RSC INTEREST GROUP
CHEMICAL INFORMATION AND
COMPUTER APPLICATIONS GROUP

NEWSLETTER winter 2021-22

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.

Schoolchildren exploring how we can help tackle the climate challenge by playing the new Net Zero game at Catalyst Science Discovery Centre & Museum, Widnes, Cheshire.

Chemical Information & Computer Applications Group Websites

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

YouTube https://www.youtube.com/c/RSCCICAG

LinkedIn https://www.linkedin.com/groups/1989945/

https://Twitter.com/RSC_CICAG

Table of Contents

Contributions to the CICAG Newsletter are welcome from all sources - please send to the Newsletter Editor Dr Helen Cooke MRSC: 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

All of the planned CICAG activities for 2021 had to be converted to virtual meetings; whilst this gives access to a wider geographical audience, it does have a negative impact on the impromptu interactions that take place at physical meetings.

Social media has become an increasingly important way for communicating with members (and non-members) during lockdown, with Twitter (@RSC_CICAG) with over 1370 followers and LinkedIn with 510 followers gaining popularity. The importance of these platforms is evidenced by the fact that Tweets of the posters from AI in Chemistry meeting contributed to 77,000 Twitter impressions.

CICAG's YouTube channel now has 650 subscribers and contains the 13 video presentations from AI4proteins meetings in addition to all 16 of the Open-Source Tools for Chemistry Workshops. These workshop videos have proved to be very popular and at the time of writing have been watched nearly 13,500 times (for a full list see *RSC CICAG Open-Source Tools for Chemistry Workshops* on p. 53). In addition, the lightening posters from the AI in Chemistry meeting were also uploaded.

The workshops covered a wide variety of topics, from open-source applications like DataWarrior and ChimeraX, to cheminformatics tutorials, and open resources like the RCSB protein databank. The two workshops by the PDB staff were timed to coincide with the PDB 50th anniversary and we were delighted to join in the celebration. We are looking for potential suggestions for 2022.

The CICAG website is often updated and we would be very interested to hear suggestions for additional content.

CICAG has collaborated with the AI3SD Network+ to create the AI4 Proteins Seminar Series 2021 which culminated in the AI3SD & RSC-CICAG Protein Structure Prediction Symposium on 16-17 June. This two-day event brought together many experts to discuss and debate the current challenges in predicting protein structure. Wendy Warr's report of the event is available here.

RSC CICAG and RSC BMCS organised the 4th AI in Chemistry meeting which was held as a virtual event, and kept up the high standards of the previous meetings. A full report on the meeting is on p. 21. The 5th AI in Chemistry meeting is planned for 1-2 September 2022 at Churchill College, University of Cambridge.

CICAG were also delighted to support the Skills4Scientists Posters & Careers Symposium in collaboration with AI3SD, providing important early career advice.

This Newsletter also includes contributions from Connor W. Coley & Steven Kearnes, Kevin Theisen, Martin White, and Wendy Warr. CICAG is extremely grateful to all our contributors. Once again, I'd like to invite readers to suggest contributions that would be of interest to the CICAG community.

Whilst RSC members can join interest groups free of charge, 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 online.

- -

CICAG Planned and Proposed Future Meetings

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

Meeting	Date	Location	Further Information
Open-Source Software Workshops	2022	Virtual	An ongoing series of workshops. Details to follow.
5th Artificial Intelligence in Chemistry Meeting	1-2 Sep 2022	Virtual/TBD	Joint event from RSC-CICAG and RSC-BMCS division
CICAG/AI3SD Webinars	Autumn 2022	Virtual/TBD	Joint event from RSC-CICAG and AI3SD
Computational Tools for Drug Discovery	Nov 2022	TBD	Details to follow
Python for Chemists	TBD	TBD	Details to follow
Chemical Information during Covid	TBD	TBD	Details to follow
Ultra large chemical databases	TBD	TBD	Details to follow

- -

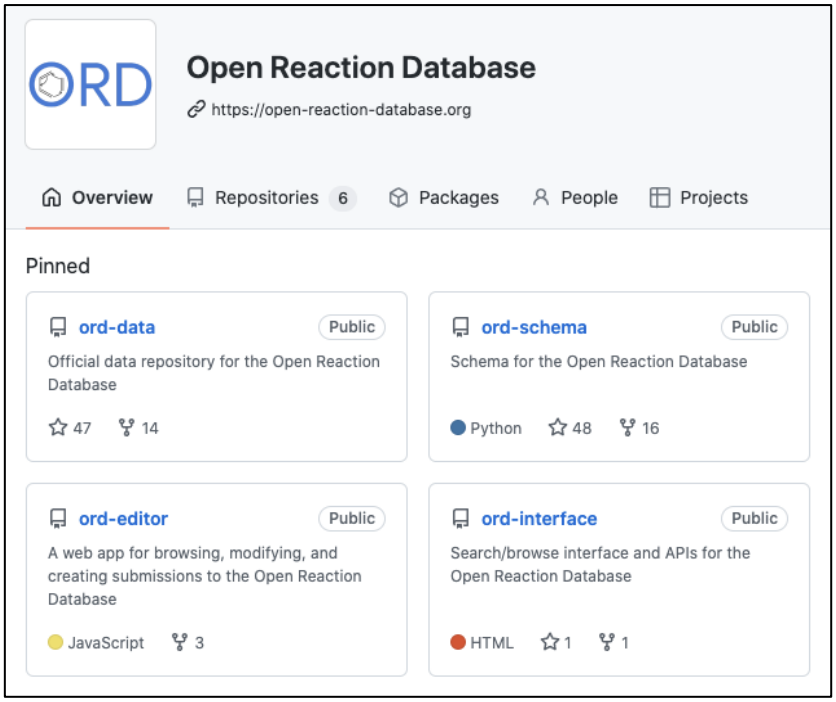
The Open Reaction Database

Contribution by Connor W. Coley, email: ccoley@mit.edu and Steven Kearnes, email: skearnes@relaytx.com

The Open Reaction Database (ORD) is a new open-access initiative to support the *structuring* and *sharing* of organic reaction data. Its primary goals are to support machine learning (ML) and related efforts in reaction prediction, chemical synthesis planning, and experiment design. We aim to:

- Provide a structured data format for chemical reaction data;
- Provide an interface for easy browsing and downloading of data;
- Make reaction data freely and publicly available for anyone to use; and
- Encourage sharing of precompetitive proprietary data, especially HTE data

The origins of the ORD connect back to the DARPA Make-It program, which sought to develop technology for accelerating the synthesis and manufacture of new organic molecules through a combination of computational planning and robotic execution. In several discussions within and beyond the program, it became clear to many of us that the current state of data sharing and standardisation in synthetic chemistry could be improved to accelerate algorithm development and the development of predictive tools to support the chemical sciences.



The infrastructure for the effort is organised into four parts: the schema that describes how chemical reactions are defined, a graphical editor for defining and downloading new reactions, a client for searching/viewing data with a database deployment, and the data itself. All four exist on GitHub under various open-source licences (Apache 2.0 for code, CC BY-SA for the data). All project development is done on GitHub and has been supported by volunteer effort so far, including generous contributions of time and computing resources from Google and Relay Therapeutics.

At present, most procedural details about chemical reactions are reported in unstructured supporting information documents as Word files exported to a PDF. That information, which includes even basic details of quantitative amounts of reactants, is not fully captured by current databasing efforts to our knowledge. The ORD schema is designed to capture the most important fields of a chemical reaction to support downstream ML tasks. It is written using Google's protocol buffer technology, so that the schema itself can be easily understood by users with little programming familiarity. In developing the schema, our goal was to address the most common types of contributions we expect to receive (based on a survey of roughly 170 users in late 2019). These types include single-step batch reactions as performed by hand or in a high-throughput experimentation workflow.

```
241  message Compound {
242    // Set of identifiers used to uniquely define this compound.
243    // Solutions or mixed compounds should use the NAME identifier
244    // and list all constituent compounds in the "components" field.
245    repeated CompoundIdentifier identifiers = 1;
246    Amount amount = 2;
247    ReactionRole.ReactionRoleType reaction_role = 3;
248    // Whether this species is intended to be a limiting reactant.
249    optional bool is_limiting = 4;
250    repeated CompoundPreparation preparations = 5;
251    message Source {
252      // Name of the vendor or supplier the compound was purchased from.
253      string vendor = 1;
254      // Compound ID in the vendor database or catalog.
255      string id = 2;
256      // Batch/lot identification.
257      string lot = 3;
258    }
259    Source source = 6;
260    // Compounds can accommodate any number of features. These may include simple
261    // properties of the compound (e.g., molecular weight), heuristic estimates
262    // of physical properties (e.g., ClogP), optimized geometries (e.g., through
263    // DFT), and calculated stereoelectronic descriptors.
264    map<string, Data> features = 7;
265    // Compounds may be assayed for quality control; analytical data should be
266    // defined in the analyses map.
267    map<string, Analysis> analyses = 8;
268  }
```

A snippet from the protocol buffer defining a reaction.

One of the ways to define data is through the graphical user interface, where users may provide all details of their reaction (including features like structure drawing and name lookups). It's also natural for the digital data record to be "born digital" and be defined programmatically. The schema can be compiled to Python classes, which in combination with numerous helper functions we have written, makes it easy to define hundreds or thousands of results (e.g., from high-throughput screening) programmatically. An enumeration script allows users to combine a template reaction with a spreadsheet to produce a full dataset of many reactions. Example Jupyter notebooks showing programmatic definitions are in the ord-schema repository, and a YouTube tutorial is available.

Additional details about the effort can be found in a recent Perspective published in the *Journal of the American Chemical Society* (DOI 10.1021/jacs.1c09820) and at the ORD website.

Our next goals are to continue to drive community adoption. We have a terrific team of collaborators on the governing committee and advisory board and have already received some institutional commitments from pharmaceutical companies to share high-throughput screening datasets. We are eager to work with additional synthetic chemistry groups from academia or industry to help them better understand the workflow for contributing a dataset and the overall value of the effort. We also intend to support translation scripts to make the reformatting from electronic laboratory notebooks (ELNs) to the ORD straightforward when there are overlapping structured fields.

Chemical Data Recovery 2: Chemical Image Recovery

Contribution from Kevin Theisen, President, iChemLabs, email: kevin@ichemlabs.com

This article is the second part of a three-part series on chemical data recovery written by Kevin Theisen, President of iChemLabs.

1. Embedded Chemical Data Recovery
2. Chemical Image Recovery
3. Legacy Chemical Data Recovery

We launched ChemDoodle 2D v11.4 on 2 April 2021. A new chemical image recovery function was included for automatically rebuilding chemical drawings from an image, which this article discusses in detail.

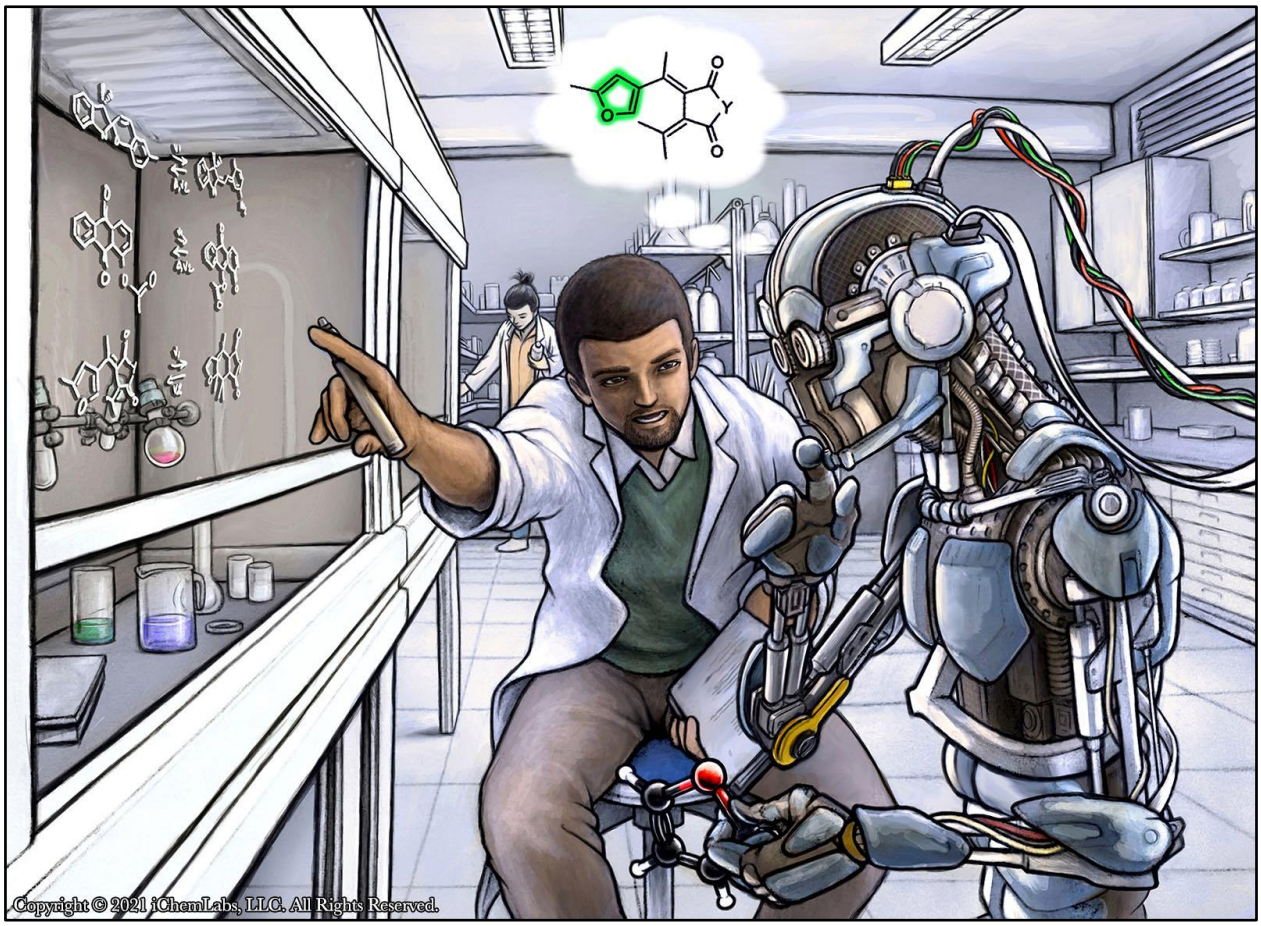


Figure 1. In this laboratory setting, an android is using its ability to see and understand molecule drawings and communicate with a scientist. New chemical image recovery features in ChemDoodle 2D make this future a possibility.

What is chemical image recovery?

When we communicate as chemists, we often use images of molecules because a picture is the most effective way to communicate information to visual creatures such as us. For well over a century, images of molecules have been created for use in documents, databases, notebooks, websites, etc. But an image is just an image; we cannot do anything more than look at it. Maybe we can enlarge it, copy it or print it, but the original chemical data it represents is only realised in our minds.

Chemical image recovery (CIR) is the process of taking an image of a chemical drawing, with no provided information other than the defined pixels, and using a computer to recreate the original chemical data to be used or edited further. For instance, take the following image of galanthamine. The image on the left is the input image, and the image on the right is the result of the CIR function in ChemDoodle 2D.

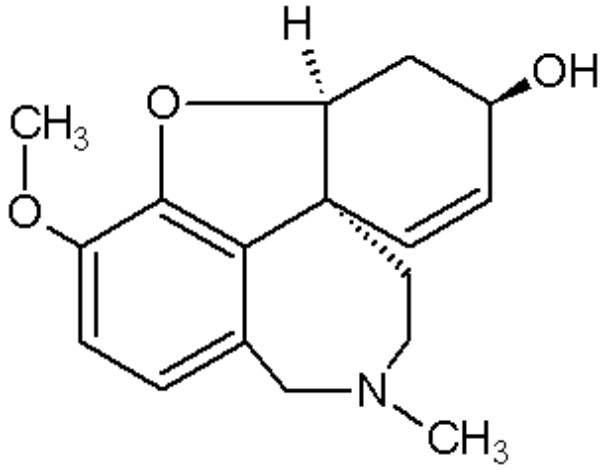


Figure 2a. An image of the molecule galanthamine.

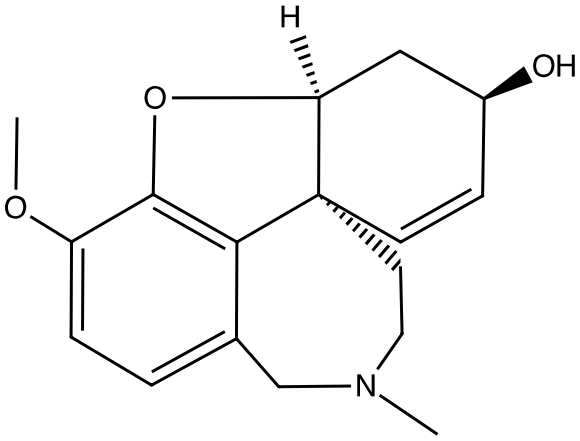


Figure 2b. The recovered chemical drawing of galanthamine.

The first impression may be, "Great! I now have a less blurry image." Yet, the result is much more significant. The actual chemical data, the arrangement of atoms and bonds, is digitised. We can now further process this information. For instance, we can produce a molecular formula, resolve the CIP stereochemical configurations, and change the graphical style to ACS 1996. We may even optimise the molecular structure in 3D and calculate a distance for the hydrogen bond. All of this output is easily produced from the result of the CIR function on the input image. Without the CIR function, a person would be required to redraw this molecular structure in a program to perform any computational chemistry task. One image is work by itself, imagine having to transcribe thousands of chemical drawings.

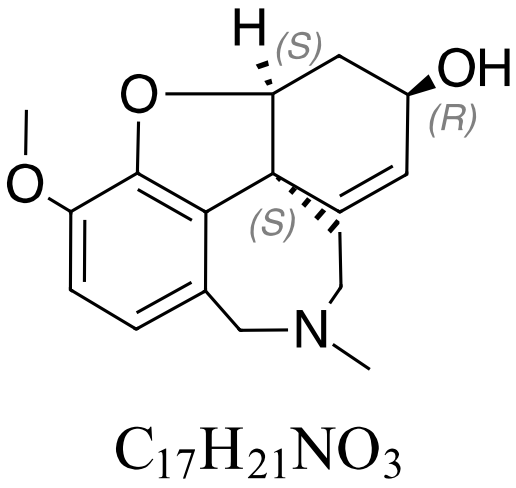


Figure 3a. The recovered chemical drawing is further processed; we are able to change styles, resolve stereochemical configurations and produce a molecular formula.

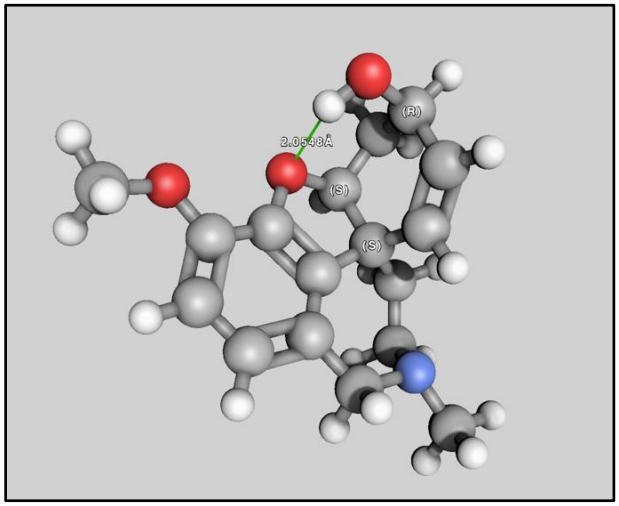


Figure 3b. Even further processing on the recovered chemical drawing, optimizing a 3D structure using the MMFF94 force field and measuring the hydrogen bond.

By recreating the chemical data originally lost in images, CIR makes it possible to produce many solutions for scientists. You can have a program automatically catalogue drawings from laboratory notebooks. Students can simply point their camera at a chemical structure on a poster and get the associated IUPAC name. An assistive tool can be produced to help vision-impaired chemists. A researcher can snap a picture of a molecule from a publication he/she is reading and immediately find more information from a chemical search engine. We may even be able to produce androids for our labs with the ability to observe and understand chemical drawings and then complement scientists so they can get their work done faster. The android could also protect the chemist if safety becomes a concern.

If you would like to try ChemDoodle's CIR function, you may use the **File>Recover from Image...** menu item in ChemDoodle 2D. You may also use the ChemDoodle Web Components CIR demo.

Background

Chemical image recovery is not a new concept. In fact, a few solutions already exist, known in the cheminformatics industry as Optical Structure Recognition (OSR) tools. I am not a fan of this name. It has been called OSR because an algorithm called Optical Character Recognition (OCR, for the computer recognition of character glyphs) already exists, and since we cannot call it Optical Chemical Recognition because the abbreviation is the same, “chemical” is replaced with "structure". This is unfortunate. If you say “Optical Structure Recognition” to a chemist, it will be very unlikely they know what you are referring to. “Chemical Image Recovery” allows a chemist to get closer to the meaning; we are attempting to recover the chemical information from an image.

Probably the earliest, most impactful literature for CIR algorithms is the Optical Recognition of Chemical Graphics paper out of IBM Almaden in 1993. Their work began in 1988 and they outline a general procedure for a CIR algorithm: you break down the input pixels into shape-based features, after which you build the chemical information from the ground up by interpreting those shapes. I call this procedural CIR. Since the IBM Almaden paper, many other procedural CIR solutions have come and gone. In 2008, NIH OSRA was introduced, providing an open-source (GPLv2+) CIR solution, and is the most well-known CIR solution today. More recently, machine learning (ML) has matured and become more applicable to solving problems. ML CIR may also be effective.

So, this brings us to the present, and at iChemLabs we are building our own CIR solution. Since OSRA already exists, some may ask “This is good, why reinvent the wheel?” I certainly encourage everyone to make use of open-source resources and contribute to those projects. At iChemLabs, we also produce open-source projects. But I think we can build something better in the ChemDoodle ecosystem where OSRA is not an ideal solution because (a) the licence is not compatible, and (b) OSRA is a C++ tool, while ChemDoodle is Java. I would also say, "If you want to create, just create!". Just have fun creating, and that is what iChemLabs excels at. In the next section, I go over our algorithm and how we are attempting to create the best CIR solution.

Implementation

We developed a procedural CIR tool. ML is certainly impressive, but it requires a high throughput of data to be generally applicable. Our algorithm should handle chemical images as generically as possible, without having seen a similar style of image. The procedural approach is optimal. Let's take a closer look at the example from the introduction.

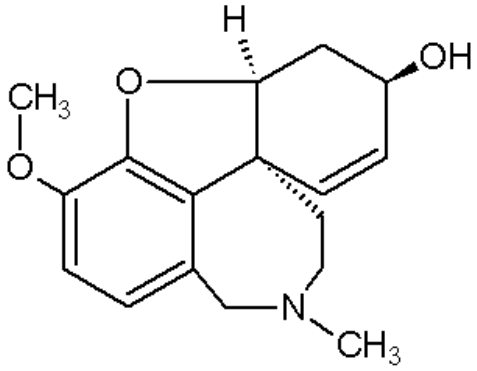


Figure 4. The input image of the molecule galanthamine.

The first step is to categorise and normalise the image, to recognise what must be done to understand the graphics as a chemical drawing. Screenshots of a chemical drawing from a publication at different scales need to all be handled differently and will be very different from a high-resolution picture of the same chemical drawing taken on your phone's camera. We invented a very creative solution allowing ChemDoodle to see the molecule in its entirety and make decisions about handling it from the image pixels before any processing begins, similar to how a human would be able to look at all these different images and recognise the molecule drawings within. The image is then intelligently normalised using image scaling, binarization and thinning functions.

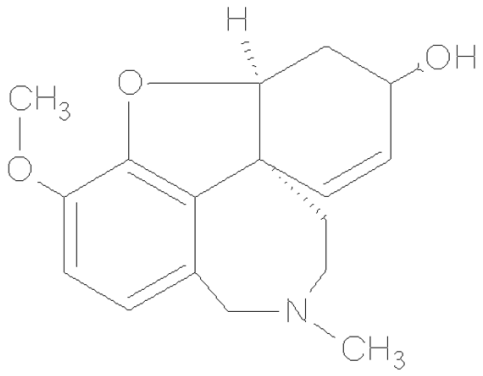


Figure 5. The result of analyzing the image and normalizing it for digestion.

A custom vectorisation is employed to break the features into shapes. Those shapes are then analysed to further break them down while grouping and categorising.

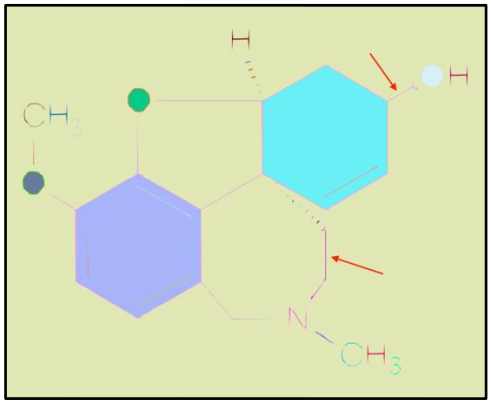


Figure 6. The pixels are separated and characterised into shapes.

All procedural CIR algorithms perform these steps, with some level of success. The remainder of the algorithm is the most important and ChemDoodle CIR excels here. The interpretation of the shapes is not trivial. Take a look at the two shapes pointed out by arrows. The top one looks like a fork coming off of a complex ring system and the bottom one looks like bent arm with a hand. How are these to be perceived? Our goal was to produce an algorithm to match how a human's mind would perceive a chemical drawing and mimic those decisions. Mathematical models are used for all of the interpretation, we do not rely on any arbitrary distance comparisons. Relying on distances is where most CIR algorithms fail, because there is no standard defined chemical structure distance, and every image may be unique in this aspect, from size and resolution, to atom label spacing, to bond thickness, to object congestion.

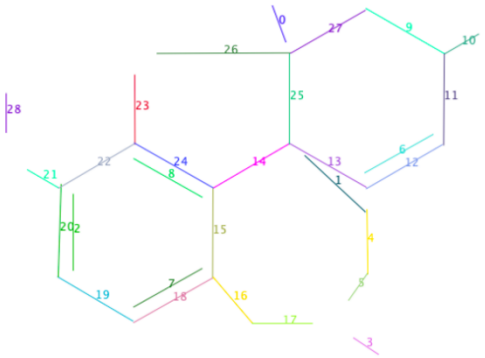


Figure 7a. The perceived bonds.

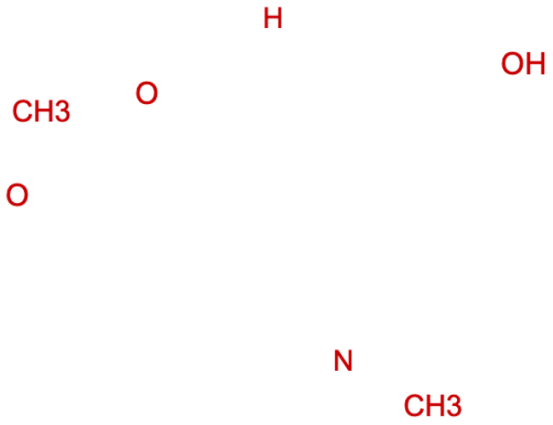


Figure 7b. The perceived atom labels.

Everything is pieced back together to recover the original drawing. The final structure is then analysed in a chemistry sense for any flaws. This is another step where the ChemDoodle algorithm excels, as ChemDoodle is one of the best and most thorough cheminformatics systems in existence.

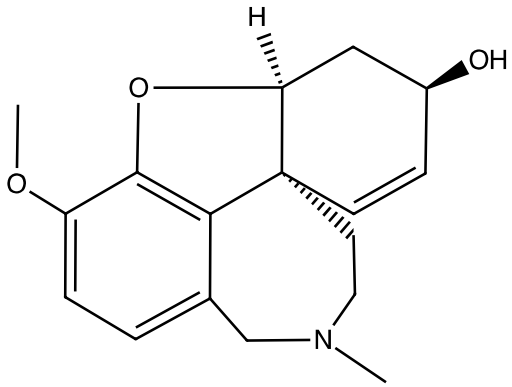


Figure 8. The recovered chemical drawing of galanthamine.

One last unique thing about our implementation. We wrote the entirety of the algorithm from scratch and did not rely on any 3rd party libraries. Image processing, thinning, scaling, normalization, vectorization, optical character recognition, our mathematical models, etc. were all developed in-house at iChemLabs. This allows us to specifically focus any part of the process to chemical drawing information. Another upside is the absence of any licence obligations and restrictions. Many existing CIR tools are dependent on Microsoft's proprietary OCR products. All you need to run ChemDoodle's CIR tool is ChemDoodle.

Performance

Our goal is to generically handle any image of a chemical drawing. To evaluate our success, we created a large testing suite of random images of chemical drawings from various sources: the internet, articles, books, cameras, software and more. Here is a selection of some of the varied images. ChemDoodle's CIR algorithm handles all these images perfectly.

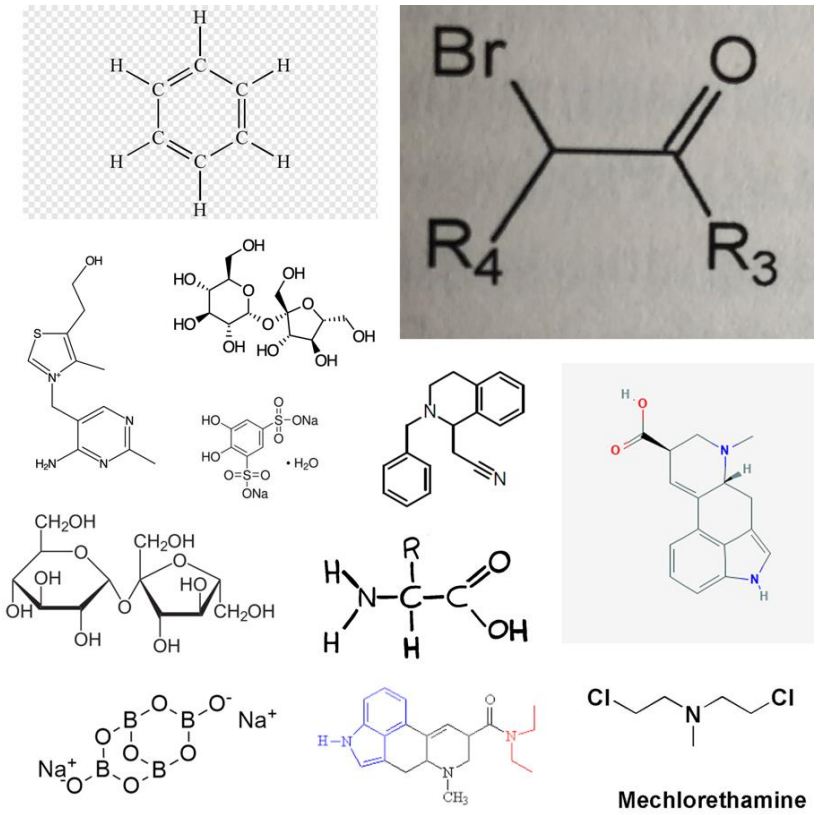


Figure 9. A sample of the testing suite variety used to evaluate ChemDoodle's CIR algorithm.

Our initial goal was to handle 100 of these random images perfectly, which we achieved. But we didn't want to just handle complex images, we also wanted to recognise simplistic edge cases other CIR tools would overlook. Take the following examples in Figure 10 overleaf. Again, ChemDoodle's CIR algorithm handles all of them perfectly.

To summarise, ChemDoodle is able to recognise complex atom labels consisting of elements, numbers, abbreviations, formulae and more. Text may be formatted including superscripts and subscripts. Bonds will be detected and overlapping bonds resolved, both where there is a break in the bond and those where the two bonds cross. Single through sextuple, as well as wedges and bold bond types are understood. Rings will be automatically created based on perceived atoms and bonds including the interpretation of aromatic circles. Charges, radicals and isotopes are recognised. ChemDoodle is also able to identify isolated non-chemical text and output it as a label shape. Multiple structures in the same image will be properly handled, even if they have different styles.

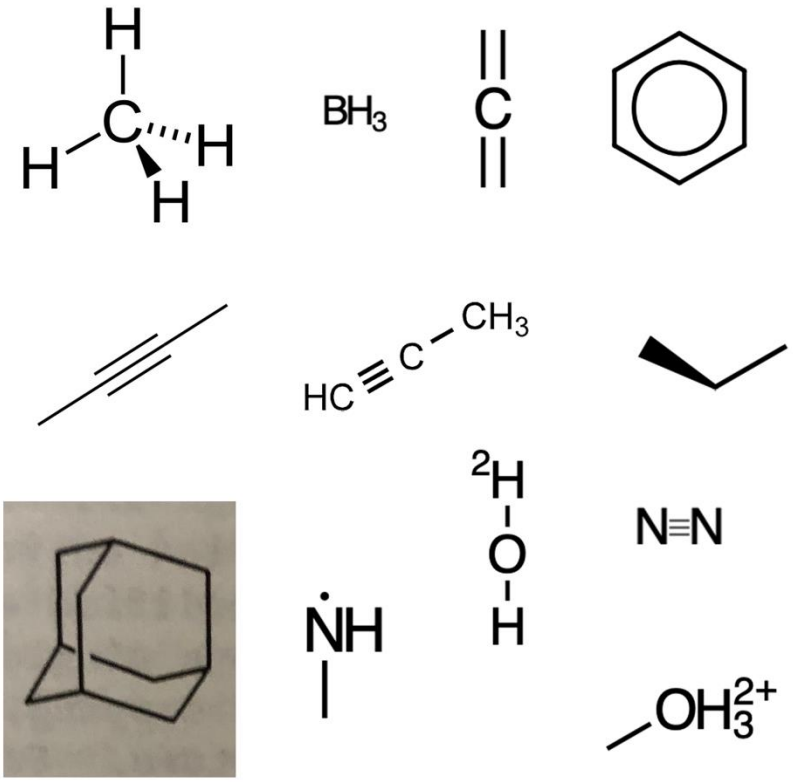


Figure 10. Simplistic tests often overlooked by other CIR algorithms.

Runtime is also an important consideration. ChemDoodle's CIR algorithm processes the galanthamine image with an average runtime of 76.4ms. The most time-intensive part of the algorithm is optical character recognition. The longest runtime we have found is for an image we created with 13 complex atom labels resulting in an average runtime of 115.1ms. So, expect the performance to scale with the amount of text in the image requiring recognition. All benchmarks were performed on a 2017 iMac running macOS 11.2.2 with a 4.2 GHz Quad-Core Intel Core i7 CPU. Each image was recovered 20 times, with the first iteration disregarded as a warm-up. The remaining 19 iterations were averaged. Java version 11.0.8 was used to compile and run the tests.

Finally, there are some limitations, as to be expected with a CIR project. CIR will work on computer-generated, skeletal images of chemical structures. Hand-drawn images may not work well. Clearer images at a crisp resolution will have the best results. The messier or blurrier the image is, the more ChemDoodle will have to use intuition to resolve the chemical structure, similar to a human looking at an unclear image. ChemDoodle may interpret graphics differently than you may. Our goal is to have the CIR features perfectly match what you perceive in the image, but you should expect to perform some level of post-editing on some images.

The future

I hope this discussion has provided an in-depth view into our current CIR work in ChemDoodle. The initial results are excellent and ChemDoodle's CIR out-performs many competing CIR solutions. We really hope this feature will help eliminate the effort you spend transcribing chemical structures from images. If you are not happy with the results, please send us the image so we may improve the algorithm. We will continuously develop it. As always, ChemDoodle subscribers, Site and Lifetime licensees are entitled to our latest

ChemDoodle features, and our customers will continue to benefit from our work. Thank you for your support, as you make these projects possible.

Moving forward, our focus includes the recognition of more bond types, the ability to dissect overlapping features (such as a bad graphic where a bond incorrectly intersects an atom label), and the perception of reaction arrows.

And if you are trying to create androids with the ability to see and perceive chemical drawings, please reach out. We would be happy to work with you! I am looking forward to seeing what our partners build with this capability.

- -

G. Malcolm Dyson (1902-1978), Chemist and Information Scientist

Contribution from Martin White, FRSC, FBCS, HonFCLIP, Visiting Professor, Information School, University of Sheffield, email: Martin.white@intranetfocus.com

George Malcolm Dyson 27 January 1948
Photograph by Bassano Ltd. CC BY-NC-ND 3.0. ©National Portrait Gallery.

1. Introduction

If you walk into the offices of any professional association, you will usually see a list of the Presidents of the association on a board in the main reception area. That is certainly the case with CILIP, where there are display boards listing Presidents from the foundation of the Library Association (LA) in 1877. In 2002 the LA merged with the Institute of Information Scientists (IIS) to create CILIP, but there is no similar record of IIS Presidents dating back to its foundation in 1958.

If there was such a list the name of the first President (1958-1961) would be G. Malcolm Dyson. It is doubtful if any information scientist would recognise the name, if only because the formation of the IIS is widely (and in most respects correctly) attributed to the efforts of Jason Farradane. When Sandra Ward, Charles Oppenheim

and I (all former Presidents of the IIS) compiled a history of the Institute in the course of the last three years, one of our objectives was to compile profiles of all the Presidents. However, we could find very little information about Dyson and his role in the establishment of the IIS. After his period as President, he seemed to play no further role in the IIS's activities. The only public source of information was a short entry in the German-language version of Wikipedia.

With the work on the history complete, I decided to research his life and achievements hoping to gain a better understanding of why he was invited to be President. Six months of research has resulted in an 18,000 word/140 citation profile of Dyson which will be released early in 2022 to mark the 20th anniversary of the formation of CILIP. This pre-publication summary only considers Dyson's work relating to chemical information science and hopefully illustrates how appropriate his achievements made him the ideal choice to be the first President of the Institute.

2. Oxford, Manchester and Loughborough

Malcom Dyson (he never used his initial given name of George) was born on 5 April 1902 in Plumstead, South London. Dyson went up to Jesus College, Oxford, on a full scholarship in October 1921. Jesus College was one of six Oxford colleges to have its own chemistry laboratory, with David Leonard Chapman FRS as Senior Fellow. Dyson not only gained a First in 1925 but records show that he also gained a First in chemistry in 1923, and then in 1925 a PhD from the University of London as an external student. The reasons for his sitting for degrees in both Oxford and London remain unclear.

From Oxford, Dyson immediately moved to Manchester to join his mother, who by this time had moved to Stockport with his brother John. He used the facilities at the recently established Laboratory of Applied Pathology and Preventive Medicine to continue his research into organometallic antimony compounds with chemotherapeutic potential.

In 1927 Dyson published his first major contribution to the chemical literature. This was a 57-page review of the chemistry of phosgene, with 189 references to research papers and patents dating back to the 1850s in English, French, German and Italian. At the same time, he was working on his first book, *The Chemistry of Chemotherapy*, which was published in 1928. John died from abdominal tuberculosis in 1927 aged just 21 and there is a touching reference to John thanking him for his work on the book. In total there are over 800 citations in these two publications, a substantial number of them to German research. The book runs to 260 pages and several hundred chemical structures and would have to have been either typed by Dyson himself or by a secretary from his manuscript. This must have been an immense task over a period of perhaps just three years.

In 1928 Dyson was invited to become a lecturer in the School of Pure and Applied Science at the Loughborough College. The College had emerged from the Loughborough Institute, which from 1915 was directed by Herbert Schofield. Schofield was a visionary educator who embarked on a project to create an engineering college and was Principal from 1915-1950. In 1932, aged just 30, Dyson was appointed Head of the School. He married his wife Bertha in 1930.

From 1934 onwards (whilst he was still at Loughborough College) Dyson continued his work in chemical synthesis, leading to patents assigned initially to British Dyestuffs and then to Parke Davis, at that time the largest pharmaceutical company in the USA. The focus of his work was again pentavalent antimonial compounds. Around this time Dyson was appointed as Technical Director to Genatosan, a German health care and pharmaceutical company that Fisons acquired in 1937, a position he retained until his appointment as Research Director at Chemical Abstracts Service in 1959.

3. Chemical nomenclature and the Dyson notation

From his authorship of *The Chemistry of Chemotherapy*, his decade as an academic and his research interests, Dyson was well aware of the challenges of creating logical and consistent names and structural diagrams for organic compounds. In addition, he had a very good understanding of the use of the chemical literature and of the needs of research chemists.

It is important to appreciate that Dyson regarded his work on notation as means of giving each chemical compound a unique 'cipher'. An example he gives is that $CSCl_4$ could be described as:

- Perchloromethyl mecapatan
- Thiocarbonyl tetrachloride
- Trichloromethyl sulfur chloride
- Tetrachloromethyl thiol
- Trichloromethyl sulfenyl chloride

To make matters worse there were British, French, German and American naming conventions.

The first announcement of what would become known as the Dyson Notation was a letter by Dyson dated 24 June 1944 and published in *Nature* on 22 July 1944. In the letter he mentions that he would be publishing a book on the systematic notation that he was developing. He stated the objective as establishing a database (though he did not use this term) of codes, each of which represented the structure of a unique chemical entity. The notation was based around determining and then supplementing the longest carbon chain.

The first public presentation by Dyson of his notation for organic compounds was at a meeting of the Royal Institute of Chemistry in 1946. The Institute was so impressed it circulated a copy of his lecture to its members. The first edition of his book *A New Notation and Enumeration System for Organic Compounds* was published by Longmans in 1947. Then on 3 February 1948 he gave a lecture to the British Society for International Bibliography that was reprinted in the inaugural issue of *Aslib Proceedings* along with the discussion which followed his presentation. A second edition of his book was published in 1949. The major change between the editions is a final chapter on the potential of punched cards for managing chemical information.

A number of notation schemes were in the process of being developed at that time, including the Wiswesser notation based on functional groups. During the 1950s there was a substantial amount of discussion and research about which of the nine schemes that had by then been published was the 'best' but no clear winner emerged. An important merit of Dyson's notation was the availability of a handbook for encoding and deciphering the notation. Although Wiswesser published his code in 1949 he did not write a user handbook until 1951. (Incidentally the Wikipedia entry on Wiswesser's notation claims it to be the first such notation.)

In the event, IUPAC, which had adopted the Dyson notation on a provisional basis in 1951, confirmed its adoption in 1961. However, by 1965 linear coding schemes were being overtaken by the connection-table approach to structural encoding developed initially by Du Pont.

An example of Dyson's notation is A6C135Q2, which represents l,3,5-trimethyl-2-hydroxycyclohexane (molecular formula $C_9H_{18}O$). [Dyson, G.M.; Riley, E.F. Mechanical storage and retrieval of organic chemical data. *Chemical & Engineering News Archive* **1961**, *39* (16), 72-77. DOI: 10.1021/cen-v039n016.p072.]

4. From chemist to information scientist

At the core of Dyson's interest was the importance of being able to access research and be certain about the identity and structure of organic compounds. His work on a notational scheme was one element of this interest. His first public presentation of his scheme in 1946 was entitled *Lecture on a new notation for organic chemistry and its application to library and indexing problems*.

The challenges facing the science community in accessing research information were discussed in depth at the landmark 1948 Royal Society Conference on Scientific Information, by which time the rapid growth in research publications was being recognised. Dyson was actively involved in the development of the Conference, in particular the exhibits area where there were two demonstrations of punched-card sorting equipment.

The publication in 1951 of Dyson's book *A Short Guide to the Chemical Literature* was very timely, and as far as I can determine was the first such book published in the UK. In September of the same year Dyson gave a paper on the preservation and availability of chemical knowledge at the XII International Congress of Pure and Applied Chemistry in New York. Probably the most remarkable aspect of this paper is his proposal for a code (in effect a semantic schema) for chemical reaction processes.

A second edition of his book was published in 1958 and Dyson contributed a paper on searching the older chemical literature to a collection of papers presented at the symposium on searching the chemical literature organised by the American Chemical Society (ACS) in 1961.

Working with Chemical Abstracts Service (CAS) 1947 - 1959

Dyson's correspondence with chemists in the USA during the preparation of the first edition of his book setting out his coding scheme was undoubtedly the reason for him being invited by the ACS to be the luncheon speaker at the 1949 Annual Meeting in San Francisco. This was a considerable honour, with Dyson being one of the few British chemists ever to have been invited to do so.

However, this would probably not have been Dyson's first visit to the USA. In the development of his notation Dyson had built up a friendship with James Perry, a highly respected chemist working in the library at MIT. Both could see the potential to manage chemical information using punched cards. This led to Dyson and Perry meeting with Thomas Watson, the President of IBM, though sources differ as to whether this meeting took place in 1948 or 1949. Watson was impressed with their vision and arranged for H.P. (Pete) Luhn to work with them on developing punched-card devices for information retrieval.

It seems that Dyson started to work as a consultant to CAS in 1956. He was present at the 1957 ACS annual conference in Atlantic City at which H.P. Luhn reported on his development of KWIC (Key Word in Context) indexes. Dyson and Perry both recognised the value of KWIC as a means of enabling chemists to scan the current literature.

CAS Director of Research 1959-1962

In 1959 Dale Baker, Director of CAS, hired Dyson to take over the position as Director of Research in succession to Karl Heumann. Dyson worked in periods of three months on site and then three months back in the UK. It is worth noting that at the time of his appointment Dyson was aged 57 and probably starting to think of retirement. Instead, he found himself taking on a huge technical and intellectual challenge working away from home for the first time in his career.

As soon as Dyson arrived in Columbus, he arranged a meeting with Luhn to begin a discussion about using the KWIC concept for a current-awareness product that would eventually be named *Chemical Titles*. This publication turned out to be a very considerable success for CAS and was the first computer-based current-awareness service. Three pilot issues were produced but the reception was so positive that the service was started after the second of these pilots. The technical development of *Chemical Titles* was only made possible by Dyson's insistence that CAS (at that time very short of investment funds) gain sponsorship from the National Science Foundation.

Dyson's vision for his notation was that it could form the basis of a chemical registry of unique compounds. However, it did not scale and it was not until several years later (and after Dyson had left CAS) that the Chemical Abstracts Registry was launched.

In 1961 Dyson was spending less time in Columbus and by 1962 Chemical Abstracts recognised that they needed a full-time director of research, a position that Dyson was not willing to consider as his wife did not wish to move to the USA. Despite only being with Chemical Abstracts for three years, he had made a very significant contribution to the organisation, providing the impetus to introduce computers more widely within CAS, which was subsequently managed by Fred Tate.

His position was taken by Karl Zabriskie, with Michael Lynch eventually being appointed as Head of the Basic Research Department. Lynch returned to the UK in 1965 to become a teacher and researcher in what was then the Postgraduate School of Librarianship and Information Science (now the Information School) at the University of Sheffield, a post that he held until his retirement in 1995.

5. IIS Founding President 1958–1961

Dyson and Farradane probably met for the first time in the build-up to the 1948 Royal Society Conference and then at the conference itself. One of the lasting outcomes of the conference was the appreciation of the need to train information workers in the skills of information management, with a recommendation that Aslib and the LA should work together to develop courses. This failed to take place and gave Farradane the opportunity to establish the IIS with a strong emphasis on professional accreditation and training.

In the period from 1956-1958 Dyson was still working in the UK and was widely recognised for his work on chemical notation and on handbooks on the use of the chemical literature. Dyson's support would have given the initiative a great deal of credibility, and it is of note that many of the early presidents were chemists and undoubtedly aware of the contributions that Dyson had made during his career. In 1961 Dyson and Farradane co-authored two papers at the ACS Annual Meeting, one on the aims of the IIS and the other on the initial development of the Institute's criteria. It has not been possible to determine if Farradane attended the meeting.

6. The final chapter

Dyson continued to publish research papers (several co-authored with CAS staff) up to 1968. In his retirement he remained closely engaged with Loughborough University and was awarded an Honorary DSc in 1972. He died on 27 December 1978. His wife Bertha died in 1984. They had no children.

7. On reflection

Michael Lynch, who worked with Dyson and Fred Tate from 1961, gives a valuable perspective on the role of Dyson at CAS.

"I had the good fortune to work closely with Malcolm Dyson during my apprenticeship years in this area, also with Fred Tate – characters as different as you could wish, each contributing in major ways to information

science: Dyson through his vision of what the emergent technology could mean for a chemical information database spanning all of the years, and Tate's preoccupation being with using the technology to get the show on the road, to continue to provides services which had outgrown printer's ink and leading in time to a cumulative database. Each was right in his own vision of the future, differing mainly about the means by which they were to be achieved. Dyson and Tate stood above the technology of their time – the fact that they started with an 8K IBM 1401 did not limit their thinking – rather, they were confident that when the need arose the technology would arise to the occasion."

In the last two decades of his life Dyson provided CAS with a vision of how computers could be used to create indexes of organic chemicals, developed the first computer-based current-awareness service, wrote two books on searching the literature of chemistry, made a major contribution to the success of the 1948 Royal Society Conference, supported Jason Farradane in setting up the Institute of Information Scientists and (albeit indirectly) led to the establishment of what would become a world-respected cheminformatics research group at the University of Sheffield.

In his transition from a highly regarded research chemist to these achievements in chemical information management Dyson was arguably the first 'information scientist'.

Acknowledgements
Robin Darwell-Smith (Archivist, Jesus College, Oxford), Evan Hepler-Smith (Andrew W. Mellon Assistant Professor of History at Duke University, Durham, North Carolina), David Allen (Librarian, Royal Society of Chemistry), Andrew Dalke (Dalke Scientific Software AB) and Peter Willett (Emeritus Professor, Information School, University of Sheffield) have made substantial contributions to my research.

Citations to papers authored or co-authored by Malcolm Dyson
This list includes only papers authored or co-authored by Dyson in the area of what might now be referred to as cheminformatics. There are also research papers, patents and a number of chemistry monographs, most of which can be found on Google Scholar under GM Dyson.

1944	G.M. Dyson**.** A notation for organic compounds. *Nature* 1944, *154*, 144.
1946	G.M. Dyson**.** Lecture on a new notation for organic chemistry and its application to library and indexing problems. Delivered at a joint meeting of the Chemical Society, the Royal Institute of Chemistry, the Society of Chemical Industry and the Bureau of Abstracts at the London School of Hygiene and Tropical Medicine, 21 October 1946.
1947	G.M. Dyson**.** A New Notation and Enumeration System for Organic Compounds, 1st ed.; Longmans, Green & Co.: London, 1947.
1947	G.M. Dyson**.** Application to medicinal chemicals of a new notation for organic compounds. *Manuf. Chem. Aerosol News* 1947, *18* (3), 109-114.
1948	G.M. Dyson**.** Classification in science – exhibits presented at the Royal Society Information Conference, 1948. Conference report, pp 228, 230.
1949	G.M. Dyson. A new notation and enumeration system for organic compounds, 2nd ed.; Longmans, Green & Co.: London, 1949.
1949	G.M. Dyson**.** International chemical abstracts and the new notation for organic chemistry. *Aslib Proceedings* 1949, *1* (1) 5-21 (paper given on 3 February 1948).
1949	G.M. Dyson. Some applications of the Dysonian notation of organic compounds. *J. Chem. Educ.* 1949 (June 1), *26* (6), 294, https://doi.org/10.1021/ed026p294.
1951	G.M. Dyson. A Short Guide to the Chemical Literature, 1st ed.; Longmans, Green & Co.: London, 1951.

1952	G.M. Dyson. The preservation and availability of chemical knowledge. *J. Chem. Educ.* 1952, *29* (5), 239. (Presented at the XII International Congress of Pure and Applied Chemistry, New York, September 1951.)
1955	G.M. Dyson. Advances in classification. *J.Doc.* 1955, *11* (1) 12-18. (Presentation to an Aslib meeting 17 December 1954.)
1958	G.M. Dyson. A Short Guide to the Chemical Literature, 2nd ed. Longmans, Green & Co., 1958.
1961	G.M. Dyson**.** Current research at Chemical Abstracts. *J. Chem. Doc*. 1961, *1* (1), 24–28. (13 September 1960.)
1961	G.M. Dyson. Searching the older chemical literature. *Advances in Chemistry 4*, Chapter 15, 96-103.
1961	G.M. Dyson**.** Indexing scientific progress by computer. *New Scientist,* 1961 (30 March), *2283,* 817-819.
1962	G.M. Dyson & E.F. Riley. Use of machine methods at Chemical Abstracts Service. *J. Chem. Doc.* 1962, *2* (1), 19-22.
1962	C.L. Bernier, G.M. Dyson & H.J. Friedman. Correlative Indexes VII. Trope vocabularies and trope indexes for chemistry. *J. Chem. Doc*., 1962, *2* (2), 93–102.
1962	G.M. Dyson & J.E.L. Farradane. Education in information work: the syllabus and present curriculum of the Institute of Information Scientists Ltd. *J.Chem Doc.,* 1962, *2,* 74-76. (Presented before the Division of Chemical Literature, ACS National Meeting, Chicago, Ill., September, 1961.)
1962	G.M. Dyson & J.E.L. Farradane, The aims of the Institute of Information Scientists Ltd. *J.Chem.Doc* 1962, *2*, 72–74, https://doi.org/10.1021/c160005a008. (Presented before the Division of Chemical Literature, ACS National Meeting, Chicago, Ill., September, 1961.)
1962	G.M. Dyson**.** The detection and orientation of substructures in organic compounds, Defense Technical Information Center, 1962, Accession Number AD0274358 (no longer accessible).
1963	R.R. Freeman & G.M. Dyson**.** Development and production of *Chemical Titles*, a current awareness index publication prepared with the aid of a computer. *J. Chem. Doc.* 1963, *1*, 16-20.
1963	G.M. Dyson**,** W.E. Cossum, M.F. Lynch & H.L. Morgan. Mechanical manipulation of chemical structure: Molform computation and substructure searching of organic structures by the use of cipher-directed, extended and random matrices. *Inform. Stor. Retr*. 1963, *1*, 67-99.
1964	G.M. Dyson. A cluster of algorithms relating the nomenclature of organic compounds to their structure matrices and ciphers. *Inform. Stor. Retr*. 1964, *2*, 159-199.
1964	G.M. Dyson**.** Generic (or Markush) groups in notation and search programs, with particular reference to patents. *Infom. Stor. Retr*. 1967, *2*, 59-71.
1967	G.M. Dyson**.** Computer input and the semantic organization of scientific terms. *Inform. Stor. Retr*. 1967, *3*, 35-115.
1968	G.M. Dyson**,** M.F. Lynch & H.L. Morgan. A modified IUPAC-Dyson notation system for chemical structures. *Inform. Stor. Retr.* 1968, *4*, 27-83.

- -

Meeting Report: 4th RSC Artificial Intelligence in Chemistry

27 – 28 September 2021

A report by Dr Wendy A Warr, https://www.warr.com/

Introduction

The ("virtual") symposium was organised by the Royal Society of Chemistry's Biological and Medicinal Chemistry Sector (RSC BMCS) and the Royal Society of Chemistry's Chemical Information and Computer Applications Group (RSC CICAG).

AI for molecular design, past, present and future

Ola Engkvist, AstraZeneca, Gothenburg, Sweden

AI-based drug design can reduce the time to deliver a clinical candidate, by helping chemists select the most efficient synthetic route (increasing speed), and making information-rich compounds in each design, analyse, make, test (DMTA) cycle (maximising learning). This could not have been done five years ago but increased computational power, advances in neural network (NN) algorithms, and the availability of open-source software have now made it possible.

We can take advantage of progress in natural language processing (NLP) by representing molecules as SMILES. NLP can then be used in synthesis prediction and molecular optimisation, and text generation can be used in chemical-space exploration. We can move from rule-based models to data-driven ones. AI-generated ideas from the whole relevant chemical space can be used for scaffold hopping and hit finding. Fast molecular optimisation is possible through AI-designed libraries. Progress in free-energy perturbation affinity prediction improves scoring of AI-generated molecules. Better prediction of synthetic routes is possible through new algorithms and there are novel and more flexible ways of predicting molecular properties.

Single-layer NNs have been used in modelling of Quantitative Structure-Activity Relationships (QSAR) for years, but recent applications use more complex networks such as multi-layer, feed-forward NNs, convolutional NNs, auto-encoder NNs, and recurrent neural networks (RNNs), trained using maximum likelihood estimation to maximise the likelihood of the next character. Recurrent NNs sample the whole chemical space in hit finding and scaffold hopping. A focused chemical space can be sampled with a transformer for molecular optimisation. Generative AI in pharma is still on the ascent in the Gartner Hype Cycle for Artificial Intelligence, 2021 and will peak in 2-5 years.

The chemical space for a file of size 41 GB is 10^9 structures if it is traditionally enumerated, whereas generative models can sample practically unlimited chemical space. They do not contain any explicit molecules but generate them probabilistically. An RNN learns the rules of chemistry, not the training examples. The trained RNN can then generate druglike molecules: SMILES are sampled and a probability distribution for each token (character) is used to generate a physicochemical property, or a structure (in which case, training is harder).

In a collaboration with Jean-Louis Reymond's team, Engkvist and his co-workers explored whether it is possible to show that a deep learning based molecular generator is sampling the whole relevant chemical space and only that chemical space. They trained an RNN with a subset of SMILES from the enumerated GDB-13 database of 975 million molecules. They showed that a model trained with 1 million structures reproduces 68.9% of the entire database after training, when sampling 2 billion molecules. An analysis of the generated chemical space showed that complex molecules with many rings and heteroatoms are more difficult to sample.[1]

The teams then performed a benchmark on models trained with subsets of GDB-13 of different sizes with different SMILES variants, recurrent cell types, and hyperparameter combinations.[2] New metrics were developed that define how well a model has generalised the training set. The generated chemical space was evaluated with respect to its uniformity, closedness and completeness. The results showed that models that use long short-term memory (LSTM) cells trained with 1 million randomised SMILES are able to generalise to larger chemical spaces than the other approaches and they represent more accurately the target chemical space.

Using reinforcement learning (RL), an RNN can be tuned to target a particular section of chemical space with optimised desirable properties using a scoring function but ligands generated by some RL methods tend to have relatively low diversity, and sometimes even result in duplicate structures. Engkvist's team has developed a new method to address this issue: memory-assisted RL introduces a memory unit and a scaffold penalty assures that diverse scaffolds are identified.[3]

Mixed improvements with novel deep learning methods have been reported: there has been no "AlphaFold moment" in blind bioactivity prediction competitions. Gradient descent NNs are approximately kernel machines. Large improvements would imply a novel way of assessing molecular similarity. Pre-training can improve prediction capacity. The data used are more important factors than molecular representation and machine learning (ML) algorithms. Uncertainty quantification and interpretability have to be considered. Most models in the future are likely to be based on deep learning because of their flexibility.

The machine learning ledger orchestration for drug discovery (MELLODDY) project aims, over three years, to enhance predictive ML models on the decentralised data of 10 pharmaceutical companies, without exposing proprietary information. A multi-task approach across partners aims to improve predictive performance and applicability. Compound and activity data and assay-specific models remain locked on the server of the pharma company that owns them. Lower-level model components are securely exchanged and trained over the network. Pre-agreed access arrangements are strictly enforced. In year two, a study showed that multi-partner modelling yields superior predictive models in drug discovery.

The MegaMolBART drug discovery model being developed by NVIDIA and AstraZeneca will be used in reaction prediction, molecular optimisation and *de novo* molecular generation. It is based on AstraZeneca's MolBART transformer model and is being trained on ZINC[4] using NVIDIA's Megatron framework to enable massively scaled-out training on a supercomputing infrastructure.

Engkvist's team have demonstrated the utility of a 3D shape and pharmacophore similarity scoring component in molecular design with a deep generative model trained with reinforcement learning (REINVENT).[5] Using dopamine receptor type 2 (DRD2) as an example and its antagonist haloperidol 1 as a starting point in a ligand-based design context, they have shown in a retrospective study that a 3D similarity-enabled generative model can discover new leads in the absence of any other information. It can be used for scaffold hopping and generation of novel series. 3D similarity-based models were compared against ones based on 2D QSAR, indicating a significant degree of orthogonality of the generated outputs, with the former having a more diverse output.[6]

A major obstacle of generative models is producing active compounds in which predictive QSAR models have been applied to enrich target activity. QSAR models are inherently limited by their applicability domains. A structure-based scoring component for REINVENT overcomes these limitations. DockStream[7] is a flexible, stand-alone molecular docking wrapper that provides access to a collection of ligand embedders and docking back-ends.

Nevertheless, AI alone cannot transform drug design. High-throughput data generation, automatisation in the DMTA cycle, and combining AI with physics (e.g., to predict physicochemical properties and estimate binding affinity) can add value to AI approaches. Combining AI with big data can transform synthesis prediction.[8] In AstraZeneca, chemists have access to data on 20 million reactions in the ReactionConnect database, from which predictive models can be built and used to automate synthesis. ReactionConnect is populated with data from AstraZeneca reaction sources and ELNs, a USPTO database, and Reaxys and Pistachio flat files.[9] AiZynthFinder can be used in retrosynthetic planning. The algorithm is based on a Monte Carlo tree search that recursively breaks down a molecule to purchasable precursors. The tree search is guided by an artificial neural network policy that suggests possible precursors by using a library of known reaction templates.[10] A "Ring Breaker" algorithm[11] improves the route-finding. It uses a data-driven approach to enable the prediction of ring-forming reactions, useful in establishing the synthetic accessibility of unprecedented ring systems. Another improvement, RAscore,[12] is an ML-based method able to classify whether a synthetic route can be identified or not for a particular compound.

Engkvist summarised the lessons that AstraZeneca has learned. The needs of workers in discovery chemistry and process chemistry are very different. Extracting and integrating reaction data is hard work. It is challenging to assess the utility of different tools such as advanced building block look-up. The impact on AI approaches on synthetic routes has mainly been from specialised tools such as Ring Breaker. Software from the Molecular AI department at AstraZeneca is openly available. iLAB is AstraZeneca's automated synthesis platform.

Engkvist is optimistic about the future of AI in drug design because of increased computational power, increased automation which provides large and consistent datasets, and advances in computational algorithms such as those that merge physics-based modelling and ML. Metrics such as time-saving cannot be used to measure success because they are the results of success not the success itself. Success can be measured by trust in the AI-designed molecules in the same way as, for instance, X-ray crystal structures are trusted. There must be trust in the predictions for individual molecules and trust that the AI-generated molecules are the best molecules to take the project most efficiently to a clinical candidate.

There remain some challenges for AI-driven drug design. They include scaling ML and AI solutions for drug design to a whole drug discovery project portfolio including projects with low data volume. Binding affinity and solubility predictions are major bottlenecks. The "Cambrian revolution" of new AI methods makes it difficult to assess progress. Flexibility of chemistry automation is another challenge. There are also educational, cultural and logistical challenges besides scientific ones. The bar is set high to transform drug design.

Driving lead optimisation with BRADSHAW

Ian Wall, Richard Lonsdale, David Marcus, Darren Green, Stephen Pickett, David Hirst, GlaxoSmithKline (GSK), Stevenage, UK

A *de novo* design program generates molecular structures which satisfy a set of constraints. Classic problems with *de novo* design algorithms are nonsense structures, structures with intrinsic liabilities, and structures that cannot be made. Biological Response Analysis and Design System using an Heterogenous, Automated Workflow (BRADSHAW), GSK's automated molecular design platform (Figure 1),[13] takes a dual approach, using cheminformatics methods to generate plausible structures based on what has been done before,[14-17] and deep learning algorithms trained on relevant GSK chemistry space including novel methods.[5,18]

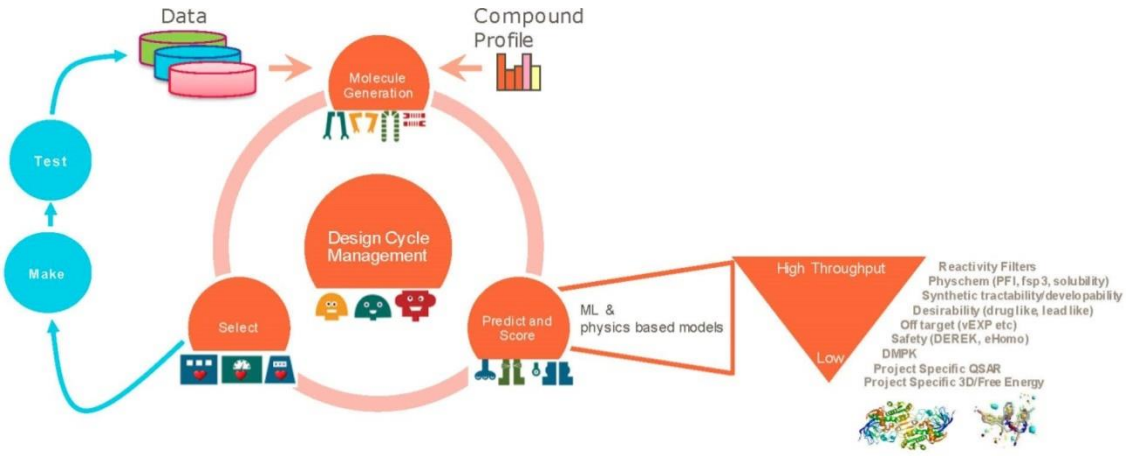


Figure 1. GSK's BRADSHAW.

A GSK team has reported[19] three Turing-inspired tests designed to evaluate the performance of three molecular generators: BioDig, a matched molecular pair-based algorithm,[16] BRICS (a fragment replacement based algorithm),[15] and RG2Smi,[18] which translates a reduced graph input to a SMILES output. BioDig performed excellently against all tests.

Currently, BRADSHAW is limited to cheminformatics and ML models. There are no 3D or docking methods, or physics-based methods such as free energy perturbation (FEP+),[20] but they can be included in a design workflow as an additional step. A multi-parameter optimisation (MPO) approach is used, in automated workflows, to design molecules with a balanced profile. MPOs can be built for predicted values and confidence in them, allowing an active learning approach with algorithmic definition of "explore and exploit".

Wall presented a case study in an active drug discovery programme. The process maximised efficiency by moving the synthetic chemistry resource between two series, allowed updating of models with new data whilst chemists moved onto alternative series, and minimised the number of compounds made without information from previous compounds. The computational chemistry workflow began with molecular generation from seed compounds, followed by building, filtering and rebuilding QSAR models, docking and scoring, and removal of undesirable compounds (by medicinal chemists). FEP calculations were then carried out and the data were collated in Spotfire for review by medicinal chemists. More than 2 million molecules were generated, 2822 FEP calculations were made, and 38 local models were built, in over more than 30,000 GPU hours.

The technologies used in BRADSHAW are modern ML, active learning, gated recurrent unit cell recurrent neural network (GRU RNN, a new molecule generator which increases the ability to make changes at multiple positions, giving better coverage of chemists' ideas),[21] BRICS, BioDig, Matsy, and RG2smi. In addition, library enumeration, Free-Wilson analysis, pK_a prediction, MetaSite, and protein-ligand interaction fingerprints are used.

Chemists selected compounds for synthesis and viewed their profiles against a range of parameters. This technique was used in conjunction with an active learning explore-exploit plot, where with MPO score on the x axis and MPO confidence on the y axis, the top right quadrant is compounds for exploitation and the bottom right quadrant is compounds for exploration. Wall showed graphs illustrating the rapid increase in "zero-risk" compounds for the two series since BRADSHAW was introduced in February 2020. The successful outcome of this pilot project was two leads with *in vivo* activity. Wall displayed some of the interesting range of structures

(core and R-groups) resulting from exploration of the chemical space, showing some simple structures but with different R-groups and complexity starting to appear.

Close interactions were needed among computational, and medicinal and synthetic chemists, including those in high-throughput chemistry (HTC), to get maximum value from the technology. Many other functions were also essential. Wall outlined some pros and cons from the medicinal chemist's viewpoint. From molecule generation, interesting, novel ideas were produced, with a good synthetic success rate, but matched molecular pairs were lacking and there were incomplete enumerations. Scoring and selection were an improvement over the subjective methods used previously, but robustness of pharmacokinetic (PK) predictions and inefficiency in selection meetings were cons. Iterative cycles provided focus but lack of design input, freedom to explore, and medicinal chemistry intuition were criticised. There was an excellent working relationship between medicinal chemists and computational chemists. Unfortunately, data generation has been challenging and restrictive.

Learnings from this pilot project are driving improvements in the system. GRU RNN, improved structural filters and visualisation have been implemented. So has DISCONNECT dHTCscore, a system to identify automatically compounds that are synthesisable from available reagents and possible arrays. Medicinal chemists and computational chemists working together have learnt a huge amount about logistics, technology and ways of working.

Efficient ML strategies to explore chemical reactivity
Fernanda Duarte, University of Oxford, UK

The Duarte group have applied computational methods to design new catalysts and study reaction mechanisms. Their open-source tool, cgbind can be used to generate and analyse metallocage structures.[22] Another tool, autodE is an open-source Python package capable of locating transition states and minima and delivering a full reaction energy profile from 1D (SMILES) or 2D chemical representations.[23] It combines graph theory and chemical knowledge in order to reduce the size of the chemical space required for sampling. It is compatible with multiple electronic structure packages, is broadly applicable and requires minimal user expertise.

Realistic simulations of chemical or biochemical reactions require the inclusion of the chemical environment where they occur (e.g., solvent and/or enzyme). Two main approaches have been historically used to account for these complex environments. The first is empirical reactive force fields (e.g., EVB), in combination with molecular dynamics (MD) or Monte Carlo (MC) simulations, which sample a reaction's potential energy surface but are limited in accuracy and transferability. Second are *ab initio* and quantum mechanics/molecular mechanics (QM/MM) which are accurate but computationally costly. ML force fields have the potential to revolutionise force-field based simulations, aiming to provide the best of both worlds.

Duarte's team[24] has used the Gaussian Approximation Potential (GAP)[25-27] framework with smooth overlap of atomic positions (SOAP)[25] descriptors to generate inexpensive potentials for solution phase reactions. GAPs have been applied to organic molecules,[28] and elemental materials[27,29] but this was the first example demonstrating its use to study chemical reactions.

Starting with solute and solvent structures, they developed a training strategy and devised a prospective error metric to assess the accuracy of the potentials. Active learning, where new training data are added based on the current state of the potential, is used for generating databases and accelerating the fitting process. The strategy

used by Duarte's team starts from a small number of randomly selected points in the configuration space, from which active learning training of intra- and inter-molecular components of the energy and forces is carried out. The CUR algorithm[27,30] is applied.

Splitting the database into training and test sets and using a standard retrospective validation strategy is not practical in the current application so a temporal cumulative error metric was used based on the time required for the cumulative error to exceed a given threshold. This does not require *a priori* knowledge of the region of configuration space likely to be sampled during a simulation with the potential. The user can specify an acceptable margin of error. The method samples regions not accessible to direct evaluation, ensures stable dynamics, and penalises large errors resulting in instabilities.

For bespoke ML potentials to be routinely developed for molecular systems, one would hope to complete the data generation and model training, and know the accuracy of the resulting potential within a matter of hours to days. With this in mind, the team trained GAP models to simulate bulk water, aiming to minimise the number of required ground truth evaluations as well as the required human intervention, while maximising stability (measured by the new prospective error metric). Only when the relevant length and energy scales of the system are decomposed by treating intra- and inter-molecular components separately was it possible to obtain a potential that is stable for picoseconds.

The model fitted using this approach yields radial distribution functions (RDFs) in good agreement with the ground-truth method, considering both the location and intensities of the peaks corresponding to the first and second coordination shells. The real significance is in moving to more accurate ground-truth methods, for which a full MD simulation would not be straightforward: indeed, using the same method, a hybrid DFT-quality water model can be generated within a few days, which would be inaccessible with other methods. The results suggest that the training strategy (and hyperparameter selection) is suitable independent of the reference method.

To demonstrate the transferability of the models, Duarte briefly presented results of successful application to aqueous Zn(II); to metallocage dynamics;[31,32] to an S_N2 reaction in gas phase and in explicit solvent (where, in both cases, with only hundreds of evaluations of the reference method, reactive ML dynamics is possible); and to a Diels-Alder reaction in the gas phase.

Duarte concluded that Gaussian Approximation Potentials can be trained in a day for reactive molecular systems; prospective model validation is crucial; general potentials must be more than pairwise additive; accuracy beyond density functional theory (DFT) can be approached; and training can be fully automated.

ML models to support risk assessment of small molecules
Andrea Volkamer, Charité Universitätsmedizin Berlin, Germany

In the risk assessment of novel compounds, regulatory agencies require *in vivo* testing for several toxic endpoints. Alternative (*in silico*) strategies include read-across,[33] structural alerts,[34] and ML and QSAR. In this talk, Volkamer addressed computational methods for holistic risk assessment,[35] and in particular, KnowTox,[36] CalUpdate,[37] ChemBioSim,[38] and cytotoxicity maps.[39]

KnowTox, developed in collaboration with BASF, has three different approaches to allow prediction of potentially toxic effects of query compounds: ML models for 88 endpoints, alerts for 919 toxic substructures,[40] and support for read-across in the form of similarity search with RDKit Morgan fingerprints, MACCS keys and physicochemical descriptors with the Tanimoto similarity coefficient.[41]

When deriving a robust and predictive *in silico* model it is important to examine not only the statistical quality of the model but also the estimate of its predictive boundaries. Key factors are applicability, reliability and decidability.[42] Conformal prediction (CP) is a method for confidence estimation in predictions.[43] The model must be statistically valid at a given confidence level and additional calibration step is that the CP framework compares predictions to those previously seen. In a binary classification, validity is the percentage of correct classifications and efficiency is the percentage of single class predictions (SCPs). Volkamer's team built 88 CP models (using RF as the underlying ML model) in KnowTox and the ToxCast dataset of about 8000 compounds and 1000 endpoints.

They then tested, in collaboration with BASF, how the model performed on one of the company's proprietary antiandrogen activity (AA) datasets. The three datasets used were ToxCast AA (for training and testing) and two external AA datasets, from BASF[36] and Norinder *et al.*[44] Results are shown in Table 1. Firstly, the CP technique was deployed (Table 1a, where accuracy of SCPs corresponds to the ratio of correct SCPs divided by all SCPs). Secondly, to improve validity and information efficiency, two adaptations were suggested: k-nearest neighbour (*k*NN) normalisation and balancing the dataset during training (Table 1b). While, initially, valid cross-validation models were obtained, validity and accuracy dropped on in-house data. The implemented adaptions restored validity and improved accuracy at the cost of efficiency but from a toxicologist's point of view, it is better to have no prediction for a compound than a wrong one.

Table 1a. KnowTox Case Study

Dataset	Efficiency			Accuracy (SCPs)			# toxic/ non-toxic
	all	cl.1	cl.0	**all**	**cl.1**	**cl.0**	
ToxCast AA	0.87	0.89	0.87	0.78	0.80	0.78	868/5842
Norinder	0.79	0.77	0.81	0.68	0.70	0.67	160/201
BASF	0.94	0.98	0.91	**0.56**	**0.97**	**0.07**	280/254

Table 1b. After *k*NN Normalization and Balancing

Norinder	0.43	0.33	0.52	0.74	0.67	0.78	160/201
BASF	0.20	0.18	0.23	**0.75**	**0.80**	**0.71**	280/254

CalUpdate[37] (developed in conjunction with workers at University College London and the Universities of Uppsala and Stockholm) assesses model calibration and suggests strategies to update models to account for predictivity drops when training and test data do not stem from the same distribution. Here, CP is used to assess the calibration of the models. Using the chronologically released Tox21 subsets Tox21Train, Tox21Test and Tox21Score, the researchers observed that while internally valid models could be trained using cross-validation on Tox21Train, predictions on the external Tox21Score data resulted in higher error rates than expected. To improve the external predictions, a strategy exchanging the calibration set with more recent data, such as Tox21Test, was introduced. The proposed improvement strategy, exchanging the calibration data only, is convenient as it does not require retraining of the underlying model.

In ChemBioSim, workers at BASF, Örebro and Vienna Universities and in Volkamer's team have enhanced the performance of CP models for *in vivo* endpoint predictions by combining molecular descriptors (RDKit Morgan fingerprints and physicochemical properties) with predicted bioactivity ones.[38] Biological fingerprints, describing the activity profile of a molecule, are more mechanistic descriptors, independent of molecular structure. These are actual assay measurements, but since they are not necessarily available at scale and would need to be measured for new compounds as well, the researchers chose to predict them by training CP models for 373 biological assays. The method was exemplified on three *in vivo* endpoints capturing genotoxic (MNT),

hepatic (DILI), and cardiological (DICC) issues. The incorporation of bioactivity descriptors increased the mean F1 scores of the MNT model from 0.61 to 0.70 and for the DICC model from 0.72 to 0.82 while the mean efficiencies increased by roughly 0.10 for both endpoints. In contrast, for the DILI endpoint, no significant improvement in model performance was observed. An analysis of the most important bioactivity features allowed detection of novel and less intuitive relationships between the predicted biological assay outcomes used as descriptors and the *in vivo* endpoints.

Finally, Volkamer's team have studied cytotoxicity prediction,[45] one of the earliest handles in drug discovery, using a deep learning approach trained on a dataset of over 34,000 compounds, fewer than 5% of which were cytotoxic. The dataset was from collaborators at the Leibniz-Forschungsinstitut für Molekulare Pharmakologie in Berlin. The encoding involved RDKit Morgan fingerprints. A deep NN with parameter optimisation, balancing and 10-fold nested cross-validation were used. The model reached a balanced accuracy of over 70%, similar to previously reported studies using RF or CP, but different underlying cytotoxicity datasets and activity shares.[46] NNs are often described as a "black boxes". To overcome this absence of interpretability, a deep Taylor decomposition method with layer-wise relevance propagation (LRP)[47] was investigated to identify toxicophores. A forward path of the trained model is used to get a prediction score which is interpreted as relevance. A backward path of the trained model is used to get decompositions of relevance on input. Toxicophores are identified by mapping the relevance back to atom environments, namely the bits in the Morgan fingerprints. The study also introduced cytotoxicity maps which provide a visual structural interpretation of the relevance of these toxicophore substructures.

About 2.8 million laboratory animals were used in Germany in 2018; establishment of alternative methods could lead to a reduction of animal testing. To this end, Volkamer's team have used CP models and deep learning to predict compounds likely to be ineffective or toxic and exclude them *a priori* from animal testing. Holistic and combined approaches with proven applicability, reliability and interpretability, demonstrated by predictive power and prospective studies will increase acceptance by regulatory authorities.

Exploring molecular space and accelerating drug discovery with Clara Discovery and MegaMolBART
Michelle Gill, NVIDIA, Santa Clara, CA, USA

To extract scientific insights from today's massive datasets we need methods that take advantage of the complexity of the data and can scale efficiently. The increased degree of parallelism afforded by GPUs has made them ideal for the acceleration of analysis and visualisations. Such applications can be combined with methods derived from deep learning to create analysis pipelines that are both faster and more accurate than the existing state of the art. Clara Discovery is a collection of frameworks, applications, and AI models that together accelerate drug discovery, supporting research in genomics, microscopy, virtual screening, computational chemistry, visualisation, clinical imaging and NLP. Gill concentrated on RAPIDS and MegaMolBART.[48]

One example is an interactive clustering and visualisation workflow in which RDKit-derived Morgan fingerprints from ChEMBL (or another database) are used in principal component analysis (PCA), *k*NN clustering, and UMAP visualisation. This pipeline is implemented using cuML and can be performed in real-time due to the acceleration afforded by GPUs. The plotly interface can be customised.

MegaMolBART[48] is mentioned in Engkvist's talk earlier in this report. Pre-training is performed on a subset of ZINC15. SMILES molecules are masked and enumerated (randomised) during training. NVIDIA carries out training on a DGX SuperPOD (4-8 nodes x 8 A100 GPUs). AstraZeneca is concurrently training on Cambridge-1. The pre-trained model is wrapped into a service (Figure 2). The interactive explorer provides a framework

for visualising and customising workflows. Deep learning derived features from MegaMolBART can enable analyses that previously required hours to be completed in seconds.

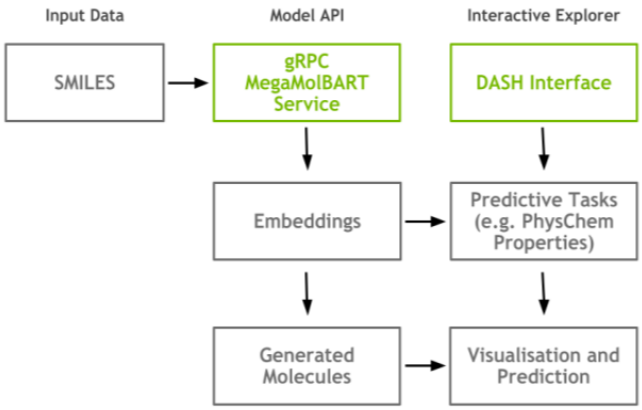


Figure 2. MegaMolBART model service.

In future, NVIDIA will investigate the limits of model size of MegaMolBART and will develop novel model architectures for improved molecule generation. Predictive tasks such as physicochemical properties, reaction prediction and retrosynthetic synthesis could be based on model embeddings. The user experience will be improved by automation of data processing, pre-training and downstream tasks.

Challenges and opportunities for machine learning in drug discovery

W. Patrick Walters, Relay Therapeutics, Cambridge, MA, USA

Over the last few years there has been a dramatic growth in the application of ML in drug discovery. It is impacting numerous areas including image analysis, organic synthesis planning, predictive models, quantum chemistry and molecule generation but there are significant challenges. AI predictions are typically treated as a "black box" which supplies no explanation, yet interpretable models could drive discovery by providing a rationale that convinces people to perform experiments, allowing scientists to gain insights that drive compound design, and enabling efficient debugging of model performance.

Matveieva and Polishchuk[49] have published benchmarks for interpretation of QSAR models. Feature attribution techniques are popular choices for explainability tools, as they can help elucidate which parts of the provided inputs used by an underlying supervised-learning method are considered relevant for a specific prediction, but Jimenez-Luna *et al.*[50] found that none of the feature attribution methods they tested generalised well when confronted with unseen examples. One interesting approach to explainability is the use of "counterfactuals".[51] They are used in credit card approval applications because the law demands that credit card denial be explained. These methods look at the small differences between two people, one whose card is declined and the other whose card is approved. Walters presented an example of the use of counterfactuals for prediction of the solubility of imatinib. He generated analogues, predicted their solubility, sorted them by similarity and evaluated the counterfactuals, looking for small differences. This method seems to work (Figure 3).

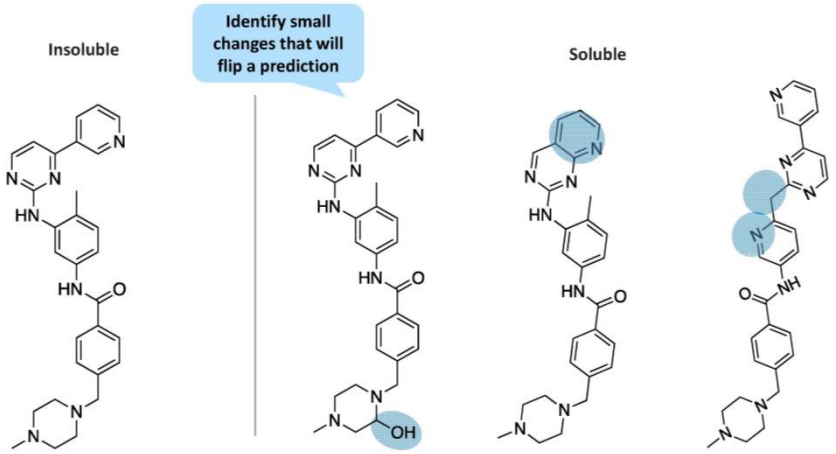


Figure 3. Predicted soluble analogues of imatinib.

Another issue is impossible molecules emerging from generative models. GuacaMol[52] benchmarking for *de novo* molecular design employs Walters' earlier metrics for compound quality but the filters do not detect a number of "chemically impossible" features such as triple bonds in aromatic rings, so Walters has written "silly walks" code.

Another issue is molecular representations. For many years, ML models have been constructed using standard molecular fingerprints. More recently, a number of groups have published methods that use neural networks to generate targeted molecular representations.[53,54] To determine if learned representations are better, Walters has written "Yet another ML method comparison" to compare a number of commonly used molecular representations and algorithms. In these tests a standard XGBoost method using molecular fingerprints tends to outperform the learned representations on smaller datasets (less than 2000 molecules).

Why do we not use 3D descriptors more often in ML? Traditional ML methods map one object to one label but molecules can have many 3D conformations. To tackle the relationship between multiple instances and a single label, specialised multiple instance ML methods must be used. Recent papers[55,56] examine whether 3D multiple-instance approaches will work. Results vary across datasets but 3D multiple-instance models do appear to be competitive with 2D ones.

Finally, Walters discussed uncertainty and model applicability. A number of methods have been tried to determine when a model is applicable but none of them is ideal. There is a pitfall in scaffold-based[57] cross-validation, training on one scaffold and testing on another. The idea feels good but why should it work? Different chemotypes often make different interactions. The model must implicitly learn these interactions. Walters found that 12 inhibitors of p38 have a very wide variety of interactions in ATP binding pockets. It is important to evaluate your model in context.

ML is impacting many aspects of drug discovery and there are many issues to address, including explainability, representation, model applicability, and multi-objective optimisation. Whilst we have made progress on parts of the puzzle, we are still far from a complete solution. To succeed we need the overlapping domains of "hacking skills", mathematics and statistics knowledge, and substantive domain expertise.

Molecular Transformer-aided biocatalysed synthesis planning

Daniel Probst, Matteo Manica, Yves Gaëtan Nana Teukam, Alessandro Castrogiovanni, Federico Paratore, Teodoro Laino, IBM Research Europe, Rüschlikon, Switzerland

Enzyme biocatalysts are an integral part of green chemistry strategies towards a more sustainable and resource-efficient chemical synthesis. Most are proteins, one third of them require one or more cofactor in the form of inorganic ions, and others require complex molecules as cofactors. Enzymes affect only the reaction rates, not the equilibria, and rate enhancements brought about by enzymes are in the range of 5-17 orders of magnitude. Enzymes have industrial uses in fermentation (e.g., in antibiotic production and brewing) and in enzyme technology (e.g., paper pre-bleaching, food processing and enantionmerically pure amino acids).

They are stereo-, regio- and chemoselective, highly efficient, reusable and biodegradable, and moderate temperatures and pH are required, but they are unstable at high temperatures or extreme pH, require expensive co-substrates, are potential allergens, and generate metabolic by-products. Unfortunately, a narrow substrate scope is documented in enzyme databases and synthetic chemists have difficulties in identifying patterns within enzyme classes that allow them to extend those patterns to unreported substrates. In addition, other domain-specific knowledge factors such as stereo- and regioselectivity are lacking.

Biocatalytic retrosynthesis has recently been automated by creating expert-curated reaction rules based on available literature, creating a network of molecules connected by enzymes and reaction rules, and applying the rules to arbitrary query molecules in order to find both a matching enzyme and a precursor that can be purchased. RetroBioCat[58] is an example. Unfortunately, the creation of expert-curated reaction rules does not scale.

Kreutter *et al.*[59] have tackled this issue by using multi-task transfer learning to train the molecular transformer[60], a sequence-to-sequence ML model, with 1 million reactions from the USPTO database combined with 32,181 enzymatic transformations annotated with a text description of the enzyme. This translates the substrates and enzyme into products. The resulting enzymatic transformer model predicts the structure and stereochemistry of enzyme-catalysed reaction products with remarkable accuracy. The researchers combined the reaction SMILES language of only 405 atomic tokens with thousands of human language tokens describing the enzymes, such that the enzymatic transformer not only learned to interpret SMILES, but also the natural language as used by human experts to describe enzymes and their mutations.

Probst *et al.* have, in addition to the forward model, introduced a retrosynthesis model using a class token based on the Enzyme Commission (EC) number classification scheme that allows them to capture catalysis patterns among different enzymes belonging to the same hierarchical families.[61] Data sources are BRENDA, MetaNetX, PathBank, and Rhea, leading to 62,222 deduplicated, biocatalysed reactions. Probst showed TMAP[62] visualisations of the substrates and products (using MAP4 fingerprints).[63] Modified cofactors are removed from the products. The dataset is not balanced: transferases are over-represented. Tokenisation includes the first three parts of the EC number. An EC number (e.g., 2.6.1.2) has four levels: class, sub-class, sub-sub-class, and serial number (SN). The SN is not used because adding it causes a drop in performance. Performance is limited by dataset size, diversity and quality. The forward prediction model achieves a top-5 accuracy of 62.7%, while the single-step retrosynthetic model shows a top-1 round-trip accuracy of 39.6%. As regards accuracy across classes, class 2, a big class, pushes up accuracy whereas class 1 is poorer because there are too few training data.

Attention weights learned by a transformer encode atom rearrangement information between products and reactants. Attention-weight analysis unboxes the forward model to understand how enzyme information is utilised. The IBM team has shown that the EC tokens relate to the centres of the enzymatic reaction and that the

forward model captures enzymatic reaction rules based on the EC number. The model mimics the expert-curated reaction rules in automated retrosynthesis.

A resident chemist has tried the system out and has been able to replace a traditional reaction with an enzyme-catalysed one. Anyone can try the system for free at IBM RXN for Chemistry. Stereochemistry is included for all reactions. The enzymatic data and the trained models are available through the RXN for Chemistry network and on GitHub.

Highly accurate protein structure prediction with AlphaFold

Alexander Pritzel, DeepMind, London, UK

A central part of DeepMind's mission is to solve fundamental scientific problems with AI. Predicting the 3D structure of a protein from its amino acid sequence is one such challenge. AlphaFold[64] is DeepMind's model that aims to solve this problem. Proteins consist of chains of amino acids that fold into a 3D structure and the exact 3D shape is important for a protein's function. Experimental structure determination takes months to years; structure prediction can provide actionable information faster.

Critical Assessment of Structure Prediction (CASP) is an organisation that conducts double-blind, community-wide experiments to determine the state of the art of computational methods for modelling protein structures. The CASP assessment involves predicting recently solved structures that are not yet public. In the 14th biennial CASP (CASP14) across a wide range of difficult targets AlphaFold was the top-ranked method: assessors judged its predictions to be at an accuracy "competitive with experiment" for approximately two thirds of proteins.

A key question in the design of neural network architectures is the question of inductive bias, which controls which kind of functions are easy or hard to model. In convolutional networks (used for computer vision, for example) the data are in a regular grid and information flows to local neighbours. AlphaFold 1 used this inductive bias. In recurrent networks (e.g., for language) data are in an ordered sequence and information flows sequentially. In graph networks (e.g., for recommender systems or molecules) data are in a fixed graph structure and information flows along fixed edges. In an attention module (e.g., for language) data are in an unordered set and information flow is dynamically controlled by the network (*via* keys and queries).

High-throughput sequencing technologies have enabled the construction of a multiple sequence alignment (MSA), and accurate coevolution signals can be disentangled. Detected coevolved pairs can be used as residue-residue contact constraints in protein structure modelling and prediction of protein-protein interactions.[65,66] In AlphaFold physical and geometric insights are built into the network structure, and are not just a process around it. This is an end-to-end system directly producing a structure instead of inter-residue distances. Inductive biases reflect knowledge of protein physics and geometry. The positions of residues in the sequence are de-emphasized. Instead, residues that are close in the folded protein need to communicate. The network iteratively learns a graph of which residues are close, while reasoning over this implicit graph as it is being built. In co-evolution, residues in contact must mutate together (mutation of a single residue breaks the contact and the organism with the mutated protein does not survive). Evolution conserves some properties such as hydrophobic and hydrophilic amino acids being on the "inside" or "outside" of a protein.

Figure 4 presents an outline of how AlphaFold works. A key input is the MSA, containing sequences evolutionarily related to the target. Related sequences are found using standard tools and public databases. The input sequence is used to create an array of representations representing all residue pairs. AlphaFold can also use template structures from the Protein Data Bank (PDB) but it often produces accurate predictions without a

template. The Evoformer blocks extract information about the relationship between residues. The MSA representation can update the pair representation and *vice versa*. The Structure Module predicts a rotation and translation to place each residue. A small network predicts side chain chi angles. The final structure is run through a relaxation process. Feeding certain outputs back through the network again improves performance.

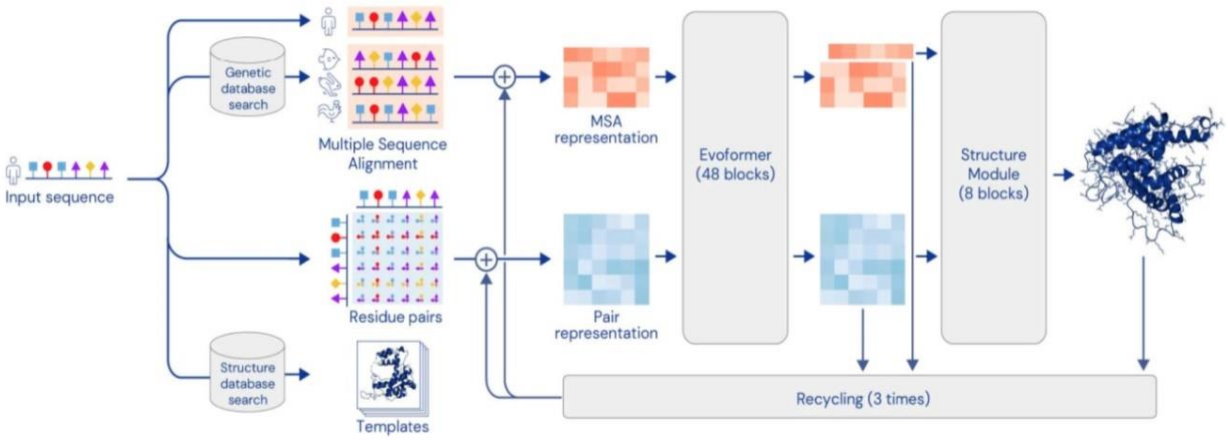


Figure 4. AlphaFold overview.

As well as a predicted structure, the Evoformer blocks produce two confidence estimates: per-residue confidence (for predicted local Distance Difference Test, plDDT)[67] and pairwise confidence (predicted aligned error, PAE). Further detail of the Evoformer architecture is given in Figure 5. In triangular attention, consider three points A, B and C. If distances AB and BC are known, the triangle inequality places a strong constraint on the distance AC. Evolution and sequence give information about relations between residues and pair embedding encodes the relations. The update for pair AC should depend on BC and AB. In the graph, edges represent pairs of residues. Since the graph is unknown it has to be inferred. There is a triplet relation in this language with cycles of a length of three in the graph. The update applied by the layer is based on all cycles involving the edge. More abstractly this can be viewed as a transitivity inductive bias that encodes the transitivity of relations (e.g., triangle inequality and loop closure).

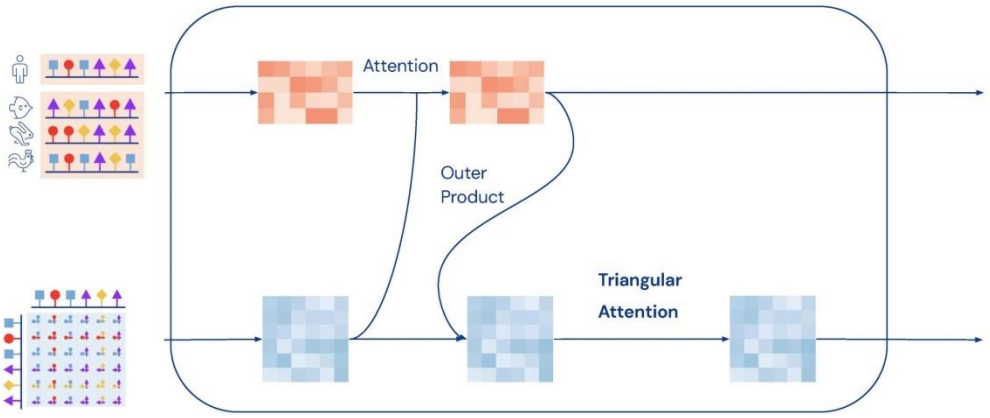


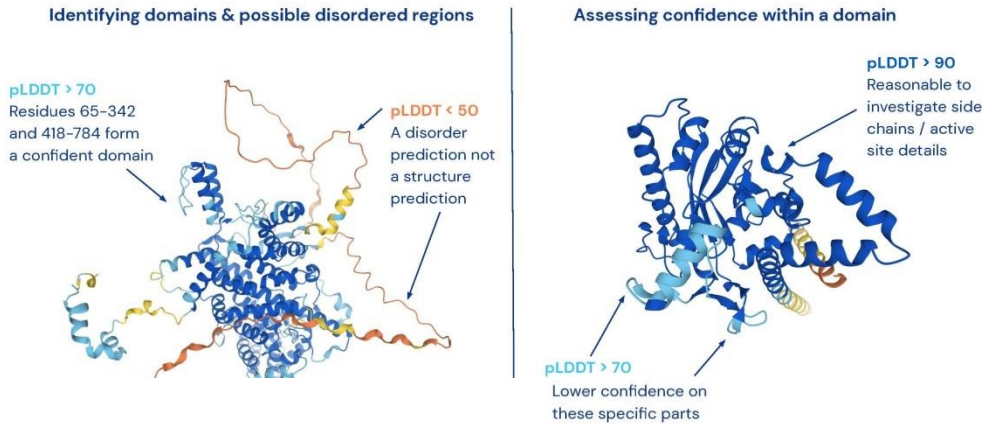
Figure 5. Evoformer.

The structure module performs end-to-end folding instead of gradient descent. Here the protein backbone is modelled as a gas of independent 3D rigid bodies. The spatial structure of the amino acid chain is not built into the model but emerges through learning. A 3D equivariant transformer architecture updates the rigid bodies

modelling the backbone and also builds the side chains by predicting torsion angles. The AlphaFold architecture can be trained to high accuracy using only supervised learning on PDB data, but accuracy can be further enhanced using an approach similar to noisy student self-distillation.[68] This is the way AlphaFold makes use of unlabelled sequences. The AlphaFold model is first trained on PDB data alone. This first model is used to predict structures on a large set of unlabelled sequences and then a second model is trained where the training set is enriched by confidently predicted structures of the first model.

Computational structure prediction is typically underspecified, for example as regards oligomeric state, ligands, DNA-binding, experimental conditions, multiple conformations etc. The AlphaFold network implicitly models this missing context using a variety of physical and evolutionary information. Movies of model interpretability for SARS-CoV-2 ORF8 (T1064, one of the hardest examples in CASP14) and a RNA polymerase with over 2000 amino acids (T1044 in CASP14) were shown to illustrate how the model can be interrogated.

Predictions can be interpreted using plDDT and PAE. Roughly speaking, lDDT measures the percentage of correctly predicted interatomic distances, not how well the predicted and true structures can be superimposed. It rewards locally correct structures, and getting individual domains right. plDDT is a measure of local confidence (Figure 6) but high plDDT on all domains does not imply AlphaFold is confident of their relative positions. Assessing inter-domain confidence requires the PAE metric. This is AlphaFold's prediction of the position error at residue x, if the predicted and the true structures are aligned on residue y. PAE aims to measure confidence in the relative positions of pairs of residues. It is mainly used to assess relative domain positions, but is applicable whenever pairwise confidence is relevant. PAE is displayed as a 2D plot. If residue y is aligned to the true structure and the position error at residue x is measured, the colour at (x, y) is AlphaFold's prediction of that error.



Uniprot ID Q5VSL9 Uniprot ID Q96PD7

Figure 6. Use of plDDT.

The AlphaFold protein structure database is a website developed by DeepMind and EMBL-EBI that contains pre-run predictions for 21 model organisms. The AlphaFold colab is a website hosting a pre-written Python program to be executed on a machine in the cloud; you enter a sequence and hit "play" at each step. There are also several other community-developed colabs for structure prediction. You can also download the code and run AlphaFold on your own machine. AlphaFold has been received with excitement by the biology community and incorporated in other tools. It has been used in accelerating structure determination, in docking, in predicting disorder and in finding new insights from the AlphaFold database. There is much exciting work ahead for the structural biology field: complexes, disorder and conformational change, etc. DeepMind is very

excited to see what others are building on top of the AlphaFold database. There is great potential in AI for science as a whole.

"Attending" to co-crystals in the Cambridge Structural Database

Aikaterini Vriza,[1] Angelos B. Canaj,[1] Rebecca Vismara,[1] Laurence J. Kershaw Cook,[1] Troy D. Manning,[1] Michael W. Gaultois,[1] Peter A. Wood,[2] Vitaliy Kurlin,[1] Neil Berry,[1] Matthew S. Dyer,[1] Matthew J. Rosseinsky.[1] (1) University of Liverpool, UK (2) Cambridge Crystallographic Data Centre, Cambridge UK

A co-crystal is a crystalline single-phase material composed of two or more different molecular compounds in a specific stoichiometry. They are connected *via* non-covalent interactions, such as hydrogen bonding, π–π stacking, halogen bonds and charge transfer interactions. Co-crystals have been particularly useful in improving the physicochemical properties of potential drugs but the current work was focused on the design of co-crystals with electronic functionalities. Polycyclic aromatic hydrocarbons (PAHs) self-assemble *via* π–π interactions and are considered promising candidates for electronic materials. Vriza and her co-workers aimed not only to detect some weakly bound PAH co-crystals but also to understand the important factors contributing to their formation.

The aim is to find molecular pairs which are more likely to form a co-crystal. The problem is that we know which combinations can form co-crystals but we have no information for those that do not. The workflow for co-crystal prediction is a closed loop of database analysis, ML, optimisation, and experimentation. Two datasets were created starting with eight electron-rich PAHs with distinct geometry by carrying out similarity searches and removing molecules with H-bonding. The sets contained 1722 known molecular combinations from the Cambridge Structural Database (CSD) and 21,736 possible ones from ZINC15, forming labelled (training) and unlabelled datasets respectively. Dragon descriptors were used as features of the two datasets. Each molecular pair was represented as a concatenation of the molecular descriptors.

Most co-crystal prediction research has focused on generating negative data for training binary classifiers. The current work, involving one-class classification, focuses only on the positive data and trying to define a reliable area where novel pairs can exist.[69] The aim of Deep Support Vector Data Description (DeepSVDD) is to find a data-enclosing hypersphere of minimum size, such that the normal data points will be mapped near the centre of the hypersphere whereas anomalous data are mapped further away. The objective of DeepSVDD is to learn the network parameters and minimise the volume of the hypersphere. The deep learning protocol is a two-step process. The first (pre-training) step uses a convolutional autoencoder to capture the representation of the data. During this step the centre of the hypersphere is calculated and is fixed as the mean of the network representations of the known data. During the second step, the latent dimension of the encoder is connected to a feed-forward NN to minimise the loss function (the distance from the centre of the hypersphere). In the Deep One Class method of Vriza *et al.* the convolutional autoencoder was substituted with a Set Transformer autoencoder which is capable of handling the order invariance of the molecular pairs.

The algorithms implemented for one-class classification were separated into eight traditional ones and one NN. Vriza showed the overlapping score distribution of both the labelled and unlabelled data for all the algorithms. The unlabelled data consist of both positive and negative examples in an unknown proportion. Consequently, a certain part of the unlabelled data is expected to belong to the known class (i.e., are inliers). Moreover, in the labelled data there is a small proportion of examples that significantly differs from the rest of the data and is regarded as noise of the normal class (i.e., outlier examples). The impact of the class noise is mitigated using one class classification, as a proportion of the labelled data is regarded as outliers during the hyperparameter optimisation process. A clearer and more definite separation among the two different datasets can be observed

for both the Ensemble and Deep One Class methods, with Deep One Class covering a bigger range of scores and thus enabling a better separation.

Vriza showed learning curves of all the algorithms showing the performance of the models while the size of the training set increases. The validation metric used was the true positive rate (the number of correctly predicted inliers divided by the total size of the training set in each fold of the five-fold cross validation). The learning model outperforms the traditional algorithms as it has higher accuracy and low standard deviation.

Scatterplots showing the distribution of representative descriptors among the molecular pairs on the labelled dataset indicate that the deep learning model can effectively learn the trends of the labelled data and is able to score the unlabelled data based on the significant patterns of the labelled data. Focusing on the highest-ranking pairs predicted, the team tried to optimise the selection by targeting molecules with similarity to 7,7,8,8-tetracyanoquinodimethane (TCNQ) which is extensively studied for its interesting electronic properties. Pyrene:benzochromenone (CSD: EHUFIZ) and pyrene:dicyanoanthracene (CSD: EHUFEV) were identified and experimentally validated, both containing molecules which have not previously been reported as co-formers in the CSD. These were two unlabelled inlier co-crystals lying in the densest area of the scatterplots regarding the polarity and electronic descriptors. Although shape, size and polarity are key factors, the rules that dominate co-crystal formation are far more complex than just some general properties.

The researchers then looked at molecular representations in Set Transformer and evaluated those using publicly available benchmarks. The representations were Mordred descriptors,[70] RDKit Morgan fingerprints, graph embeddings (GNN fingerprints) and representations used in NLP such as Molecular Transformer.[60] Vriza *et al.* tuned the hyperparameters to reduce the reconstruction error and found that Morgan and GNN fingerprints performed best on all the validation data in terms of total accuracy (specificity, area under the receiver operating characteristic curve (ROC AUC) and recall). The two types of fingerprint also performed well in a head-to-head comparison on co-crystal screening data for 18 active pharmaceutical ingredients.

It has been said that there are some tasks for which there are simply not enough labelled data so we need to focus on ML methods that do not rely on labels. The applicability of the one class unsupervised approach to all CSD co-crystals has been validated in real case scenarios. Currently there are several ML models for co-crystal screening. The Liverpool team has provided a large amount of external validation data and carried out extensive testing against several methods. The workers focused on AI model development: permutation invariant neural networks, attention to extract relations, hyperparameter tuning and reconstruction error minimisation. They also tested several types of distinct inputs and found that Morgan and GNN fingerprints described the molecular pairs better than other inputs.

PyPEF, an integrated framework for data-driven protein design and engineering

Niklas E. Siedhoff,[1] Alexander-Maurice Illig,[1] Ulrich Schwaneberg,[1,2] **Mehdi D. Davari**.[3] (1) RWTH Aachen University, Aachen, Germany (2) DWI-Leibniz Institute for Interactive Materials, Aachen, Germany (3) Leibniz Institute of Plant Biochemistry, Halle, Germany

Davari's group is interested in enzymes involved in catalysis in cells. Establishing protein sequence, structure, and function relationships is a grand challenge for experiment and computation. There has been progress on structure-sequence links, on design of sequences based on function, and on prediction of function based on sequences, but the dynamics linking structure to function is still a big challenge.

Directed evolution (for which Frances H. Arnold won half a Nobel Prize in 2018) depends on generating a large gene library, needing lots of costly effort. Rational, computer-aided design techniques might never be able to sample through the entire protein sequence space and benefit from nature's full potential for the generation of better enzymes. There is a clear trend to combine the rational design and directed evolution approaches. Semi-rational design generates small, functionally rich, mutant libraries using rationally pre-selected target sites. Knowledge-driven approaches navigate sequence space intelligently. Recently, ML methods have been increasingly applied to find patterns in data that help predict protein structures, improve enzyme stability, solubility, and function, predict substrate specificity, and guide rational protein design.[71-74]

In evolutionary biology, fitness landscapes are used to understand the relationship between genotypes and reproductive success. It is assumed that every genotype has a well-defined replication rate (fitness). This fitness is the "height" of the landscape. Genotypes which are similar are said to be close to each other, while those that are very different are far from each other. The set of all possible genotypes, their degree of similarity, and their related fitness values is then called a fitness landscape. The size of the protein sequence space is huge and the fitness landscape is complex. Current challenges are screening throughput (leading to limited exploration, information gaps and local maxima); the combinatorial problem of epistasis (a phenomenon in which the effect of a gene mutation is dependent on the presence or absence of mutations in one or more other genes); and cost and time.

Combining next generation sequencing (high-throughput analysis of DNA and RNA sequences) with high-throughput screening of 10^4-10^8 variants per day is a powerful strategy (deep mutational scanning) for comprehensively analysing sequence-function relationships.[72] ML-guided directed evolution reduces experimental effort and mutates multiple positions simultaneously, combining directed evolution and rational design (Figure 7).[71,73]

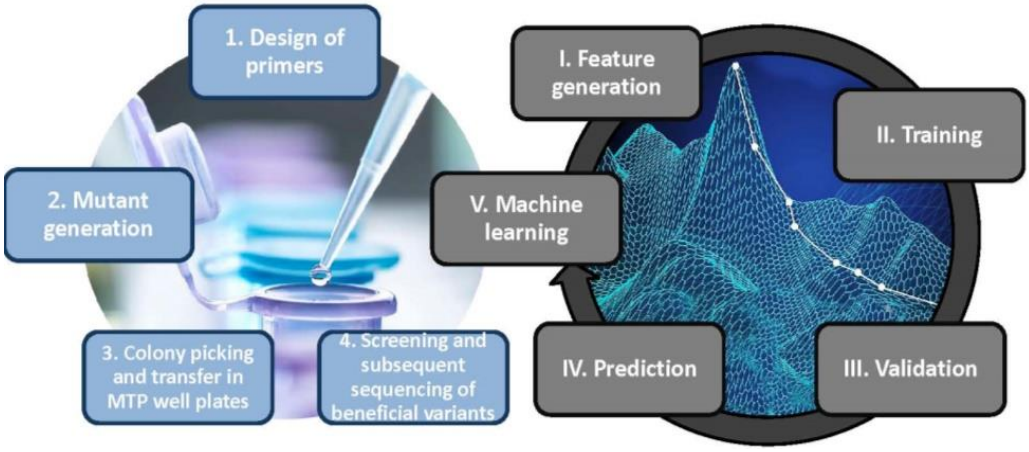


Figure 7. ML-guided directed evolution.

Pythonic Protein Engineering Framework (PyPEF, Figure 8) is a general-purpose framework for data-driven protein engineering by combining ML methods (partial least squares (PLS), RF, support vector regression (SVR), and multilayer perceptron (MLP) based regression) with signal processing (fast Fourier transform, FFT) and statistical physics (Metropolis-Hastings algorithm) techniques.[75] It assists in the identification and selection of beneficial proteins in the sequence space by either systematically or randomly exploring the fitness of protein variants and by sampling random evolution pathways. It applies featurisation by Fourier-transforming

numerical indices, which represent physicochemical and biochemical properties for each amino acid, taken from the amino acid index (AAindex).

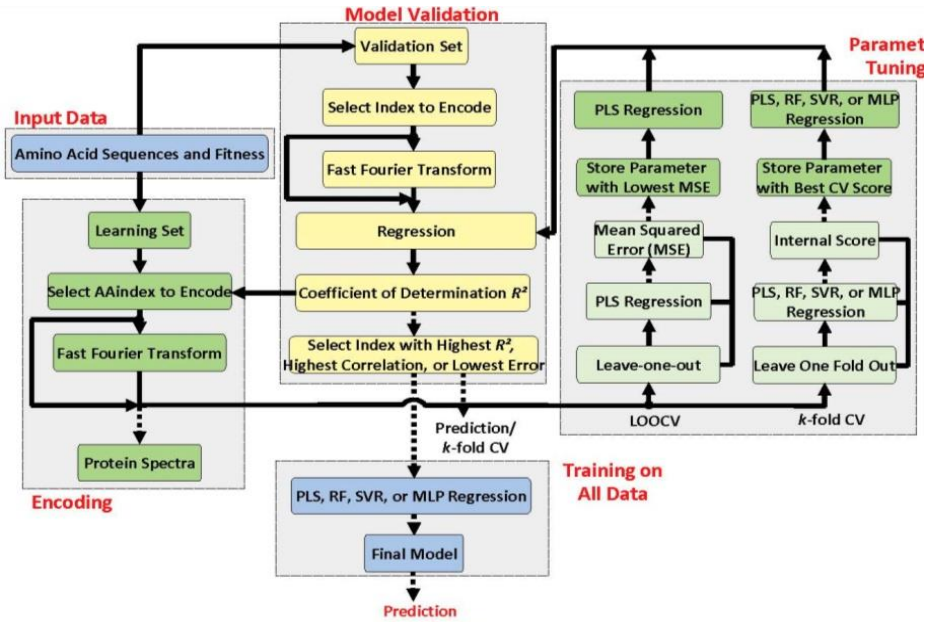


Figure 8. PyPEF framework.

The predictive accuracy and throughput performance of the framework was evaluated based on four publicly available datasets of proteins and enzymes and their properties, using common regression models. PyPEF learned on datasets of small-to-medium-size, derived by diverse evolution strategies, and demonstrated potential to generate predictive models consistently, by accounting for either additive effects only (AAindex encoding and linear models) or non-additive effects within the range of values learned during modelling (AAindex encoding and non-linear models) as well as both inside and outside the range of values learned during modelling, while providing effective *in silico* screening capabilities.

The framework could efficiently predict the fitness of protein sequences for different target properties with R^2 using PLS regression and FFT encodings ranging from 0.58 to 0.92. It enabled more than half a million protein sequences to be screened for various functions in only a few minutes on a standard PC. Data-driven models generated by PyPEF with significant accuracies on four public datasets highlighted the potential for predicting the fitness of variants with high accuracy or capturing the general trend of introduced mutations on the fitness in directed protein evolution campaigns. PyPEF code is publicly available.

Best practice for chemical language model *de novo* design of GPCR ligands: datasets, scoring functions and optimisation algorithms

Morgan Thomas,[1] Noel M. O'Boyle,[2] David Araripe,[1] Rob T. Smith,[2] Chris de Graaf,[2] Andreas Bender.[1] (1) University of Cambridge, UK (2) Sosei Heptares, Cambridge, UK

There has been significant interest in *de novo* molecular design recently. Thomas discussed some aspects of the practical use of chemical language models (e.g., SMILES with recurrent neural networks) which are popular due to their simplicity, performance (by benchmarking works GuacaMol,[52] Molecular Sets (MOSES),[76] and Smina and Therapeutic Data Commons), code availability and support. Both structure-based and ligand-based design can be used. In the latter case prior ligand knowledge may not be available and if it is, it may bias molecule generation towards known chemotypes. Structural data are difficult to acquire (though they are

increasingly available) but structure-based design is not biased by prior ligand knowledge. G Protein Coupled Receptors (GPCRs) are a particular target class where structural data can have a significant impact.[77]

Thomas and his co-workers[78] have assessed the use of molecular docking *via* Glide (a structure-based approach) as a scoring function to guide the deep generative model REINVENT[5,79] and compare model performance and behaviour to a ligand-based scoring function. The case study involved dopamine receptor D2 (DRD2). The approach taken is depicted in Figure 9, where data sources are coloured blue and scoring functions orange. The REINVENT framework (in grey) consists of two recurrent neural networks, a prior and an agent. The main steps in the current work are (1) removing known DRD2 active molecules from the ZINC training data; (2) training the prior model on druglike molecules from ZINC; (3) initializing the agents as a copy of the prior; (4) preparing the scoring functions to evaluate *de novo* molecules; (5) iteratively training both agents *via* reinforcement learning; and (6) evaluating the structure- and ligand-based approaches with respect to different quantitative, chemical and structural aspects of the generated molecules.

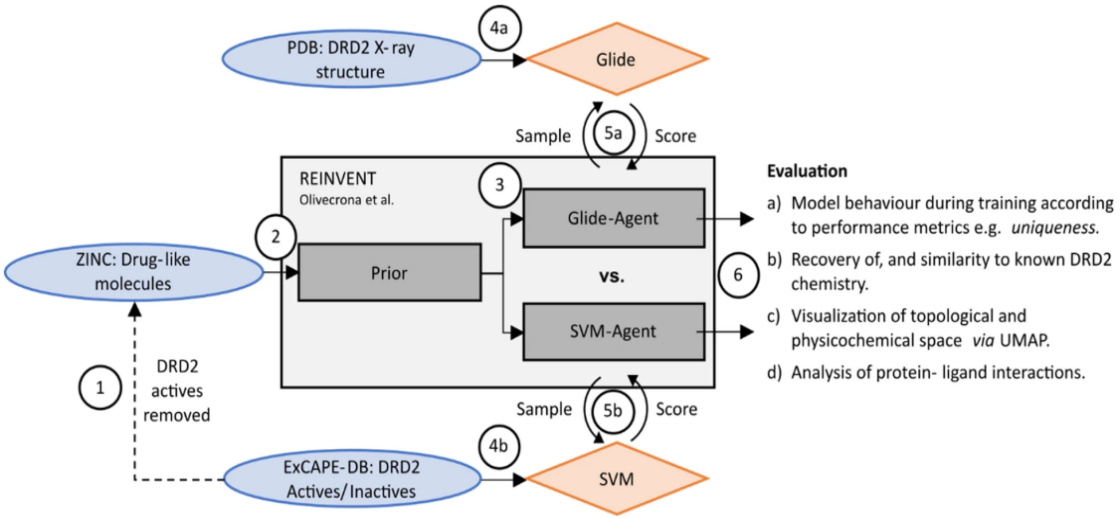


Figure 9. Comparison of structure- and ligand-based scoring functions.

The structure-based approach improved uniqueness and molecular diversity during training, produced higher similarity to the training set and provided a greater coverage of known active ligands than the ligand-based approach, despite having no prior ligand knowledge, as more clusters were shared between Glide-Agent and known actives than were shared between the SVM-Agent results and known actives. Glide-Agent generates high-scoring molecules that are more novel than the SVM-Agent ones and generates more novel areas of physicochemical space, consistent with the prior. Moreover, Glide-Agent learns to satisfy a crucial interaction with D114^{3x32} which is associated with better docking scores and is a prerequisite for experimental affinity.[80]

Unfortunately, docking score optimisation is slow (each run takes about 1 week on about 30 CPUs) and it is system dependent. Can the computational expense associated with model optimization be minimised? The REINVENT loss function (augmented likelihood) includes a value sigma used to scale up the scoring function and lower the prior contribution.[79] Comparison of REINVENT[5] with REINVENT 2[79] shows that sigma variation has a small effect on a short time scale. When rewards are sparse, loss drives agent back towards prior. This can be circumvented by using the hill-climb algorithm[81] to focus learning on the best molecules. Thomas *et al.* found that a hybrid, augmented hill-climb, is more efficient at optimising docking score and is more sensitive to sigma values and hence more tunable. Augmented hill-climb has the propensity to undergo mode collapse (drop in uniqueness). Mode collapse can be rescued by using a diversity filter (DF)[3] to penalise non-unique or similar

molecules. DF stabilises optimisation augmented hill-climb plus DF is seven times more efficient than Glide-Agent in the original work on the short time scale, and up to 100 times more efficient on the long time scale whilst maintaining similar chemical behaviour, and runtime is reduced to about 2-5 hours compared to one week on about 30 CPUs.

DF would rescue mode collapse but it would not address the issue of generating unrealistic molecules. Benchmarking datasets are either too restrictive (as in the case of MOSES) or too broad (as in the case of GuacaMol) but the ChEMBL$_{potent}$ subset of ChEMBL provides a dataset rich with druglike properties. SMILES outperforms alternative grammars in the prior dataset. Surprisingly, DeepSMILES[82] suffer lower validity. SELFIES[83] are more diverse but fewer of them pass standard drug-likeness filters. SELFIES are least like the training set and the use of them results in many more "unusual" compounds. The prior dataset is still generating relatively featureless structures compared with risperidone (a DRD2 inverse agonist), regardless of the chemical grammar.

Thomas' final topic was the effect of scoring function protocol on failures of docking and of QSAR functions. We know, for example in DockStream,[7] that different protocols lead to variable enrichment in docking and that adding constraints such as particular residue interactions increases performance[84] and can outperform ML.[85] It has been observed that ligand protonation is important in docking and that similar chemotypes can have inconsistent docked poses. There is also a trend for certain physicochemical properties to be violated. To study the effect of scoring function protocol, Thomas *et al.* chose MPO against Adenosine 2A (A2A).[86,87] They analysed a diverse range of known chemotypes.[88] They increased prior contribution by decreasing sigma from 60 to 30, protonated only the most likely states, and introduced a more difficult optimisation problem using constrained docking score, retrosynthetic accessibility score (RAscore),[12] TPSA ≥ 40 and number of rotatable bonds ≤ 6. The added constraints worsen docking optimisation but improve molecule quality. The A2A MPO recovered more of, and a wider range of known A2A chemotypes.[88] The added constraints avoid full occupation of the cavity.

As for failures of QSAR functions, we know that generative models can overfit QSAR functions[89] and that QSAR models with similar performance select different prospective candidates in virtual screening.[90] Thomas *et al.* compared the molecules designed *de novo* for three targets using several different molecular representations, QSAR models and generative models. Both descriptor and QSAR method have a significant impact on generative model behaviour, such as molecular diversity and similarity to the training set. The vast majority of molecules are unique to a particular replicate and a particular method. The best way to incorporate synthesizability has not been considered. Work is ongoing on prospective validation, interaction fingerprints and alternative scoring functions.

Machine learning models for predicting human *in vivo* PK parameters using chemical structure and dose
Olga Obrezanova,[1] Filip Miljković,[2] Anton Martinsson,[2] Beth Williamson,[1] Martin Johnson,[1] Andy Sykes,[1] Andreas Bender,[1] Nigel Greene.[3] (1) AstraZeneca, Cambridge, UK (2) AstraZeneca, Gothenburg, Sweden (3) AstraZeneca, Waltham, MA, USA

Animal and human pharmacokinetic (PK) data are routinely used in drug discovery to understand absorption, disposition, metabolism, and excretion (ADME) of candidate drugs. AstraZeneca has a suite of over 40 global ADME and safety models to guide virtual compound generation, enable compound selection and prioritisation, design compounds with good ADME and safety profiles, improve speed and efficiency in the DMTA cycle and reduce the number of *in vitro* experiments. The ultimate goal is to enable human PK prediction at the point of design.

Prediction of rat PK is a stepping stone towards modelling human PK. An AI model predicts rat PK parameters from chemical structure and measured *in vitro* ADME properties. The chemical structure is encoded by a graph convolutional neural network (GCN). Properties used as input features are solubility, Caco2 (colorectal adenocarcinoma cell) intrinsic permeability and efflux, intrinsic clearance (CL_{int}) in human liver microsomes (HLM), rat hepatocytes, intrinsic clearance and fraction unbound, and rat and human plasma protein binding (PPB). Properties predicted are clearance (CL), bioavailability (%F, the fraction of an oral administered drug that reaches systemic circulation), C_{max} (the maximum serum concentration that a drug achieves in a specified test area of the body after the drug has been administrated and before the administration of a second dose), $t_{1/2}$ (elimination half-life, the time required for the concentration of the drug in the plasma to reach half of its original value) and V_{ss} (volume of distribution at steady-state).

The method uses message passing NNs for molecular property prediction ("chemprop") from the Machine Learning for Pharmaceutical Discovery and Synthesis (MLPDS) consortium, a collaboration between industry and the Massachusetts Institute of Technology. The rat PK model achieved good accuracy on key PK parameters (Table 2). CL was predicted within 2-fold error for 75% of compounds and within 3-fold for 90% of compounds.

Table 2. Rat PK model: test set performance.

	R^2	RMSE	Experimental variability
CL	0.57	0.28	0.18
%F	0.48	0.72	0.55
V_{ss}	0.50	0.28	0.21

RMSE = root mean square error

The use of *in vitro in vivo* extrapolation (IVIVe) from human hepatocyte and HLM stability assays, typically the "well stirred model" (WSM)[91] is a widely accepted predictive methodology for human metabolic clearance. The rat PK CL prediction results were compared with those of WSM IVIVe. The *in vivo* rat CL model has higher accuracy (RMSE= 0.28, R^2= 0.57 as opposed to RMSE= 0.43, R^2= -0.11) and is not limited by liver blood flow (LBF). It provides insight into potential additional routes of elimination when compared to WSM (which is restricted by LBF). To test if predictions could be made for virtual compounds purely from chemical structure, the researchers went back to "old" *in silico* models for ADME properties for a training set. The test set was made by a 10% temporal split. RMSE results proved that the rat PK models are useful at the point of design.

To build a human PK model,[92] PK data were extracted from PharmaPendium and curated based on expert opinion. The final dataset contained 1001 SMILES and 4491 compound-dose combinations for 12 PK parameters. For each compound–dose combination median values were calculated per PK parameter. Levels of data completeness for each PK parameter varied from 3.5% to 67%. The data are biased towards optimised compounds with good PK profiles. Doses covered a wide range.

Obrezanova *et al.* built on the rat PK model to predict human PK. The feature set consisted of dose, chemical structure, predicted *in vitro* ADME data and *in vivo* rat PK data. Random forest was used as modelling technique. The split into validation and test sets was random by compound with dose stratification. Varying data distributions and data availability had an impact on the ability to model endpoints. Three endpoints had satisfactory models: oral area under the plasma time–concentration curve (AUC PO, R^2_{test}= 0.63; $RMSE_{test}$ = 0.76), maximum plasma drug concentration *per* oral (C_{max} PO, R^2_{test} = 0.68; $RMSE_{test}$ = 0.62), and volume of distribution

intravenous (V_d IV, R^2_{test} = 0.47; $RMSE_{test}$ = 0.50).[92] Dose is one of the most important features to model AUC PO and C_{max} PO. Predictions of *in vivo* rat PK parameters and *in vitro* ADME properties are also important.

Performance of the models was additionally investigated using an internal AstraZeneca compendium of first-time-in-human measurements in the 2000–2020 period.[93] In addition, drug metabolism and pharmacokinetics (DMPK) prediction values are provided allowing for a side-by-side performance comparison with machine learning models. Despite the different sample sizes and chemical composition of the hold-out test set and internal clinical candidates, the model performance was comparable for both datasets. The accuracy of the ML models was lower than that of pre-clinical prediction (DMPK). Nevertheless, the ML models are fit-for-purpose to be used in early drug discovery and are complementary to current pre-clinical predictions.

The *in vivo* rat and human PK models increase the efficiency of the DMTA cycle allowing scientists to design compounds with better safety and PK properties early in the drug discovery process. The models can drive prioritisation for *in vivo* testing and reduction in animal experiments and guide *de novo* generative models to build in good PK. They can also inform safety-related models of the therapeutic window: predicted human C_{max} can be used to enable safety risk assessment at earlier stages. In future Obrezanova and her colleagues will expand the in-house and commercial datasets, use dog and rat PK models built on larger datasets to improve the human model, and explore transfer learning and multitask learning deep learning architectures.

References

(1) Arús-Pous, J.; Blaschke, T.; Ulander, S.; Reymond, J.-L.; Chen, H.; Engkvist, O. Exploring the GDB-13 chemical space using deep generative models. *J. Cheminf.* **2019,** *11*, 20.
(2) Arús-Pous, J.; Johansson, S. V.; Prykhodko, O.; Bjerrum, E. J.; Tyrchan, C.; Reymond, J.-L.; Chen, H.; Engkvist, O. Randomized SMILES strings improve the quality of molecular generative models. *J. Cheminf.* **2019,** *11*, 71.
(3) Blaschke, T.; Engkvist, O.; Bajorath, J.; Chen, H. Memory-assisted reinforcement learning for diverse molecular de novo design. *J. Cheminf.* **2020,** *12* (1), 68.
(4) Irwin, J. J.; Tang, K. G.; Young, J.; Dandarchuluun, C.; Wong, B. R.; Khurelbaatar, M.; Moroz, Y. S.; Mayfield, J.; Sayle, R. A. ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *J. Chem. Inf. Model.* **2020,** *60* (12), 6065-6073.
(5) Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular de-novo design through deep reinforcement learning. *J. Cheminf.* **2017,** *9*, 48.
(6) Papadopoulos, K.; Giblin, K. A.; Janet, J. P.; Patronov, A.; Engkvist, O. De novo design with deep generative models based on 3D similarity scoring. *Bioorg. Med. Chem.* **2021,** *44*, 116308.
(7) Guo, J.; Janet, J. P.; Bauer, M. R.; Nittinger, E.; Giblin, K. A.; Papadopoulos, K.; Voronov, A.; Patronov, A.; Engkvist, O.; Margreittera, C. DockStream: A Docking Wrapper to Enhance De Novo Molecular Design. http://chemrxiv.org/engage/chemrxiv/article-details/6107fc3340c8bd01539a36f4 (accessed November 5, 2021).
(8) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018,** *555* (7698), 604-610.
(9) Thakkar, A.; Kogej, T.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. J. Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain. *Chem. Sci.* **2020,** *11* (1), 154-168.
(10) Genheden, S.; Thakkar, A.; Chadimová, V.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. *J. Cheminf.* **2020,** *12* (1), 70.
(11) Thakkar, A.; Selmi, N.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. J. "Ring Breaker": Neural Network Driven Synthesis Prediction of the Ring System Chemical Space. *J. Med. Chem.* **2020,** *63* (16), 8791-8808.
(12) Thakkar, A.; Chadimová, V.; Bjerrum, E. J.; Engkvist, O.; Reymond, J.-L. Retrosynthetic accessibility score (RAscore) – rapid machine learned synthesizability classification from AI driven retrosynthetic planning. *Chem. Sci.* **2021,** *12*, 3339-3349.

(13) Green, D. V. S.; Pickett, S.; Luscombe, C.; Senger, S.; Marcus, D.; Meslamani, J.; Brett, D.; Powell, A.; Masson, J. BRADSHAW: a system for automated molecular design. *J. Comput.-Aided Mol. Des.* **2020,** *34* (7), 747-765.
(14) Lewell, X. Q.; Judd, D.; Watson, S.; Hann, M. RECAP-Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **1998,** *38* (3), 511-522.
(15) Degen, J.; Wegscheid-Gerlach, C.; Zaliani, A.; Rarey, M. On the art of compiling and using 'drug-like' chemical fragment spaces. *ChemMedChem* **2008,** *3* (10), 1503-1507.
(16) Hussain, J.; Rea, C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. *J. Chem. Inf. Model.* **2010,** *50* (3), 339-348.
(17) Free, S. M., Jr.; Wilson, J. W. A mathematical contribution to structure-activity studies. *J. Med. Chem.* **1964,** *7* (4), 395-399.
(18) Pogany, P.; Arad, N.; Genway, S.; Pickett, S. D. De Novo Molecule Design by Translating from Reduced Graphs to SMILES. *J. Chem. Inf. Model.* **2019,** *59* (3), 1136-1146.
(19) Bush, J. T.; Pogany, P.; Pickett, S. D.; Barker, M.; Baxter, A.; Campos, S.; Cooper, A. W. J.; Hirst, D.; Inglis, G.; Nadin, A.; Patel, V. K.; Poole, D.; Pritchard, J.; Washio, Y.; White, G.; Green, D. V. S. A Turing Test for Molecular Generators. *J. Med. Chem.* **2020,** *63* (20), 11964-11971.
(20) Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J.; Romero, D. L.; Masse, C.; Knight, J. L.; Steinbrecher, T.; Beuming, T.; Damm, W.; Harder, E.; Sherman, W.; Brewer, M.; Wester, R.; Murcko, M.; Frye, L.; Farid, R.; Lin, T.; Mobley, D. L.; Jorgensen, W. L.; Berne, B. J.; Friesner, R. A.; Abel, R. Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *J. Am. Chem. Soc.* **2015,** *137* (7), 2695-2703.
(21) Amabilino, S.; Pogany, P.; Pickett, S. D.; Green, D. V. S. Guidelines for Recurrent Neural Network Transfer Learning-Based Molecular Generation of Focused Libraries. *J. Chem. Inf. Model.* **2020,** *60* (12), 5699-5713.
(22) Young, T. A.; Gheorghe, R.; Duarte, F. cgbind: A Python Module and Web App for Automated Metallocage Construction and Host-Guest Characterization. *J. Chem. Inf. Model.* **2020,** *60* (7), 3546-3557.
(23) Young, T. A.; Silcock, J. J.; Sterling, A. J.; Duarte, F. autodE: Automated Calculation of Reaction Energy Profiles- Application to Organic and Organometallic Reactions. *Angew. Chem., Int. Ed.* **2021,** *60* (8), 4266-4274.
(24) Young, T. A.; Johnston-Wood, T.; Deringer, V. L.; Duarte, F. A transferable active-learning strategy for reactive molecular force fields. *Chem. Sci.* **2021,** *12* (32), 10944-10955.
(25) Bartok, A. P.; Csanyi, G. Gaussian approximation potentials: A brief tutorial introduction. *Int. J. Quantum Chem.* **2015,** *115* (16), 1051-1057.
(26) Bartok, A. P.; Payne, M. C.; Kondor, R.; Csanyi, G. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Phys. Rev. Lett.* **2010,** *104* (13), 136403.
(27) Bernstein, N.; Csányi, G.; Deringer, V. L. De novo exploration and self-guided learning of potential-energy surfaces. *npj Comput. Mater.* **2019,** *5* (1), 99.
(28) Cole, D. J.; Mones, L.; Csanyi, G. A machine learning based intramolecular potential for a flexible organic molecule. *Faraday Discuss.* **2020,** *224*, 247-264.
(29) Deringer, V. L.; Proserpio, D. M.; Csanyi, G.; Pickard, C. J. Data-driven learning and prediction of inorganic crystal structures. *Faraday Discuss.* **2018,** *211*, 45-59.
(30) Mahoney, M. W.; Drineas, P. CUR matrix decompositions for improved data analysis. *Proc. Natl. Acad. Sci. U.S.A.* **2009,** *106* (3), 697-702.
(31) Young, T. A.; Marti-Centelles, V.; Wang, J.; Lusby, P. J.; Duarte, F. Rationalizing the Activity of an "Artificial Diels-Alderase": Establishing Efficient and Accurate Protocols for Calculating Supramolecular Catalysis. *J. Am. Chem. Soc.* **2020,** *142* (3), 1300-1310.
(32) Spicer, R. L.; Stergiou, A. D.; Young, T. A.; Duarte, F.; Symes, M. D.; Lusby, P. J. Host-Guest-Induced Electron Transfer Triggers Radical-Cation Catalysis. *J. Am. Chem. Soc.* **2020,** *142* (5), 2134-2139.
(33) Raies, A. B.; Bajic, V. B. In silico toxicology: comprehensive benchmarking of multi-label classification methods applied to chemical toxicity data. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2018,** *8* (3), e1352.

(34) Sushko, I.; Salmina, E.; Potemkin, V. A.; Poda, G.; Tetko, I. V. ToxAlerts: A Web Server of Structural Alerts for Toxic Chemicals and Compounds with Potential Adverse Reactions. *J. Chem. Inf. Model.* **2012,** *52* (8), 2310-2316.
(35) Lang, A.; Volkamer, A.; Behm, L.; Röblitz, S.; Ehrig, R.; Schneider, M.; Geris, L.; Wichard, J.; Buttgereit, F. In silico methods - Computational alternatives to animal testing. *ALTEX* **2018,** *35* (1), 124-126.
(36) Morger, A.; Mathea, M.; Achenbach, J. H.; Wolf, A.; Buesen, R.; Schleifer, K.-J.; Landsiedel, R.; Volkamer, A. KnowTox: pipeline and case study for confident prediction of potential toxic effects of compounds in early phases of development. *J. Cheminf.* **2020,** *12* (1), 24.
(37) Morger, A.; Svensson, F.; Arvidsson McShane, S.; Gauraha, N.; Norinder, U.; Spjuth, O.; Volkamer, A. Assessing the calibration in toxicological in vitro models with conformal prediction. *J. Cheminf.* **2021,** *13* (1), 35.
(38) Garcia de Lomana, M.; Morger, A.; Norinder, U.; Buesen, R.; Landsiedel, R.; Volkamer, A.; Kirchmair, J.; Mathea, M. ChemBioSim: Enhancing Conformal Prediction of In Vivo Toxicity by Use of Predicted Bioactivities. *J. Chem. Inf. Model.* **2021,** *61* (7), 3255-3272.
(39) Webel, H. E.; Kimber, T. B.; Radetzki, S.; Neuenschwander, M.; Nazare, M.; Volkamer, A. Revealing cytotoxic substructures in molecules using deep learning. *J. Comput.-Aided Mol. Des.* **2020,** *34* (7), 731-746.
(40) Ji, C.; Svensson, F.; Zoufir, A.; Bender, A. eMolTox: prediction of molecular toxicity with confidence. *Bioinformatics* **2018,** *34* (14), 2508-2509.
(41) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998,** *38* (6), 983-996.
(42) Hanser, T.; Barber, C.; Marchaland, J. F.; Werner, S. Applicability domain: towards a more formal definition dol. *SAR QSAR Environ. Res.* **2016,** *27* (11), 865-881.
(43) Norinder, U.; Carlsson, L.; Boyer, S.; Eklund, M. Introducing Conformal Prediction in Predictive Modeling. A Transparent and Flexible Alternative to Applicability Domain Determination. *J. Chem. Inf. Model.* **2014,** *54* (6), 1596-1603.
(44) Norinder, U.; Rybacka, A.; Andersson, P. L. Conformal prediction to define applicability domain - A case study on predicting ER and AR binding. *SAR QSAR Environ. Res.* **2016,** *27* (4), 303-316.
(45) Svensson, F.; Norinder, U.; Bender, A. Modelling compound cytotoxicity using conformal prediction and PubChem HTS data. *Toxicol. Res. (Cambridge, U. K.)* **2017,** *6* (1), 73-80.
(46) Mervin, L. H.; Cao, Q.; Barrett, I. P.; Firth, M. A.; Murray, D.; McWilliams, L.; Haddrick, M.; Wigglesworth, M.; Engkvist, O.; Bender, A. Understanding Cytotoxicity and Cytostaticity in a High-Throughput Screening Collection. *ACS Chem. Biol.* **2016,** *11* (11), 3007-3023.
(47) Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Mueller, K.-R.; Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One* **2015,** *10* (7), e0130140/1.
(48) Irwin, R.; Dimitriadis, S.; He, J.; Bjerrum, E. Chemformer: A Pre-Trained Transformer for Computational Chemistry. http://chemrxiv.org/engage/chemrxiv/article-details/60ee8a3eb95bdd06d062074b (accessed November 17, 2021).
(49) Matveieva, M.; Polishchuk, P. Benchmarks for interpretation of QSAR models. *J. Cheminf.* **2021,** *13* (1), 41.
(50) Jimenez-Luna, J.; Skalic, M.; Weskamp, N. Benchmarking molecular feature attribution methods with activity cliffs. http://chemrxiv.org/engage/chemrxiv/article-details/613b21fe27d906d4c183cfc1 (accessed November 25, 2021).
(51) Wellawatte, G. P.; Seshadri, A.; White, A. D. Model agnostic generation of counterfactual explanations for molecules. http://chemrxiv.org/engage/chemrxiv/article-details/613268f0d5f0803706ba0c79 (accessed November 25, 2021).
(52) Brown, N.; Fiscato, M.; Segler, M. H. S.; Vaucher, A. C. GuacaMol: Benchmarking Models for de Novo Molecular Design. *J. Chem. Inf. Model.* **2019,** *59* (3), 1096-1108.
(53) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular graph convolutions: moving beyond fingerprints. *J. Comput.-Aided Mol. Des.* **2016,** *30* (8), 595-608.
(54) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019,** *59* (8), 3370-3388.

(55) Zankov, D. V.; Matveieva, M.; Nikonenko, A. V.; Nugmanov, R. I.; Baskin, I. I.; Varnek, A.; Polishchuk, P.; Madzhidov, T. I. QSAR Modeling Based on Conformation Ensembles Using a Multi-Instance Learning Approach. *J. Chem. Inf. Model.* **2021,** *61* (10), 4913-4923.
(56) Nikonenko, A.; Zankov, D.; Baskin, I.; Madzhidov, T.; Polishchuk, P. Multiple Conformer Descriptors for QSAR Modeling. *Mol. Inf.* **2021,** *40* (11), 2060030.
(57) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996,** *39* (15), 2887-2893.
(58) Finnigan, W.; Hepworth, L. J.; Flitsch, S. L.; Turner, N. J. RetroBioCat as a computer-aided synthesis planning tool for biocatalytic reactions and cascades. *Nat. Catal.* **2021,** *4* (2), 98-104.
(59) Kreutter, D.; Schwaller, P.; Reymond, J.-L. Predicting enzymatic reactions with a molecular transformer. *Chem. Sci.* **2021,** *12* (25), 8648-8659.
(60) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **2019,** *5* (9), 1572-1583.
(61) Probst, D.; Manica, M.; Teukam, Y. G. N.; Castrogiovanni, A.; Paratore, F.; Laino, T. Molecular transformer-aided biocatalysed synthesis planning. http://chemrxiv.org/engage/chemrxiv/article-details/60c75919842e6599a7db4990 (accessed November 26, 2021).
(62) Probst, D.; Reymond, J.-L. Visualization of very large high-dimensional data sets as minimum spanning trees. *J. Cheminf.* **2020,** *12*, 12.
(63) Capecchi, A.; Probst, D.; Reymond, J.-L. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *J. Cheminf.* **2020,** *12* (1), 43.
(64) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021,** *596* (7873), 583-589.
(65) Ovchinnikov, S.; Park, H.; Varghese, N.; Huang, P.-S.; Pavlopoulos, G. A.; Kim, D. E.; Kamisetty, H.; Kyrpides, N. C.; Baker, D. Protein structure determination using metagenome sequence data. *Science* **2017,** *355* (6322), 294-298.
(66) Cong, Q.; Anishchenko, I.; Ovchinnikov, S.; Baker, D. Protein interaction networks revealed by proteome coevolution. *Science* **2019,** *365* (6449), 185-189.
(67) Mariani, V.; Biasini, M.; Barbato, A.; Schwede, T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **2013,** *29* (21), 2722-2728.
(68) Xie, Q.; Luong, M. T.; Hovy, E.; Le, Q. V. Self-Training With Noisy Student Improves ImageNet Classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: 2020; pp 10684-10695.
(69) Vriza, A.; Canaj, A. B.; Vismara, R.; Kershaw Cook, L. J.; Manning, T. D.; Gaultois, M. W.; Wood, P. A.; Kurlin, V.; Berry, N.; Dyer, M. S.; Rosseinsky, M. J. One class classification as a practical approach for accelerating π-π co-crystal discovery. *Chem. Sci.* **2021,** *12* (5), 1702-1719.
(70) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: a molecular descriptor calculator. *J. Cheminf.* **2018,** *10*, 4.
(71) Yang, K. K.; Wu, Z.; Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* **2019,** *16* (8), 687-694.
(72) Mazurenko, S.; Prokop, Z.; Damborsky, J. Machine Learning in Enzyme Engineering. *ACS Catal.* **2020,** *10* (2), 1210-1223.
(73) Siedhoff, N. E.; Schwaneberg, U.; Davari, M. D. Machine learning-assisted enzyme engineering. *Methods Enzymol.* **2020,** *643*, 281-315.
(74) Wittmann, B. J.; Johnston, K. E.; Wu, Z.; Arnold, F. H. Advances in machine learning for directed evolution. *Curr. Opin. Struct. Biol.* **2021,** *69*, 11-18.
(75) Siedhoff, N. E.; Illig, A.-M.; Schwaneberg, U.; Davari, M. D. PyPEF-An Integrated Framework for Data-Driven Protein Engineering. *J. Chem. Inf. Model.* **2021,** *61* (7), 3463-3476.

(76) Polykovskiy, D.; Zhebrak, A.; Sanchez-Lengeling, B.; Golovanov, S.; Tatanov, O.; Belyaev, S.; Kurbanov, R.; Artamonov, A.; Aladinskiy, V.; Veselov, M.; Kadurin, A.; Johansson, S.; Chen, H.; Nikolenko, S.; Aspuru-Guzik, A.; Zhavoronkov, A. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. 2018, arXiv e-print archive. http://arxiv.org/abs/1811.12823 (accessed January 20, 2021).
(77) Congreve, M.; de Graaf, C.; Swain, N. A.; Tate, C. G. Impact of GPCR Structures on Drug Discovery. *Cell* **2020,** *181* (1), 81-91.
(78) Thomas, M.; Smith, R. T.; O'Boyle, N. M.; de Graaf, C.; Bender, A. Comparison of structure- and ligand-based scoring functions for deep generative models: a GPCR case study. *J. Cheminf.* **2021,** *13* (1), 39.
(79) Blaschke, T.; Arus-Pous, J.; Chen, H.; Margreitter, C.; Tyrchan, C.; Engkvist, O.; Papadopoulos, K.; Patronov, A. REINVENT 2.0: An AI Tool for De Novo Drug Design. *J. Chem. Inf. Model.* **2020,** *60* (12), 5918-5922.
(80) Kaczor, A. A.; Silva, A. G.; Loza, M. I.; Kolb, P.; Castro, M.; Poso, A. Structure-Based Virtual Screening for Dopamine D2 Receptor Ligands as Potential Antipsychotics. *ChemMedChem* **2016,** *11* (7), 718-729.
(81) Neil, D.; Segler, M.; Guasch , L.; Ahmed, M.; Plumbley, D.; Sellwood, M.; Brown, N. Exploring deep recurrent models with reinforcement learning for molecule design. http://openreview.net/pdf?id=Bk0xiI1Dz (accessed December 1, 2021).
(82) O'Boyle, N. M.; Dalke, A. DeepSMILES: an adaptation of SMILES for use in machine-learning of chemical structures. http://chemrxiv.org/engage/chemrxiv/article-details/60c73ed6567dfe7e5fec388d (accessed November 30, 2021).
(83) Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Mach. Learn.: Sci. Technol.* **2020,** *1* (4), 045024.
(84) Kooistra, A. J.; Vischer, H. F.; McNaught-Flores, D.; Leurs, R.; de Esch, I. J. P.; de Graaf, C. Function-specific virtual screening for GPCR ligands using a combined scoring method. *Sci. Rep.* **2016,** *6*, 28288.
(85) Tran-Nguyen, V.-K.; Bret, G.; Rognan, D. True Accuracy of Fast Scoring Functions to Predict High-Throughput Screening Data from Docking Poses: The Simpler the Better. *J. Chem. Inf. Model.* **2021,** *61* (6), 2788-2797.
(86) Congreve, M.; Andrews, S. P.; Dore, A. S.; Hollenstein, K.; Hurrell, E.; Langmead, C. J.; Mason, J. S.; Ng, I. W.; Tehan, B.; Zhukov, A.; Weir, M.; Marshall, F. H. Discovery of 1,2,4-Triazine Derivatives as Adenosine A2A Antagonists using Structure Based Drug Design. *J. Med. Chem.* **2012,** *55* (5), 1898-1903.
(87) Borodovsky, A.; Barbon, C. M.; Wang, Y.; Ye, M.; Prickett, L.; Chandra, D.; Shaw, J.; Deng, N.; Sachsenmeier, K.; Clarke, J. D.; Linghu, B.; Brown, G. A.; Brown, J.; Congreve, M.; Cheng, R. K.; Dore, A. S.; Hurrell, E.; Shao, W.; Woessner, R.; Reimer, C.; Drew, L.; Fawell, S.; Schuller, A. G.; Mele, D. A. Small molecule AZD4635 inhibitor of A2AR signaling rescues immune cell function including CD103+ dendritic cells enhancing anti-tumor immunity. *J. Immunother. Cancer* **2020,** *8* (2), e000417.
(88) Weiss, D. R.; Bortolato, A.; Tehan, B.; Mason, J. S. GPCR-Bench: A Benchmarking Set and Practitioners' Guide for G Protein-Coupled Receptor Docking. *J. Chem. Inf. Model.* **2016,** *56* (4), 642-651.
(89) Renz, P.; Van Rompaey, D.; Wegner, J. K.; Hochreiter, S.; Klambauer, G. On failure modes in molecule generation and optimization. *Drug Discovery Today: Technol.* **2019,** *32-33*, 55-63.
(90) Jiang, D.; Wu, Z.; Hsieh, C.-Y.; Chen, G.; Liao, B.; Wang, Z.; Shen, C.; Cao, D.; Wu, J.; Hou, T. Could graph neural networks learn better molecular representation for drug discovery A comparison study of descriptor-based and graph-based models. *J. Cheminf.* **2021,** *13* (1), 12.
(91) Yang, J.; Jamei, M.; Yeo, K. R.; Rostami-Hodjegan, A.; Tucker, G. T. Misuse of the well-stirred model of hepatic drug clearance. *Drug Metab. Dispos.* **2007,** *35* (3), 501-502.
(92) Miljkovic, F.; Martinsson, A.; Obrezanova, O.; Williamson, B.; Johnson, M.; Sykes, A.; Bender, A.; Greene, N. Machine Learning Models for Human In Vivo Pharmacokinetic Parameters with In-House Validation. *Mol. Pharmaceutics* **2021,** *18* (12), 4520-4530.
(93) Davies, M.; Jones, R. D. O.; Grime, K.; Jansson-Lofmark, R.; Fretland, A. J.; Winiwarter, S.; Morgan, P.; McGinnity, D. F. Improving the Accuracy of Predicted Human Pharmacokinetics: Lessons Learned from the AstraZeneca Drug Pipeline Over Two Decades. *Trends Pharmacol. Sci.* **2020,** *41* (6), 390-408.

- -

News from Catalyst Science Discovery Centre & Museum

Contribution from RSC-CICAG Treasurer Dr Diana Leitch MBE, FRSC, email: diana.leitch@googlemail.com

Catalyst Science and Discovery Centre is open from 10am – 5pm from Tuesday – Sunday every week.

Double celebration at Catalyst!

There was cause for a double celebration at Catalyst when friends and relatives of the famous Brunner family gathered to witness the naming of the Henry and John Brunner room in the Catalyst archives, and the presentation of two artefacts for permanent display in the museum.

Born and educated in Liverpool, Henry and John Brunner came to Widnes in 1861 to work for John Hutchinson whose offices were based in the building that today houses Catalyst Science Discovery Centre and Museum and where they later met Ludwig Mond with whom John subsequently formed the chemical company Brunner Mond and Co. in Northwich.

The Henry and John Brunner room was formally opened by George Windsor, Earl of St Andrews whose mother Katharine, Duchess of Kent is descended from the Brunner family; her mother was the granddaughter of Sir John Brunner. Also in attendance was Robert Mee the High Sheriff of Cheshire, Derek Twigg MP for Halton, Alex Cowan one of the founders of Catalyst, and descendants of both Henry and John Brunner.

Opening the room, the Earl of St Andrews said: "I was delighted to be asked to open the Henry and John Brunner Room at the Catalyst Centre. Catalyst does fantastic work in showing young people (and not just young people) the fascination, excitement and importance, both today and in the past, of chemistry and the chemical industry, while not ignoring its negative aspects such as damage to both health and the environment…We need chemistry more than ever to meet the challenges of today's world, from clean and renewable energy to carbon capture and storage."

The two artefacts presented to Catalyst by Sir Hugo Brunner, great grandson of Sir John Brunner, are family heirlooms and comprise a ceremonial key that was presented to Sir John Brunner when he opened the Widnes-Runcoun Transporter Bridge in 1905 and a silver table bell, also presented to Sir John Brunner on the occasion of the re-opening of the Bridge in 1913 after its generator was replaced by mains power. These artefacts will be on permanent display at Catalyst.

On presenting the artefacts, Sir Hugo Brunner said: "It is wonderful for my family and I to be here at Catalyst today to see the great changes that have taken place throughout the pandemic. Catalyst is a very special place as both an educational establishment, encouraging young people to study science, and as a museum, preserving the rich heritage of this area. We feel that it is fitting for these artefacts to have a permanent home at Catalyst, alongside other Brunner memorabilia and so close to the site of the Transporter Bridge."

CEO of Catalyst, Martin Pearson commented "It has been a delight to host the extended Brunner family here at Catalyst. Their family is directly linked to our local heritage and their continued support to everything we do at Catalyst is very much appreciated. The gift of the Transporter bell and key will form a significant addition

to our archive and one I'm sure our visitors will enjoy seeing and reading about the history of the Transporter bridge in our Brunner gallery."

Chair of Catalyst Trustees, Dr Diana Leitch MBE said: "Although we have named rooms after some of the greats of the early chemical industry world in Widnes – William Gossage, Henry Deacon and John Hutchinson, we have never had a named room commemorating the work of the brothers Henry and John Brunner. This was a great opportunity to do that in the presence of assembled descendants of both brothers – Henry the qualified and knowledgeable scientist who spoke German from his student days in Switzerland and John the gifted entrepreneurial clerical and financial person. Henry's work with German-speaking Ludwig Mond led to the first-ever adoption of Mond's sulphur recovery process at Hutchinson's Works where he went on to be Works Manager. In 1881 he was a founder member of the Society of Chemical Industry (SCI) which started in Widnes as the Lancashire Chemical Society. The room is a fitting tribute to these greats of the Victorian 'Northern Powerhouse' which Widnes was and is rising to be again. Catalyst is very proud to be part of that."

Sir Hugo Brunner and Martin Pearson with the ceremonial key and the silver table bell presented to Catalyst.

New hydrogen exhibit at Catalyst Science Discovery Centre

Young people will be able to take on the climate change challenge thanks to a new interactive 'Net Zero' game at the Catalyst Science Discovery Centre and Museum in Widnes. Year 6 pupils at Lunt's Heath Primary School, Widnes, were the first to test run the exhibit, made possible by the North West Hydrogen Alliance (NWHA), which represents over 30 of the UK's most influential organisations driving forward a hydrogen economy in the region. They attended a launch event with Cllr Gill Wood, Deputy Portfolio Holder for Climate Emergency & Renewable Energy at the Liverpool City Region Combined Authority.

Local schoolchildren exploring the new Net Zero game at Catalyst.

The three-level, fun-filled game teaches all ages about the ways in which we can help tackle the climate challenge and deliver a net zero future. It takes players on a journey from reliance on fossil fuels and gas to a greener future in 2050 where hydrogen is used to power industry, fuel our cars and heat our homes. The game is based on the north west and some of the projects anticipated to be delivered in the region over the coming years, including carbon capture and storage beneath Liverpool Bay as part of HyNet North West.

Professor Joe Howe, Chair of the NWHA and Executive Director, Energy Research Institute at the University of Chester, said: "Young people understand the simple fact that climate change is a massive risk to society and that we must make as many changes as we can to tackle it. We've got to empower the young generations to have a passion for delivering a greener future and equip them with knowledge of how that can happen. Our fantastic Catalyst exhibition brings that to life in way that is both educational and fun. When it comes to hydrogen, there'll be so many opportunities and jobs in the north west as we roll it out as a clean energy source. The children visiting the museum now could be the ones taking up these skilled roles in the future and we hope the new exhibit inspires them to think about a career in net zero."

The donation from the NWHA follows swiftly on from COP26, the UN climate change conference, held in Glasgow. The north west has ambitions to become the UK's first net zero region by 2040, with the rollout of hydrogen being a key contributor to reducing emissions.

Cllr Gill Wood, Deputy Portfolio Holder for Climate Emergency & Renewable Energy at the Liverpool City Region Combined Authority, said: "The Liverpool City Region Combined Authority was the first combined authority to declare a climate emergency and then put in place ambitious targets to become net zero. It's fantastic to have this facility that gets children really thinking about the decisions and investment needed to reach our climate change goals. Hydrogen will play such a significant role in our low carbon future and will really transform the green opportunities for those in school now. Our young people today will be the leaders and innovators of tomorrow, so it's great to see them so enthused about tackling the climate change challenge."

Martin Pearson, Catalyst CEO, said: "It was fantastic to see the next generation of scientists engaging with our new interactive 'Net Zero' exhibit sponsored by the North West Hydrogen Alliance…Catalyst is proud to play its part in the science communication of this important technology in a fun and educational way."

- -

AI3SD News and 2022 Conference

Contribution from AI3SD Network+ Coordinator Dr Samantha Kanza, email: s.kanza@ai3sd.org

Save the date for our Network Conference: 1-3 March 2022.

We would be delighted if you would join us for our AI3SD Network Conference on 1-3 March 2022. We are hoping to run this as a hybrid event at the Best Western Chilworth Manor, although it may be moved to fully online depending on the Covid situation. The conference will be a mixture of keynote talks, discussion sessions and networking opportunities.

If you would be interested in submitting a short talk abstract please complete our AI3SD Conference 2022 Abstract Submission Form. There will also be some musical entertainment in the evening! We will be opening up registration for this conference closer to the time. Further details can be found on our website.

If you would be interested in finding out how to sponsor this event please contact us info@ai3sd.org and we can tell you more.

New network management resource

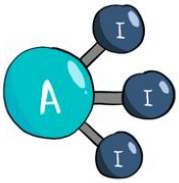
AI3SD has been working as part of the Network of Networks group to combine our shared knowledge on how to run a Network+. This resource has been produced by a group of diverse research management professionals, representing different disciplines and organisations to aid network managers and investigators in the creation and management of research communities. Please check it out here and share with your contacts.

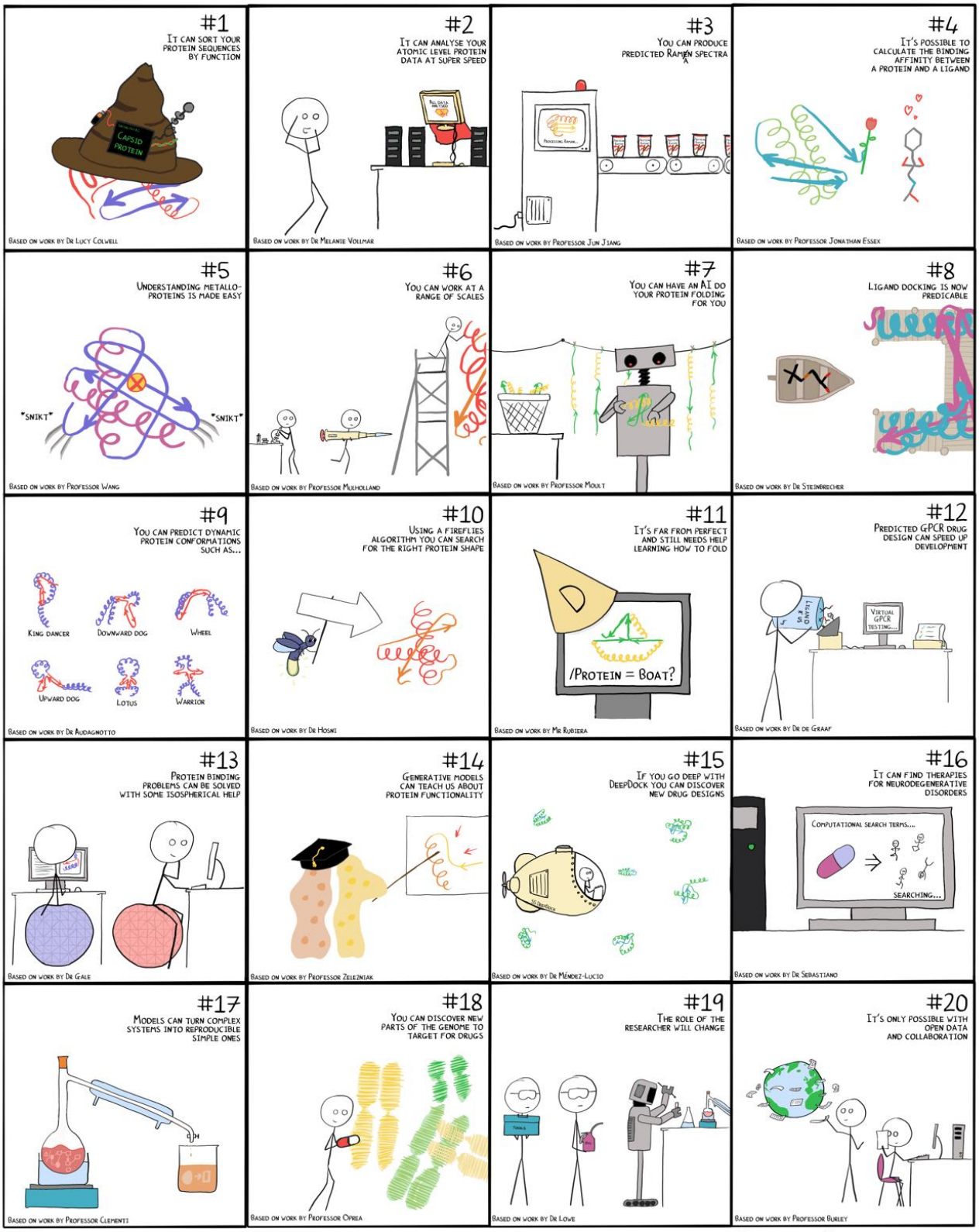
AI4Proteins

Back in April, AI3SD and RSC-CICAG launched our AI4Proteins Seminar Series. The videos from this series can be found on our AI4Proteins Playlist. This entire series has been captured in an incredibly detailed report by the wonderful Dr Wendy Warr, and can be downloaded from here.

WHY YOU SHOULD LOVE MACHINE LEARNING AND PROTEINS

CARTOONS BASED ON TALKS GIVEN BY PARTICIPANTS IN THE 2021 AI 4 PROTEINS SEMINAR SERIES





AI3SD.ORG

ERRANTSCIENCE.COM

AI3SD & PSDS Skills4Scientists series 2021 and internship programme

As detailed in CICAG's summer newsletter, AI3SD teamed up with the Physical Sciences Data-science Service (PSDS) to create a Skills4Scientists series for our summer interns. This included events to educate scientists on: research data management, Python, version control, LaTeX, creating posters and presentations, ethics and valuable careers skills, including a two-day virtual event run in conjunction with RSC-CICAG on careers and posters. All of the material from these sessions are now available on our YouTube Channel: Skills4Scientists Playlist.

Here is a list of our intern projects with links to their reports:

Project Title	Project Student	Project Supervisor
A deep neural network for generation of functional organic molecules	Rhyan Barrett (University of Warwick)	Dr Reinhard Maurer & Dr Julia Westermayr (University of Warwick)
Interactive Knowledge-Based Solvent Selection Tool (report embargoed)	Hewan Zewdu (University of Nottingham)	Professor Jonathan Hirst (University of Nottingham)
X-ray Coherent Imaging using AI-based Phase Reconstruction (report coming soon)	Gavin Man (University of Oxford)	Dr Bill Brocklesby (University of Southampton)
Optimising Ag/Au Alloyed Nanoparticle Catalysts in Continuous Flow; Discrete vs. Continuous Variable Optimisation	Louis Greenhalgh (University of Leeds)	Dr Thomas Chamberlain (University of Leeds)
Machine Learning Physics Models for Materials Self-Assembly	Aspen Fenzl (University of Sheffield)	Professor Nigel Clarke (University of Sheffield)
Relative Structural Analysis on Molecular Perovskite	Kevin Daniel Calvache (Queen Mary University of London)	Dr Anthony Phillips (Queen Mary University of London)
Curating a chemical dataset to train recurrent neural network models to predict IUPAC names from InChI's	Thomas Allam (University of Southampton)	Professor Simon Coles (University of Southampton)
Learning the Crystallographic Phase Problem	Sarah Scripps (University of Edinburgh)	Dr James Cumby (University of Edinburgh)
Latent Space Encoding of Molecular Crystal Structure	King Wong (University of Southampton)	Professor Graeme Day (University of Southampton)
High-throughput generation of structural isomers for fast development of molecular datasets to train machine learning algorithms	Anna Catton (Swansea University)	Dr Francisco Martin-Martinez (Swansea University)
Nearer the nearsightedness principle: Large-scale quantum chemical calculations	Andras Vekassy (University of Southampton)	Professor Chris-Kriton Skylaris (University of Southampton)
Computer Vision in High Throughput Chemistry (report embargoed)	Jamie Longino (University of Strathclyde)	Dr Marc Reid (University of Strathclyde)
Structure-activity relationship analysis of supramolecular antimicrobials (report embargoed)	Kaylee Patel (Unviersity of Manchester)	Dr Cally Hayes (UCL)

Bayesian Optimisation in Chemistry (report embargoed)	Rubaiyat Khondaker (University of Cambridge)	Dr Stephen Gow (University of Southampton)
Creating a merged dataset and investigation of correlations in the data with SOM and VAE models (report embargoed)	Maximilan Hoffman (Freie Universität Berlin)	Mr Samuel Munday (University of Southampton)

Machine learning for materials and chemicals 2021
Most of the videos from our Machine Learning for Materials and Chemicals (ML4MC) Summer School run with the Directed Assembly Network are now available on our YouTube Channel: ML4MC Playlist.

AI3SD Autumn Seminar Series 2021
We have finished our Autumn Seminar Series. This was a 10 part series with 21 talks covering topics on:

- Linked Data, Ontologies & Deep Learning
- Explainable AI & ML
- Data Science 4 Chemistry
- AI & ML 4 Drugs & Materials
- Quantum Machine Learning
- Medicinal Chemistry
- Digital Twins
- Molecules, Graphs & Networks
- Large Spaces
- Molecules & Data

All the videos from this series can be found on our AI3SD YouTube Channel: AI3SD Autumn Seminar Series.

- -

RSC CICAG Open-Source Tools for Chemistry Workshops

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

In 2020 RSC CICAG ran a five-day virtual meeting on Open-Source Chemical Sciences, this event had three streams Open Data, Open Publishing and Open-Source tools for Chemistry. The Open-Source tools for Chemistry workshops proved to be enormously popular and so CICAG held a series of monthly workshops through 2021. These workshops covered a variety of open-source tools and resources ranging from visualisation tools, data analysis using cheminformatics toolkits, and online resources like the PDB.

The workshops were all recorded and are available on the CICAG YouTube channel. As we plan for this year's workshops, I thought it might be timely to remind everyone what is now available and also to thank all the presenters and developers who made the workshops possible. I've included links to all the workshops below:

PDB workshop 2 using Mol*
PDB workshop 1 Registration system
Clustering using KNIME
Web apps for fragment-based drug discovery

Introduction to Cheminformatics and Machine Learning
Oxford Protein Informatics Group antibody modelling tools
Advanced DataWarrior
GNINA
Chemical Structure validation/standardisation
ChimeraX
DataWarrior
ChEMBL
UsingGoogleCoLab workshop
Fragalysis workshop
Knime workshop
PyMOL workshop

These workshops have now been viewed nearly 13,500 times and some of the comments are worth highlighting:

"I am not sure why this software is not famous. This is presumably the best chemoinformatics software I have seen, Great presentation!"

"Two hours of distilled pure science."

CICAG is extremely grateful to our sponsor, Liverpool Chirochem, for enabling these workshops to take place.

News from CAS

Contribution from Dr Anne Jones, email: ajones2@acs-i.org

As we wrap up and reflect on 2021, CAS is proud to share some accomplishments and highlights of our activities over the past year. While the pandemic continues to have long-reaching impacts around the globe, CAS has remained dedicated to supporting our customers and helping them accelerate scientific breakthroughs and discoveries.

In April, CAS launched our new brand, reflecting the broadening scope of our solutions and capabilities critical to advancing scientific discovery. This new brand mirrors the organisation's continued goal to curate, connect, and analyse scientific knowledge in ways that help innovators find new connections and accelerate innovation.

COVID-19

From the onset of the unprecedented COVID-19 pandemic in 2020, CAS has made a wide range of resources and expertise openly available to support the search for a cure. From anti-viral candidate SAR datasets, to assay techniques and diagnostic test development methods, the curated and published CAS content have been some of the most downloaded material ever offered by the organisation.

Webinars & 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.

Blog

The CAS blog features deep insights into a variety of issues impacting the scientific community. Recently covered topics include bioorthogonal reactions, applications, and trends in the CAS Content Collection™; emerging trends in targeting "undruggable" RAS proteins for cancer treatment; assessing structural novelty of the first AI-designed drug candidates; and more. Visit and follow the CAS blog to stay up to date on new insights and resources.

CAS SciFinder Discovery Platform

In 2021, CAS launched the CAS SciFinder Discovery Platform, an enterprise-wide platform solution with workflow tools and capabilities designed to support multiple scientific research requirements. The CAS SciFinder Discovery Platform includes CAS SciFinder^n, CAS Formulus, CAS Analytical Methods, as well as all the new enhancements to Retrosynthetic Planning, and our newest capabilities in Biosequences.

CAS SciFinder^n

From small workflow adjustments based on input of our customers to far-reaching expansions to core capabilities, our team of technology and scientific experts collaborate daily to ensure that CAS SciFinder^n continues to meet the ever-increasing needs of the scientific community. Several notable recent enhancements include:

- Workflow enhancements to improve the retrosynthetic planning process
- Improvements that allow more flexibility when performing reference searches
- Numerous filter and downloading options to get users to their best result more quickly
- The introduction of NCBI sequences in the biosequences interface, taking CAS SciFinder^n beyond the bounds of chemistry

For more detailed information on the latest enhancements, visit the What's New portion of the help section within CAS SciFinder^n, 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

2021 was a great year in CAS Formulus advancements. CAS released our exciting Formulation Designer feature allowing creation of a prototype formulation with suggested ingredients, alternative ingredients, functional roles, and associated regulatory information.

A new compare feature was deployed in which up to three formulations can be viewed side-by-side for easy analysis of differences and similarities in composition.

A new Commonly Used As feature was released, which provides an overview of what functional roles an ingredient typically plays in formulations. To enable further discovery of interesting ingredients, we released a new feature identifying similar ingredients with regulatory information.

Finally, to better support the formulation workflow, we enhanced exporting capabilities to XLSX file format and extended exporting to include formulation results, formulation details, and ingredient results.

STN IP Protection Suite

CAS continues to provide solutions for customers – from scientists to professional searchers – to identify and navigate the IP landscape. CAS launched the STN IP Protection Suite, comprised of STNext®, CAS Scientific Patent Explorer™ (a new, intuitive tool to help searchers navigate the competitive landscape), and CAS Search Guard™.

New enhancements this year include:

- Development of proprietary AI-driven prior art searching capabilities to support STNext users
- Enrichment of the Biosequences offering
- Extended reporting capabilities in STNext
- Extension of dedicated IP-focused Customer Success team across EMEA and North America.

CAS-FIZ partnership expanded, a more efficient path to new discoveries with the continued growth and enhancements to STNext.

News from the RSC

Contribution from Richard Kidd, email: KiddR@rsc.org

The RSC has collected its Text and Data Mining (TDM) tools together, to help organisations make the most of TDM's potential. You can now request the following resources from the RSC:

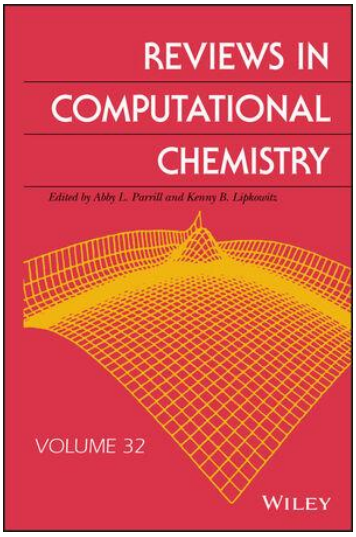
- A full text XML article sample
- Tokenised full text articles from *Chemical Science*

You'll find a range of open-source TDM resources to use on the TDM: free resources and open source software web page, including ChemListem, Chemtok and chemical ontologies. Head to the above page to learn more and request samples.

Chemical Information / Cheminformatics and Related Books

Contributed by Stuart Newbold, email: stuart@psandim.com

Reviews in Computational Chemistry, Volume 32

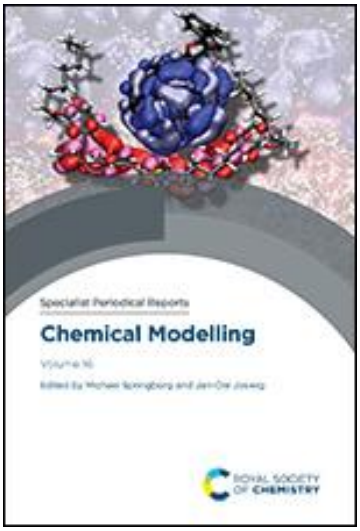


The latest volume in the *Reviews in Computational Chemistry* series, the invaluable reference to methods and techniques in computational chemistry.

Reviews in Computational Chemistry reference texts assist researchers in selecting and applying new computational chemistry methods to their own research. Bringing together writings from leading experts in various fields of computational chemistry, Volume 32 covers topics including global structure optimisation, time-dependent density functional tight binding calculations, non-equilibrium self-assembly, cluster prediction, and molecular simulations of microphase formers and deep eutectic solvents. In keeping with previous books in the series, Volume 32 uses a non-mathematical style and tutorial-based approach that provides students and researchers with easy access to computational methods outside their area of expertise.

Wiley
Abby L. Parrill, Kenny B. Lipkowitz
ISBN: 978-1-119-62589-6 (available April 2022)

Chemical Modelling: Volume 16

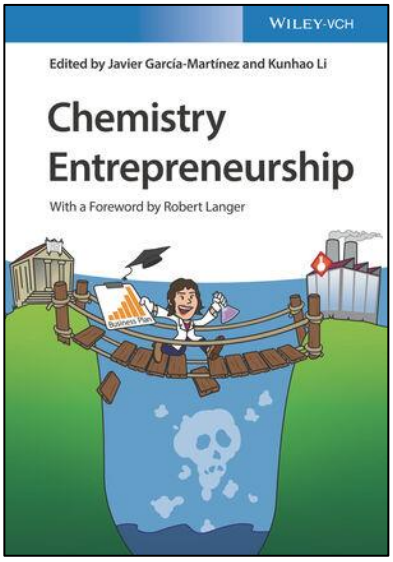


Chemical modelling covers a wide range of disciplines and this book is the first stop for any materials scientist, biochemist, chemist or molecular physicist wishing to acquaint themselves with major developments in the applications and theory of chemical modelling.

Containing both comprehensive and critical reviews, it is a convenient reference to the current literature. Coverage includes, but is not limited to, isomerism in polyoxometalate chemistry, modelling molecular magnets, molecular modelling of cyclodextrin inclusion complexes and graphene nanoribbons heterojunctions.

RSC Publishing
Editors: Michael Springborg, Jan-Ole Joswig
ISBN: 978-1-83916-170-4

Chemistry Entrepreneurship



Chemistry Entrepreneurship is a step-by-step guide that is specifically devoted to understanding what it takes to start and grow a new company in the chemistry sector. Comprehensive in scope, the book covers the various aspects of the creation of a new chemical enterprise including: the protection of the invention, the business plan, the transfer from the research center or university, the financing, the legal setup, the launching of the company and its growth and exit strategies.

This hands-on book contains the information needed to help to determine if you have what it takes to be a chemistry entrepreneur, explains how to take an ideas out of the lab and into the real world, reveals how to develop your burgeoning business, and shows how to sustain and grow your business. This much-needed resource also includes interviews with founding scientists who created their own successful chemical companies. This important book:

- Provides the practical information on how to start a company based on a scientific breakthrough
- Offers information on the mindset it takes to become, and remain, successful in the marketplace
- Presents case studies from world-renowned and highly experienced professionals who have successfully started a company

Written for chemists in industry, chemists, materials scientists, chemical engineers, Chemistry Entrepreneurshipis a guide for becoming a founder of a successful chemical company.
Wiley
Javier García-Martínez, Kunhao Li
ISBN: 978-3-527-81987-4

Other Chemical Information News

Contributed by Stuart Newbold, email: stuart@psandim.com

Dr Wendy Warr's latest **AI3SD report**, AI 4 Proteins: Protein Structure Prediction, is now available on open access at https://eprints.soton.ac.uk/452733/. This is a report on a series of virtual meetings organised by the AI3SD Network (Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery) and the Royal Society of Chemistry Chemical Information and Computer Applications Group (RSC-CICAG) on April 14, May 5, May 26, and June 16-17, 2021. Direct link: https://eprints.soton.ac.uk/452733/1/AI3SD_Event_Series_Report_23_AI4Proteins.pdf

£1.02 Million Research Funding awarded for Robotic Chemist Project
Dstl (Defence Science and Technology Laboratory) is pioneering the use of autonomous mobile robotics within the scientific research laboratory.
https://www.gov.uk/government/news/102-million-research-funding-awarded-for-robotic-chemist-project
Source: UKRI

AI in Drug Discovery Market to Expand by 36.1% CAGR as Popularity of Mindful AI Offers Breakthrough Opportunities

Artificial intelligence (AI) is expected to be a lucrative technology in the healthcare industry. The implementation of AI reduces research and development gap in the drug manufacturing process and helps in targeted manufacturing of drugs. Hence, biopharmaceutical companies are turning to AI to enhance market share. This is a major factor fuelling the growth of the global AI for drug discovery market.

https://www.biospace.com/article/ai-in-drug-discovery-market-to-expand-by-36-1-percent-cagr-as-popularity-of-mindful-ai-offers-breakthrough-opportunities-notes-tmr/

Source: BioSpace

Nominations are open for International Information Manager and Tony Kent Strix awards

UKeIG announces that nominations are open for the International Information Manager of the Year Award 2021 and for the Tony Kent Strix Award 2021.

https://www.infotoday.eu/Articles/News/Featured-News/Nominations-are-open-for-International-Information-Manager-and-Tony-Kent-Strix-awards-148233.aspx

Source: Information Today

Digital Science launches Dimensions Life Sciences & Chemistry

Digital Science has announced the launch of a new version of its popular Dimensions platform – Dimensions Life Sciences & Chemistry (Dimensions L&C) – focused on life sciences and chemistry research activities. Dimensions L&C analyses more than 120 million scientific publications, millions of patents, grants and clinical trial documents. It is both larger than other databases, and unlike traditional manually curated tools, applies up-to-the-minute semantic text analysis tools and ontologies, providing powerful up-to-date discovery functionality previously unavailable at such scale.

https://www.digital-science.com/press-release/digital-science-launches-dimensions-life-sciences-chemistry/

Source: Digital Science

Machine Learning reveals how Glucose helps the SARS-CoV-2 virus

The École Polytechnique Fédérale de Lausanne's (EPFL) Blue Brain Project has deployed its simulation technology and expertise in cellular and molecular biology to develop a better understanding of the severity of Covid-19.

https://www.scientific-computing.com/news/machine-learning-reveals-how-glucose-helps-sars-cov-2-virus

Source: Scientific Computing World

April19 and CAS partner to fast-track real-world application of AI to discovery of novel Mental Health Therapeutics

April19 Discovery (April19), an artificial-intelligence-enabled drug discovery company and CAS have announced a collaboration to accelerate the identification of novel small-molecule therapeutic drug candidates. As part of this collaboration, CAS is using scientist-curated chemical substance and patent data from the CAS Content Collection™, as well as specialised chemical analytics, to prioritise leads generated by April19's advanced AI-led drug discovery approach. By helping April19 advance the lead compounds that are most innovative and most likely to be patentable, this collaboration accelerates the company's aspiration to address the growing need for targeted therapeutics to treat mental health conditions by capturing the untapped potential of psychedelics.

https://www.knowledgespeak.com/news/april19-and-cas-partner-to-fast-track-real-world-application-of-ai-to-discovery-of-novel-mental-health-therapeutics/

Source: Knowledgespeak

Taking the Research Journal in a New Direction

A new journal concept from Cambridge University Press will bring researchers from different fields together around the fundamental questions that cut across traditional disciplines. By focussing research on finding answers to such questions, this unique approach will speed discovery by fostering collaboration and knowledge sharing between subject communities. It will also provide opportunities to publish research from areas that are not well served by traditional, discipline-specific journals.

https://www.cambridge.org/news-and-insights/news/Taking-the-research-journal-in-a-new-direction

Source: CUP

EMBL-EBI selects Google Cloud as Strategic Partner to accelerate the pace of Research

Hosting the world's most comprehensive set of freely available and up-to-date molecular data resources, EMBL's European Bioinformatics Institute (EMBL-EBI) has announced it has chosen Google Cloud as a strategic cloud partner. As part of a new, comprehensive, five-year partnership, EMBL-EBI will tap Google Cloud's innovative technologies and global infrastructure to accelerate the pace of service delivery to its global user community, which includes research labs, pharmaceutical companies, academic institutions, and more.

The partnership between Google Cloud and EMBL-EBI aims to:

- Improve access to biomedical research through the use of cloud technology as an exemplar to the global community.
- Use analytics and machine learning to glean better insights from data to help speed up the pace of scientific discovery and to distribute these insights globally.
- Support EMBL-EBI's multicloud and hybrid-cloud strategy by providing a flexible platform to develop new cloud tools and technologies.
- Train EMBL-EBI staff in building, deploying, and using cloud-native applications to accelerate cloud adoption within the life sciences community.

https://www.scientific-computing.com/news/embl-ebi-selects-google-cloud-strategic-partner-accelerate-pace-research

Source: Scientific Computing World

GitHub's AI Programming Assistant can introduce Security Flaws

A neural network that automatically generates source code to help human programmers complete projects has been found to include bugs or security flaws in up to 40 per cent of the code it outputs.

https://www.newscientist.com/article/2288699-githubs-ai-programming-assistant-can-introduce-security-flaws/#ixzz7H7Hjdmii

Source: NewScientist

Is AI the Future of Scholarly Publishing?

AI, Data Science and other emerging technologies are revolutionising the creation, dissemination and consumption of academic, research and professional content.

https://www.stm-publishing.com/is-ai-the-future-of-scholarly-publishing/

Source: STM Publishing News

World's leading Academic Journals join pledge to fight bias in Publishing

Global giants Springer Nature, De Gruyter and Taylor & Francis join Wiley, Elsevier and SAGE Publishing to sign up to the Royal Society of Chemistry's commitment for equality in publishing.

https://www.stm-publishing.com/worlds-leading-academic-journals-join-pledge-to-fight-bias-in-publishing/

Source: STM Publishing News

MEDLINE 2022 initiative - Transition to Automated Indexing

As part of the efforts of the National Library of Medicine (NLM) to transform and accelerate biomedical discovery and improve health and healthcare, NLM is transitioning to automated MeSH indexing of MEDLINE citations in PubMed. Automated indexing will provide users with timely access to MeSH indexed metadata and allow NLM to scale MeSH indexing for MEDLINE to the volume of published biomedical literature. Human indexers have been and will continue to be involved in the refinement of automated indexing algorithms and will play a significant role in the quality assurance approaches for automated indexing.

https://www.knowledgespeak.com/news/medline-2022-initiative-transition-to-automated-indexing/

Source: Knowledgespeak

An AI Revolution

Robert Roe finds that the use of AI is driving new areas of research and increasing the competitiveness of early adopters. The confluence of data, compute power and advances in the design of algorithms for AI (artificial intelligence) and ML (machine learning) are driving new approaches in the laboratory. This gives scientists access to additional tools that can open new avenues for research or accelerate existing workflows.

https://www.scientific-computing.com/feature/ai-revolution

Source: Scientific Computing World

It's time to stop Excluding Disabled Scientists

Science and society are not built to welcome disabled people – and that's a fault that we have to start fixing, says RSC disability and accessibility specialist Emrys Travis.

https://www.rsc.org/news-events/opinions/2021/jul/time-to-stop-excluding-disabled-scientists/

Source: RSC News

AAAS launches new Science Partner Journal, Intelligent Computing

The American Association for the Advancement of Science has announced the launch of new Science Partner Journal, *Intelligent Computing*, published in affiliation with Zhejiang Lab. Intelligent Computing's mission is to build an open communication and cooperation platform for global science and technology professionals in the field of intelligent computing. It is interested in presenting the latest research outcomes and technological breakthroughs in intelligent computing in order to boost the development of intelligent computing science and technology, promote global academic communication and cooperation, and enhance human well-being. Categories of manuscripts include Research Articles, Review Articles, and Perspectives. Intelligent Computing is currently open for submission and will publish under a Creative Commons Attribution License (CC BY).

https://www.knowledgespeak.com/news/aaas-launches-new-science-partner-journal-intelligent-computing/

Source: Knowledgespeak

New Journal to push the Boundaries of Biological Imaging

A new Open Access journal from Cambridge University Press will provide a home for interdisciplinary research in the fast-growing field of quantitative and computational imaging in the life sciences.

https://www.cambridge.org/news-and-insights/news/New-journal-to-push-the-boundaries-of-biological-imaging

Source: CUP

Five Online Tools that aim to save Researchers Time and Trouble

From investigating a lab's publication history to scanning manuscripts for statistical errors, these apps can help streamline some of the most time-consuming tasks.

https://www.natureindex.com/news-blog/five-online-tools-that-aim-to-save-researchers-time-and-trouble

Source: Nature

Wikipedia tests AI for Spotting Contradictory Claims in Articles
Artificial intelligence can be used to scour the crowdsourced encyclopedia for contradictory information and flag it to human editors.
https://www.newscientist.com/article/2298169-wikipedia-tests-ai-for-spotting-contradictory-claims-in-articles/#ixzz7H7DtnG1S
Source: NewScientist

Digital Science partners with OntoChem GmbH to better support Life Sciences industry R&D
Digital Science has announced a new partnership with OntoChem GmbH. The partnership allows OntoChem and Digital Science to join forces for mutual clients, particularly in the Life Sciences industry, through OntoChem's powerful semantic indexing capabilities. OntoChem has more than 14 years' experience creating innovative technologies. The German-based life sciences company develops cognitive computing solutions, indexing intranet and internet data and applying semantic search solutions for pharmaceutical, material science and technology-driven businesses.
https://www.digital-science.com/press-release/digital-science-partners-with-ontochem/
Source: Digital Science

Karger Publishers Launches Trial with Writefull Language Check
Karger Publishers has started offering its authors an automated proofreading service from the company Writefull. Writefull uses AI-based language models to suggest language edits, enabling authors to improve the language of their manuscript before submission.
https://www.stm-publishing.com/karger-publishers-launches-trial-with-writefull-language-check/
Source: STM Publishing News

Wiley acquires eJournalPress
Global research and education leader Wiley has announced the asset purchase of eJournalPress (EJP), a leading provider of software and support services for scholarly publishing. With this investment, Wiley plans to drive the evolution of the technology and platforms that underpin research publishing and advance the future of research communication. EJP's online manuscript submission, peer review, and journal production tracking systems deliver a comprehensive service for authors, editors and publishers to create, review and manage scholarly content as it moves through peer-review and gets published online. With the purchase of EJP, Wiley seeks to further its mission to support researchers and enable discovery.
https://www.knowledgespeak.com/news/wiley-acquires-ejournalpress/
Source: Knowledgespeak

OpenAthens Launches MyAthens Plus with CORE Open Access
Newly redesigned OpenAthens' library-owned information portal *MyAthens Plus* has just announced its integration with *CORE*, the world's largest collection of open access full texts, which are used and referenced by people globally, including researchers, libraries, software developers, funders and many more.
https://www.stm-publishing.com/openathens-launches-myathens-plus-with-core-open-access/
Source: STM Publishing News

Writefull's tech disrupts AI-based Proofreading
Writefull, provider of automated language editing services, has announced the release of a new set of language models that offer next-generation automated editing to the world – signalling a significant change in how AI can support authors and editors. Writefull's new models, made available to all users through a mode called 'Full Edit', make language edits of unprecedented quality, with the ability to rewrite entire sentences where

needed. This is the first time an AI-based language tool demonstrates an understanding of sentences that goes beyond grammar.
https://www.digital-science.com/press-release/writefulls-tech-disrupts-ai-based-proofreading/
Source: Digital Science

AIP Publishing to Launch two Fully Open Access Journals in 2022
AIP Publishing (AIPP) has announced the addition of two new titles to the growing family of Open Access (OA) journals. *APL Energy* and *APL Machine Learning* will join a portfolio that also includes *APL Materials, APL Photonics, and APL Bioengineering* – three high-impact Gold OA journals that advance open science while preserving the diversity, quality, and financial sustainability of the peer-reviewed publishing upon which PLOS research community depends. The journals will open for submissions in mid-2022 and start publishing by the end of 2022.
https://www.stm-publishing.com/aip-publishing-to-launch-two-fully-open-access-journals-in-2022/
Source: STM Publishing News

CCC Exceeds 1,000 Institutions and Funders on its RightsLink for Scientific Communications Platform
CCC, a leader in advancing copyright, accelerating knowledge, and powering innovation, has announced that it has surpassed 1,000 institutions and funders on its RightsLink for Scientific Communications (RLSC) platform.
https://www.stm-publishing.com/ccc-exceeds-1000-institutions-and-funders-on-its-rightslink-for-scientific-communications-platform/
Source: STM Publishing News

Back to the Future
Phil Gooch looks back at the history of semantic enrichment, and how it will be used going forward.
https://www.scientific-computing.com/analysis-opinion/back-future
Source: Scientific Computing World

Springer Nature expands its eBook portfolio in Artificial Intelligence, Electrical Engineering and Computer Science.
Springer Nature has acquired *the Synthesis Digital Library of Engineering and Computer Scienc*e from Morgan & Claypool Publishers, a pioneer in online publishing of concise books on the newest areas of engineering and computer science.
https://www.stm-publishing.com/springer-nature-expands-its-ebook-portfolio-in-artificial-intelligence-electrical-engineering-and-computer-science-with-the-purchase-of-the-synthesis-digital-library/
Source: STM Publishing News

SAGE Publishing founder Sara Miller McCune passes control of SAGE to the SAGE-SMM Trust
SAGE Publishing founder and owner Sara Miller McCune has signed over her voting shares and control of the company to the independent SAGE-SMM Trust. The move takes an irrevocable step towards her long-standing estate plan goal of ensuring SAGE remains an independent company focused on its mission to build bridges to knowledge through educational and research publishing.
https://www.stm-publishing.com/sage-publishing-founder-sara-miller-mccune-passes-control-of-sage-to-the-sage-smm-trust/
Source: STM Publishing News

Synthace unveils its Life Sciences R&D Cloud

Synthace has developed a 'no-code' cloud platform for life sciences R&D which addresses complexity, speed and reproducibility for scientists while lowering the barrier to automated biological experimentation.

https://www.scientific-computing.com/news/synthace-unveils-its-life-sciences-rd-cloud

Source: Scientific Computing World

AJE Launches New AI-Based English-language Translation Service

AJE (American Journal Experts) announces the launch of its Standard Translation service, which combines AJE's highly advanced artificial intelligence (AI) software and its own US-trained editors to return high-quality technical English-language translations in no more than five business days.

https://www.stm-publishing.com/aje-launches-new-ai-based-english-language-translation-service/

Source: STM Publishing News

ACS to pilot transparent peer review model in ACS Central Science and The Journal of Physical Chemistry Letters

As a part of its commitment to open science, the Publications Division of ACS is piloting a new peer review process, called transparent peer review, in two of its journals, *ACS Central Science* and *The Journal of Physical Chemistry Letters*. For authors and reviewers who choose to participate, transparent peer review makes the reviewers' comments and the authors' response to the reviewers visible to readers of a published article. Traditionally, ACS journals have used a model of peer review in which the reviewers' comments and the authors' response were visible only to those who participated in the review of the manuscript. Under this new format, readers will benefit by seeing how peer review shaped and strengthened the final article. For a manuscript to undergo the new process, all authors and reviewers must opt into transparent peer review. If all parties agree, their reports and responses will be published alongside an article as Supporting Information.

https://www.knowledgespeak.com/news/acs-to-pilot-transparent-peer-review-model-in-acs-central-science-and-the-journal-of-physical-chemistry-letters/

Source: Knowledgespeak

Newly launched searchRxiv builds search community to foster easier, quicker research

CABI launched searchRxiv (pronounced 'search archive'), its new open access platform. The website is designed to let researchers report, store and share their searches, thus helping with the review and re-use of existing searches to make research quicker and easier.

https://www.infotoday.eu/Articles/News/Featured-News/Newly-launched-searchRxiv-builds-search-community-to-foster-easier-quicker-research-150706.aspx

Source: Information Today

Taylor & Francis launches its first Open Research Publishing Platform with F1000

Taylor & Francis will launch its first Open Research Publishing platform for the Materials Science community, utilising the publishing model, technology and knowledge pioneered by their open research publishing partner F1000, which they acquired in 2020.

https://www.stm-publishing.com/taylor-francis-launches-its-first-open-research-publishing-platform-with-f1000/

Source: STM Publishing News

A Digital Journey

Sophia Ktori takes a look at the role software companies play in driving digital transformation in the laboratory.

https://www.scientific-computing.com/feature/digital-journey

Source: Scientific Computing World

Karger Publishers Selects Silverchair to Connect and Advance Health Sciences

Karger Publishers is a globally active independent publisher dedicated to serving the information needs of the scientific community, clinicians, and patients. The Karger publishing program encompasses more than 100 peer-reviewed journals (including a number of Gold and Platinum Open Access journals) and over 9,000 books, as well as video and interactive content for visualisation and education.

https://www.stm-publishing.com/karger-publishers-selects-silverchair-to-connect-and-advance-health-sciences/

Source: STM Publishing News

Octopus platform 'will change research culture'

Funding has been agreed to help develop a ground-breaking global service which aims to 'positively disrupt research culture for the better'. Octopus Publishing Community Interest Company (CIC), in collaboration with Jisc, will receive £650,000 over three years from Research England's emerging priorities fund. The money will support development of a new platform for the scientific community. Called Octopus, it will provide a new 'primary research record' for recording and appraising research 'as it happens'.

https://www.researchinformation.info/news/octopus-platform-will-change-research-culture

Source: Research Information

New Research4Life Content Portal now Live

Research4Life has launched a new user portal for access to our content with improved features and functionality. The new Research4Life user interface has a modern look and feel and intuitive navigation.

https://www.stm-publishing.com/new-research4life-content-portal-now-live/

Source: STM Publishing News

The Future of Librarianship

Digital-first strategies are driving a change in how librarians can support research.

https://www.natureindex.com/news-blog/the-future-of-librarianship

Source: Nature

Giving Drug Researchers Control of their Data

A transition empowering scientists is underway in data management for drug discovery.

https://cen.acs.org/business/informatics/Giving-drug-researchers-control-data/99/i41

Source: Chemical & Engineering News

CliniSys to provide a Single Laboratory Information System for Northern Ireland

CliniSys has won a major contract to deploy a single laboratory information system across Northern Ireland as part of a modernisation programme to create a world-class pathology service for the country.

https://www.scientific-computing.com/news/clinisys-provide-single-laboratory-information-system-northern-ireland

Source: Scientific Computing World

New Computational approach predicts Chemical Reactions at high temperatures

Method combines quantum mechanics with machine learning to accurately predict oxide reactions at high temperatures when no experimental data is available; could be used to design clean carbon-neutral processes for steel production and metal recycling.

https://www.sciencedaily.com/releases/2021/12/211201085150.htm

Source: Science Daily

BenevolentAI achieves second major collaboration milestone with novel Idiopathic Pulmonary Fibrosis target selected for AstraZeneca's portfolio

Novel target for idiopathic pulmonary fibrosis was discovered using BenevolentAI's AI-drug discovery platform and experimentally validated by AstraZeneca.

https://www.prnewswire.co.uk/news-releases/benevolentai-achieves-second-major-collaboration-milestone-with-novel-idiopathic-pulmonary-fibrosis-target-selected-for-astrazeneca-s-portfolio-824300885.html

Source: Cision PR Newswire

2021 STM Report highlights rapid transformation to Open Access

STM (the Association of Scientific, Technical and Medical Publishers) has published the latest edition of '*The STM Report*', the organisation's comprehensive overview of the scientific and scholarly publishing market. The revised report, which adopts a new supplement format to be issued in regular thematic updates, reveals significant publisher-driven growth in Open Access (OA) and continued dynamism in the scholarly communication ecosystem.

https://www.stm-publishing.com/2021-stm-report-highlights-rapid-transformation-to-open-access/

Source: STM Publishing News

Optibrium adds a 3D Ligand-based Design Module to its Chemistry Software

The latest version of Optibrium's StarDrop software helps scientists target high-quality compounds faster due to a collaboration with BioPharmics and the integration of its eSim and ForceGen software.

https://www.scientific-computing.com/news/optibrium-adds-3d-ligand-based-design-module-its-chemistry-software

Source: Scientific Computing World

PLOS Announces a New Policy on Inclusion in Global Research

The Public Library of Science (PLOS) announced its aim to bring equity to publishing by launching a new policy to improve transparency in the reporting of research that is conducted in other countries or communities.

https://www.stm-publishing.com/plos-announces-a-new-policy-on-inclusion-in-global-research/

Source: STM Publishing News

Artificial Intelligence Being Used to Accurately Predict Synergistic Cancer Drug Combinations

https://www.prnewswire.co.uk/news-releases/artificial-intelligence-being-used-to-accurately-predict-synergistic-cancer-drug-combinations-860981323.html

Source: Cision PR Newswire

ALPSP Copyright Committee responds to UKRI Open Access Policy

As an international trade association, ALPSP supports and represents not-for-profit organisations that publish scholarly and professional content, as well as those that work with them. ALPSP members are very supportive of open access and have already taken and continue to take significant steps to ensure that as much content as possible is published on an open access basis.

https://www.stm-publishing.com/alpsp-copyright-committee-responds-to-ukri-open-access-policy/

Source: STM Publishing News

ACS celebrates its 2021 Heroes of Chemistry

Chemists are honoured for their contributions to treatments for bloodstream infections, cancer, diabetes, and heart failure.

https://cen.acs.org/people/awards/ACS-celebrates-its-2021-Heroes-of-Chemistry/99/i44

Source: Chemical & Engineering News

How the National Library of Scotland continued to provide a quality service during lockdown
Library professionals at the National Library of Scotland decided to see the pandemic as an opportunity. Revised work processes and innovative service provision to the public and to staff when library buildings were closed and physical access denied were the result. Craig Statham, Maps Reading Room Manager, explains.
https://www.infotoday.eu/Articles/Editorial/Featured-Articles/How-the-National-Library-of-Scotland-continued-to-provide-a-quality-service-during-lockdown-148762.aspx
Source: Information Today

PerkinElmer integrates Scientific Data Silos and enhances Collaboration in the Cloud
PerkinElmer has announced the launch of its Signals Research Suite, a fully cloud-based solution, deployed on Amazon Web Services. The suite is a secure, informatics platform, providing integrated, end-to-end scientific data and workflow management for pharmaceutical and industrial customers. Designed to help drive more informed and accelerated decision making around drug, compound and formulation candidates, the offering brings together PerkinElmer's leading informatics technologies across data access, processing, enhanced analytics and collaboration.
https://www.scientific-computing.com/news/perkinelmer-integrates-scientific-data-silos-and-enhances-collaboration-cloud
Source: Scientific Computing World

Tiny 'Maniac' Robots could Deliver Drugs Directly to Central Nervous System
A new study investigates tiny tumbling soft robots that can be controlled using rotating magnetic fields. The technology could be useful for delivering drugs to the nervous system. In this latest study, researchers put the robots through their paces and showed that they can climb slopes, tumble upstream against fluid flow and deliver substances at precise locations to neural tissue.
https://blog.frontiersin.org/2021/08/11/frontiers-robotics-ai-tiny-maniac-robots-targeted-drug-delivery/
Source: Frontiers Science News

Clarivate Successfully Completes Acquisition of ProQuest
Clarivate plc has completed its acquisition of ProQuest – one of the leading global software, content, data and analytics providers.
https://www.stm-publishing.com/clarivate-successfully-completes-acquisition-of-proquest/
Source: STM Publishing News

AI can quickly identify Structure of Drugs Designed for 'Legal Highs'
An artificial intelligence can identify designer drugs that have similar effects to substances such as cocaine and heroin, but which can't be detected by current tests
https://www.newscientist.com/article/2297611-ai-can-quickly-identify-structure-of-drugs-designed-for-legal-highs/#ixzz7H7GkvgNV
Source: NewScientist

Artificial Intelligence Applications for Libraries
Artificial Intelligence (AI) pervades our everyday lives, whether we realise it or not. AI technologies power our digital personal assistants, show us traffic flows on our roads, monitor our health and diagnose diseases, guide our product selections when we shop online, and give financial advice. Libraries also benefit from AI technologies and AI applications are increasing.
https://www.infotoday.eu/Articles/Editorial/Featured-Articles/Artificial-Intelligence-applications-for-libraries-150740.aspx
Source: Information Today

Wision A.I. Achieves CE-MDR Mark Approval for AI-assisted Diagnostic Software Medical Device Supporting Colonoscopy

Wision A.I. Ltd, a startup in the field of artificial intelligence assisted diagnostics for gastrointestinal endoscopy, announced it received the European CE Mark approval for EndoScreener, its AI-assisted polyp detection software during colonoscopy. It is the first CE Mark class II certificate under the new Medical Devices Regulation.

https://www.prnewswire.co.uk/news-releases/wision-a-i-achieves-ce-mdr-mark-approval-for-ai-assisted-diagnostic-software-medical-device-supporting-colonoscopy-863753851.html

Source: Cision PR Newswire

Creating Effective Videos for Libraries

One of Internet Librarian International's most popular speakers, Ned Potter, looks at the different types of videos that librarians can create and shares his tips for making library videos. It's not as hard as you might think, as his examples illustrate.

https://www.infotoday.eu/Articles/Editorial/Featured-Articles/Creating-effective-videos-for-libraries-147787.aspx

Source: Information Today

Auto articles: an Experiment in AI-generated Content

AI-generated summaries of three articles selected from a data set of 175 Springer Nature publications.

https://www.natureindex.com/news-blog/auto-articles-an-experiment-in-ai-generated-content

Source: Nature

Google Tries for Transparency

Two Google initiatives over the past few months resonate particularly well with information professionals. Google Scholar added a "Public access" section to track and manage public access mandates and Google Search can now detect when a topic is rapidly evolving and warn people to check back late.

https://www.infotoday.eu/Articles/News/Featured-News/Google-Tries-for-Transparency-147786.aspx

Source: Information Today

Clarivate Identifies the One in 1,000 Citation Elite with Annual Highly Cited Researchers List

Clarivate has unveiled its 2021 list of Highly Cited Researchers. The methodology that determines the "who's who" of influential researchers draws on the data and analysis performed by bibliometric experts and data scientists at the Institute for Scientific Information™ at Clarivate.

https://www.stm-publishing.com/clarivate-identifies-the-one-in-1000-citation-elite-with-annual-highly-cited-researchers-list-2/

Source: STM Publishing News

Insilico Medicine Announces the Nomination of Two Preclinical Candidates for PHD2, 12 Months After Program Initiation

Insilico Medicine, a global clinical-stage biotechnology company specialising in the applications of end-to-end AI for drug discovery and development, has announced that the company has nominated preclinical candidates (PCC) for ISM012-077 and ISM012-042 for the treatment of anemia of chronic kidney disease (CKD) and inflammatory bowel disease (IBD), respectively.

https://www.prnewswire.co.uk/news-releases/insilico-medicine-announces-the-nomination-of-two-preclinical-candidates-for-phd2-12-months-after-program-initiation-825254978.html

Source: Cision PR Newswire

Libraries 'can support Researchers more Effectively'

Ex Libris has announced the publication of its annual study on the challenges that academic researchers face, the priorities of research office leaders, and key opportunities for libraries and research offices to advance scholarship at their institution. Commissioned by Ex Libris, the study was conducted by Alterline, an independent research agency. The report presents findings from a survey of more than 400 researchers and research office leaders across a range of disciplines in the USA, the UK and Australia.

https://www.researchinformation.info/news/libraries-can-support-researchers-more-effectively

Source: Research Information

DOGS and CATS help design new Natural Product-Based Drugs

A new algorithm helps researchers search out new molecules for applications in medicine, keeping their synthesis quick and cost-effective.

https://www.advancedsciencenews.com/dogs-and-cats-help-design-new-natural-product-based-drugs/

Source: Advanced Science News

Clarivate and KAIST Innovation Strategy and Policy Institute release report on the Global AI Innovation Landscape

Clarivate Plc and the KAIST Innovation Strategy and Policy Institute (ISPI) has launched a report in Korea on the global innovation landscape of artificial intelligence. The report shows that AI has become a key technology and that cross-industry learning is an important AI innovation. It also stresses that the quality of innovation, not volume, is a critical success factor in technological competitiveness.

https://www.knowledgespeak.com/news/clarivate-and-kaist-innovation-strategy-and-policy-institute-release-report-on-the-global-ai-innovation-landscape/

Source: Knowledgespeak

Chemistry Breakthrough leads way to more Sustainable Pharmaceuticals

Chemistry researchers have developed a new method using blue light to create pharmaceuticals in a more sustainable way, significantly reducing the amount of energy needed and the chemical waste created in the manufacture process.

https://www.sciencedaily.com/releases/2021/11/211117161359.htm

Source: Science Daily

Research Literacy for International Internet Librarians

Information professionals have long been charged with educating people about information literacy, information management, copyright and licensing. Although librarians understand today's information landscape, non-information savvy people frequently don't. Even the word "research" has taken on different connotations. Mary Ellen Bates, internationally renowned independent information professional, consultant, speaker, and Online Searcher columnist, considers how we can better explain these concerns.

https://www.infotoday.eu/Articles/Editorial/Featured-Articles/Research-Literacy-for-International-Internet-Librarians-148167.aspx

Source: Information Today

The Lab of the Future is Now

Recent demonstrations of AI-directed automation may herald a new world for drug and materials discovery.

https://cen.acs.org/business/informatics/lab-future-ai-automated-synthesis/99/i11

Source: Chemical & Engineering News

The Electrochemical Society launches two new gold open access journals with IOP Publishing

The Electrochemical Society (ECS), together with IOP Publishing, is launching two new, fully open access (OA) journals. *ECS Advances* and *ECS Sensors Plus* add to the Society's journal family and provide the research community with a diverse suite of interconnected journals sharing impactful research across the world. *ECS Advances* delivers a platform for research across all areas of electrochemical and solid-state science and technology research with the broadest dissemination of all journals in the field. *ECS Sensors Plus* offers a specialised outlet for all content related to sensors technology. The journal will lead and promote scholarly communication and interactions among scientists, engineers, and technologists whose primary interests focus on materials, structures, properties, performance, and characterisation of sensing and detection devices and systems, including sensor arrays and networks.

https://www.knowledgespeak.com/news/theelectrochemicalsocietylaunches-two-newgoldopen-access-journalswith-iop-publishing/

Source: Knowledgespeak

A new Tool to Assess Researchers for Promotion and Recruitment

Model measures qualities that evaluation metrics usually miss, developers say.

https://www.natureindex.com/news-blog/new-tool-assess-researchers-promotion-recruitment

Source: Nature

OpenAIRE sets up a customised portal that allows users to search, browse and access Canadian research outcomes

Canadian Association of Research Libraries (CARL) entered into a collaboration with OpenAIRE. The aim of the collaboration is to use the OpenAIRE services to identify Canadian research outputs. The work will provide the community with a better understanding of what Canadian-funded publications are openly available. This collaboration involves working with local repositories and journals, as well as the Canadian research funders, to ensure that the necessary information about authors, funders and institutions is included in the metadata of related publications.

https://www.knowledgespeak.com/news/openaire-sets-up-customized-portal-allowing-users-to-search-browse-and-access-canadian-research-outcomes/

Source: Knowledgespeak

Pistoia Alliance SEED project unlocks the value of data in Electronic Lab Notebooks

The Pistoia Alliance has announced the second phase of its Semantic Enrichment of ELN Data (SEED) project.

https://www.scientific-computing.com/news/pistoia-alliance-seed-project-unlocks-value-data-electronic-lab-notebooks

Source: Scientific Computing World

The State of Open Data 2021 - Survey Results

Figshare, Digital Science, and Springer Nature's annual State of Open Data report finds increasing concern among researchers about misuse of data as well as a lack of credit and acknowledgement for those who do openly share their data. Among the key findings, 55 per cent feel they need support in regard to copyright and licenses when making research data openly available, and 73 per cent strongly or somewhat support the idea of a national mandate for making research data openly available.

https://www.researchinformation.info/news/state-open-data-2021-survey-results

Source: Research Information

IBM Research Europe and Thieme Chemistry Collaboration Accelerates Discovery in Organic Chemistry
https://www.scientific-computing.com/news/ibm-research-europe-and-thieme-chemistry-collaboration-accelerates-discovery-organic-chemistry
Source: Scientific Computing World

Buttoned up Biomolecules: A Click Reaction for Living Systems
Bioorthogonal hydroamination of activated linear alkynes now suitable in living cells.
https://www.advancedsciencenews.com/buttoned-up-biomolecules-a-click-reaction-for-living-systems/
Source: Advanced Science News

Springer Nature Publishes one Million Open Access articles
Springer Nature has become the first publisher to immediately publish one million gold open access (OA) primary research and review articles - testament to the company's long commitment to making research immediately available for all to read, share, use, and reuse to advance discovery. This means that 25% of all articles Springer Nature has published since 2005 are gold OA. In 2020 alone, such open access articles accounted for 34% of all articles published by Springer Nature.
https://www.knowledgespeak.com/news/springer-nature-publishes-one-million-open-access-articles/
Source: Knowledgespeak

ACS Publications appoints Ombudsperson to support Authors in the Peer Review Process
The Publications Division of ACS has announced the appointment of Dr. Kathleen H. Canul as the ACS Publications' ombudsperson. Her appointment fulfils a crucial element of ACS' commitment to confronting racism and discrimination within its journals and strengthens efforts to increase diversity throughout its author and reviewer communities. As an ombudsperson, Canul will serve as an impartial, independent and confidential channel for concerns regarding the peer review process. In the event that an author or reviewer has concerns over issues such as editor bias, editorial advisory board member or reviewer misconduct, or editorial professionalism, Canul would act as a resource, providing guidance, and as an outlet for the expression of concerns. The ombudsperson will not, however, be conducting formal investigations into complaints. In the event that a complaint requires an investigation, Canul will be available to provide guidance to the relevant journal or institution.
https://www.knowledgespeak.com/news/acs-publications-appoints-ombudsperson-to-support-authors-in-the-peer-review-process/
Source: Knowledgespeak

Exscientia enhances AI with acquisition of Allcyte
https://cen.acs.org/business/informatics/Exscientia-enhances-AI-acquisition-Allcyte/99/i23
Source: Chemical & Engineering News

IOPP announces deal with MyScienceWork
IOP Publishing has announced the creation of one of the largest collections of academic journals, books and conference series in physical sciences on the MyScienceWork (MSW) platform. The research management tech provider will index IOPP's content, making it available to the academic community.
https://www.researchinformation.info/news/iopp-announces-deal-mysciencework
Source: Research Information

Entos raises $54 million for AI-based drug discovery
https://cen.acs.org/business/informatics/Entos-raises-54-million-AI/99/i26
Source: Chemical & Engineering News

Panasas storage to support Australian cryo-EM research
Panasas and the University of Wollongong (UOW) in Australia have announced a five-year strategic alliance to support medical and scientific research initiatives that deploy cryogenic electron microscopy.
https://www.scientific-computing.com/news/panasas-storage-support-australian-cryo-em-research
Source: Scientific Computing World

How to write an Abstract that Stands Out
A well-written abstract helps to attract readership.
https://www.natureindex.com/news-blog/how-to-write-good-abstract-scientific-research-paper
Source: Nature

New Springer Nature White Paper reveals Gold Open Access is best for Authors and Researchers
Springer Nature has published a new white paper that builds on the growing body of evidence that shows that Gold open access (OA) is best for authors and researchers. Springer Nature's 2018 white paper, Assessing the open access effect for hybrid journals, highlighted 'the OA effect' and showed that OA articles in hybrid journals achieve greater impact, usage and reach than comparative non-OA articles. *Going for gold: exploring the reach and impact of Gold open access articles in hybrid journals* offers important additional analysis. In looking at the specific non-OA subset of subscription articles where an earlier version (such as a Green OA accepted manuscript) exists in an OA repository, it shows that there is no significant corresponding 'Green OA effect'; the availability of a 'Green' version is not sufficient to match the benefits of Gold OA given that the Version of Record (VOR) of the article it is attached to remains behind a paywall.
https://www.knowledgespeak.com/news/new-springer-nature-white-paper-reveals-gold-open-access-is-best-for-authors-and-researchers/
Source: Knowledgespeak

BIO-ISAC Founded to protect Bioeconomy Infrastructure
Dotmatics has announced that Charles Fracchia, vice president of data and founder of BioBright, a Dotmatics company, has been appointed to the Bioeconomy Information Sharing and Analysis Center (BIO-ISAC) board of directors and the executive board.
https://www.scientific-computing.com/news/bio-isac-founded-protect-bioeconomy-infrastructure
Source: Scientific Computing World

The must-have Multimillion-dollar Microscopy Machine
Cryo-EM facilities are exponentially improving protein resolution for structural biologists. Access to this vastly expensive equipment is subject to the availability of highly skilled operators.
https://www.natureindex.com/news-blog/must-have-multimillion-dollar-microscopy-machine-cryo-em
Source: Nature

Collaboration aims to Accelerate Chromatography Method Development
Thermo Fisher Scientific and ChromSword, a provider of innovative software products, have collaborated to launch an automated high performance liquid chromatography (HPLC) and ultra-high performance liquid chromatography (UHPLC) method development system.
https://www.scientific-computing.com/news/collaboration-aims-accelerate-chromatography-method-development
Source: Scientific Computing World

AACR issues call for papers for new OA journal, Cancer Research Communications

The American Association for Cancer Research has announced the opening of the submission site for its new open access journal, *Cancer Research Communications*, signalling the official call for papers. This new journal is the 10th in the AACR's portfolio of scientific publications, adding an open access publishing option to this high-quality, trusted journal collection. *Cancer Research Communications* welcomes research spanning the full breadth of cancer science and medicine. In addition to expanding the AACR's portfolio of peer-reviewed publications, the new journal will further stimulate the exchange of innovative ideas and approaches in cancer research and will provide a rapid publication outlet that serves the cancer field.

https://www.knowledgespeak.com/news/aacr-issues-call-for-papers-for-new-oa-journal-cancer-research-communications/

Source: Knowledgespeak

Pandemic Has Cultivated New Segment of Online Learners

The 'Voice of the Online Learner' Report Finds Pool of Online Learners Expanded During the Pandemic, and New Cohort Emerged that Skews Younger, More Likely to Pursue Online Undergraduate Degrees.

https://newsroom.wiley.com/press-releases/press-release-details/2021/Pandemic-Has-Cultivated-New-Segment-of-Online-Learners-According-to-New-Wiley-Report/default.aspx

Source: Wiley

Recent ownership changes of three influential News Publications pose interesting questions for Librarians and Researchers

Axel Springer signed an agreement to acquire POLITICO, Nexstar acquired digital media platform The Hill, and Forbes merged with the special interest acquisition company Magnum Opus Acquisition Ltd.

https://www.infotoday.eu/Articles/News/Featured-News/Recent-ownership-changes-of-three-influential-news-publications-pose-interesting-questions-for-librarians-and-researchers-148794.aspx

Source: Information Today

A better-fitting Molecular 'Belt' for making New Drugs

Chemistry invention could affect the most common medications.

https://www.sciencedaily.com/releases/2021/11/211116131733.htm

Source: Science Daily

Loughborough's new £1m High-Performance Computer will Transform Research

https://www.scientific-computing.com/news/loughborough-s-new-1m-high-performance-computer-will-transform-research

Source: Scientific Computing World

Research Square reaches 100,000 Preprint Milestone

Fewer than three years after the first preprint was posted on Research Square, the multidisciplinary preprint platform has surpassed 100,000 preprints.

https://www.knowledgespeak.com/news/research-square-reaches-100000-preprint-milestone/

Source: Knowledgespeak

Dialog Solutions Adds ClinicalTrials database

Dialog Solutions announced it was adding the ClinicalTrials.gov database to its platform. Since ClinicalTrials.gov is a free database on the web, what is the advantage to searching it on Dialog?

https://www.infotoday.eu/Articles/News/Featured-News/Dialog-Solutions-Adds-ClinicalTrials-database-149298.aspx

Source: Information Today

Clarivate acquires Bioinfogate

The Bioinfogate OFF-X™ portal is a cutting-edge safety intelligence solution aimed at empowering pharmaceutical organisations to identify toxicology and safety signals, mitigate safety liabilities and de-risk early-stage assets. It is one of the largest translational safety and toxicity portals, featuring over 1,200,000 safety alerts corresponding to over 23,000 drugs and biologics and more than 15,000 targets of pharmacological interest. As a leading provider of trusted information and insights to accelerate innovation, Clarivate offerings include a comprehensive suite of research intelligence solutions coupled with deep domain expertise. The acquisition of Bioinfogate will fill a critical need for drug toxicity data and translational safety intelligence across all stages of drug R&D. This follows a previous acquisition from the Prous family of companies.

https://www.knowledgespeak.com/news/clarivateacquiresbioinfogate/

Source: Knowledgespeak

Qlucore and Clarivate Partnership aims to Expedite Omics Data Analysis

Qlucore and Clarivate have announced a partnership that aims to provide researchers with powerful and flexible omics data analysis and visualisation.

https://www.scientific-computing.com/news/qlucore-and-clarivate-partnership-aims-expedite-omics-data-analysis

Source: Scientific Computing World

PLOS Expands Footprint in the European Union with a Publishing Agreement in Germany

The Public Library of Science (PLOS) has announced an agreement with Sachsen Consortia to facilitate unlimited publishing across all 12 PLOS titles with no fees for researchers. This agreement encompasses PLOS' three innovative publishing models, ensuring researchers from 9 Saxon institutions benefit from frictionless, fee-free publishing with PLOS. This agreement represents PLOS' second major consortia deal in the European Union.

https://www.stm-publishing.com/plos-expands-footprint-in-the-european-union-with-a-publishing-agreement-in-germany/

Source: STM Publishing News

Cabells' Predatory Reports Passes 15,000 Predatory Journals Listed

Despite the increasing awareness of the perils of predatory journals and librarians' warnings about their dangers, the number of journals is rising not falling as the academic community had hoped.

https://www.infotoday.eu/Articles/News/Featured-News/Cabells-Predatory-Reports-Passes-15000-Predatory-Journals-Listed-149299.aspx

Source: Information Today