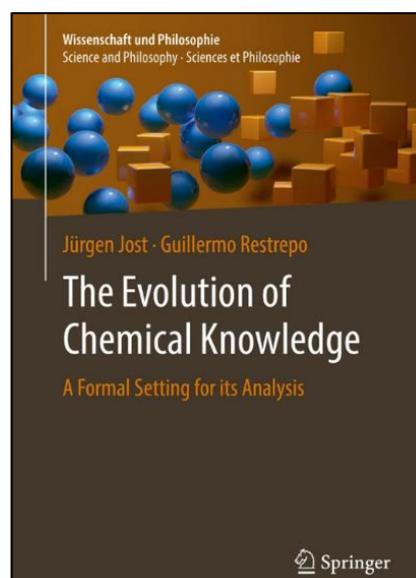
Cheminformatics and Chemical Information Books

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

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

[The Evolution of Chemical Knowledge](#)

Chemistry shapes and creates the disposition of the world's resources and provides novel substances for the welfare and hazard of our civilisation at an exponential rate. Can we model the evolution of chemical knowledge? This book not only provides a positive answer to the question, it provides the formal models and available data to model chemical knowledge as a complex dynamical system based on the mutual interaction of the social, semiotic and material systems of chemistry. These systems, which have evolved over the history, include the scientists and institutions supporting chemical knowledge (social system); theories, concepts and forms of communication (semiotic system) and the substances, reactions and technologies (material system) central for the chemical practice. These three systems, which have traditionally been mostly studied in isolation, are brought together in this book in a grand historical narrative, on the basis of comprehensive data sets and supplemented by appropriate tools for their formal analysis. We thereby



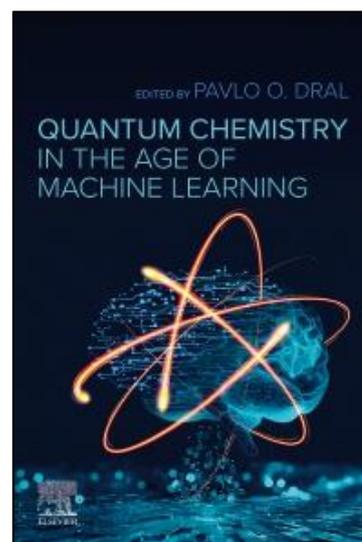
develop a comprehensive picture of the evolution of chemistry, needed for better understanding the past, present and future of chemistry as a discipline. The interdisciplinary character of this book and its non-technical language make it an ideal complement to more traditional material in undergraduate and graduate courses in chemistry, history of science and digital humanities.

Jürgen Jost, Guillermo Restrepo

Springer, 2022. <https://doi.org/10.1007/978-3-031-10094-9>

[Quantum Chemistry in the Age of Machine Learning](#)

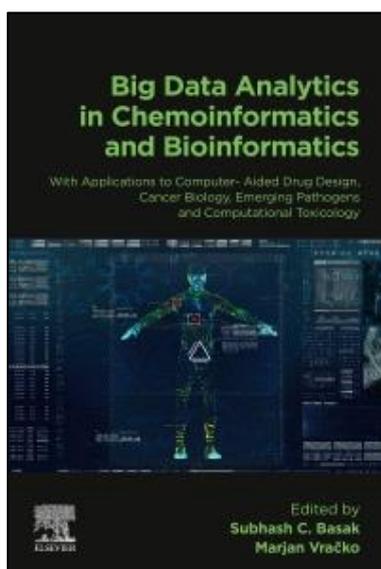
Quantum chemistry is simulating atomistic systems according to the laws of quantum mechanics, and such simulations are essential for our understanding of the world and for technological progress. Machine learning revolutionises quantum chemistry by increasing simulation speed and accuracy and obtaining new insights. However, for nonspecialists, learning about this vast field is a formidable challenge. *Quantum Chemistry in the Age of Machine Learning* covers this exciting field in detail, ranging from basic concepts to comprehensive methodological details to providing detailed codes and hands-on tutorials. Such an approach helps readers get a quick overview of existing techniques and provides an opportunity to learn the intricacies and inner workings of state-of-the-art methods. The book describes the underlying concepts of machine learning and quantum chemistry, machine learning potentials and learning of other quantum chemical properties, machine learning-improved quantum chemical methods, analysis of Big Data from simulations, and materials design with machine learning. Drawing on the expertise of a team of specialist contributors, this book serves as a valuable guide for both aspiring beginners and specialists in this exciting field.



Pavlo Dral (ed.)

Elsevier, 2022, ISBN: 9780323900492

[Big Data Analytics in Chemoinformatics and Bioinformatics](#)



This book provides an up-to-date presentation of big data analytics methods and their applications in diverse fields. The proper management of big data for decision-making in scientific and social issues is of paramount importance. This book gives researchers the tools they need to solve big data problems in these fields. It begins with a section on general topics that all readers will find useful and continues with specific sections covering a range of interdisciplinary applications. Here, an international team of leading experts review their respective fields and present their latest research findings, with case studies used throughout to analyse and present key information.

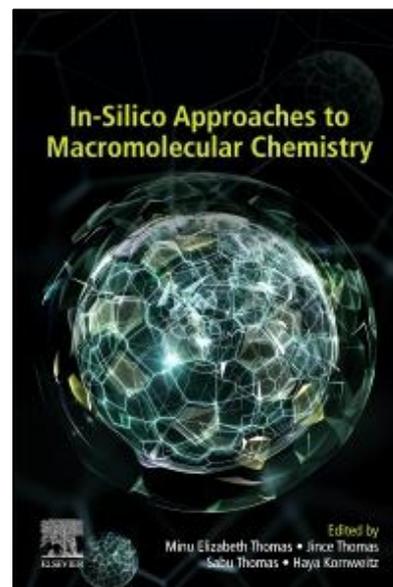
Subhash Basak, Marjan Vračko (eds)

Elsevier, August 2022, ISBN: 9780323857130

[In-silico Approaches to Macromolecular Chemistry](#)

This book helps students, researchers and industry professionals gain a clear overview of the field, giving users the knowledge needed to understand and select the most appropriate tools for conducting and analysing computational studies. With applications across a broad range of areas, many different methods have been developed for exploring macromolecules in silico, making it difficult for researchers to select the most appropriate for their specific needs. Covering both biopolymers and synthetic polymers, this book familiarises readers with the theoretical tools and software appropriate for such studies. In addition to providing essential background knowledge on both computational tools and macromolecules, the book presents in-depth studies of in silico macromolecule chemistry, discusses and compares these with experimental studies, and highlights the future potential for such approaches.

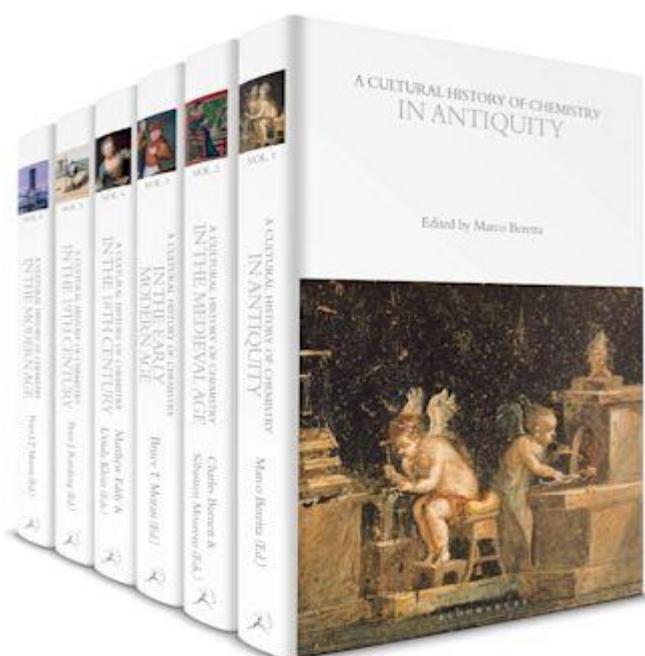
Minu Thomas, Jince Thomas, Sabu Thomas, Haya Kornweitz (eds)
Elsevier, August 2022, ISBN: 9780323909952



[A Cultural History of Chemistry](#)

From prehistoric metal extraction to medieval alchemy to modern industry, chemistry has been central to our understanding and use of the physical world as well as to trade, warfare and medicine. In its turn, chemistry has been shaped by changing technologies, institutions and cultural beliefs. *A Cultural History of Chemistry* presents the first detailed and authoritative survey from antiquity to today, focusing on the West but integrating key developments in Egypt, Mesopotamia, and the Arabic-Islamic and Byzantine empires.

The six volumes cover: 1 Antiquity (3,000 BCE-600 CE); 2 Medieval Age (600 to 1500); 3 Early Modern (1500-1700); 4 Eighteenth Century (1700-1815); 5 Nineteenth Century (1815-1914); 6 Modern Age (1914 to present).



Peter J. T. Morris & Alan Rocke (anthology editors)
Bloomsbury Academic, 2022. ISBN: 9781474294928

2022 Reflections on Life at the Catalyst Science and Discovery Centre and Museum in Widnes

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



On 8 November 2022, I as Chair of Trustees, and the other trustees and staff were joined by patrons, sponsors, local schools and dignitaries to celebrate the completion of the Wellcome/ISF funded project to revamp many of the educational areas within the [Catalyst Science Discovery Centre and Museum](#) in Widnes. It was the culmination of a £1 million project which had started before the COVID pandemic in 2019 but had survived all the traumas of that period and come out successfully at the end. It was an opportunity to thank everyone who had supported us through this work when we had to raise £330,000 ourselves as part of the overall sum and to show the work that had been done – the INEOS theatre, the new interactive gallery, the new exhibition area, new education studios, the new café and shop, and the changing spaces toilets for the disabled. It was also an opportunity to inform our visitors about our plans for the future.



Left to right: Diana Leitch, Leader of Halton Borough Council
Cllr Mike Wharton, Sir Phil Redmond, Lady Redmond the
Lord Lieutenant of Cheshire, Dr Lee Juby, CEO Catalyst,
Mayor of Halton.



Lady Redmond cuts the ribbon to open the
Interactive Gallery.

Local dignitaries included the Lord Lieutenant of Cheshire, Lady Redmond, accompanied by her husband Sir Phil Redmond, the Metro Mayor of Liverpool City Region, Steve Rotheram, the Mayor of Halton and the Leader of Halton Borough Council. Lady Redmond declared our new facilities officially open and symbolically cut the ribbon to the entrance to the new interactive gallery. It was a pleasure to welcome our own Chair of CICAG, Chris Swain, on his first visit to Catalyst. Sadly the train strikes prevented three of our very supportive patrons – Sir Hugo Brunner, Mr Peter Gossage and Professor David Philips, representing the RSC, from joining us.

Our concerns that visitors would not come back to Catalyst after the pandemic have been negated and we have had an excellent year with the highest number of public visitors and families being recorded for many years. Schools and sleepover groups (uniformed organisations and schools) have all come back and our targets for usage have been met. The financial business plan we had created was on target and we were jubilant until 8 August when our new CEO, Dr Lee Juby CEng, FIET, who joined us on 1 August, rang me to tell me that he had received critical financial information that our energy bill for our gas usage was going up by five times more than what we had been paying before. We, like all other charities and organisations, were devastated.

How were we going to cover this increase of tens of thousands of pounds and as a charity retain financial sustainability? I appeared in *The Guardian* newspaper in an article organised by the Museums Association to talk about the difficulties. With other increasing costs and difficult financial times facing several of our industrial sponsors the last few months have not been easy but we are nevertheless still providing quality science education and fund raising continues unabated as we receive no government or local funding.

So what of the future? We have been awarded a grant by the National Heritage Lottery Fund (NLHF) of up to £1 million to revamp the museum and heritage areas of the Centre and have recently appointed a Heritage Manager, Paige Halliday, who starts on 6 January 2023 and will manage the initial phase of this project. She will work with a co-curation team and local community groups to present and interpret the heritage of the local chemical industry and the chemical sciences in a new way. So an exciting new phase in the life of Catalyst which will be completed towards the end of 2024. Again more fund raising.



Archivist Judith Wilde shows visitors some of the archives and artefacts in the Gossage Room.

Our collection archivist, Judith Wilde, continues to receive enquiries and visitors from all over the world to use our archives of the major chemical companies of the NW England (ICI, Brunner Mond, Peter Spence, Solvay) and the digitisation of our archival material continues with a dedicated team of seven volunteers.

Currently we are hosting an exhibition created by the British Geological Survey on core sample testing on Ince Marshes in Cheshire. It was at the Glasgow Science Centre and has replaced the previous exhibition about the Space Sapling which we have growing outside our front door. It is an apple tree grown from one of the seeds from Newton's garden in Lincolnshire that went to space with Tim Peake.

So 2023 will be interesting and we hope to work closely with other local museums such as Nantwich Museum, where Helen Cooke is a Trustee, on various science activities. Not of course forgetting chemistry – we will be celebrating 'Chemistry at Work Week' from 23-27 January 2023 and are fully booked with pupils from local secondary schools coming to meet representatives from local industrial companies. Thanks to the RSC Outreach fund for a helpful grant.

Other Chemical Information News

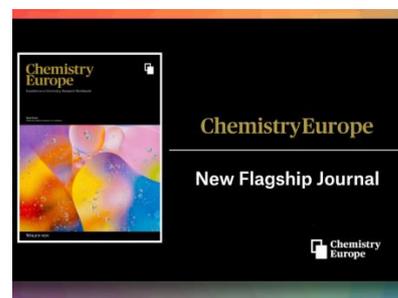
Contribution from Stuart Newbold, email: stuart@psandim.com

Top Journal Made by Scientists for Scientists

Chemistry Europe is expanding its journal portfolio with ChemistryEurope, a high-quality, high-impact Gold Open Access journal that disseminates work by international scientists in all areas of chemistry. The goal of the journal is to disseminate the highest-quality content from all areas of chemistry. Original reports on developments that are intended to attract a broad readership in the chemical community and related sciences, as well as review articles on current topics of societal importance or breakthrough chemical concepts that are of general interest to the scientific community will be published.

<https://www.chemistryviews.org/top-journal-made-by-scientists-for-scientists/>

Source: ChemistryViews



Benefits of Semantic Enrichment Across the Drug Development Pipeline

White paper discussing practical use cases using semantic enrichment technology in various stages of drug development to show the benefits that leveraging semantic search can bring to the table.

<https://www.copyright.com/wp-content/uploads/2021/09/CCC-SciBite-White-Paper-Benefits-of-Semantic-Enrichment-Across-the-Drug-Development-Pipeline.pdf>

Source: Copyright Clearance Center

Insilico Medicine Nominates Potential First-in-Class Preclinical Candidate with Novel AI-Designed Structure for Novel AI-Discovered Target in Immuno-oncology

The company has nominated ISM4312A as a preclinical candidate targeting DGKA with an AI-identified target and AI-designed structure for immuno-oncology therapeutics. It is another program fully discovered and designed by AI that Insilico has delivered from target identification to preclinical candidate nomination by leveraging its proprietary end-to-end AI platform, Pharma.AI.

<https://www.biospace.com/article/releases/insilico-medicine-nominates-potential-first-in-class-preclinical-candidate-with-novel-ai-designed-structure-for-novel-ai-discovered-target-in-immuno-oncology/>

Source: BioSpace

Three reasons why the Periodic Table needs a Redesign

Chemists can't agree on the best way to arrange the elements, prompting proposals of everything from spiral-shaped alternatives to radically elongated versions.

<https://www.newscientist.com/article/mg24132190-400-three-reasons-why-the-periodic-table-needs-a-redesign/>

Source: New Scientist

Announcing the World's First Curation of Scientific Conference Presentations: Underline Science, Streaming the World's Best Science

<https://www.stm-publishing.com/announcing-the-worlds-first-curation-of-scientific-conference-presentations-underline-science-streaming-the-worlds-best-science/>

Source: STM Publishing News

Chemical Computer can be Programmed to solve Hard Problems

Programmable computers that use chemical reactions to process information could solve problems faster than conventional computers. And they may better mimic the brain than their electronic counterparts.

<https://www.newscientist.com/article/2319242-chemical-computer-can-be-programmed-to-solve-hard-problems/>

Source: *New Scientist*

AAAS Selects Cadmore Media to Host Streaming Content on Science.org

In line with recent undertakings to improve the user experience and accessibility, The American Association for the Advancement of Science (AAAS) – publisher of the Science Family of Journals – has selected Cadmore Media as its streaming partner on science.org. This will result in improved accessibility and discoverability of streaming content on the site, as well as a more efficient workflow for internal production staff.

<https://www.stm-publishing.com/aaas-selects-cadmore-media-to-host-streaming-content-on-science-org/>

Source: *STM Publishing News*

Automated Chemical Reaction Prediction: Now in Stereo

Automated reaction path search method predicts accurate stereochemistry of pericyclic reactions using only target molecule structure.

<https://www.sciencedaily.com/releases/2022/12/221201082154.htm>

Source: *ScienceDaily*

Patent Knowledge for Research: Start of DFG project 'Patents4Science'

Together with its partners in research, FIZ Karlsruhe develops an innovative information infrastructure for patent information to be used by scientists. Entirely new: a patent-centered knowledge graph.

<https://www.knowledgespeak.com/news/patent-knowledge-for-research-start-of-dfg-project-patents4science/>

Source: *Knowledgespeak*

UK AI Tech-Bio Company, Biorelate, and CCC Announce Strategic Integration

Biorelate and CCC have announced a strategic integration. Biorelate's customers will immediately have the option to access subscription-only scholarly article content in XML format licensed through and delivered by CCC's RightFind XML platform. With this XML integrated into Biorelate's Galactic AI™ platform, Biopharma companies can now work with the combined services to automatically curate biomedical data from millions of additional full-text articles, revealing previously unseen biomedical insights.

<https://www.copyright.com/media-press-releases/uk-ai-tech-bio-company-biorelate-and-ccc-announce-strategic-integration/>

Source: *Copyright Clearance Center*

Reactions

Reactions is a video series produced by the ACS and PBS Digital Studios. Subscribe to Reactions at <http://bit.ly/ACSReactions>.

<https://www.eurekalert.org/news-releases/963761>

Source: *AAAS EurekAlert!*

Unique insights afforded to us by Computational Chemistry

Though experimentation is still king in most chemists' minds, computational chemistry has the potential to transform the field.

<https://www.advancedsciencenews.com/unique-insights-afforded-to-us-by-computational-chemistry/>

Source: *Advanced Science News*

Using Computers to Understand Diels–Alder Reactions

Computational methods can play an important role in the chemical sciences. Thanks to molecular simulations, it is often possible to obtain information on chemical reactions at a much lower cost compared with experiments. In this context, it is crucial to find suitable methods for the problem at a hand. They need to be as accurate as necessary, but as fast as possible.

<https://www.chemistryviews.org/using-computers-to-understand-diels-alder-reactions/>

Source: *ChemistryViews*

Cambridge Journals see leap in Open Access Research

The amount of new research published open access in Cambridge's Transformative Journals leaped by almost 70 per cent in 2021.

<https://www.cambridge.org/news-and-insights/news/Cambridge-journals-see-leap-in-open-access-research>

Source: *CUP*

PLOS unveils a new way to measure Open Science Practices

The Public Library of Science has announced the release of the first results of a new initiative to measure researchers' Open Science practices across the published literature, 'Open Science Indicators' (OSI). These initial results show increasing prevalence of good practices in research data and code sharing, and increasing use of preprints by researchers. PLOS is releasing the OSI dataset, and the associated framework used to develop it, to support Open Science initiatives in the wider community.

<https://www.eurekalert.org/news-releases/974106>

Source: *AAAS EurekAlert!*

Why creating a Chemical Brain will be how we understand Consciousness

Unorthodox chemist Lee Cronin is leading a radical quest to use chemistry to explain consciousness and create artificial life.

<https://www.newscientist.com/article/mg23931950-100-why-creating-a-chemical-brain-will-be-how-we-understand-consciousness/>

Source: *New Scientist*

Integrated Platform promises to accelerate Drug Discovery Process

Many successful drugs have their origins in natural sources such as plants, fungi, and bacteria, but screening natural products to identify potential drugs remains a difficult undertaking. A new approach using molecular biology, analytical chemistry, and bioinformatics to integrate information from different screening platforms addresses some of the biggest challenges in natural products drug discovery.

<https://www.sciencedaily.com/releases/2022/12/221201082138.htm>

Source: *ScienceDaily*

Robotic Chemist may be able to recreate Earth's Primordial Soup

Recreating the mix of compounds and experimental conditions that interacted over billions of years to create life on Earth is impossible in the lab. But an autonomous robot can shorten the time it takes to test each possible mixture, which could help reveal the precise combination that let proteins, DNA and enzymes emerge from the prebiotic soup on early Earth.

<https://www.newscientist.com/article/2280573-robotic-chemist-may-be-able-to-recreate-earths-primordial-soup/>

Source: *New Scientist*

Insilico Medicine receives Company of the Year Award at BioCentury-BayHelix East-West Summit

Insilico Medicine was named Company of the Year at the BioCentury-BayHelix East-West Summit on November 15 this year. The clinical stage artificial intelligence (AI)-driven drug discovery company is among over 50 companies presenting at the event in Redwood City, California, which is focused on the globalisation of drug discovery and development, as well as cross-border collaborations between the East and West.

<https://www.eurekalert.org/news-releases/971576>

Source: AAAS EurekAlert!

Access more Online Content

The British Library has added over 20 million items that can be accessed remotely with a Reader Pass.

<https://www.bl.uk/news/2022/july/access-more-content-online>

Source: BL

Top 3 Challenges When Using Scientific Articles in AI & Machine Learning Projects

Here are the three primary challenges we hear when companies build a collection of articles (or “corpus”) for their text mining projects, with tips to overcome them.

<https://www.copyright.com/blog/top-3-challenges-when-using-scientific-articles-in-ai-machine-learning-projects/>

Source: Copyright Clearance Center

Chemical.AI to expand to the Indian Market

Chemical.AI has announced a partnership with an Indian company, ChemIntel Technologies. Together with ChemIntel Technologies' resources, Chemical.AI will accelerate the expansion of Indian business and lay out the global market, thereby promoting the development of AI (artificial intelligence) in pharmaceuticals and bringing a positive impact on human welfares.

<https://www.eurekalert.org/news-releases/971851>

Source: AAAS EurekAlert!

C&EN's Year in Chemistry 2022

C&EN cover 2022's exciting chemistry trends, quirky molecules, and remarkable discoveries.

<https://cen.acs.org/education/science-communication/CENs-Year-Chemistry-2022/100/i44>

Source: Chemical & Engineering News

Clarivate Names World's Influential Researchers with Highly Cited Researchers 2022 List

The 2022 list of Highly Cited Researchers™ details individuals at universities, research institutes and commercial organisations who have demonstrated a disproportionate level of significant and broad influence in their field or fields of research. The methodology draws on data from the Web of Science™ citation index, together with analysis performed by bibliometric experts and data scientists at the Institute for Scientific Information (ISI)™ at Clarivate.

<https://clarivate.com/news/clarivate-names-worlds-influential-researchers-with-highly-cited-researchers-2022-list/>

Source: Clarivate

PLOS and DataSeer expand Partnership to better understand researchers' Open Science Practices.

<https://www.eurekalert.org/news-releases/964339>

Source: AAAS EurekAlert!

Springer Nature completes acquisition of Research Square Company

RSC comprises American Journal Experts (AJE), which provides best-in-class AI-powered and professionally delivered author solutions, and Research Square, the world's number one multi-disciplinary preprint platform.

<https://group.springernature.com/gp/group/media/press-releases/springer-nature-completes-acquisition-of-research-square-company/23768186>

Source: Springer Nature

Automation enables Modular Synthesis of new Molecules for Lasers

An automated synthesis platform called Chemspeed reduces time and labor when searching for organic molecules as gain mediums in lasers.

<https://www.advancedsciencenews.com/automation-enables-lego-like-synthesis-of-new-molecules-for-lasers/>

Source: Advanced Science News

ResearchGate and the International Union of Crystallography announce Content Partnership

<https://www.eurekalert.org/news-releases/961417>

Source: AAAS EurekAlert!

Taylor & Francis trials Proofig software to help detect image duplication

Taylor & Francis has announced the beginning of a 6-month pilot of Proofig image integrity software, as part of an ongoing program to prevent image duplication and manipulation in academic journal articles. Image duplication or manipulation is a serious form of misconduct which involves inappropriately duplicating, manipulating, or fabricating images. Even where image duplication has occurred due to honest error, this still damages the integrity of the scholarly record.

<https://newsroom.taylorandfrancisgroup.com/taylor-francis-trials-proofig-software-to-help-detect-image-duplication/>

Source: Taylor & Francis

Chemists develop reactions for the general synthesis of promising unexplored compounds

Chemists at Scripps Research have devised the first general method for synthesising a family of compounds called 1,2,3,5-tetrazines, which hold great promise for making pharmaceuticals, biological probes and other chemical products.

<https://www.sciencedaily.com/releases/2022/12/221206204929.htm>

Source: ScienceDaily

Conditions for Suzuki-Miyaura Coupling Optimised with Machine Learning

The Suzuki–Miyaura reaction is widely used to form new C–C bonds in organic chemistry. Such reactions that combine two fragments can be employed in automated syntheses to generate a wide range of products from a library of different building blocks. For such automated syntheses, reaction conditions that are general, i.e., applicable to a wide range of substrates and high-yielding, are useful. However, optimising the reaction conditions is challenging due to the very large “search space” of possible substrate combinations and conditions.

<https://www.chemistryviews.org/conditions-for-suzuki-miyaura-coupling-optimized-with-machine-learning/>

Source: ChemistryViews

ResearchGate and Royal Society partner to increase accessibility of Research

<https://www.eurekalert.org/news-releases/955473>

Source: AAAS EurekAlert!

UK's National Synchrotron Facility reports £2.6 Billion Impact

An updated report reveals that the UK's national synchrotron facility, Diamond Light Source, has had a £2.6 billion impact on UK science and economy since 2007.

<https://www.ukri.org/news/uks-national-synchrotron-facility-reports-2-6-billion-impact/>

Source: UKRI

GigaScience celebrates its first decade in publishing

The first biological and biomedical "data" journal [GigaScience](https://www.gigascience.com/) recently celebrated its 10th year of being at the forefront of open scientific publishing. GigaScience was launched on July 12, 2012, at the Intelligent Systems for Molecular Biology (ISMB) conference in Long Beach, CA.

<https://www.knowledgespeak.com/news/gigascience-celebrates-its-first-decade-in-publishing/>

Source: Knowledgespeak

Clarivate Correctly Predicted Seven 2022 Nobel Prize Winners

Clarivate is celebrating "the seven new Nobel Laureates across the fields of science and economics who were accurately identified as potential Nobel Prize recipients. Each individual was awarded the designation of Citation Laureate several years before being named by the Nobel Assembly – four, more than a decade – thanks to expert interpretation of high-quality citation data at Clarivate.

<http://newsbreaks.infotoday.com/Digest/Clarivate-Correctly-Predicted-Seven-2022-Nobel-Prize-Winners-155384.asp>

Source: Information Today Inc

Five Artificial Intelligence and Data Predictions for 2023

Ryan Welsh, founder and CEO of Kyndi, recently compiled the following list of artificial intelligence (AI) and data-related topics that he believes will be important in 2023.

<http://newsbreaks.infotoday.com/Digest/Five-Artificial-Intelligence-and-Data-Predictions-for-2023-155930.asp>

Source: Information Today Inc

IUPAC Announce Top Ten Emerging Technologies in Chemistry 2022

The International Union of Pure and Applied Chemistry (IUPAC) has announced the 2022 Top Ten Emerging Technologies in Chemistry finalists. With this project, started in 2019, IUPAC aims to showcase the value of chemistry and inform the general public as to how the chemical sciences contribute to the well-being of society and sustainability. The jury selects emerging technologies, i.e., those at a stage between a new scientific discovery and a fully commercialised technology, and highlights those with the greatest capacity to provide new opportunities and transform our world.

<https://www.chemistryviews.org/iupac-announced-top-ten-emerging-technologies-in-chemistry-2022/>

Source: ChemistryViews

Nature authors can now seamlessly share their Data

In a further move to support open research, more journals in the Nature Portfolio – including Nature itself – will now provide authors with the opportunity to openly share their data, thanks to an integration with Figshare.

<https://www.researchinformation.info/news/nature-authors-can-now-seamlessly-share-their-data>

Source: Research Information

Springer Nature launches new AI-led Service

Springer Nature has launched a new AI-led service to help research decision-makers from academic, government, and corporate organisations make informed data-driven funding and strategy decisions. Nature

Research Intelligence is powered by Nature's 150 years of editorial and research expertise and builds on the existing success of Nature Index.

<https://www.knowledgespeak.com/news/springer-nature-launches-new-ai-led-service/>

Source: *Knowledgespeak*

R Discovery partners with Springer Nature

R Discovery, a Cactus Communications (CACTUS) brand, has partnered with Springer Nature, global academic publisher, to help broaden the reach of open access (OA) content to the global researcher community.

<https://www.researchinformation.info/news/r-discovery-partners-springer-nature>

Source: *Research Information*

Artificial Intelligence and Molecule Machine join forces to generalise Automated Chemistry

Artificial intelligence, building-block chemistry and a molecule-making machine teamed up to find the best general reaction conditions for synthesising chemicals important to biomedical and materials research - a finding that could speed innovation and drug discovery as well as make complex chemistry automated and accessible. Researchers at the University of Illinois Urbana-Champaign and collaborators in Poland and Canada reported their findings in the journal *Science*.

<https://www.sciencedaily.com/releases/2022/10/221028180848.htm>

Source: *ScienceDaily*

AAAS Survey reveals how open access publishing trends and costs are affecting the scientific enterprise

To better drive scientific innovation and apply lessons learned from the COVID-19 pandemic, federal policymakers are exploring methods to increase access to published scientific research and data that is federally funded.

<https://www.knowledgespeak.com/news/aaas-survey-reveals-how-open-access-publishing-trends-and-costs-are-affecting-the-scientific-enterprise/>

Source: *Knowledgespeak*

Insilico Medicine demonstrates value of AI approach to Drug Discovery

Partnership with Deerfield Discovery and Development aims to break ground on new targets.

<https://www.eurekalert.org/news-releases/962749>

Source: *AAAS EurekAlert!*

Empower your Data with Infographics

Infographics give an added dimension to textual information, as taxonomist and trainer Joyce van Aalten explains. Visuals provide powerful elements to get your message across--and they are not that hard to create.

<https://www.infotoday.eu/Articles/Editorial/Featured-Articles/Empower-your-data-with-infographics-151083.aspx>

Source: *Information Today*

Artificial Intelligence Used to Predict Millions of Protein Structures

Predicting the 3D structure of a protein from its amino acid sequence is no easy task due to the large number of possible structures. It is considered an important challenge in computational biology and chemistry. Large tech corporations have started to tackle this problem using artificial intelligence.

<https://www.chemistryviews.org/artificial-intelligence-used-to-predict-millions-of-protein-structures/>

Source: *ChemistryViews*

Springer Nature expands AI-driven Digital Editing Services

Springer Nature has announced a new pilot with American Journal Experts (AJE),* which will see its AI-driven editing services being made available to book authors and editors. It will enable authors to have their manuscripts improved by grammatical errors being corrected as well as improving phrasing and word choice. The service will allow them to spend less time preparing their work for publication and have more time doing the research that drives society forwards.

<https://www.researchinformation.info/news/springer-nature-expands-ai-driven-digital-editing-services>

Source: *Research Information*

Driving Bioscience Innovation with Transformative Technology

The development of new technology is vital in advancing bioscience research and innovation.

<https://www.ukri.org/news/driving-bioscience-innovation-with-transformative-technology/>

Source: *UKRI*

ACS journals achieved the highest Journal Impact Factor

ACS journals have once again sustained consistent growth, reporting impressive levels of impact, citations, and output through 2021. The 2022 Journal Citation Reports™ from Clarivate, released on June 28, 2022, substantiate the commitment, quality, and innovation of the ACS journals portfolio and its impact.

<https://www.knowledgespeak.com/news/american-chemical-society-journals-achieved-the-highest-journal-impact-factor/>

Source: *Knowledgespeak*

ResearchGate and EDP Sciences announce Content Partnership

ResearchGate, the professional network for researchers, and EDP Sciences, an international academic publisher specialising in scientific, technical and medical disciplines, have announced a content syndication partnership that will see the addition of content from over 30 open access (OA) journals to ResearchGate.

<https://www.researchinformation.info/news/researchgate-and-edp-sciences-announce-content-partnership>

Source: *Research Information*

ACS and scite partner on the development of Smart Citations

scite, an award-winning tool that helps students and researchers discover and understand research findings more efficiently through Smart Citations, has partnered with the ACS to enhance the citation experience for its globally diverse readership.

<https://www.knowledgespeak.com/news/the-american-chemical-society-and-scite-partner-on-the-development-of-smart-citations/>

Source: *Knowledgespeak*

Where is Open Access Publishing Heading?

One of the first Gold Open Access (OA) titles published by Wiley, ChemistryOpen, has turned 10 years old! We are celebrating this milestone by taking the opportunity to reflect on the role of Gold OA in the current STEM (science, technology, engineering, and mathematics) publishing landscape.

<https://www.chemistryviews.org/where-is-open-access-publishing-heading/>

Source: *ChemistryViews*

Lab Leaders wrestle with paucity of Postdocs

Even high-profile scientists are struggling to recruit qualified postdoctoral researchers.

<https://www.nature.com/articles/d41586-022-02781-x>

Source: *Nature*

Fully OA Group launches Fully OA blog

A brand-new initiative, the Fully OA blog, has been launched by the [Fully OA Group](#). Born out of the OASPA Interest Group of Fully OA journal organisations, the group provides a forum for exchange of ideas and, where appropriate, collaboration among publishers that only publish open access.

<https://www.researchinformation.info/news/fully-oa-group-launches-fully-oa-blog>

Source: *Research Information*

UKRI unveils detailed plans for Research and Innovation

UKRI councils show how they will play their part to support world-class research and innovation and drive economic, social, environmental and cultural benefits.

<https://www.ukri.org/news/ukri-unveils-detailed-plans-for-research-and-innovation/>

Source: *UKRI*

Increasing visibility of Research Resources on bioRxiv and medRxiv

In scientific studies, key research resources, such as antibodies, are underspecified by authors, making it unclear which resource was used in a study.

<https://www.researchinformation.info/news/increasing-visibility-research-resources-biorxiv-and-medrxiv>

Source: *Research Information*

Plan S Journal Comparison Service now open for libraries and library consortia to register and access price and service data

cOAlition S has released the end-user portal of the Journal Comparison Service (JCS). This secure, free, online service aims to shed light on publishing fees and services and enable those who procure publishing services to better understand how journals and publishers compare.

<https://www.knowledgespeak.com/news/plan-s-journal-comparison-service-now-open-for-libraries-and-library-consortia-to-register-and-access-price-and-service-data/>

Source: *Knowledgespeak*

Is Big Tech draining AI talent from Academia?

Movement to industry is raising concerns about the future researcher workforce and maintaining ethical expertise.

<https://www.nature.com/articles/d41586-022-03214-5>

Source: *Nature*

RSC commits to 100% Open Access

The RSC has announced that it aims to make all fully RSC-owned journals open access within five years, making it the first chemistry publisher and one of the first society publishers to commit to a fully open access (OA) future.

<https://www.researchinformation.info/news/royal-society-chemistry-commits-100-open-access>

Source: *Research Information*

CCC's 'RightFind Cite It' now available for Mac users

CCC has announced that RightFind Cite It is now available for Mac users. RightFind Cite It is CCC's easy-to-use reference management tool that automatically formats bibliographies directly within Microsoft Word.

<https://www.knowledgespeak.com/news/cccs-rightfind-cite-it-now-available-for-mac-users/>

Source: *Knowledgespeak*

Life Sciences centre aims to accelerate drug development and enable precision medicine

Atos and the Wellcome Genome Campus in Cambridgeshire, UK, have announced the official opening of its global Life Sciences Centre of Excellence. The facility will provide scientists on campus, and global genome and bio-data institutes worldwide, with early access to emerging technologies to support their research. Helping to accelerate the process of bringing new drugs to market and delivering tangible societal benefits.

<https://www.scientific-computing.com/news/life-sciences-centre-aims-accelerate-drug-development-and-enable-precision-medicine>

Source: *Scientific Computing World*

Getting answers directly from Research Articles

The capability for [scite](#) users to ask research questions in plain language and get answers directly from the full text of research articles is the premise of scite's new "Ask a Question" feature.

<https://www.infotoday.eu/Articles/News/Featured-News/Getting-answers-directly-from-research-articles-156160.aspx>

Source: *Information Today*

Dow and CAS partner to accelerate R&D capabilities and efficiency

Materials science company Dow, and CAS have established a strategic collaboration focused on digital capabilities that accelerate research and development and identify new opportunities within key growth areas around the world. Dow and CAS have been long-time collaborators with a shared vision to facilitate innovation and discovery. Their latest collaboration leverages a tailored CAS Custom Services solution designed on a foundation of specialised CAS technologies, data science, and the CAS Content Collection™ to empower efficiencies in Dow's research processes that improve efficiency and productivity.

<https://www.knowledgespeak.com/news/dow-and-cas-partner-to-accelerate-rd-capabilities-and-efficiency/>

Source: *Knowledgespeak*

CCC acquired U.K.-based Ringgold, a leading provider of Organisation Identifiers in Scholarly Communications, making it a wholly owned Subsidiary

The acquisition, said CCC, reflects its ongoing commitment to promoting interoperability, addressing market friction and collaborating with stakeholders. It also further solidifies its transformation from focusing on copyright compliance and content to becoming involved with the entire lifecycle of content creation, research management, and information analysis as part of organisation's workflow. Ringgold has long been known for its leadership in Persistent Identifiers (PIDs) for organisation and institutions. PIDs are globally unique and are associated with accurate metadata about an article, a grant, a person, a project or an organisation.

<https://www.infotoday.eu/Articles/News/Featured-News/CCC-acquired-UK-based-Ringgold-a-leading-provider-of-organization-identifiers-in-scholarly-communications-making-it-a-wholly-owned-subsidiary-153248.aspx>

Source: *Information Today*

EBSCO Information Services introduces BiblioGraph

EBSCO Information Services (EBSCO) is introducing BiblioGraph, a linked data technology that allows users to explore, use and access library catalogs from anywhere on the web. BiblioGraph is the next step in EBSCO's long-standing commitment to developing linked data-driven technologies and is a direct result of EBSCO's acquisition of Zepheira in 2020.

<https://www.researchinformation.info/news/ebSCO-information-services-introduces-bibliograph>

Source: *Research Information*

CAS launches CAS Insights, a content hub to track scientific innovation trends and opportunities

CAS has launched CAS Insights™, a new content hub at the intersection of science, technology, and innovation. Offering business and research leaders actionable perspectives on the latest developments across science and technology, CAS Insights draws on the human-curated data collection and deep scientific expertise from CAS to highlight emerging trends, unseen connections, new applications, and future opportunities across disciplines.

<https://www.knowledgespeak.com/news/cas-launches-cas-insights-an-content-hub-to-track-scientific-innovation-trends-and-opportunities/>

Source: Knowledgespeak

Exploring the World of Data Science: A Primer for Librarians

<https://newsbreaks.infotoday.com/NewsBreaks/Exploring-the-World-of-Data-Science-A-Primer-for-Librarians-154727.asp>

Source: Information Today Inc

Publisher Collaboration to showcase Research about AI

Kudos, the platform for showcasing research, has announced the launch of a new showcase and associated outreach campaign to help the public, media, industry, policymakers, educators and others understand the current and future role and capabilities of artificial intelligence.

<https://www.researchinformation.info/news/publisher-collaboration-showcase-research-about-ai>

Source: Research Information

Using Machine Learning to better understand how water behaves

New research uses machine learning models to better understand water's phase changes, opening more avenues for a better theoretical understanding of various substances. With this technique, the researchers found strong computational evidence in support of water's liquid-liquid transition that can be applied to real-world systems that use water to operate.

<https://www.sciencedaily.com/releases/2022/12/221216200854.htm>

Source: ScienceDaily

AI System not yet ready to help Peer Reviewers assess Research Quality

Machine-learning tool needs to be more accurate before it can replace or aid human assessment in the UK Research Excellence Framework.

<https://www.nature.com/articles/d41586-022-04493-8>

Source: Nature

Over 2000 journals share price and service data via Plan S's Journal Comparison Service

cOAlition S has announced that 27 publishers – who publish more than 2000 journals – have embraced the Journal Comparison Service (JCS) and shared their service and price data, responding to the call for transparent pricing of publishing services.

<https://www.knowledgespeak.com/news/over-2000-journals-share-price-and-service-data-via-plan-ss-journal-comparison-service/>

Source: Knowledgespeak

Partnership aims to enhance digital lab connectivity

Scitara has announced a partnership with Agilent to integrate its Scientific Integration Platform SIP with Agilent's Software and Informatics Division portfolio of products, including chromatography software and lab workflow management solutions. Scitara's SIP provides a connectivity solution in a cloud-native infrastructure

that allows scientific labs to realise the benefits of digital transformation. Data mobility plays a critical role as lab automation and workflow management continue to take centre stage in the digital laboratory debate.

<https://www.scientific-computing.com/news/partnership-aims-enhance-digital-lab-connectivity>

Source: *Scientific Computing World*

OCLC and Google now connect Web Searchers directly to Library Collections

Google now links search results directly to records of print books in hundreds of libraries using WorldCat data.

<https://www.infotoday.eu/Articles/News/Featured-News/OCLC-and-Google-now-connect-web-searchers-directly-to-library-collections-152690.aspx>

Source: *Information Today*

'Ignored and not appreciated': Women's research contributions often go unrecognised

Data reveal that to earn credit on scientific articles, women need to work harder than men.

<https://www.nature.com/articles/d41586-022-01725-9>

Source: *Nature*

Oncotarget is again on MEDLINE

Oncotarget was accepted again for indexing by MEDLINE. Oncotarget is now indexed by MEDLINE/PubMed and PubMed Central/PubMed.

<https://www.eurekalert.org/news-releases/969383>

Source: *AAAS EurekAlert!*

CAS introducing Transformational Capabilities to accelerate Life Sciences Innovation

CAS is collaborating with Chemotargets to leverage Clarity® to serve as its technology foundation for rapid development, with future investments driving integration and expansion to cover the end-to-end pharmaceutical R&D workflow and accelerate therapeutic innovation.

<https://www.knowledgespeak.com/news/cas-introducing-transformational-capabilities-to-accelerate-life-sciences-innovation/>

Source: *Knowledgespeak*

CAS unveils Innovation Incubator to assist early-stage scientific organisations

CAS is launching the CAS Innovation Incubator™ to assist early-stage scientific organisations accelerating new ventures, particularly those creating unique applications that could benefit future research and development. CAS will provide early-stage innovators with access to the CAS Content Collection™ and its exclusive technologies that reveal unseen connections among disparate data, as well as CAS expertise in mapping and managing intellectual property in multiple scientific disciplines, and in some cases, financial support. Eligible organisations may apply via the CAS Innovation Incubator portal. Eligible applicants should be start-ups seeking to advance scientific innovation solutions, or participants in the angel investing or venture capital space. Individual entrepreneurs in the commercial or academic sectors may also be eligible. Awards could be in the form of access to CAS solutions, and/or advice from CAS experts in key areas, such as scientific content management, AI, machine learning, etc. In some instances, funding may also be possible.

<https://www.knowledgespeak.com/news/cas-unveils-innovation-incubator-to-assist-early-stage-scientific-organizations/>

Source: *Knowledgespeak*

Ground-breaking Open Research Platform to Launch

Octopus aims to enable fast, free and fair publishing of research that is open to all, and which focusses on the intrinsic quality of research.

<https://www.researchinformation.info/news/ground-breaking-open-research-platform-launch>

Source: *Research Information*

How Can Chemists Make Their Data FAIR?

The National Research Data Infrastructure (NFDI; Nationale Forschungsdaten Infrastruktur) is an initiative by the German federal and state governments to create an interdisciplinary network that enables the sustainable handling of research data according to the FAIR guiding principles (findable, accessible, interoperable, reusable).

<https://www.chemistryviews.org/how-can-a-chemist-make-their-data-fair/>

Source: *ChemistryViews*

Open research platform, Octopus, Launched

Octopus aims to enable fast, free, and fair publishing of research that is open to all. The emphasis is on speed, openness, fairness, and ease of use, to prioritise the pure, intrinsic quality of research. Octopus will provide a primary research record for publishing and research as it happens. It will offer the research community a place to record full details of ideas, methods, data, and analyses to be peer-reviewed and assessed for quality. Octopus will allow faster results sharing with credit given to individual work at all stages of the research process – including peer review. Unlike a traditional research-publishing model, Octopus breaks down research publication into eight smaller modules or elements: Problem; Hypothesis/theoretical rationale; Methods/protocol; Data/results; Analysis; Interpretation; Real-world implementation; and Peer review. These elements are linked together to form branching chains, but each can be authored by different people.

<https://www.knowledgespeak.com/news/open-research-platform-octopus-to-be-launched-on-29-june-2022/>

Source: *Knowledgespeak*

ResearchGate and EDP Sciences announce Content Partnership

ResearchGate, the professional network for researchers, and EDP Sciences, an international academic publisher specialising in scientific, technical, and medical disciplines, have announced a content syndication partnership.

<https://www.eurekalert.org/news-releases/964592>

Source: *AAAS EurekAlert!*

ACS appoints Albert G. Horvath as new CEO

The ACS Board of Directors has selected Albert G. Horvath, Treasurer and CFO at ACS, as the Society's next CEO, effective January 1, 2023. He succeeds Thomas Connelly Jr., retiring after nearly eight years with ACS.

<https://www.knowledgespeak.com/news/american-chemical-society-appoints-albert-g-horvath-as-new-ceo/>

Source: *Knowledgespeak*

Infosys to Acquire BASE

Infosys has announced a definitive agreement to acquire BASE life science, a European technology and consulting firm in the life sciences industry. The acquisition reaffirms Infosys commitment to help global life sciences companies realise business value from cloud-first digital platforms and data, to speed-up clinical trials and scale drug development.

<https://www.scientific-computing.com/news/infosys-acquire-base>

Source: *Scientific Computing World*

Increasing visibility of research resources on bioRxiv and medRxiv

Identifiers for research resources are a new way to link papers and improve resource findability and reproducibility. A new project by SciScore, funded by the Chan Zuckerberg Initiative (CZI), will seek to implement a Resource table for each preprint.

<https://www.knowledgespeak.com/news/increasing-visibility-of-research-resources-on-biorxiv-and-medrxiv/>

Source: *Knowledgespeak*

Springer Nature becomes first publisher to partner with CiteAb - helping researchers with their experiments by streamlining access to high quality life science data

Integrating its Protocols and Methods Portfolio data into the CiteAb platform will add to the resources already available to help researchers make more informed decisions in the planning and carrying out of their experiments.

<https://group.springernature.com/gp/group/media/press-releases/partners-with-citeab/23440320>

Source: *Springer Nature*

Optimise your Academic Writing with new AI tool for Researchers

To enhance the often challenging manuscript writing process, [Paperpal](#) has introduced a new Microsoft Word add-in that provides real-time writing assistance for researchers. The tool uses state-of-the-art machine learning (ML) and AI models to make grammar, punctuation, style and readability suggestions while the user writes, ensuring academics can deliver a high-quality research manuscript. The tool is simple to use, minimises the need for long hours spent editing and proofreading manuscripts and reduces the risk of desk rejection.

<https://www.researchinformation.info/news/optimize-your-academic-writing-new-ai-tool-researchers>

Source: *Research Information*

A new Publishing Model from eLife: It's all about Peer Review

eLife decided to end accept/reject decisions following peer review, opting to emphasise the public peer review of preprints, restoring author autonomy and promoting the assessment of scientists based on what rather than where they publish.

<https://www.infotoday.eu/Articles/News/Featured-News/A-new-publishing-model-from-eLife-Its-all-about-peer-review-155655.aspx>

Source: *Information Today*