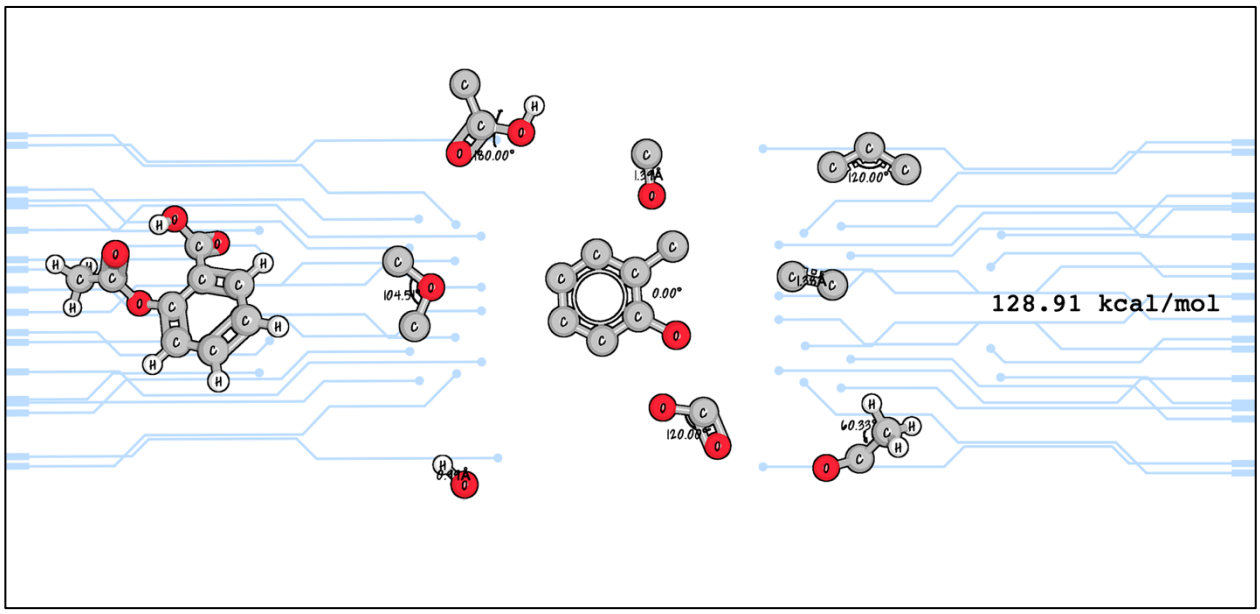
RSC INTEREST GROUP
CHEMICAL INFORMATION AND
COMPUTER APPLICATIONS GROUP

NEWSLETTER winter 2020-21



Above: A representation of a parallel process for calculating the energy of a molecule (see page 17)

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.

Chemical Information & Computer Applications Group Websites:

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

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

LinkedIn http://www.linkedin.com/groups?gid=1989945

https://Twitter.com/RSC_CICAG

QR Code

Table of Contents

Contributions to the CICAG Newsletter are welcome from all sources - please send to the Newsletter Editor:
Stuart Newbold, FRSC, email: stuart@psandim.com

Chemical Information & Computer Applications Group Chair's Report

Contributed by RSC CICAG Chair Dr Chris Swain, email: swain@mac.com

The current COVID pandemic has cast its shadow over the world in 2020 and are thoughts are with all those who have been affected. On a more practical note, the RSC has a community fund that can be used for support in these difficult times.

Social media is becoming an increasingly active way for us to communicate with members during lockdown, with our Twitter account now having nearly 1200 followers, and LinkedIn reaching 426 followers. We have recently added a YouTube channel, where you can view the lightening posters from the AI in Chemistry meeting and the workshops from the Open Chemical Sciences meeting. In addition, the CICAG website is often updated and we would be very interested to hear suggestions for additional content for all channels.

The Artificial intelligence in Chemistry meeting (#AIChem20) in September was converted from a physical meeting to a virtual event. As ever this proved to be a very popular meeting, with 500 registrations and an extensive waiting list. One advantage of the virtual events is the geographical reach is extended, with participants from over 40 countries attending. The meeting also involved a lightening poster session that proved very popular, these are hosted on our YouTube channel as Day1 and Day 2 videos, which have been viewed over 1,000 times to date.

The Open Chemical Science meeting (#OpenChem20) in November was a 5-day virtual event with three intertwined themes, Publishing, Data and Workshops. Once again moving to a virtual event enabled much greater geographical participation with delegates from 45 different countries joining. The highlight was perhaps the workshops, where we were fortunate to have outstanding contributions demonstrating six critical open-source software packages, PyMOL, KNIME, Fragalysis, Google CoLab, ChEMBL and DataWarrior. These workshops were all recorded and are all on the YouTube channel and have already been viewed over 1000 times in total.

These workshops highlighted a clear unmet need and CICAG will be looking to host more in the future. If you have any suggestions, we would be delighted to hear from you.

This newsletter also includes contributions from David Ball, Neil Hammond, Garrett Morris, Carlos Outeiral, Nathan Price, Kevin Theisen and Wendy Warr, and 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, in practice many members do not take up this opportunity. You can make a request to join a group via email (membership@rsc.org) or telephone (01223 432141).

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
Big Data	TBD	Virtual/TBD	Proposed joint meeting with the SCI)
In-silico Protein Structure Prediction	Q2	Virtual event	Details to follow

Open-source Software Workshops	Q2/3 2021	Virtual event	Details to follow
4th Artificial Intelligence in Chemistry Meeting	Sep 2021	Virtual event	Joint event from RSC-CICAG and RSC-BMCS division
Chemical Information during Covid	TBD	Virtual/TBD	Details to follow
Python for Chemists	TBD	Virtual/TBD	Details to follow
Storing Chemical Data, why use a UDM?	TBD	Virtual/TBD	Details to follow

Memories of Dr Angus McDougall, 1934-2020

Contribution from CICAG Committee member Dr Helen Cooke, email: helen.cooke100@gmail.com

It was with great sadness that I and other members of the CICAG Committee learnt of the death on 14 March 2020 of Dr Angus McDougall. I got to know Angus, when he was a Senior Lecturer in the Chemistry Department at UMIST where I also worked. He taught physical chemistry and his research focus was electrochemistry, with fuel cells being a speciality. Passionate about education, he was appointed Dean of Undergraduate Studies and was awarded a UMIST prize for excellence in undergraduate teaching.

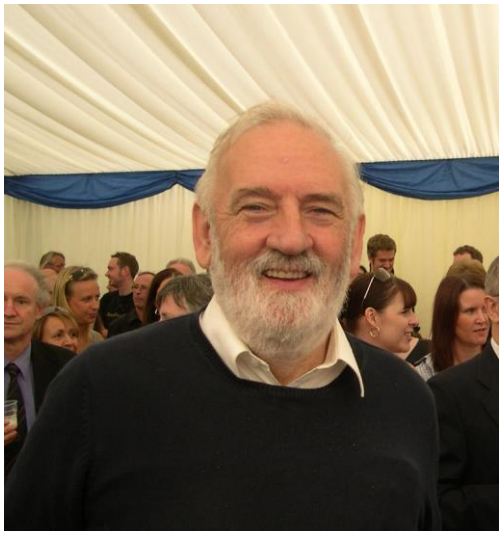


Photo: Manchester Literary and Philosophical Society https://www.manlitphil.ac.uk/news/dr-angus-mcdougall

Angus also had a strong interest in libraries and was a member of the Library Committee. He and I shared an interest in chemical information and it was in this capacity that we worked together, developing course work in this field, including practical classes dedicated to chemical information and in particular structure searching. As a consequence, UMIST was one of the first universities in the UK to introduce these skills to chemistry undergraduates.

Upon Angus's suggestion, in 1998, we published a paper together, entitled "*Introduction to the use of the chemical literature: an innovative library workbook*" in the RSC journal University Chemistry Education, Vol.2(1), 5-9. The workbooks were a library teaching resource integrated into chemistry degree programmes at UMIST and were tailored to assist with the location of information of direct relevance to practical, essay and project work, including "new developments in databases".

Angus was consistently supportive throughout the 20 years of my career at the university. We also shared an interest in the history of chemistry, and he was very helpful when I was studying for my PhD on the history of chemical information while working at UMIST. We kept in touch regularly after I left the university.
After retiring from UMIST, Angus joined the Manchester Literary and Philosophical Society, becoming a Member of Council in 2014. He was actively engaged almost until the time of his death. Angus is greatly missed and is remembered with fondness by colleagues and students alike.

CASP14: What Google DeepMind's AlphaFold 2 really achieved, and what it means for Protein Folding, Biology and Bioinformatics

Contribution from Carlos Outeiral, PhD candidate at the Oxford Protein Informatics Group, University of Oxford, email: carlos.outeiral@gtc.ox.ac.uk (note this contribution has been adapted from a blog post published after the CASP14 conference)

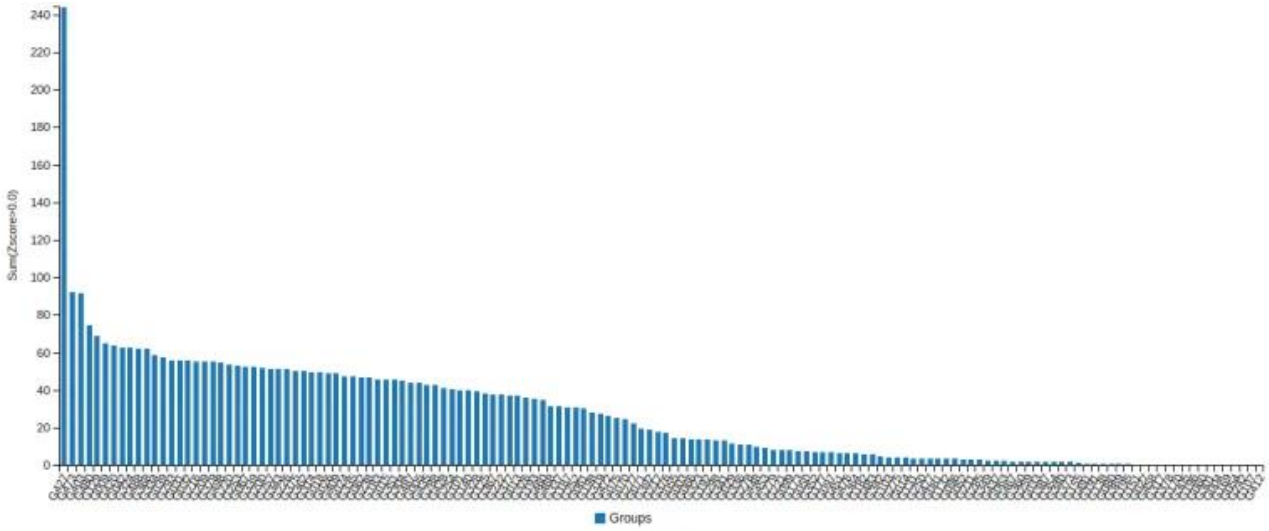
You might have heard it from the scientific or regular press, perhaps even from DeepMind's own blog. Google's AlphaFold 2 indisputably won the 14th Critical Assessment of Structural Prediction (CASP) competition, a biannual blind test where computational biologists try to predict the structure of several proteins whose structure has been determined experimentally — yet not publicly released. Their results are so incredibly accurate that many have hailed this code as the solution to the long-standing protein structure prediction problem.

How much of the press release is true, what has actually happened, and how significant is it? There has been endless discussion about this topic in multiple forums. Frankly, I wasn't able to think about anything else for much of the week following the announcement. In an attempt to clear my own thoughts, I wrote a long blog post which I have now adapted for this newsletter. I hope this is useful for my fellow protein chemists who could not attend CASP14, but also to anyone who wants to hear a little bit more about this topic.

Please bear in mind that my report of the CASP14 assessment and conference will necessarily be interspersed with conjecture. The details of how AlphaFold 2 works are still unknown, and we may not have full access to them until their paper is peer-reviewed (which may take more than a year, based on their CASP13 paper). The magnitude of the breakthrough is undeniable — but we need more details to gauge its potential impact.

How good is AlphaFold 2, exactly?
Astoundingly so.

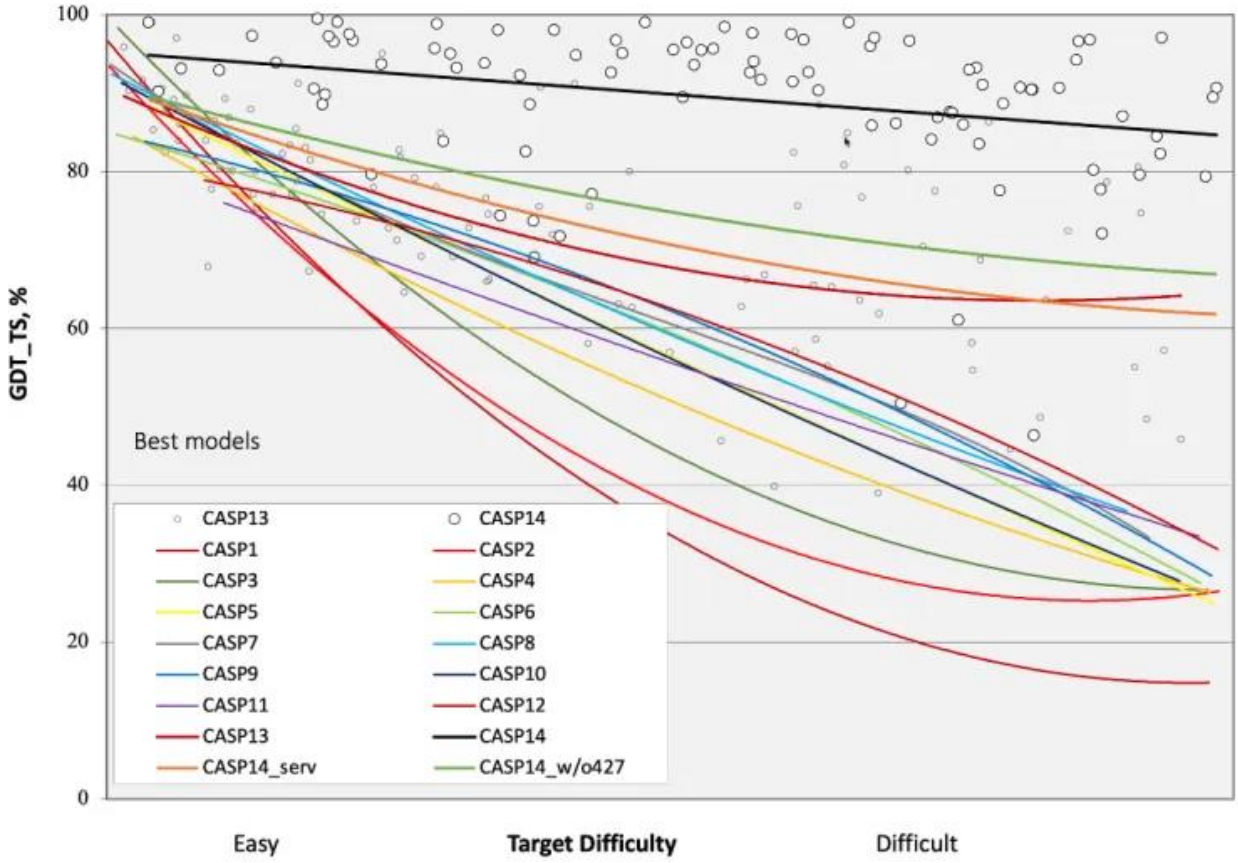
Hours before the CASP14 conference started, one image started to make the rounds around Twitter:



Ranking of participants in CASP14, as per the sum of the Z-scores of their predictions (provided that these are greater than zero). One group, 427, named AlphaFold 2, shows an incredible improvement with respect to the second-best group, 473 (BAKER). This figure was obtained from the official CASP14 webpage on Tuesday 1st December 2020.

This bar plot describes the sum of Z-scores representing the predictions from the different groups. One group performs much better than the rest: Group 427, whose average Z-score was around 2.5 when considering all targets; and rose to 3.8 in the hardest ones. If the relative comparison is astounding, the actual performance is just as impressive: about a third (36%) of Group 427's submitted targets were predicted with a root-mean-square deviation (RMSD) under 2 Å, and 86% were under 5 Å, with a total mean of 3.8 Å.

We may also consider a more widely used metric, the global distance test (GDT_TS). As a rule of thumb, a GDT_TS around 60% represents a "correct fold", meaning that we have an idea of how the protein folds globally; and over 80% we start seeing side chains that closely resemble the model. As you can see, AlphaFold 2 achieves this objective for all but a small fraction of the tasks.



Combined results of all the CASP competitions. The dark orange line (CASP14_serv) corresponds to the predictions made by fully automated servers, the olive-green line (CASP14_w/o427) includes all predictions assisted by humans except for the highest performing group; and the black line (CASP14) represents the predictions by the best performing team: Group 427, or AlphaFold 2. This plot uses the GDT_TS score, where 100 represents perfect results and 0 is a meaningless prediction.

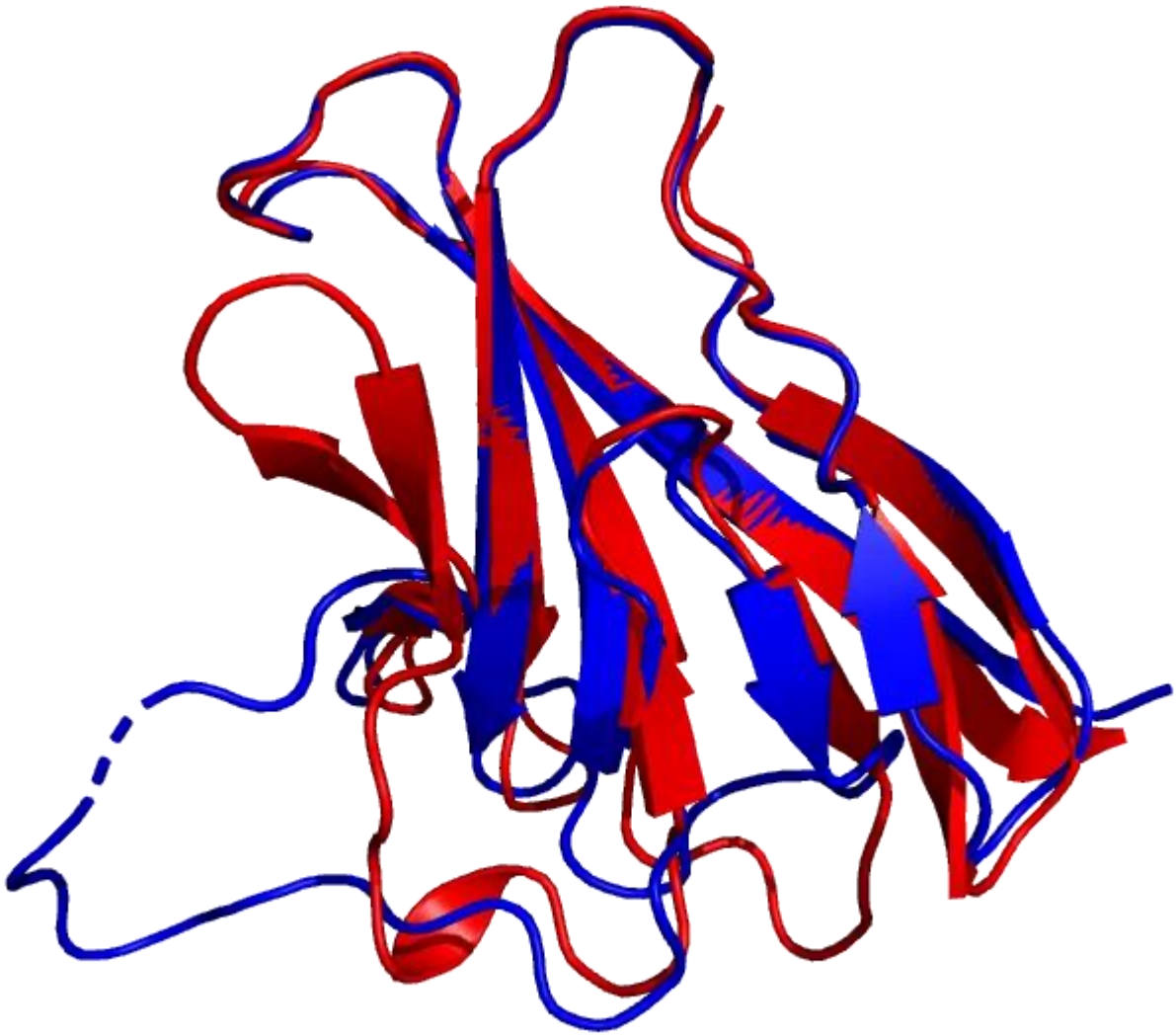
The models produced by AlphaFold 2 were so good that in some cases defied the results of the experiment. Osnat Herzberg's group, studying a phage tail protein, noticed they had assigned the wrong conformation to a cis-proline only after they appreciated DeepMind's excellent agreement to their structure. Similarly, Henning Tidow's group, studying an integral membrane protein, the reductase FoxB, was able to solve the diffraction structure via molecular replacement only after they saw AlphaFold 2's prediction — even after trying all sort of tricks, including experimental phasing, for two years.

There are many interesting points of discussion. Many will argue that the set of targets studied at CASP14 is not representative of all interesting structural prediction problems — and they will be right. And, yes, certainly there are some problems where AlphaFold 2 hasn't performed that well. But let us be clear on something: AlphaFold 2 is a tool that can solve the protein structure prediction problem for a very significant number of targets.

How does this compare to other methods?

I am going to have a closer, albeit brief look at two of the targets in the competition, comparing AlphaFold 2 with two of the best ranked groups: David Baker's and Yang Zhang's. They have both (1) consistently performed really well in past CASP competitions, and (2) given fantastic talks at CASP14, so I have a decent idea of what is happening under the hood.

I am going to look at the ORF8 protein, a viral protein involved with the interaction between SARS-CoV-2 and the immune response (PDB: 7JTL, preprint available on bioRxiv). Let's have a glance at how the structure predicted by AlphaFold (red) compares to the crystal structure (blue).

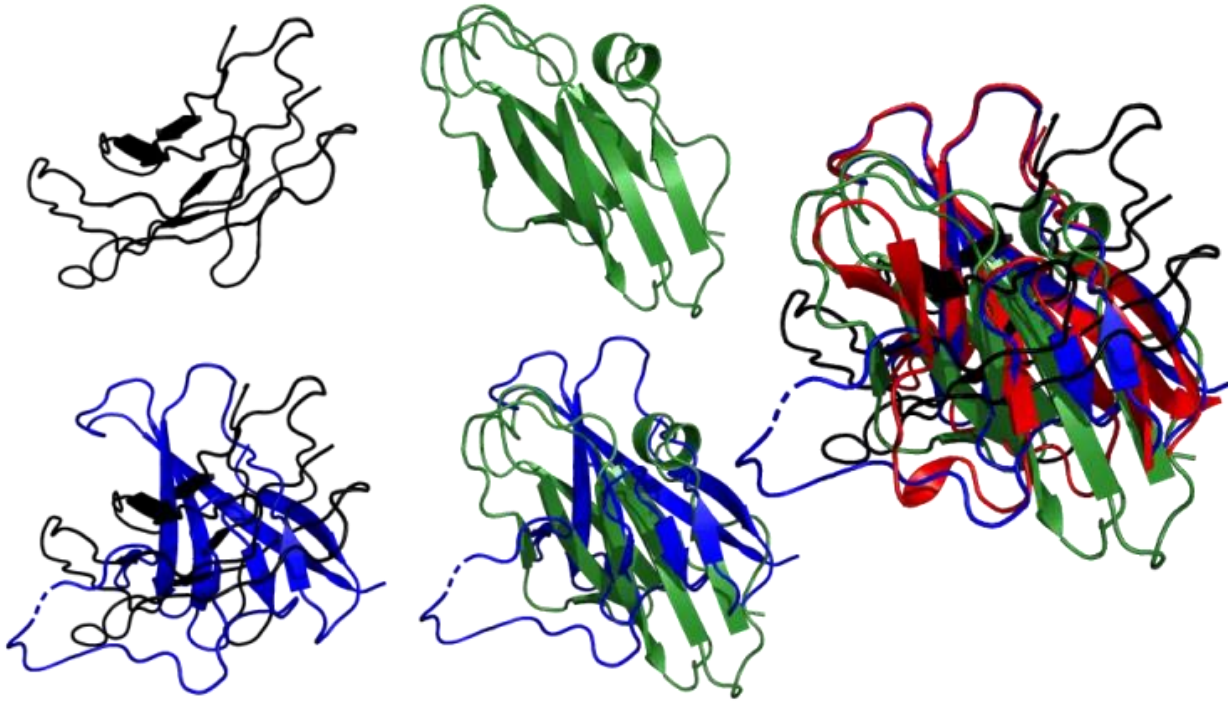


Top group 427 model for the T1064 target (red), superimposed onto the 7JTL_A structure (blue). DeepMind's structure was obtained from the CASP14 webpage on Tuesday 1st December 2020.

The prediction of the core of the protein is in excellent agreement with the experiment, closely reproducing the structure of the antiparallel β-sheets, and more impressively, the loops that connect them. Notice, however, that there is a large loop region, on the bottom left corner of the image, that is very significantly different from the crystal structure. Initially, I thought (and so did the AlphaFold 2 team, following their talk at CASP14) it was a failure of the problem, until one of the readers of my original post, Juliette Martin, from the Institut de Biologie et Chimie des Protéines, pointed out to me that a new structure for the SARS-CoV2 ORF8 protein, where the loop that was badly predicted by AlphaFold 2 actually displays a much similar structure to the prediction. AlphaFold 2 made no mistake – it just knew that the loop was highly flexible!

How did the other groups do? Both the Baker and Zhang groups used a similar pipeline, which incorporates many of the ideas of CASP13's AlphaFold: build a multiple sequence alignment, potentially incorporating metagenomics sequences; predict a potential using deep learning and find a minimum using their lab-branded

method (ROSETTA for Baker group, I-TASSER for Zhang group), and apply some refinement, potentially also using deep learning.



Top: highest-ranked models for the target T1064 submitted by the Zhang (black) and Baker (green) human groups. Bottom: models aligned with the crystal structure. Right: all three models (Zhang, Baker and AlphaFold 2) aligned with the crystal structure. The submissions were obtained from the CASP14 webpage on Tuesday 1st December, 2020.

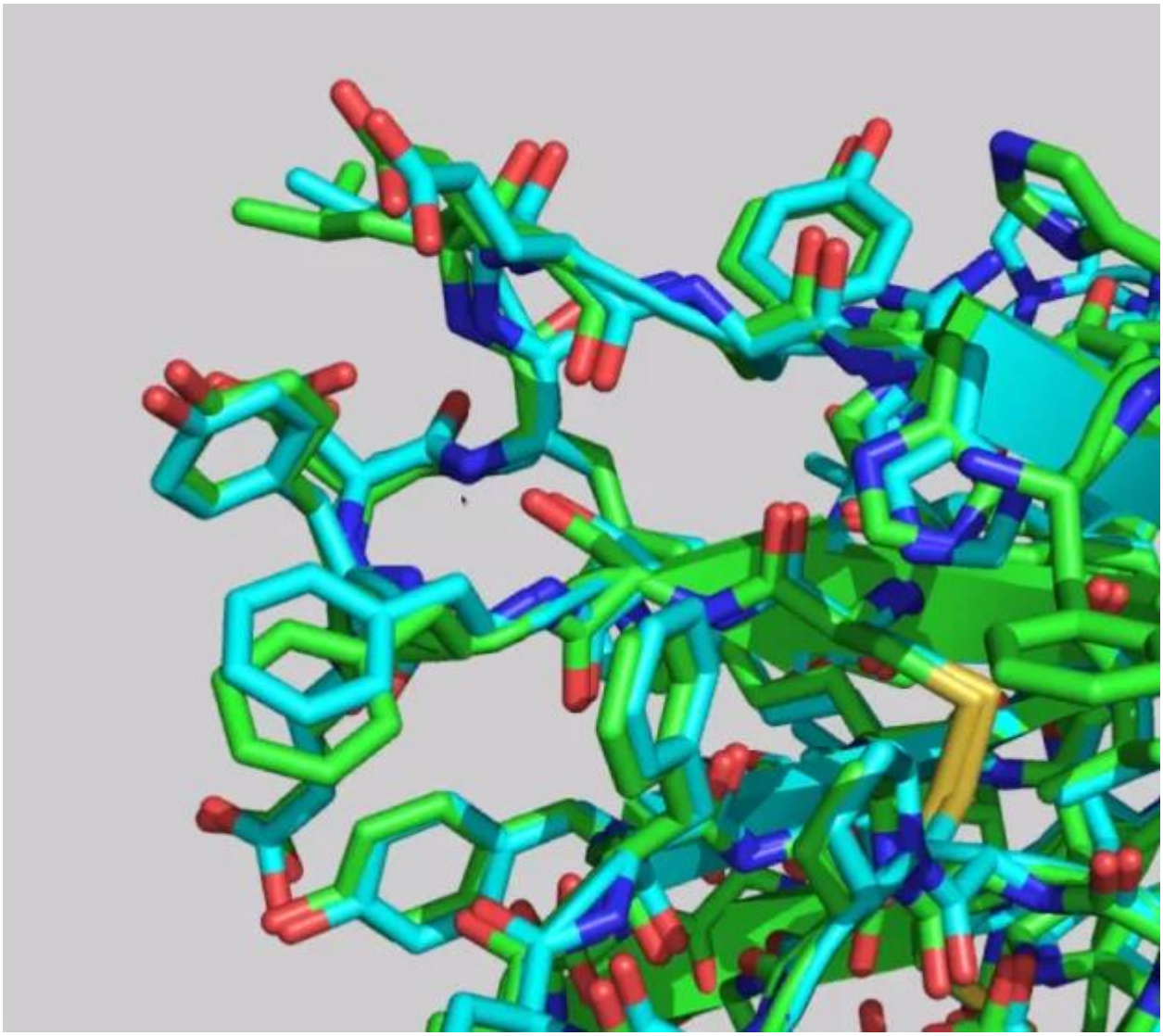
We can see clear differences between the models and the crystal structure. Both models get the topology of the core wrong: Baker's group shows more β-sheets than the crystal structure, and their topology is wrong, combining parallel and antiparallel sheets; Zhang's group barely captures the structure of the core. In both cases, the loops connecting the β-sheets are all around the place, and the large 30-residue loop region discussed before is modelled even worse by these two submissions.

Don't get me wrong — this is a difficult target, Baker's and Zhang's work has been excellent, and their predictions would be state-of-the-art in any other CASP. However, when we compare against the best performing groups in the protein structure prediction community, AlphaFold 2's accuracy is simply on a whole different level. One of the assessors, Nick Grishin, summarized this uncanny performance in a quote that goes more or less like this: "*What did AlphaFold 2 get right, that other models did not? The details*". In fact, the agreement is so good that it extends to side chains:

How did they do it? Part 1: Technical details

DeepMind's description of their protocol in the CASP14 book of abstracts is sparing in detail, and while their talk did unveil some interesting information, much is still unknown. We won't know exactly what they did until they release the corresponding paper, which will take months if not more than a year. However, I can tell you what they have said so far, and we can try to speculate what is going on under the hood.

AlphaFold 2 relies, like most modern prediction algorithms, on a multiple sequence alignment (MSA). The sequence of the protein whose structure we intend to predict is compared across a large database. The underlying idea is that, if two amino acids are in close contact, mutations in one of them will be closely followed by mutations of the other, in order to preserve the structure.



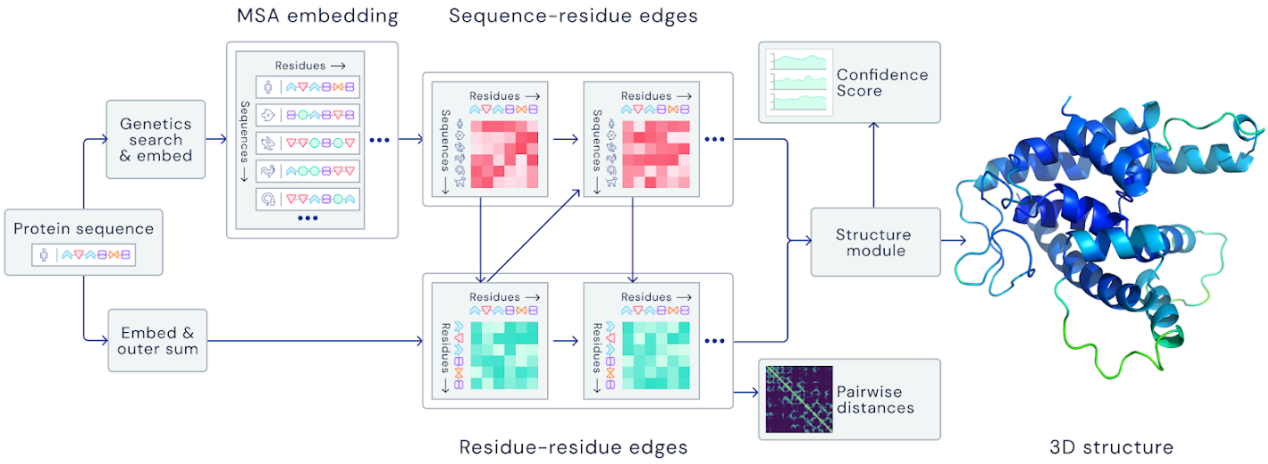
AlphaFold 2 not only predicts the global structure of the protein with high accuracy — it also produces incredibly accurate forecasts of the side chain structure. Image taken from the CASP14 slides.

Suppose we have a protein where an amino acid with negative charge (say, glutamate) is near to an amino acid with positive charge (say, lysine), although they are both far away in the amino acid sequence. This Coulombic interaction stabilises the structure of the protein. Imagine now that the first amino acid mutates into a positively charged amino acid — in order to preserve this contact, the second amino acid will be under evolutionary pressure to mutate into a negatively charged amino acid, otherwise the resulting protein may not be able to fold. Of course, real situations are rarely as clear-cut as this example, but you get the idea.

This principle has inspired very many algorithms to predict structural properties of proteins, from contacts to secondary structure. AlphaFold's own success in CASP13 did in fact use deep learning to predict an interresidue potential from a MSA (and many other features, that included the output of some co-evolution software). This time, however, DeepMind decided to develop an end-to-end model. Instead of using the MSA to predict constraints, they produced a deep learning architecture that takes the MSA as input (plus some template information, but that is another story) and outputs a full structure at the end.

Why did they do this? Given that the ~170,000 structures available in the PDB constitute a small dataset, they wanted to make the most of inductive bias — introducing constraints into the model architecture that ensure that the information is assimilated efficiently. This is similar to the way in which convolutional neural networks (CNNs) learn the importance of local information in an image directly from their architecture, instead of spending time and data inferring it from the training set.

How they use inductive bias is less clear. We know that the input information is embedded into an embedding space, of which we know little. John Jumper, who represented DeepMind at the conference, explained it "*learns sequence-residue edges and residue-residue edges*", and mentioned that the model "*employs an attention-based network, that identifies which edges are important*". We know however that an important piece is a 3D equivariant transformer — a deep learning architecture widely known for its role in GPT-3 and BERT — which are in charge of updating the protein backbone and building the side chains.



DeepMind's diagram (taken from their blog) provides an overview of the architecture of AlphaFold 2 but lacks the details that would be required to reproduce it.

The predictive process proceeds in an iterative fashion, "*passing information back and forth between the MSA and the internal representation of the protein*". I guess this means that the information obtained from a forward pass through the network is somehow fed back to the input features, and then rerun until convergence — but that is, of course, a conjecture. From the graphs shown at the conference, the first predictions are often very good (around 70-80 GDT_TS) and after a few iterations converge to the impressive 90+ GDT_TS predictions that we have seen in CASP14. The final result is not guaranteed to obey all stereochemical constraints, so the final structure is relaxed by coordinate-restrained gradient descent using the Amber ff99SB force field and OpenMM.

How did they do it? Part 2: Not-so-technical details

Of course, the success of the DeepMind team is not only related to deep learning. There is more, a lot more. First of all, DeepMind's blog post states about their model that it "*uses approximately 128 TPUv3 cores (roughly equivalent to ~100-200 GPUs) run over a few weeks, which is a relatively modest amount of compute in the context of most large state-of-the-art models used in machine learning today*".

Tensor Processing Units (TPUs) are a proprietary integrated circuit developed by Google to accelerate the training of neural networks. There is not a clear equivalence between TPUs and GPUs, since performance depends on the problem, but in the right hands they can deliver quite a speedup. Perhaps more importantly, an 8-core TPU v3 chip has 128 GB of vRAM, which is necessary for some architectures that have high memory cost — like attention models. They are also quite expensive – renting 128 TPUv2 cores has an annual cost of half a million dollars as per Google Cloud's pricing page. Replicating DeepMind's experiment using cloud services would take anywhere between $25,000 and $200,000, and the total computational cost is likely to be around several million dollars.

Is this massive computational power the only factor behind DeepMind's success? I don't think so. The work of this talented team displays novel ideas and creative problem-solving, and the differences cannot be attributed merely to computational muscle. At the same time, the massive compute power means not only that they can work with larger models — they can also achieve a much higher throughput than any academic group would. Looking forward, one can only wonder how this imbalance in resources will impact academic computational research. This might lead to a future where computational research groups require significant

investments in infrastructure to be viable — much like our colleagues in experimental biosciences, albeit with a much faster rate of equipment obsolescence.

This leads me to another point. A major piece of DeepMind's success is the availability of ~170,000 structures painstakingly collected by structural biology groups for decades. Big as DeepMind's war chest might be, the taxpayers' investment that has made their achievement possible is several orders of magnitude larger. This is no less true for the wide body of research about protein structure prediction that is available in peer-reviewed articles, which has been conducted, written, and reviewed by academics. If they have managed to see further, it is because they stood on the shoulders of giants.

This raises many interesting questions about the ethics of research, and of artificial intelligence. Consider, for example, the possibility that Alphabet decides to commercially exploit AlphaFold, for example — is it reasonable that they make profit off such a large body of research paid almost exclusively by the taxpayers? To what extent is the information created by publicly available research — made public, mind you, to stimulate further public research — belong to the public, and under what conditions could it be used in for-profit initiatives? There are many questions to ask if we want to keep science being the open, collaborative machine that it ought to be.

What will this mean for biology, and for us bio/cheminformaticians?
There are many questions that are circulating amongst most bio/cheminformaticians right now. The first question is: will they [DeepMind] make their code available, and if so, how? The second, which is probably the most worrying, is: how will this affect my work?

The first question is of utmost importance. When asked about code availability (just over a third of the questions in the virtual [CASP14] Q&A chat-box), John Jumper claimed they were having "*internal discussions*" at DeepMind, about "*making their work available to the community*" and that they were hoping to make an announcement in early January.

There are multiple ways that this could happen. Alphabet may decide to exploit AlphaFold 2 commercially — very much like OpenAI decided to do with GPT-3, the celebrated language model unveiled earlier this year. Then there is the possibility that they decide to open source the code — probably with some sort of license for commercial users. Mind you, making their code available does not mean that anyone could run it. When their Nature paper was published last January, there was a crucial piece lacking: the code to build the input features to the neural network, which the community has so far been unable to reproduce.

A related question is how long will it actually take to run this code. For that, fortunately we have an answer: Demis Hassabis (CEO of DeepMind) confirmed that the model requires "*between a few hours to a few days*" on 5-40 GPUs, depending on the protein. This is a significant computational cost, and highly variable, but not beyond the resources of most academic institutions, and could be considered modest in comparison to, say, routine computational fluid dynamics calculations.

The question most readers will be concerned with is, well, what does it mean for us? It means, generally, that we can focus on those areas previously hindered by the difficulties of obtaining protein structures.

However, while AlphaFold 2 provides a general solution for protein structure prediction, this does not mean that it is universal. Several of the CASP14 targets were not predicted successfully, suggesting that there are some protein families that require further work; and of course, these targets are not fully representative of a proteome. The model was trained on the Protein Data Bank, which has a well-known bias towards proteins that are easy to crystallize. Furthermore, since AlphaFold takes a multiple sequence alignment as input, it remains to be seen if it can tackle problems where these are shallow or not very informative, as happens for example in the very important problem of protein design, in mutated sequences or sometimes in antibody sequences.

All in all, AlphaFold 2 may prove highly beneficial for the readers of this newsletter. The increased availability of protein structures will only heighten the interest in developing areas of chemical informatics, such as proteochemometrics and protein-ligand docking, among others. Likewise, areas, such as protein folding and misfolding, protein movement (including flexibility and allostery), protein-protein interactions or protein design and engineering, as well as their pharmaceutical applications (say, allosteric modulators or drugs that

stabilise a partially unfolded state of a protein), are likely to receive increased interest. One thing I am certain: it is a good time to be in this business.

Conclusion
What Google just achieved might very well be among the most important scientific achievements this century, in terms of impact if not epistemologically. The long sought-after ability to predict the structure of a protein from its sequence (and, as of yet, availability of similar mutated sequences) will unlock applications spanning the entirety of the life and medical sciences, from basic biology to pharmaceutical applications. The prospects are truly astounding.

That said, this statement has to be taken carefully. While we have a general solution to the protein structure prediction problem, we do not yet have a universal one. Some of the structures in CASP were predicted with low accuracy even by AlphaFold 2, suggesting that further work might be required in particular target families. The Protein Data Bank, which was used for training, displays a well-known bias towards easy to crystallize proteins, and it is unclear how this will affect its usefulness for the dark proteome. Furthermore, since prediction relies on a multiple sequence alignment, it remains to be seen whether this method works when there are few or no sequences in the alignment, as might happen with designed proteins, or when it is not very informative, as in antibodies.

The success of DeepMind also raises a number of points that we, the scientific community, need to consider quite seriously. While nimbler and better funded than most individual research groups, this achievement elicits deep-rooted questions about the way we conduct and communicate research, and whether our community, which collectively has more resources and accumulated knowledge, has really been using their potential efficiently. We also need to reflect on our responsibility as scientists to ensure that science remains open, and that the research pursued with the support of the public remains useful for the public.

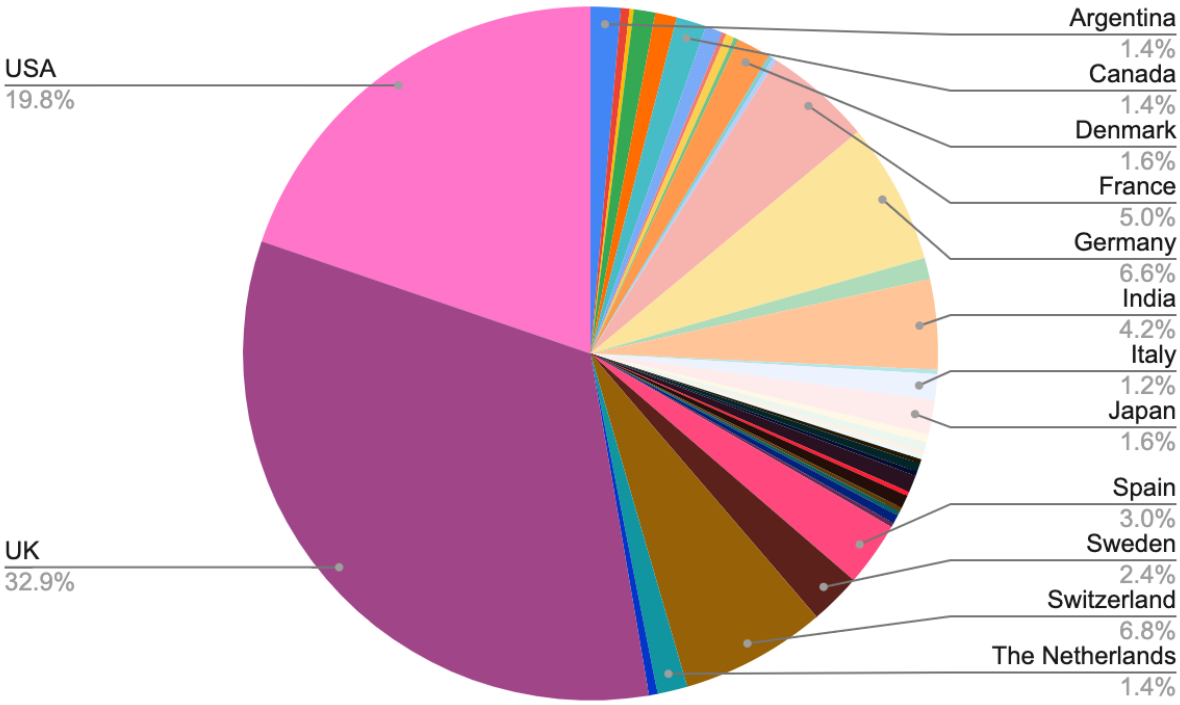
These concerns aside, the solution of the structure prediction problem will finally stimulate novel pathways of research. For too long we have focused on reproducing the static picture of protein structure that we capture through X-ray crystallography. We can now dedicate more efforts to other equally interesting questions: how do proteins fold into these fantastically complicated conformations? How do they move, and how is that movement regulated? How do they interact with other proteins, and with ligands? This is only the start of a very exciting time for protein and chemical informatics.

Meeting Report: 3rd RSC BMCS & CICAG AI in Chemistry Meeting

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

The 3rd annual RSC-BMCS / RSC-CICAG Artificial Intelligence in Chemistry meeting had originally been planned as a physical event on 28th-29th September 2020 at Churchill College in Cambridge, however due to the COVID-19 pandemic it was reorganised as a virtual event at relatively short notice. We are grateful that all speakers selected for the original meeting agreed to give their presentations remotely. In addition, the sponsorship of AstraZeneca and MSD together with the exhibitors CCDC, Concept Life Sciences, IKTOS, Liverpool ChiroChem, Mcule and o2h Discovery for this experimental virtual event is very much appreciated.

This was the first multi-day meeting we had organised as a virtual event, and the decision was made to cap attendance at 500, as it turned out registrations hit this figure a couple of days after opening. One of the advantages of a virtual meeting is that the lack of travel means delegates from a wide range of countries can attend, in total 41 countries were represented, and the breakdown is shown below.



Among the 500 registrations, the Male to Female ratio was 384:116 with a variety of organisations represented, Academic/Industry/Other 211:259:30 (Other = government bodies, consultants, etc). Only 69 out the 500 registratrants were RSC members.

The twelve presentations came from a range of countries and had a good mixture of early career and more established researchers, from both academia and industry, with a 5:7 F:M gender balance.

Poster Session

In an effort to mimic a physical meeting a formal poster session was organised, and 45 posters were presented. These were available in PDF format on the conference website, but in order to stimulate interactions, presenters were encouraged to post on Twitter, and details are shown in the table below. In addition, the majority of the poster presenters also submitted a 2-minute Lightning presentation which were then combined into a video posted on YouTube; these video presentations have been viewed over 1,000 times. At the end of each day breakout rooms were organised to enable delegates to chat to the poster presenters.

- Day 1 (Odd numbered posters) https://youtu.be/ikbrrnjOvdc
- Day 2 (Even numbered posters) https://youtu.be/tf8cGodlWfU

The use of social media platforms was felt to be important to engage with a wider community, and the #AIChem20 hashtag has been used on Twitter nearly 800 times.

Poster Number	Name	Title	Twitter link
P01	Antreas Afantitis	Enalos cheminformatics tools: development of a de novo drug design module	View on Twitter
P02	Nurlybek Amangeldiuly	Transfer learning with graph neural networks for protein_ligand binding kinetics prediction	View on Twitter

P03	Andy Sode Anker	Characterising the atomic structure of mono_metallic nanoparticles from x_ray scattering data using conditional generative models	View on Twitter
P04	Jenna Bilbrey	A look inside the black box: using graph_theoretical descriptors for the post hoc interpretation of neural networks	View on Twitter
P05	Nicolas Bosc	MAIP: a prediction platform for predicting blood_stage malaria inhibitors	View on Twitter
P06	Xiaojing Cong	Receptor ligand prediction by proteochemometric modeling: an application to G protein_coupled olfactory receptors	View on Twitter
P07	Simon Durr	EVOLVE: a genetic algorithm to predict thermostability	View on Twitter
P08	Umberto Esposito	Building a connected data pipeline to target drug development challenges	View on Twitter
P09	Benedek Fabian	MolBERT: molecular representation learning with advanced language models and useful auxiliary tasks	View on Twitter
P10	Miguel Garcia_Ortegon	Improving VAE molecular representations by tailoring them to predict docking poses and scores	View on Twitter
P11	Wenhao Gao	Can we synthesize molecules proposed by generative models	View on Twitter
P12	Helena Gaspar	Proteochemometric models using multiple sequence alignments and a SentencePiece_based masked language model: application to CYP and kinome selectivity modelling	View on Twitter
P13	Ed Griffen	An explainable AI system for medicinal chemists	View on Twitter
P14	Ed Griffen	"Chemists: AI is here, unite to get the benefits"	View on Twitter
P15	Thomas Hadfield	Explicit incorporation of structural information into a fragment elaboration model via deep reinforcement learning	View on Twitter
P16	Hans Hanley	"GENerateZ: designing anticancer drugs using transcriptomic data, genetic algorithms, and variational autoencoder"	View on Twitter
P17	Fergus Imrie	Generating property_matched decoy molecules using deep learning	View on Twitter
P18	Kjell Jorner	Uniform quantitative predictive modelling for route design	View on Twitter
P19	Itai Levin	Computationally assisted synthesis planning for hybrid chemoenzymatic pathways	View on Twitter
P20	Timur Madzhidov	Deep conditional variational autoencoder for reaction conditions prediction	View on Twitter

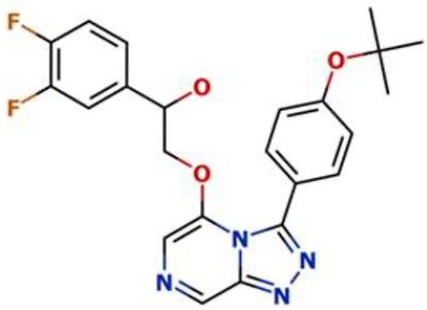
P21	Gergely Makara	AI_assisted lead optimization with derivatization design	View on Twitter
P22	Neann Mathai	Performance and scope of a similarity_based and a random forest_based machine learning approach for small_molecule target prediction	View on Twitter
P23	Janosh Menke	Enhancing molecular fingerprints using neural networks	View on Twitter
P24	Juan Carlos Mobarec	Evolutionary chemistry for the design of desired pharmacological profiles	View on Twitter
P25	Rohit Modee	Neural network potentials for representing potential energy surface and their applicability for geometry optimization	View on Twitter
P26	Joseph Morrone	Challenges and progress in combining docking programs with deep neural networks	View on Twitter
P27	Eva Nittinger	Non_additivity in public and inhouse data and its influence on ML performance	View on Twitter
P28	Ferruccio Palazzesi	Integrating multi task graph convolutional neural network with a deep generative model	View on Twitter
P29	Yashaswi Pathak	Deep learning enabled inorganic material generator	View on Twitter
P30	Quentin Perron	Integrating data_driven computer_aided synthetic planning with generative AI	
P31	Daniel Probst	Classification of chemical reactions through NLP_inspired fingerprinting	View on Twitter
P32	Mikolaj Sacha	Molecule edit graph attention network: modeling chemical reactions as sequences of graph edits	View on Twitter
P33	withdrawn		
P34	Jenke Scheen	Data_driven estimation of optimally_designed perturbation networks in relative alchemical free energy calculations	View on Twitter
P35	Philippe Schwaller	RXNMapper: unsupervised attention_guided atom_mapping	View on Twitter
P36	Matthew Segall	Imputation versus prediction and applications in drug discovery	View on Twitter
P37	Vishnu Sresht	Can generative models learn privileged substructures and identify new bioisosteres?	View on Twitter
P38	Gergely Takács	Analysis of commercial and public compound databases by self_organizing maps	View on Twitter
P39	Morgan Thomas	Towards integrating deep generative models with structure_based design	View on Twitter

P40	Hao Tian	What is hidden behind allostery? An integrated framework to decipher key components in AuLOV dimerization	View on Twitter
P41	Alain Vaucher	Learning how to do chemical reactions from data	View on Twitter
P42	Alexander van Teijlingen	Beyond tripeptides _ two_step active machine learning for very large datasets	View on Twitter
P43	James Wallace	eApps – enabling a predict first culture for computational medicinal chemistry	View on Twitter
P44	withdrawn		
P45	Yuanqing Wang	Bayesian active drug discovery via deep graph kernel learning	View on Twitter
P46	Robbie Warringham	DigitalGlassware: structuring and contextualising chemical outcomes for faster discovery	View on Twitter
P47	withdrawn		
P48	Jerome Wicker	AIScape: a machine learning platform for activity and ADME predictions	View on Twitter

The Main Programme

Gareth Conduit, (Intellegens, UK) gave an outstanding opening Keynote entitled "*Machine learning for efficient design of industrial formulations*". One of the major challenges in this area is the lack of data, many data sets are incomplete, and Gareth described their work dealing with sparse data. The ability to impute values from sparse data has now been demonstrated in a wide variety of industrial applications, ranging from drug discovery, high temperature alloys, lubricants, inks, and battery design. These methods can be used to impute missing values, but also identify outliers (potential false negatives), or suggest which missing data-points would be the most valuable to determine experimentally.

This technology was used in collaboration with Optibrium as part of a competition organised by the Open-Source Malaria Consortium and presented at a recent meeting AI & ML in Drug Discovery Meeting '*Predicting the Activity of Drug Candidates when there is no Target*' sponsored by CICAG and CDD. The full meeting report is in the CICAG Summer 2020 Newsletter.



They predicted pIC50: 6.4; the experimental measurement was pIC50: 6.2

Ten organisations/groups entered the competition to build predictive models using the Open-Source Malaria dataset, these models were then used to predict the activity of several structures for which the data was hidden. The top scoring groups from the first round were then asked to use their models to design novel molecules. The novel molecules were then made and evaluated and the results were made public. The structure of the winning entry from the Intellegens/Optibrium group is shown on the left.

References
Enhancing NEMD with improved sampling of shear rates to model viscosity and correction of systematic errors in modelling density: Application to linear and light branched alkanes
P. Santak & G.J. Conduit Journal of Chemical Physics 153, 014102 (2020) https://aip.scitation.org/doi/10.1063/5.0004377
Practical Applications of Deep Learning to Impute Heterogeneous Drug Discovery Data
Benedict W. J. Irwin, Julian Levell‖, Thomas M. Whitehead, Matthew D. Segall, Gareth J. Conduit Journal of Chemical Information and Modelling 2020, 60, 6, 2848–2857
https://doi.org/10.1021/acs.jcim.0c00443

Barking up the right tree: de novo design via searching over molecule synthesis DAGs John Bradshaw, (University of Cambridge, UK)
Recently, machine learning (ML) models for generating novel molecules have been proposed to perform de novo molecular design and to construct libraries for virtual screening [1, 2, 3]. These ML models use deep generative neural networks (NNs) to output new SMILES strings or molecular graphs. However, in creating new molecules in this manner, these ML approaches have ignored subtle constraints: synthesisability and stability. When suggesting new molecules with desirable property profiles to a chemist, it is not only important what to make but crucially how to make it. These instructions form a synthesis directed acyclic graph (DAG) (see Figure), where each node in the DAG represents a molecular graph and the edges reactions between them. Overall, a DAG describes how a large vocabulary of simple building blocks can be recursively combined through chemical reactions to create more complicated molecules of interest.

Inspired by early work on reaction driven de novo design [4], we therefore design a generative model around this process, by directly generating molecular synthesis DAGs. Our model, DoG-Gen (DAG of graphs generator), is based on recurrent neural networks and graph neural networks that are used to construct the DAG in a bottom-up manner through a series of actions. DoG-Gen can be used in combination with a scoring function, hill climbing procedures [2] or reinforcement learning to search for molecules with the desired properties. Furthermore, the DoG-Gen architecture can also be used as the decoder in an autoencoder structure [1] (resulting in an overall model which we call DoG-AE), to provide a latent space that can be used for sampling new molecules and interpolating within chemical space.

We assess our models in a variety of generation and optimization tasks. We show that our models generate a wide range of diverse molecules, covering the target chemical space well. Moreover, in the GuacaMol [5] de novo design optimization benchmark, our model can find molecules with comparable property scores to those found using methods that do not encode synthesisability constraints. However, importantly we show that most of the optimized molecules our model suggests are synthesisable and stable, which cannot be said for many of the molecules suggested by the previous unconstrained methods.

We envision that by directly predicting which molecules to make next and how to make them, our model can provide immediately actionable suggestions and leads to faster experimental feedback, and so is another step towards speeding up the search for new drugs.

References
[1] Gómez-Bombarelli et al., ACS Cent. Sci. 2018. https://doi.org/10.1021/acscentsci.7b00572
[2] Segler et al. ACS Cent. Sci. 2018. https://doi.org/10.1021/acscentsci.7b00512
[3] Simonovsky and Komodakis. ICANN 2018. 2018. pp 412-422
[4] Maarten Vinkers et al. J. Med. Chem. 2003, 46, 2765, https://doi.org/10.1021/jm030809x
[5] Brown et al. J. Chem. Inf. Model. 2019 https://doi.org/10.1021/acs.jcim.8b00839

Bayer's in silico ADMET platform: how the combination of artificial and human intelligence can make a difference in drug discovery, Andreas Göller, (Bayer AG, DE)
Over the past two decades, we have implemented a productive platform at Bayer Pharmaceuticals for the computation of absorption, distribution, metabolism, and excretion (ADMET) property profiles of compounds with the goal to generate models for a variety of pharmacokinetic and physicochemical endpoints in early

drug discovery. These tools are accessible to all scientists within the company and can be useful in assisting with the selection and design of novel leads, as well as the lead optimization process.

The talk deals with the development of machine-learning (ML) approaches with special emphasis on data, descriptors, and algorithms, always based on concrete application use cases. Evidence will be provided that data quality and quantity have the highest impact on model quality and that a thorough understanding of the experimental endpoints is key to success. On algorithms, examples for classical single task machine-learning as well as multi-task deep neural networks will be given. Finally, I will describe problem-tailored atom descriptors developed by us for the interpretation of metabolic transformation reactions that are also applied for Ames mutagenicity assessment and hydrogen bond strengths calculations.

References
AH Göller, L Kuhnke, F Montanari, A Bonin, S Schneckener, A ter Laak, J Wichard, M Lobell, A Hillisch, Bayer's in silico ADMET platform: a journey of machine learning over the past two decades, Drug Discovery Today, in press. http://dx.doi.org/10.1016/j.drudis.2020.07.001
AH Göller, The art of atom descriptor design, Drug Discovery Today: Technologies, in Press. http://dx.doi.org/10.1016/j.ddtec.2020.06.004
AR Finkelmann, AH Göller, D Goldmann, G Schneider, MetScore: Site of Metabolism Prediction Beyond CYP P450 Enzymes, ChemMedChem., 2018, 13, 2281-2289. http://dx.doi.org/10.1002/cmdc.201800309

Unassisted noise-reduction of chemical reactions data sets, Alessandra Toniato, (IBM Research Europe - Zurich, CH)
Existing deep learning models applied to reaction prediction in organic chemistry are able to reach extremely high levels of accuracy (> 90% for NLP- based ones1). With no chemical knowledge embedded than the information learnt from reaction data, the quality of the data sets plays a crucial role in the performance of the prediction models. While human curation is prohibitively expensive, the need for unaided approaches to remove chemically incorrect entries from existing data sets is essential to improve the performance of artificial intelligence models in synthetic chemistry tasks.

Here we propose a machine learning-based, unassisted approach to remove chemically wrong entries (noise) from chemical reaction collections. We applied this method to the collection of chemical reactions Pistachio, extracted from USPTO patents. Results show that models trained on the cleaned and balanced dataset improve the quality of the predictions without a decrease in performance. For the retrosynthetic models, the round-trip accuracy is enhanced by 13% and the value of the cumulative Jensen Shannon metric is lowered down to 70% of its original value, while maintaining high values of coverage (97%) and constant class-diversity (1.6) at inference.

The data is the filter is in the data: meaningful machine learning models and machine-learned pharmacophores from fragment screening campaigns, Carl Poelking, (Astex Pharmaceuticals and University of Cambridge, UK)

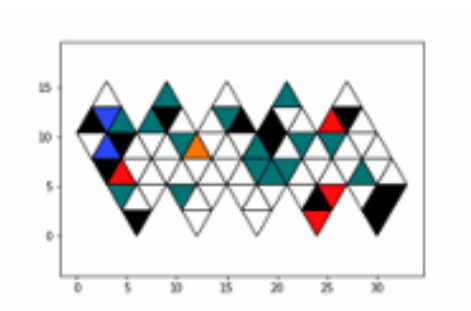
Machine learning (ML) is widely used in drug discovery to generate models that predict protein-ligand binding. These models are of great value to medicinal chemists, in particular if they provide case-specific insight into the physical interactions that drive the binding process. In this study we derive ML models from over 50 fragment-screening campaigns to introduce two important elements that we believe are absent in most -- if not all -- ML studies of this type reported to date: First, alongside the observed hits we use to train our models, we incorporate true misses and show that these experimentally validated negative data are of significant importance to the quality of the derived models. Second, we provide a physically interpretable and verifiable representation of what the ML model considers important for successful binding. This representation is derived from an attribution and filtering procedure that explains the prediction in terms of the action of chemical environments. Critically, we validate the attribution outcome on a large scale against prior annotations made independently by expert molecular modellers. We find good agreement between the key molecular substructures proposed by the ML model and those assigned manually, even when the model's performance in discriminating hits from misses is far from perfect. By projecting the attribution onto

predefined interaction prototypes (pharmacophores), we show that ML allows us to formulate simple rules for what drives fragment binding against a target automatically from screening data.

Novel chemical structure input method for machine learning to capture 3D information like chirality and rotational symmetry, Ella Gale, (University of Bristol, UK)
Introduction: Chemical structure is inherently 3D (4D including vibrations). The problems of de novo drug design, lead compound identification, reaction prediction, and retrosynthesis require the tracking of chirality, and, we posit, require 3D information to do this. However, most machine learning algorithms work with 1D (SMILES) or 2D (connection matrix) data. Although SMILES attempts to include chirality, it uses rules-based transformation and symbols (not good for some ML) and is dependent on the algorithm deducing background information about chemistry. While a human chemist can infer 3D structure from 1D formulae (i.e. C2H6) or 2D formulae (a), asking an ML algorithm to learn this transform increases the difficulty of the task. Thus, we have developed a method to input 3D information in a machine-readable form. Furthermore, part of the success of ML algorithms comes from data augmentation (which increases the size of the dataset), and this technique enables augmentation.

Method: Spherical convolutional neural networks (S-CNNs) work on 3D spherical data, usually an icosphere (b), and S-CNNs are rotationally and translationally invariant. The presented input technique involves taking a molecular formula or structure (SMILES or mol file), (a) optimising the geometry using MM, then placing the molecule inside an icosphere (b) and projecting the atoms onto the surface of that icosphere (imagine a light at the centre of the molecule and the atoms casting shadows onto the surface). As this is ray-casting, the calculations are very fast using graphical processing units. The icosphere is unwrapped to a net (c). The icosphere can be unwrapped in many different ways (60) and each unwrapping is equivalent to a fixed rotation of the molecule, augmenting the dataset and giving rotational information. Furthermore, molecules can be translated or rotated within the icosphere, or slightly deformed (via molecular dynamics or normal mode analysis) to further augment the dataset and give information of symmetry breaking and conformers. Standard CNN augmentation includes creating mirror images of inputs, we do not do this to allow the preservation of chirality. Spherical CNNs take in the net data, and map it to the icosphere, maintaining the 3D nature of the input.

Preliminary results: We develop some simple benchmark datasets and demonstrate that this approach performs substantially above chance. Work is on-going to develop optimal input procedure (there are choices as to how the atoms are encoded) and develop simple 3D benchmark datasets, before applying to more complex benchmark tests.



DNA-encoded small molecules libraries meet machine learning, Patrick Riley, (Google, US)
DNA-encoded small molecule libraries (DELs) have enabled discovery of novel inhibitors for many distinct protein targets of therapeutic value. We demonstrate a new approach applying machine learning to DEL selection data by identifying active molecules from large libraries of commercial and easily synthesisable compounds. We train models using only DEL selection data and apply automated or automatable filters to the predictions. We perform a large prospective study (~2000 compounds) across three diverse protein targets: sEH (a hydrolase), ERα (a nuclear receptor), and c-KIT (a kinase). The approach is effective, with an overall hit rate of ~30% at 30 μM and discovery of potent compounds (IC50 < 10 nM) for every target. The system

makes useful predictions even for molecules dissimilar to the original DEL, and the compounds identified are diverse, predominantly drug-like, and different from known ligands. This work demonstrates a powerful new approach to hit-finding.

References

J. Med. Chem. 2020, 63, 16, 8857–8866 https://doi.org/10.1021/acs.jmedchem.0c00452

Additional talks:

- Jessica Lanini, (Novartis, CH) Data driven representations for predicting molecular properties: benchmarking and applications in generative chemistry
- Olexandr Isayev, (Carnegie Mellon, US) Teaching a neural network about chemical reactivity
- Keynote: Olexandr Isayev, Carnegie Mellon, US
- Esben Jannik Bjerrum, (AstraZeneca, SE) Artificial neural network enhanced synthesis and retrosynthesis prediction
- Sereina Riniker, (ETH Zürich, CH) Using machine learning for molecular dynamics simulations
- Keynote: Sereina Riniker, ETH Zürich, CH
- Hannah Bruce McDonald, (MSD, UK) Free energy calculations for drug discovery with hybrid ML/MM potentials

Full details of the meeting are available on the conference website, together with copies of the slide decks where speakers have given their permission.

A few comments from Twitter

"*Still absorbing all of the mind-blowing content of #AIchem20! Humbling see what is going across the field! The talks by @Google's Patrick Riley, @BristolChem's Ella Gale and @intellegensai's Gareth Conduit were really quite fascinating! Very lucky to have been able to attend!*"

"*Amazing, @RSC_CICAG #AIChem20 meeting beats record of online engagement! 250+ people from all over the world Earth globe europe-africa at the same time #compchem*"

"*I do want to emphasise that I spent the whole of Monday and Tuesday with my jaw on the floor, EVERY talk was amazing! I am very excited for AIChem2021! Hopefully, it will be in Cambridge but even as an online event, it was brilliant! Well done to the team behind it! Clapping hands sign*"

"*Thrilled to be speaking this afternoon at AI in Chemistry, introducing a new way to represent molecules and augment small datasets. This is an unfolded projection onto an icosphere of tamiflu, ready to be input into a neural network (Hydrogen atoms are teal coloured)*"

"*Thank you for the organisers & presenters #AIChem20 for such a stimulations symposium. In 90s, when I started data-driven chemistry, people somewhat hesitated to use the terms AI and intelligent systems, but now they are normal. Simply fantastic*"

"*Blown away by last keynote speaker in #AIChem20! Thank you, Patrick Riley, (@GoogleAI) raising very good questions and opinions at the conference closing. A lot to think about!The future of drug discovery is here! And a big "Thank you" to all organisers! Merci!*"

"*Thanks to all involved in #AIChem20 for a really exciting couple of days showcasing cutting edge applications of #AI in chemistry. We're already looking forward to next year!!*

ReadMe and HowTo for Lightning Poster Presentations

Contribution from RSC CICAG Chair Dr Chris Swain and Garrett Morris, email: garrett.morris@dtc.ox.ac.uk

At the latest AI in Chemistry meeting, we had two sessions of lightning poster presentations, presenters submit a recording of a couple of minutes to describe their poster. These proved incredibly popular being viewed over 1,000 times and are viewable here – Day 1, Day 2.

During the construction of the videos, it became apparent that perhaps a little guidance might be useful, both for presenters and for those who might be stitching together the presentations in the future. With this is mind, Garrett and Chris put together a short ReadMe that it is hoped will be useful to others.

README for RSC CICAG Lightning Posters

When submitting your video, please follow these instructions:

- **File-naming convention**
  - Use the following standard file naming format: "P" followed by your poster number, using leading zeros, followed by an underscore, then your first name, then a second underscore, your last name, and finally the filename extension, ".mp4". For example, a video about Poster 5 from John Smith would be named: "P05_John_Smith.mp4".
- **Video format**
  - Use MPEG-4 movie format, i.e. filename extension ".mp4".
- **Video codec**
  - Use either the AAC or H.264 codecs.
- **Aspect ratio**
  - Use 16:9 standard, *i.e.* dimensions: 1920 x 1080 (this is also known as HD video standard; US & UK digital broadcast TV standard)
- **Sound level**
  - Make sure your microphone sensitivity or sound input level is set to be as sensitive as possible. Ideally, use a headset or get close to the microphone.
- **Ambient sound**
  - Try to record your audio in an environment without smooth, hard surfaces, to avoid echoes. Try to avoid background noise or echo.
  - Offices with hard surfaces might not be the best place to record, bedrooms with soft furnishing might be better; You can place soft items such as a blanket nearby, or record in a bedroom near curtains, to soften the acoustics.
  - For more tips, check out: https://training.npr.org/2020/03/31/professional-sound-from-a-diy-studio-it-can-be-done/.
- **Online storage**
  - These are large files so email transfer may not be possible it is better to store your video on a cloud-accessible drive such as iCloud, Google Drive, or Dropbox, *etc.*
- **Headshot**
  - You can use the "picture-in-picture" feature to include video of you presenting, if you are comfortable doing so. Viewers who are hard of hearing appreciate the option for lip-reading. In Zoom, *e.g.*, you can share your screen while presenting your slide(s), and then live video from your webcam will appear in the top right corner. If you do this, make sure to leave room in the corner of your slide(s) clear for the picture-in-picture.

HOW-TO for Merging Videos for RSC CICAG Lightning Posters

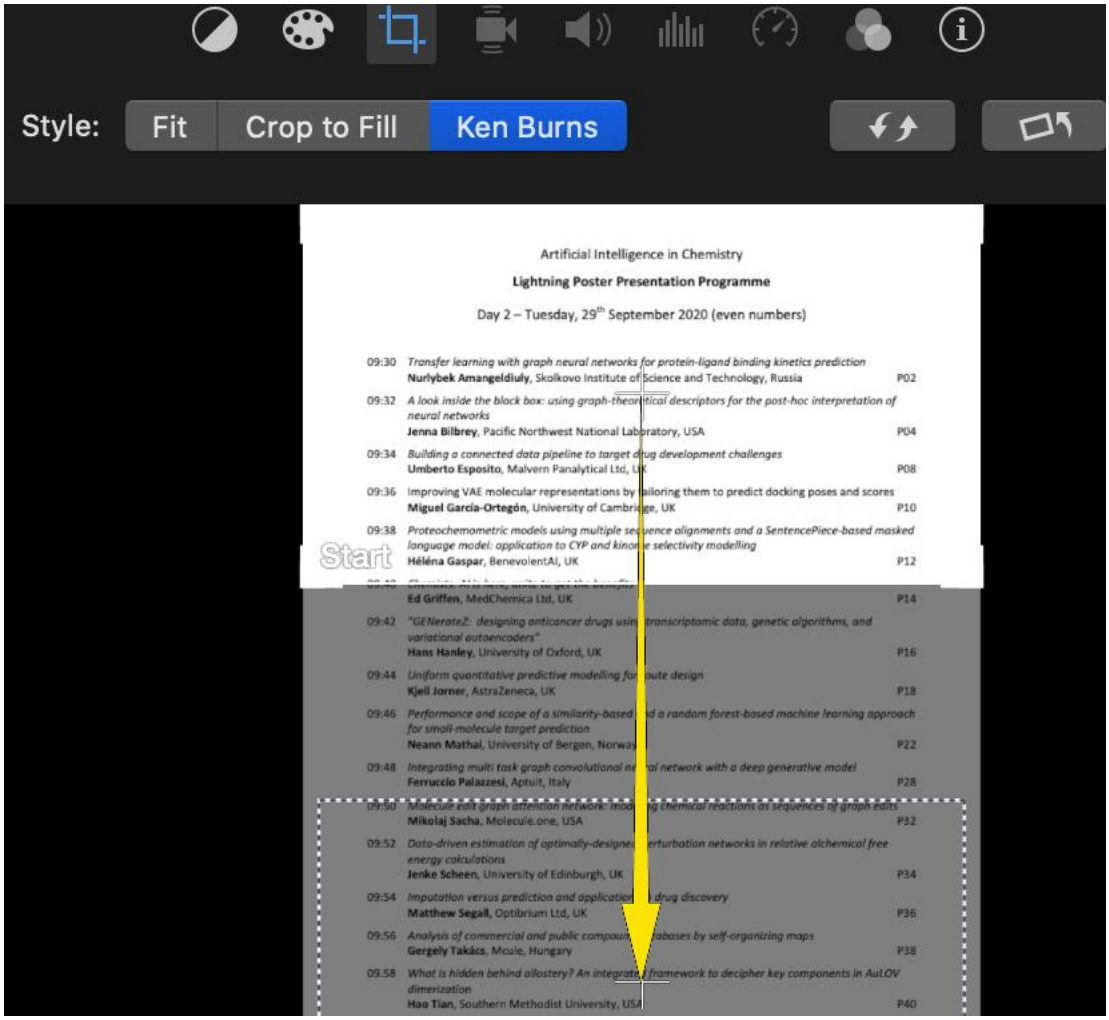
It may be necessary to use several pieces of software to merge the individual contributions. On a Mac, we recommend:

- iMovie (https://www.apple.com/uk/imovie/) - a free Movie editor for Mac and iPad.
- Quicktime Player (https://support.apple.com/downloads/quicktime) for updating a couple of older codecs.
- Handbrake (https://handbrake.fr) a tool for converting video from nearly any format to a selection of modern, widely supported codecs.

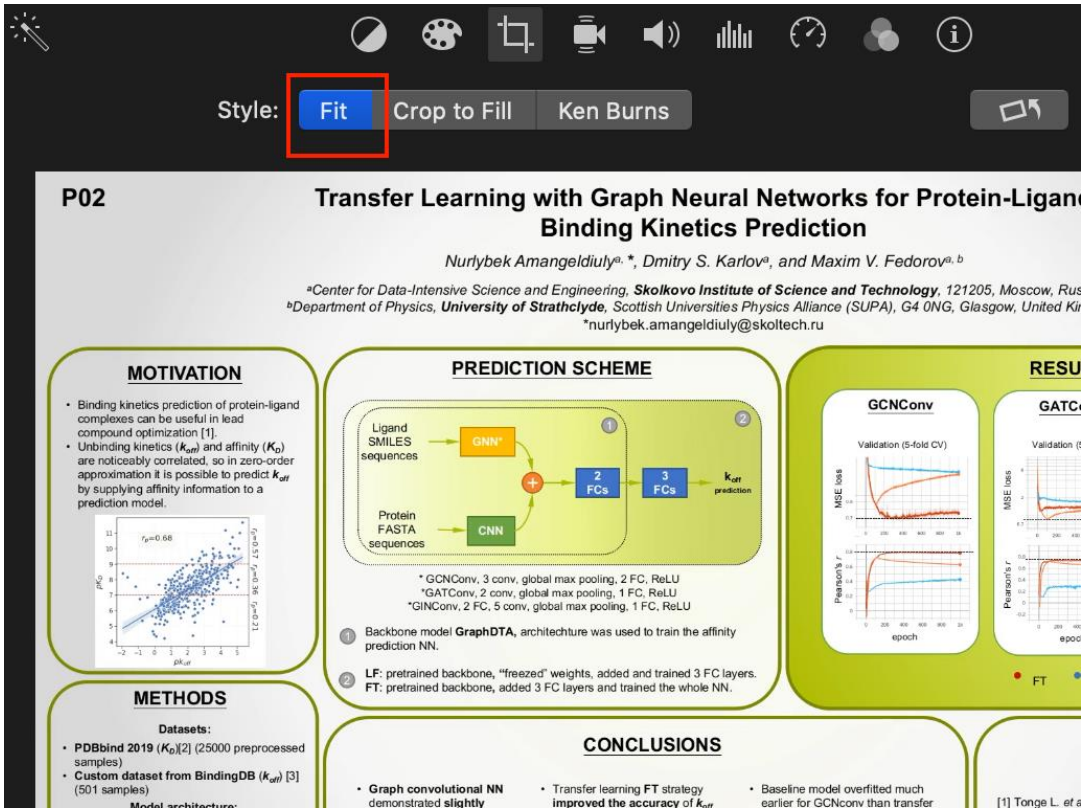
Convert all submissions to MPEG-4 before importing into iMovie.

Use separate iMovie projects for each day of the conference.

Import each day's agenda as a PDF and use the "Ken Burns" effect in iMovie to pan down the page. You can also use this to pan down any page in "Portrait" orientation.

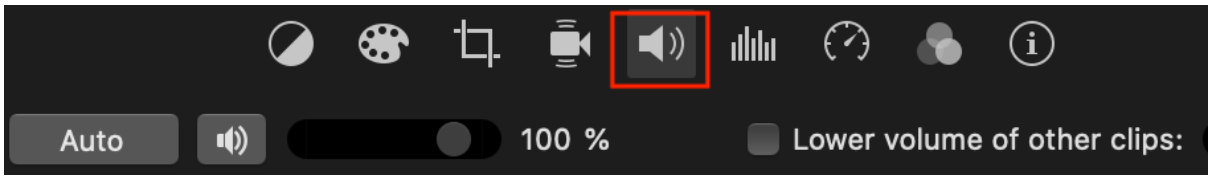


For presentations that have incorrect aspect ratio use the "Fit" option



Add the posters (or lectures) in order and then add a cross-dissolve between each.

Adjust the volume level manually for each presentation, as necessary.



iMovie also has tools for reducing background noise.

Exported the assembled videos as a single file for each day and upload to the RSC CICAG YouTube Channel. These are large files, so you need a good upload speed.

An alternative is the free and open source software Open Broadcaster Software for video recording and live streaming (https://obsproject.com).

Happy recording!

As Conferences went Online: What do we miss the most from in-person Events?

Contributed by RSC CICAG Committee member Jack Simpson, email: J.D.Simpson@liverpool.ac.uk

We all recognise the low hum of chatter that take place over a cup of coffee during a break. It is a familiar sound at any conference, one which won't be heard for the foreseeable future. As the COVID-19 pandemic continues across the world thoughts have turned to the future of conferences and how they will respond as social distancing restrictions ease.

As the pandemic hit, conferences inevitably continued online and allowed the sharing of scientific knowledge to continue, but virtual conferences just aren't the same, are they? No virtual conference can sufficiently replace an in-person conference and with this in mind, we took to Twitter to ask our followers what they miss most about in-person conferences.

From the 70 respondents, there was very little to separate the top 3 results with 33% of people missing the opportunity for unexpected collaborations the most. This was followed at 30% by people missing the social aspect of conferences the most and 29% of people picking what I thought would be the most popular choice, the opportunity to get out of the lab (and maybe an excuse for a quick holiday!).

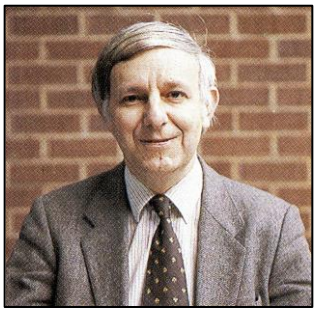
Looking forward, it is difficult to envisage a world where conferences are 100% as they were pre-covid in the "new normal". In this new normal, what will conferences look like in future? Again, we took to the expertise of our Twitter followers to gauge their thoughts. From the 51 responses of our second poll, there was a clear winner. Almost half (49%) of respondents believe that conferences will evolve into a hybrid model consisting of a physical conference that also has an accompanying stream.

Which begs the question, why did it take a pandemic for this type of conference to connect people all over the world? The unintended consequence of this shift online, is that conferences now have the ability to connect even more people, especially those who may not have otherwise have been able to attend, whether this is due to expenses or personal matters, and any extra inclusion in science can only be beneficial.

For other responses, the other half of respondents were split equally between reverting back to wholly in-person and embracing the wholly virtual approach. Whatever the approach will be looking forward, staying connected through conferences throughout various lockdowns has been a welcome normalcy in an otherwise unprecedented and tumultuous time.

Alan F Neville (1943-2020) BSc, PhD

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



One of the saddest consequences of the COVID Pandemic and lockdowns has been the breakdown in normal communications between people, and particularly the discovery in Christmas letters and cards that you had missed the death of a former colleague. That happened to me on 22nd December 2020, when I found out that Alan Frederick Neville had died on 6th May. His wife wrote to say that he had died of rapid onset vascular dementia which caused him to have lots of falls in a care home in Alderley Edge, where he broke his hip, and he had died in Macclesfield Hospital, probably of COVID. His funeral was attended by 3 people – his wife, Elisabeth, and their two sons, Timothy and Christopher, as funerals had enormous restrictions on numbers at that time.

I first met Alan in July 1973 when I started at the Science Library of the University of Manchester as the Chemistry Subject Specialist and cataloguer. I took over his work as he had been moved to the Medical Library to work on the emerging Medlars (Medline) service established at Manchester. Alan was a chemist having a BSc and PhD in organic chemistry from Manchester, and a year doing post doc work on mass spectrometry in the Chemistry Department. In the growing world of computerised chemical information retrieval Alan became a star using his depth of chemical knowledge and quirky ability to make electronic systems work when many others couldn't. LANs and WANs using CDROMs happened, amazing (to the recipients) searches of chemical databases (CAS, Beilstein and many others) followed, and led to the involvement of Manchester in national and international services. Could one really draw a chemical structure on a computer screen and get wonderful answers from a search strategy – yes you could? The problems of working around the holy days at ESA Frascati and time changes with the USA were mastered, and a service that was admired and helped the research of generations of science students and researchers in Manchester took shape. Together with Helen Schofield/Cooke at UMIST and Sean Dunne of Manchester Computing Centre, Alan won the right to provide the national Beilstein Crossfire Service to UK universities, which enabled him to establish a close working bond with Sandy Lawson and others. Alan was a pioneer in a world that is now almost commonplace.

He retired from the university library a few years after I did in 2008, and spent time walking in his beloved Derbyshire hills. I last saw him in Autumn 2019 when he came to a talk I gave to Congleton Local History Society on '*The Development of the Chemical Industry in Runcorn and Widnes*' and was his usual self. Rest in peace Alan. Your contribution to the development of chemical information retrieval is not forgotten.

Dr Diana M Leitch MBE, BSc, PhD, FRSC, former Deputy University Librarian, John Rylands University of Manchester, RSC-CICAG Treasurer

---

Parallel Processing for the Molecular Modeling Engine in ChemDoodle 3D

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

We launched ChemDoodle 3D v6.1 (chemdoodle.com/3d) on August 28, 2020. Included was parallel processing support for the molecular modeling engine, which this article discusses in detail.

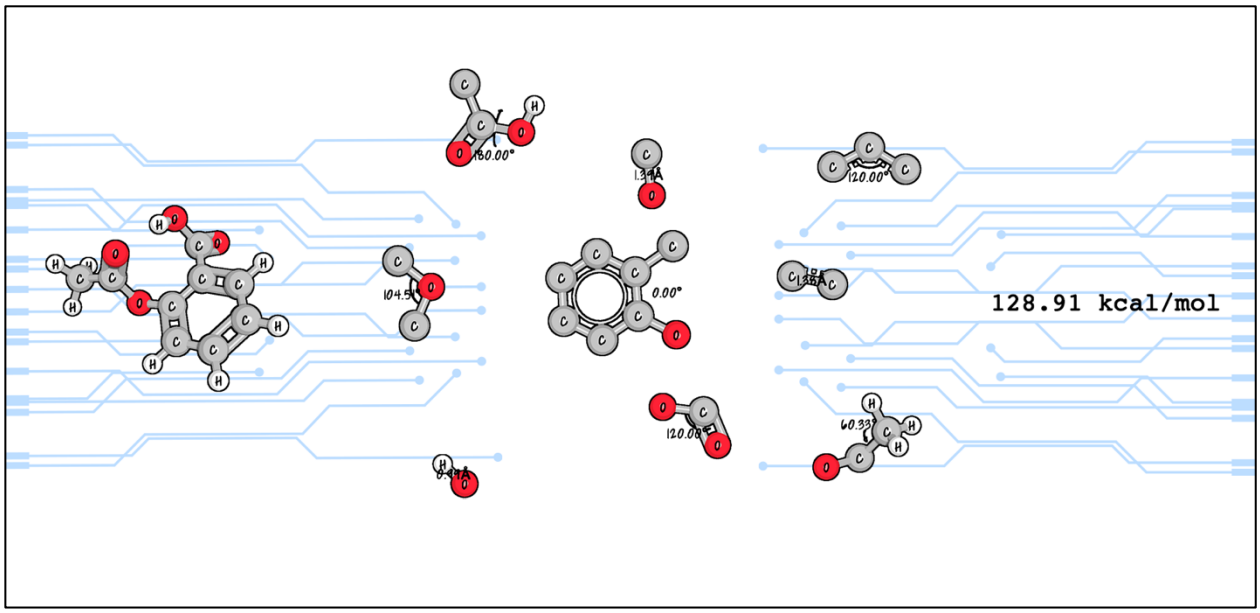


Figure 1: Abstract representation of the parallel process for calculating the energy of a molecule. The molecular graphics in the image were created in ChemDoodle 3D. The background circuit shapes are used with attribution: Circuit Vectors by Vecteezy

The complexity of molecular modeling algorithms increases exponentially as the number of atoms in the system increases. This fact forces us to endure longer and longer runtimes to optimize our molecules as they become larger. One remedy is to increase our computer's power with a faster central processing unit (CPU). And of course, a faster CPU will speed up the performance of any program. Yet, there is a limit to how fast contemporary CPUs can run at, regardless of how much effort the industry applies to increasing that ceiling, and the most powerful CPUs are priced accordingly.

There is another trick. Most modern computers are now equipped with multi-core CPUs. We can then split up the work of the algorithm across the multiple cores of the CPU and benefit from a much faster calculation. But unlike the free benefit of a faster CPU, correctly parallelizing computational tasks is not easy, and there are many pitfalls to implementing algorithms that safely benefit from parallelization. Fortunately, iChemLabs has already done all of the hard work for you in ChemDoodle 3D.

> **Did you know?**
> Just by using ChemDoodle 3D, you have already been taking advantage of parallel processing as the 3D graphics are produced on the graphics processing unit (GPU, also known as the graphics card), which is custom suited to parallelizing the creation of images. Parallelization of the molecular modeling engine in ChemDoodle 3D occurs on multi-core CPUs, instead of on the GPU.

Parallel processing for the molecular modeling engine in ChemDoodle 3D can be enabled by selecting the '**Enable Parallel Processing'** option in the '**Force Fields'** section of the '**Functions'** tab via the **Preferences** window.

Performance considerations
In ChemDoodle 3D, we parallelized our force field implementations. This means that the optimization of a single molecule will run in parallel and a single molecule will be optimized faster, as opposed to optimizing multiple molecules in parallel, which is much easier to do, but provides no benefit to those optimizing a single system. It may seem logical for an algorithm run in parallel on a quad-core CPU (for instance) to run 4 times faster than a sequentially run (non-parallelized) algorithm, but this is not the case. In fact, by enabling parallelization, you may actually decrease the performance of your task. There are a number of reasons for this, which we will discuss now.

1. **Extent of Parallelization** - The algorithm we are improving may not be able to be parallelized in its entirety. Only associative tasks may be parallelized. If we are optimizing 4 different molecules, we can expect a dramatic decrease in runtime by parallelizing the task across 4 cores, one for each molecule. The optimization of one molecule is irrelevant to the optimization of the others. But other algorithms cannot be handled so cleanly. For instance, the optimization procedure itself requires a certain number of partial steps from the unoptimized structure to the optimized structure. Each successful step towards the optimized molecular structure depends on the results of the step preceding it. It does not make sense to run all of the steps concurrently, and they must be run sequentially. If only 10% of the runtime of a function is parallelizable, well then parallelization can never improve runtimes more than 10%.

2. **Overhead** - Parallel processing does not come for free. Algorithms running in parallel must be stateless. To make an algorithm stateless, you have to produce specific data structures to store data for each task. More data is stored, and more objects are instantiated. A complex fork-join algorithm is employed to split the work and combine the results, significantly increasing the runtime of your algorithm. More memory will be used. Moving data between multiple cores is an expensive operation. For a parallel algorithm to run faster than its sequential counterpart, the runtime saved by splitting the work must overcome the runtime introduced by the additional overhead.

3. **Setup Specific** - Most computers are unique: a specific make, operating system, CPU, GPU, memory, hyper-threading, etc. The benefit of parallel processing will be heavily dependent on your hardware. More cores available to your CPU will increase the likelihood that parallelization will improve your runtime.

4. **Computation Specific** - Just because a parallel algorithm improves runtime for a specific molecule calculated for a specific force field, that does not mean another molecule or another force field will have similar results. It is all dependent on how well the computations can be divided and how complex the computations are. Many programmers use the NQ model to assess parallel processing efficacy. N is the magnitude for splitting our computations and Q is the complexity of the computations. The higher the product, the more likely parallel processing will benefit the algorithm. For instance, the optimization runtime of a water molecule, with two bond stretch contributions and one angle bend contribution, will almost certainly be negatively impacted by parallelization.

5. **Other factors** - Software is not running in a vacuum and other applications may be using computer resources. If you are running other CPU intensive applications, then parallel processing in ChemDoodle 3D will be less effective.

In ChemDoodle 3D, we have developed a powerful parallel processing system for our force fields. We micromanage the implementation with our own proprietary *MapReduce* algorithm, and perform our own chunking, forking and joining. Our goal is to minimize overhead. As we develop even better ways to reduce overhead, you will continue to see parallel processing performance in ChemDoodle 3D improve without doing anything.

Analysis

We benchmark the optimization of a few chemical structures in this section to provide a better understanding of the benefits of parallel processing and illustrate when it is appropriate to enable parallel processing. For this analysis, we optimized several chemical structures, each 21 times, discarding the first result as a warm-up. Any iterations resulting in unrealistic configurations were rerun. The 20 recorded runtimes for successful optimizations were then averaged for each molecule. Only the optimization runtimes were recorded, not any file parsing, structure loading, hydrogen enforcement, etc. The optimizations were performed using a MMFF94 force field with a conjugate gradients search direction and Newton line search to convergence. Understandably, each optimization is unique because the starting configuration is random, so the minimum and maximum runtime range may be large, but the averages were pretty consistent across multiple runs and this coarse analysis is suitable for our understanding of the parallelized force field performance in ChemDoodle 3D. The following table collects the results.

All benchmarks were performed on a 2017 iMac running macOS 10.15.6 with a 4.2 GHz Quad-Core Intel Core i7 CPU, with 8 logical cores due to hyper-threading. Java version 11.0.2 was used to compile and run the tests. No other CPU intensive applications were active.

Table 1: Optimization Runtimes - Sequential vs. Parallel Execution

Molecule (#atoms)	Sequential Runtime	Parallel Runtime	Performance (Parallel/Sequential)
Water (3)	0.46ms	3.10ms	674%
Cyclohexane (18)	43.07ms	57.87ms	134%
Aspirin (21)	157.12ms	157.58ms	100%
Sildenafil (63)	1529.27ms	746.02ms	49%
Precoxin A (103)	4465.22ms	1809.07ms	41%

The following graph illustrates the relationship between molecular complexity and runtime for optimizations.

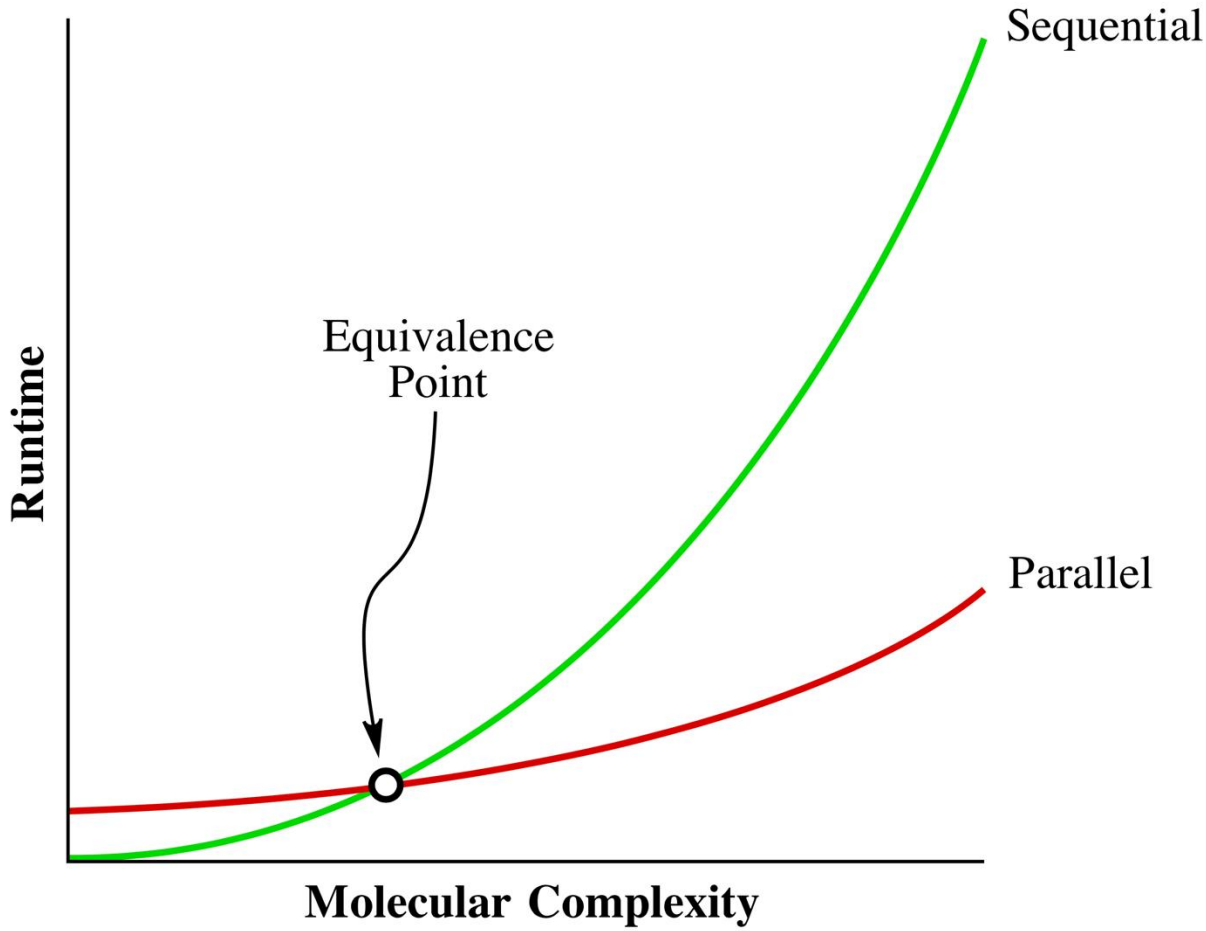


Figure 2: Relationship between molecular complexity and runtime for sequential and parallel optimization

For simple molecules, parallel processing will be a detriment to performance. At the equivalence point where both curves intersect, the optimization of a molecule will perform identically using either sequential or parallel processing. More complex molecules will see improved performance. In our testing environment, aspirin was close to the equivalence point, so molecules more complex than aspirin will benefit from parallel processing.

This equivalence point will be unique to the computer you are using ChemDoodle 3D on, and will be dependent on the force field, search direction and line search you are using and how many other applications you are running. The more complex your molecule gets, the more you will benefit from parallel processing up to the limit of dividing the work by the number of cores available. You can use this understanding to make an educated decision about whether parallel processing will benefit your work in ChemDoodle 3D.

Should Parallel Processing be Enabled?
Why not? If your molecules are taking a long time to optimize and you desire faster runtimes, turn it on. If it helps, great! If it doesn't, simply turn it back off. If you wish to reserve CPU cores for other tasks, and do not want ChemDoodle 3D to consume significant CPU resources, then keep it off.

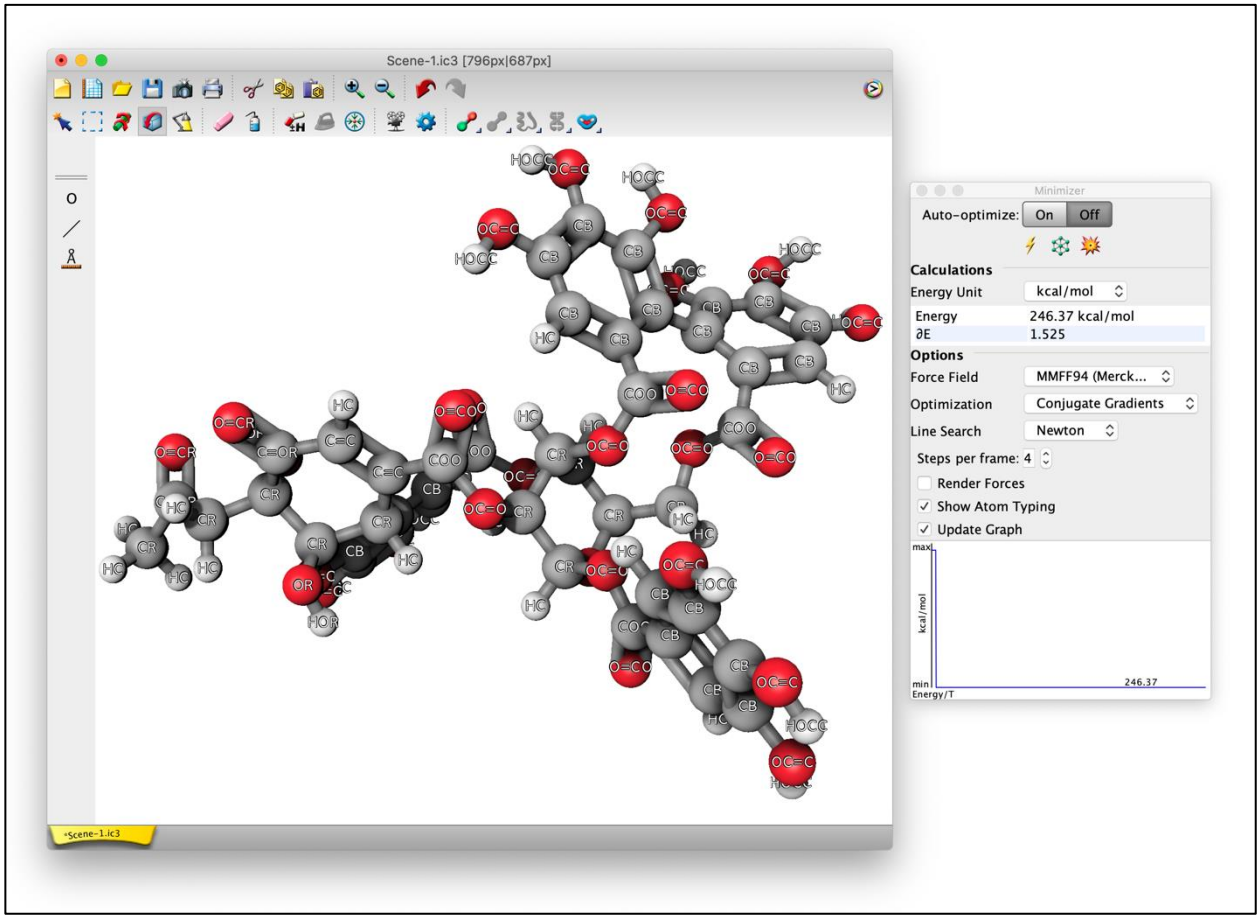


Figure 3: An optimized structure of a stereoisomer of Precoxin A in ChemDoodle 3D with MMFF94 atom types displayed

Catalyst Science Discovery Centre & Museum Trust in Widnes: A Year in the Life of RSC-CICAG's Treasurer

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

At the beginning of 2020 I had recently started my second year of being the Chair of Trustees at the Catalyst Science Discovery Centre and Museum in Widnes, Cheshire and was looking forward to building on the work that had been carried out in my first year when we had successfully bid for and won one of the UKRI/Wellcome Trust sums of money to revamp some of the very dated educational areas of Catalyst, and appointed a very good part-time Chief Executive Officer (CEO). The future looked rosy for Catalyst,which was initially established in 1981 as the Halton Museum of the Chemical Industry to celebrate the centenary of the creation of the Society of Chemical Industry (SCI), and which evolved in 1987 to be the Catalyst Science Discovery Centre and Museum. The last 33 years had not been financially easy for Catalyst, but it was the oldest and longest surviving of the UK Science Discovery Centres (SDCs), and the Trustees were determined that the excellent educational work it was doing in the chemical sciences and the role it had as the only UK Museum of the Chemical Industry and guardian of extensive archives of the NW chemical industry would continue.

Quite a few of the UKRI/Wellcome-funded changes had already occurred by the start of 2020 – a new theatre, entrance area, café, and about half of the new scientific interactives designed by a company called Whitefire had been installed. We also had the new and unique interactive Periodic Table, partially funded by RSC and RSC-CICAG money, which families loved. Our 30 year old glass lift had been overhauled. We had had to close for a few weeks in January 2020 to carry out some of this work, so we decided to reopen to public visitors for the February 2020 half term week. It was a roaring success as everyone saw the changes we had already made, loved them, and the financial coffers were boosted. What a change a few weeks/days make. A new infective virus, COVID-19, had arrived in the UK and we were placed under lockdown. Like all other SDCs and museums, we had to close our doors to the public, put all our staff, except for the CEO and our 1 day a week collections archivist, on furlough from the 20th March,and with one or two minor exceptions that is still the situation today, 11th January 2021, and ten months on, as I write.

What has happened in the meantime? Firstly we have survived so far, despite having no normal visitor revenue, as a result of paring back all our costs to the bone, and shutting down non-essential services, most staff salaries coming from government furlough monies, bidding for any national and regional funding that was available, and receiving donations from well-wishers. Behind the scenes the installation work of the remaining scientific interactives which had already been funded by Wellcome has occurred, work with groups of people with mental health problems has carried on using science as therapy, internal building work changes have taken place to improve working conditions in the future, many areas have been painted, and a new internet system has been installed. All this work has been managed and overseen by the CEO. We experimented in the summer with a take-away café, and tried booking visitors in when we were allowed to, but each time new restrictions were introduced we had to go back to complete closure.

The fact that we are both an accredited museum and an SDC has been critical to our way forward, as we made a bid for, and were successful, in getting a sum of money from the Arts Council England Cultural Recovery Fund to help us through to March 2021. This was not open to SDCs, but only to museums and we were the only science museum to receive this money. We had hoped to open again before Christmas but that became

impossible; then we hoped for 4th January 2021, but that date then moved to 3rd/4th February, and now the earliest date is likely to be Easter 2021, depending on the rollout of the vaccination programme. What we will have waiting for our visitors when we can open and they feel confident to visit is a newly equipped gallery with lots of exciting interactives, a wonderful entrance area and café, new exhibits, and a COVID-secure environment.

Administratively over the last 10 months as Trustees we have had 8 Board meetings, 6 Extraordinary General Meetings and 3 Annual General meetings as we had to take urgent financial and personnel decisions, worked through financial and personnel Subgroups, wrote funding bids, kept abreast, sometimes on a daily basis, of every change in government and charity policy, and all done by virtual meetings using Microsoft Teams software. Naturally, we could not use the staff at all as they were and are still on furlough. So the time expended has been tremendous but necessary. Next week we are bidding for more money from the ACE CRF fund to see us through to July 2021, as that may be necessary given the latest predictions on the spread of the virus and potential further lockdowns. We will not be ceasing our efforts as the work we do in STEM and spreading the word about the chemical sciences at Catalyst is essential. The value of scientists has been proved in the last year, and let us hope that that continues at a national level and they are respected.

The 6th Tony Kent Strix Annual Memorial Lecture 2020

Contribution from David Ball, email davidball1611@gmail.com

The 6th Tony Kent Strix Annual Memorial Lecture 2020, hosted by the UK electronic information Group (UKeiG), took place on Thursday, 26th November 2020, and was a great success. It was delivered by the 2019 Strix award winner Professor Ingemar J. Cox, Department of Computer Science at University College London. Martin White opened the event with a presentation on enterprise search. Recordings and slides from both presentations are now available to colleagues who were unable to attend the event.

Professor Ingemar J. Cox is Head of UCL's Media Futures Research Group and a Professor at the Department of Computer Science at the University of Copenhagen. His current research interests include information retrieval and data analytics of online social media, Twitter and query logs.

Professor Cox's 2020 Strix lecture was entitled: '*Analysing Web searches for public good: inferring the health of populations and individuals*.'

Abstract: Traditionally, information retrieval seeks to find relevant content (e.g. documents, images, videos) in response to a user's query. However, a user's queries also reveal information about the user. This information can be used for many purposes,including personalising Web search results, and facilitating personalised/targeted advertising. In this talk, I will describe ongoing work we are conducting that analyses users' Web searches to infer the health of populations and individuals. At the population level, we discuss methods to estimate prevalence and virulence of a disease, and effectiveness of national public health interventions (vaccines and changes to law). Our work has mainly focused on influenza, but we will also report recent work on covid-19.At an individual level, we discuss methods to stratify users into disease risk groups, and to predict the likelihood of specific diseases, including some forms of cancer. These methods have a variety of advantages for public health surveillance and individual health. However, these same methods raise significant privacy and ethical concerns. We discuss technical solutions to address privacy at a population level and highlight concerns at an individual level.

The video is available here. A PDF of the slides is available here.

Martin White's opening presentation was entitled '*Defining the enterprise search experience*.'

Abstract: There is a great deal of interest at present around the topic of the digital employee experience (DEX). Enterprise search is a very important discovery and integration platform and yet little attention (and no research) has been paid to the user experience of enterprise search. In this presentation Martin White will present his view of what the elements should be of the enterprise search experience based on projects he has undertaken in the last two decades.

Martin White is Managing Director of Intranet Focus Ltd and a Visiting Professor at the Information School, University of Sheffield. He started using online search services in 1973 and set up a specialist search business in 1979. He is the author of four books on enterprise search. Martin is a Fellow of both the Royal Society of Chemistry and of the British Computer Society, and worked with Dr Tony Kent when he was at UK Chemical Information Service.

The video is avaiolable here. A PDF of the slides is available here.

The Tony Kent Strix Award was inaugurated in 1998 by the Institute of Information Scientists. It is now presented by UKeiG in partnership with the International Society for Knowledge Organisation UK (ISKO UK), the Royal Society of Chemistry Chemical Informationand Computer Applications Group (RSC CICAG) and the British Computer Society Information Retrieval Specialist Group (BCS IRSG). The Award is given in recognition of an outstanding practical innovation or achievement in the field of information retrieval.

The 2020 Strix award winner will be announced shortly. For more information about the Tony Kent Strix Award please refer to:
https://www.cilip.org.uk/members/group_content_view.asp?group=201314&id=712682

Open Chemical Science Meetings and Workshops 9-13 Nov 2020 - Introduction

Introduction from CICAG Committee member Dr Helen Cooke, email: helen.cooke100@gmail.com

Open Chemical Science: Reports from the online meetings and workshops, 9-13 November 2020

Globally, significant progress has been made towards open sharing of information. This drive to share data, software, and publications (and thereby knowledge) is enabling the development of a more inclusive society. However, the myriad of solutions, repositories and information sources is at the same time complicating the landscape. So, in mid-2019, CICAG started to discuss the possibility of running a three-day meeting at Burlington House to examine the benefits, risks and likely future developments associated with the rise of open chemistry. The three themes would be Open Access Publishing, Open Data and Open-Source Software, with the first two being delivered as a series of talks, discussions and possibly posters. The Open-Source Software component would have a practical approach and be addressed through a series of workshops.

Just as our plans were coming to fruition, Covid-19 arrived, and we knew immediately that we would have to develop contingency plans in case the pandemic hadn't subsided by November. The meeting's organising committee decided to plan for a series of online events in parallel with planning face-to-face meetings, but it soon became apparent that online was the only option. We decided to schedule the events over a five-day period rather than three days, partly to break up the days into two-hour sessions to prevent 'online meeting fatigue; and also to enable increased attendance at the software workshops which previously would have overlapped in the schedule.

There were pros and cons to holding the event online. On the one hand we were able to invite more overseas speakers and session chairs, and the attendees were also from a wider range of countries than would have been possible for an in-person meeting. The events were free of charge, also assisting with attendance numbers. The downside was that the interactivity between delegates was reduced; we also found that attendance to all the sessions was only about 50% of those who had booked.

We were very pleased that Dr Wendy Warr agreed to write reports for the Open Access Publishing and Open Data sessions, while Dr Chris Swain has written up the Open-Source Software workshops.

Open Access Publishing for Chemistry – Meeting Reports

Meeting report provided by Dr Wendy Warr, email: wendy@warr.com

Open Chemical Science: Reports from the online meetings and workshops, 9-13 November 2020

Keynote: Enabling Open Science
Martin Hicks, Scientific Director, Beilstein Institut

Many reports and analyses conclude that making science more open would be beneficial to society as a whole. Core values include accessibility, reproducibility, inclusivity, transparency, equity, and diversity. A benefit is the ability to carry out text and data mining. Open science also impacts innovation. Martin gave examples from some of the institutional drivers:

- The European Open Science Cloud (EOSC)
- the European Union (EU) Horizon 2020 programme
- National Institutes of Health (NIH) mandates
- the United Nations Educational, Scientific and Cultural Organization (UNESCO) Recommendation on Open Science
- the Scholarly Publishing and Academic Resources Coalition (SPARC) "setting the default to open"
- the Open Research Funders Group (ORFG)
- the Research Data Alliance (RDA) for data sharing.

Martin stressed that open science is not just about content but also about collectivism, in that it will improve equity and inclusiveness for those less well off.

If open science has such impact and benefits why do not more researchers practise it? ORFG has listed some of the excuses. Scientists claim that making data open is an additional hassle; open access (OA) limits freedom to publish in a journal of choice; OA is expensive; "my" data will have limited or no value to others; there is no place to deposit the research data; and other workers will scoop and steal scientists' intellectual property (IP) if the research is made open.

One issue to be addressed is technical infrastructure. Digital data starts and ends with automation, and automation starts and ends with digital data. The laboratory has not changed a great deal over the years, although there have been some interesting recent developments in automation such as the chemputer, DigitalGlassware, and ElectraSyn. A major difference between chemistry and other scientific disciplines is that chemists continuously create new research entities (molecules). The language of chemistry is idiomatic, fuzzy, incomplete, and sometimes redundant. The future demands of AI and machine learning in chemistry will require not only that data collection and storage be revolutionised but also that bonding models be extended rationally and carefully.[1]

Another open science issue is the hypercompetitive research environment. Research evaluation should be holistic and not done by metrics. The work-life balance should tip towards more life and less work. Universities should create knowledge not patents. Canadian experience suggests that, despite a promise to universities that patents would triple commercial revenues, total commercialisation revenue only rose 28% over the period 2001-2009, while expenditure almost doubled. Scientific articles and data should be exempt from copyright. Publishing should involve preprints and open data not Impact Factors (IF) and transformative deals; primary journals can become overlay journals.

Grant applications should be selected for funding by lottery not by committees. In a podcast, Michael Lauer of NIH has discussed the "Powerball Revolution". It has been shown that there is no strong correlation between scoring by a panel of peers and grant productivity. In the Powerball approach to grant evaluation, applications are triaged to find those without sufficient merit and a random selection of funded proposals in made. This approach removes bias, is transparent, and saves time and money. Rejection is not a personal blow and researchers with meritorious but unfunded applications can reapply.

Michael Sandel argues that to overcome the crises that are upending our world, we must rethink the attitudes toward success and failure that have accompanied globalisation and rising inequality. He offers an alternative way of thinking about success, more attentive to the role of luck in human affairs, more conducive to an ethic of humility and solidarity, and more affirming of the dignity of work.[2]

To encourage open science, we need to change the incentives and evaluation criteria; create exceptions for scientific research in copyright and database laws; define standards to foster research integrity; define metadata and data standards; and digitise laboratories and workflows. Citing some thoughts on collectivism by the football manager Arsène Wenger, Martin concluded that creative individuals in both science and sports teams can achieve more by working collectively.

Open Access for UK Research and Innovation: Emerging Policy for 2021
Rachel Bruce, Head of Open Science, UK Research and Innovation

UK Research and Innovation (UKRI) is a non-departmental public body sponsored by the UK Department for Business, Energy and Industrial Strategy. It brings together the seven disciplinary research councils, the Innovate UK agency, and Research England, which supports research and knowledge exchange at higher education institutions in England.

Openness is the foundation for an outstanding research and innovation system that gives everyone the opportunity to contribute and to benefit. UKRI prioritises open access to publications, open research data, and an open culture, with a focus on reward, recognition and incentives for a more inclusive and diverse research system. The COVID-19 pandemic has shown all of us the vital importance of science and innovation. The UK Research and Development Roadmap, 2020 acknowledges that we must embrace the potential of open research practices.

Working nationally and internationally is important to achieve open research. Plan S is an initiative for Open Access (OA) publishing that was launched in September 2018. The plan is supported by cOAlition S, an international consortium of 26 research funding organisations. Plan S requires that, from 2021, scientific publications that result from research funded by public grants be published immediately in compliant OA journals, platforms or repositories. UKRI is a member of cOAlition S.

UKRI is carrying out an OA review to determine a single OA policy across the organisation. It is also working with UK higher education funding councils to help inform the open access policy for the Research Excellence Framework (REF) after REF 2021. The 2018-2021 review is now in its fourth phase with the aim of announcing the new policy in 2021.

UKRI proposes that after 1st January 2022 research articles must be OA immediately, *via* a journal, a publishing platform or a repository with a CC BY licence. UKRI is also considering further options such as other licences as an exception, technical standards for publishing platforms, authors retaining copyright, and public value for money. For monographs, book chapters and edited collections, UKRI proposes OA applying from January 2024 with an embargo of up to 12 months, and a CC BY-ND licence is allowed, but not preferred. Alternative routes to OA publication (e.g., crowdfunding and freemium models) are also allowable. Some further issues such as exceptions, technical standards, copyright, and costs are being considered. Wider implications of the OA policy are also being considered, for example, the impact on low and middle income countries.

UKRI is also modernising research assessment to support necessary changes to recognise all contributions to research. UKRI is a signatory to the Declaration on Research Assessment (DORA) which has been implemented across the seven research councils relating to guidance for peer reviewers of grant applicants. It is also applied in the current REF. UKRI has piloted narrative CV approaches in grant applications and is

working, collaboratively and internationally, to implement best practice in research assessment, including work with cOAlition S and the Global Research Council.

How open are Chemists? An Academic Librarian's Perspective
Clair Castle, Librarian, Department of Chemistry, University of Cambridge

OA publishing has been relatively slow to take hold in chemistry in comparison with other disciplines. Only three of the top 50 chemistry journals are OA (*Nature Communications*, *ACS Central Science*, and *Chemical Science*) according to SCImagoJR. Data from the Web of Science show that the proportion of chemistry articles published OA[3] has increased from 12% in 2010 to 26% in 2018, while according to the SJR World Report, the proportion of OA output has risen from 15.09% in 2010 to 31.77% in 2018. According to the Directory of Open Access Journals (DOAJ), of 15,231 OA journals across all disciplines, only 142 are in chemistry. The Open Science Monitor ranks by discipline the percentage of OA publications in science and technology. Chemistry is placed fourth from the bottom. In contrast, medical, clinical, biological sciences are all near the top. So why is it that chemists are not as open as researchers in other disciplines?

If as a researcher you publish OA, you will get more exposure for your work; practitioners can apply your findings; you may well get higher citation rates; your research can influence policy; the public can access your findings; you are complying with grant rules; taxpayers get value for money; and researchers in developing countries can see your work.

There are other incentives for publishing OA. Funder policies have effectively mandated compliance.[4] In addition, government policies have acted as an incentive. For example, the UK Research Councils' Policy on Open Access requires that funded research must be publicly available *via* a repository within 6-12 months, and REF 2021 will make only OA research outputs eligible. Publisher policies will now often allow the deposit of published versions of manuscripts in non-commercial repositories and on academic websites using CC BY licences. Many universities have developed OA, open data, and open research policies. The OA citation advantage, however, is not universally proven.[5]

On the other hand, there are some barriers to OA publishing. OA can be an administrative burden for researchers and librarians. The infrastructure for OA publishing at universities is not always well developed. The benefits of immediate open access upon publication also have a financial implication. There is concern from researchers over the compatibility of OA and commercial funding. It has also been found[6] that many researchers strongly object to their work being reused for commercial gain, and authors are concerned that journals will not publish until the IP situation has been resolved. Some authors are also worried that their freedom to publish where they choose is diminished if they are restricted to OA publishing. Finally, there are also concerns over lack of peer review on some OA platforms.

Researchers have had to acquire new skills in open science for which training opportunities are not yet widely offered. Moreover, in one study, researchers said that actual practice in these skills is better than doing training courses. Researchers may be more willing to learn from their peers while not feeling that they are taking any extra time away from their research. An EU report describes open science skills for library staff as well as researchers.
Finally, Clair considered future developments. There has been an increase in the number of preprints in chemistry and this is an upward trend.[7] Two main preprint servers in chemistry are ChemRxiv (see below) and Elsevier's ChemRN. Concerns about publishing on preprint servers are similar to those for publishing open access, but preprints do offer faster dissemination of publications.

The radical and controversial Plan S from cOAlition S states that from 2021, all scientific papers must be published in OA journals, or on other OA platforms, without embargo. Many libraries have OA publishing agreements ("transformative" or "read and publish") in place with publishers which mean that authors themselves often will not have to worry about paying article processing charges.

Last but not least is the increase in the number of sources of OA content, and the availability of tools to find them. Plugins are available for various browsers and OA content is flagged everywhere. Directories include DOAJ, the Directory of Open Access Repositories (OpenDOAR) and the Registry of Open Access Repositories (ROAR). Journal table of contents apps that allow filtering for OA are more prevalent now, as are tools such

as Dimensions and Symplectic for analysing OA compliance and impact. Sherpa Fact is a compliance tool for funders and authors. Open access repositories can be used for text and data mining.

Universities need to promote a culture of open research. They need to explore rewarding researchers for being open, perhaps by linking this practice to promotion and tenure policy. Culture change has to be part of a strategy that progressively makes OA possible, by providing adequate infrastructure; making it easy with a good user interface; making it normal behaviour in research communities; making it rewarding; and only then making it a requirement through policy.

Accelerating Research with Preprints
Marshall Brennan, Publishing Manager of ChemRxiv, American Chemical Society

ChemRxiv was launched as a "beta" service in August 2017 by the American Chemical Society with guidance from the German Chemical Society (GDCh) and the Royal Society of Chemistry, who, with the Chemical Society of Japan and the Chinese Chemical Society, joined as official partners in March 2018. A ChemRxiv scientific advisory board was formed in December 2018 and it is being expanded.

A preprint is a draft manuscript that has not yet been accepted for formal publication following peer review. Preprints received by ChemRxiv are checked for plagiarism and signs of abuse and the metadata are checked before the preprint is posted within two working days. Preprints and journals have complementary strengths: even a fast journal cannot publish within the week; even a few weeks is a significant delay for a postdoc seeking career advancement. Preprint servers decouple dissemination from validation by peer review. Preprints are a key part of the research life cycle.

More than 6,500 preprints from over 80 countries have been posted on ChemRxiv and they have been read more than 17 million times. Preprints have been written by hundreds of well-respected chemists and more than half of these authors have submitted more than one preprint. More than 3,200 preprints have later been published in journals. Top journal destinations for preprints are the *Journal of the American Chemical Society*, *Angewandte Chemie International Edition*, the *Journal of Physical Chemistry*, *Chemical Science*, and *ACS Central Science*. The most popular subject matter is computational chemistry. Materials science and organic chemistry follow.

You can submit your preprint in any file format. ChemRxiv preserves the original file's information and formatting. Once the preprint is posted, readers can view it through the web browser viewing tool or download the preprint in the original format. Many formats can be displayed natively in the browser. A date-stamp and citable DOI, usage metrics and altmetrics are displayed for each preprint, together with a link to the published version, if there is one.

New features since the initial launch are RSS feeds, export of curated PDFs, improved search capabilities, a citation format generator, social media sharing functions, and automated virus and plagiarism scans. There is a simple, streamlined interface, where file submission is drag and drop and everything required can fit on one screen on most devices. This is designed so that users can submit in as little time as possible.

Direct journal transfer is now being introduced: the chemist will submit a paper once to ChemRxiv, and then select from a choice of journals and transfer the article directly to the journal of choice. This facility is already available for ACS, RSC and GDCh journals and it will be extended to other journals in the near future. Note also that ChemRxiv is completely free to use.

Pros and Cons of Open Peer Review
Tony Ross-Hellauer Leader, Open and Reproducible Research Group, Graz University of Technology

Peer review is an inherently fallible process but is nonetheless usually considered the gold standard for assuring quality in the literature. Peer review is generally anonymous, opaque, and selective: it is not compatible with open science. It takes too long, lacks accountability, introduces bias, lacks incentives, and wastes effort. Tony undertook a systematic review of what *open* peer review (OPR) is. He created a corpus of 122 definitions, and proposed ways that peer review models can be adapted in line with the aims of open

science, including making reviewer and author identities open, publishing review reports and enabling greater participation in the peer review process.[8]

If authors and reviewers are aware of each other's identities, accountability and quality are increased; conflicts of interest might be avoided; and the language of interchanges could be more civil. On the other hand, without protection of anonymity, reviewers might blunt their opinions for fear of reprisals. Moreover, "blind" peer review potentially protects reviewers from social biases, and "double blind" peer review protects authors as well.

Open peer review reports are published alongside the relevant work. They contain valuable contextual information; allow wider scrutiny; perhaps increase review quality; enable credit and reward for review work; and help train young researchers in peer reviewing. On the other hand, open reports can lead to higher refusal rates amongst potential reviewers, and it takes longer to write a review. Some reviewers may fear undesirable exposure to criticism.

In OPR, the wider community is able to contribute to the review process and cross-disciplinary dialogue is supported. Cons are the difficulty of motivating self-selecting commentators; the tendency of these reviewers to leave more superficial responses; and the potential for adding "noise" to the discussion.

In 2017, Tony and co-workers carried out a survey on attitudes to OPR.[9] They found that OPR is already in the mainstream: 76.2% of respondents have practical experience and 60% believe OPR should be common practice. Reactions to most OPR traits (especially open interaction, reports and participation) were positive but 47.7% rejected open identities. Tony and his colleagues have also investigated the peer review and preprint policies of 171 major journals across disciplines.[10] Of the journals surveyed, 31.6% do not provide information on the type of peer review used; information on whether preprints can be posted is unclear in 39.2% of journals; and 58.5% offer no clear information on whether reviewer identities are revealed to authors. Around 75% of journals have no clear policy on co-reviewing, citation of preprints, and publication of reviewer identities. Less than 20% provide information on OPR practices. Transpose is a database of journal policies on peer review, co-reviewing and preprinting.

Another strand of Tony's research analysed the limitations in our understanding of peer review.[11] Open questions concern the OPR services which researchers prefer; measures that can incentivise OPR; researcher attitudes towards OPR; the impact of OPR on participant diversity; the impact of blinding on biases and review quality; and the impact of open review reports.
General advice for editors and publishers on the implementation of OPR[12] is to set goal(s); listen to research communities; plan technologies and costs; be pragmatic in the approach; further communicate the concept; and evaluate performance. Editors and publishers should devise strategies to compensate for the possibility that open identities might make it harder to find reviewers; be alert to possible negative interactions and have a workflow for dealing with them; enable credit; and consider piloting or making open identities optional. They should meet industry best-practice for publishing review reports and be aware of potential challenges in publishing them.

A Tale of two Societies: Are Differences in Open Access Policy Driving a split in UK and US Chemistry Publishing?
Cameron Neylon, Professor of Research Communications, Centre for Culture and Technology, Curtin University

The Royal Society of Chemistry (RSC) and the American Chemical Society (ACS) have very different histories but in the 1990s and 2000s their publishing operations had strong parallels. From 2010 on, each responded differently as funders and policy makers took different approaches to OA. Gold OA means immediate access to an article in an online journal; Green open access involves publishing in a traditional journal, but articles are also "self-archived" in a repository. The Budapest Open Access Initiative was launched in 2002. Cameron listed some UK milestones after that:

- July 2012 RSC "Gold for Gold"
- April 2013 UK Research Councils' Policy on Open Access

- 2014 Research Excellence Framework
- 2015 Higher Education Funding Council for England (HEFCE) mandate
- January 2017 *RSC Advances* goes open access
- 2018 RSC read and publish initiative

Some US milestones are as follows:

- 2008 NIH open access policy
- 2013 Office of Science and Technology Policy (OSTP) mandate
- 2015 *ACS Central Science* launched
- July 2016 First issue of *ACS Omega* launched
- 2020 JACS Au announced.

The Curtin Open Knowledge Initiative (COKI) has studied 1,207 institutions worldwide and found that in 2017 the top-performing universities published around 80–90% of their research as open access.[13] They have also produced a number of research funding dashboards plotting adoption of OA over time. Cameron compared ACS and RSC over time in the UK (Figure 1).

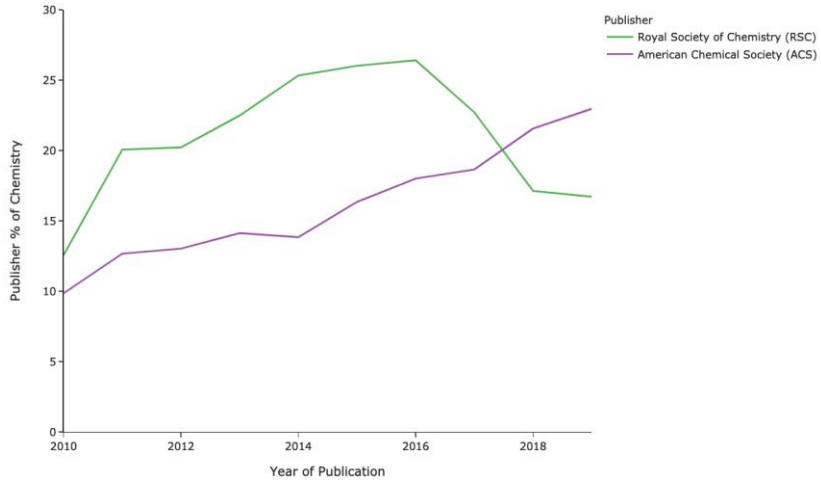


Figure 1. Choice of publisher in the UK.

In the Unites States, the curves are much flatter with ACS scoring 21-27% and the RSC rising from 6% in 2010 to about 16% in 2015, then falling to about 10% from 2018. The proportion of OA publishing is perhaps not changing very much (Figure 2). It seems as if OA policies are not driving choice of publisher. ACS had a big uptake of OA when researchers received one free OA article if they published an article in an ACS journal. The greater adoption of OA in the UK could be because the mandates come from across the disciplines and the government, not from a funder focused on one discipline such as NIH.

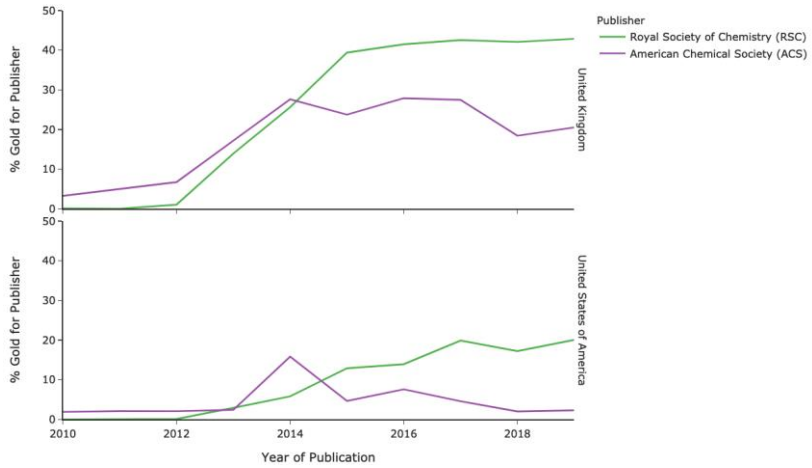


Figure 2. Proportion of OA publishing.

Cameron next presented an institutional analysis (including pharmaceutical companies). Swansea, Bath and Hull universities are outliers with a bias towards RSC in 2018. Oxford and Cambridge tend somewhat towards ACS. In the Netherlands, Eindhoven and Delft have a strong bias towards ACS in 2019, and MIT in the United States is more for RSC than for ACS with a significant shift after a transformative deal struck in 2018. National averages do not tell the full picture. Specific institutions show very different and quite specific patterns. Cameron drew two other conclusions. RSC took a significant lead in early open access provision for chemistry, particularly in the UK, but has fallen back. Despite the fact that national polices in the UK have favoured Gold OA, they have been delivered on *via* Green OA. Chemistry has lagged on open access for most of the past twenty years but is showing some signs of a shift now, particularly in the UK.

The Growth in Importance of Open Access Sources in the Pharmaceutical Industry
Helen Malone Senior Director, External Data Lead, Data & Computational Science, GlaxoSmithKline

The exponential increase in global information and data offers challenges and opportunities. Pharma companies have invested in advanced analytical capabilities such as AI and are considering all types of data including OA sources, but the data must be reliable and of high quality and value.

OA data is of value in all phases of the drug development process. In early drug discovery, ChEMBL is vital because it has both chemical structures and biological data. In Preclinical, National Center for Biotechnology Information (NCBI) resources, including PubMed and PubMed Central are used. Later on, ClinicalTrials.gov is used and even after the drug is on the market there are useful OA literature resources.

Comprehensive sources of evidence to answer critical business questions include internal company information, information from sources subscribed to by the company, and OA sources. Integration of OA information with internal and other external information, followed by harmonisation, adds value. Text and data mining adds extra value (and cost). It is important to understand the terms of use of OA data. A CC BY licence allows reuse for commercial purposes, but if there is no CC BY licence, negotiation may be required. Time will be needed to gain permissions and there might be an extra cost for text and data mining.

After a steady increase in the importance of OA sources, there has been massive acceleration as the world is devastated by COVID-19. António Guterres, UN Secretary General, has stated that science matters, cooperation matters, and misinformation kills. There is now much more collaboration and openness between academia and pharma, and between pharma companies. STM publishers have also been quick to react with free access to learned articles. The Copyright Clearance Center is providing access to relevant resources, and links to data visualisation, dashboards, and datasets. The Semantic Scholar team at the Allen Institute for AI has partnered with leading research groups to provide a search interface to CORD-19, a free, large dataset of scholarly articles about the coronavirus. Preprints are increasingly important for COVID-19 research. ChemRxiv is becoming an essential resource for chemists (e.g., in repurposing of drugs). Preprints in general can be used at all stages of the drug discovery process.

In pharma, a balance has to be achieved between paid and OA resources, but OA is an increasingly important part of the information landscape.

Open Science in an Open Access World
Egon Willighagen, Team Leader, Department of Bioinformatics, Maastricht University and Editor-in-Chief, *Journal of Cheminformatics*

The current editors of the *Journal of Cheminformatics* have published several editorials on improving the practice of cheminformatics.[14-18] The journal is an OA one but open access is not the same as open science. Egon showed some Venn diagrams of "open", "publish" and "findable, accessible, interoperable and reusable" (FAIR).[19] The overlap of "open" and "publish" allows distribution, for example by the Open Journal Systems free software created by the Public Knowledge Project. Google Scholar links "publish" and "FAIR", but it is not open. (Not all FAIR data are open.) Reuse of material demands the overlap of all three themes.

Open science concerns efficient and open sharing of research output by allowing others to reuse, modify, and redistribute that research. If a journal is to further open science, it must recognise and reward all research outputs, not just articles. There is a move away from Impact Factor (IF) as a measure. New technologies have appeared: owners of datasets and software etc. can be rewarded on repositories (e.g., GitHub or Zenodo). The Journal of Open Source Software rewards both the software and the article; reviewers have to use the software and focus on the research output.

Journals need to improve knowledge dissemination. Some articles have more figures than text[20], and the figures are not easily machine readable. Supporting information is usually rendered useless when converted to PDF. In response to a recent issue with finding data (involving DataCite), Egon agreed that the "availability" section needs to support all resource output clearly. Putting additional files on the Figshare repository is FAIRer, and several publishers support this.

The first effort by the *Journal of Cheminformatics* to support open science was to start a Twitter account as an alternative to the RSS feed of new articles.[16] The second was to start encouraging more authors to provide their ORCID identifiers.[21] The corresponding author is asked to provide his or her ORCID identifier and the editorial board page has been updated with ORCID identifiers for the board members. Starting in 2020, ORCID identifiers will be required for *all* authors in a paper.[16] Egon showed some interesting plots that can be produced using data for Barbara Zdrazil (an associate editor of the journal) using Scholia, ORCID and Wikidata. A further change made recently was starting a GitHub organisation for the journal.

The Initiative for Open Citations (I4OC) promotes the unrestricted availability of scholarly citation data. Although the Impact Factor of the *Journal of Cheminformatics* is high, the number of times an article is cited is low (eight or less). Reuse is a much more interesting measure than IF. Networks of cited/citing links can be built, but the reason for citing (as an authority, as a method etc.) is not always clear. Thus use of the Citation Typing Ontology CiTO annotation is being tested in the journal.[18] CiTO formalises a hierarchy of reasons that readers have for citing an article. CiTO is harvested by Wikidata and interesting analyses can be carried out.

An easy way of using OA to promote open science is manuscript checklists (critical appraisal tools used when reading research). Figshare is useful but more details such as SDfiles[22] or FAIR biological pathways are needed. Linked open chemical data[23] (e.g., with SciGraph) is harder to achieve. Harder still is the Semantic Web for publishers (e.g., Open PHACTS), but this is the way ahead.

Open Access: A Society Publisher's Perspective
Dr Neil Hammond, Publisher for Open Access Journals, Royal Society of Chemistry

The RSC publishes books, magazines, databases, and 46 journals (seven of them Gold and the rest hybrid). RSC publishing fulfils a primary purpose to share knowledge, but also generates revenues to support its other charitable activities. RSC supports OA publishing for many reasons: openness to research outcomes is a public good; facilitating maximum access to research aligns with RSC's purpose; OA prioritises the relationship with the author (an RSC strength); and the community wants RSC to support openness. In a survey of more than

1000 scientists worldwide, 50-64% (depending on the country) wanted societies and publishers to do more in support of OA publishing.

According to data from Science Citation Index Expanded, RSC's Gold OA share (28% in 2019), is greater than that of other chemistry publishers combined (23%). In 2012 "Gold for Gold" was introduced; in 2015 *Chemical Science* became OA; in 2016 the read and publish model was launched and *RSC Advances* became OA; and in 2018 RSC became a partner in ChemRxiv and struck the society's first North American read and publish deal (with MIT). RSC has such deals with 100 institutes to date. The impact can be seen in Table 1, where the OA profile in RSC journals for those countries with many such deals is contrasted with that of China, which has no such deals.

Table 1. RSC Open Access Publication

Country	Percentage of articles that are OA	Percentage of the OA articles in a hybrid journal	Percentage of the OA articles in *RSC Advances*	Percentage of the OA articles in *Chemical Science*
Finland	100	84	13	0
Sweden	87	78	15	4
Netherlands	85	87	4	9
China	18	1	83	11

Amongst advocates of open access publishing, publishers with a significant subscription journal business are often characterised as a barrier to change. Whilst this may be true in some cases, the situation for a mission-driven, non-profit publisher, such as the RSC, is complex, with competing incentives and pressures from various directions. Green OA does not completely solve various needs. Moreover, within the chemistry community there is a perception that OA means low quality and there is a relative lack of OA enthusiasm and advocacy. There is also geographical inequality in access to funding and divergence in funder and institutional policies. In late 2016 RSC moved *RSC Advances* to a Gold OA model with an article processing charge of £500. The average number of submissions fell by about a third, but the fall was more pronounced for some countries (the number of publications from India fell from 14% to 5%). Strong regional differences can be seen across the top five countries producing chemistry papers in 2020: China (18% OA), the United States (17% OA), India (12% OA), Germany (44% OA) and Japan (23% OA).

RSC is responding to the challenges of OA by launching and promoting high quality OA chemistry journals, promoting the benefits of OA to the community, and exploring broader open science initiatives such as transparent peer review. The society continues to develop transformative business models, to provide choice for authors, to invest in author services, and to listen to, and respond to, the community.

References

(1)Clark, T.; Hicks, M. G. Models of necessity. *Beilstein J. Org. Chem.* **2020,** *16*, 1649-1661.
(2)Sandel, M. J. *The Tyranny of Merit: What's Become of the Common Good?*; Macmillan: New York, NY, 2020.
(3)Novara, F. R. A Big Year for Open Access Chemistry Publishing. *ChemistryOpen* **2020,** *9* (1), 4-7.
(4)Lariviere, V.; Sugimoto, C. R. Do authors comply when funders enforce open access to research? *Nature* **2018,** *562* (7728), 483-486.
(5)Piwowar, H.; Priem, J.; Larivière, V.; Alperin, J. P.; Matthias, L.; Norlander, B.; Farley, A.; West, J., *et al.* The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles. *PeerJ* **2018,** *6*, e4375.
(6)Rowley, J.; Johnson, F.; Sbaffi, L.; Frass, W.; Devine, E. Academics' behaviors and attitudes towards open access publishing in scholarly journals. *J. Assoc. Inf. Sci. Technol.* **2017,** *68* (5), 1201-1211.
(7)Coudert, F.-X. The rise of preprints in chemistry. *Nat. Chem.* **2020,** *12* (6), 499-502.
(8)Ross-Hellauer, T. What is open peer review? A systematic review. *F1000Research* **2017,** *6*, 588.
(9)Ross-Hellauer, T.; Deppe, A.; Schmidt, B. Survey on open peer review: attitudes and experience amongst editors, authors and reviewers. *PLOS ONE* **2017,** *12* (12), e0189311.
(10)Klebel, T.; Reichmann, S.; Polka, J.; McDowell, G.; Penfold, N.; Hindle, S.; Ross-Hellauer, T. Peer review and preprint policies are unclear at most major journals. *PLOS ONE* **2020,** *15* (10), e0239518.

(11)Tennant, J. P.; Ross-Hellauer, T. The limitations to our understanding of peer review. *Research Integrity and Peer Review* **2020,** *5* (1), 6.
(12)Ross-Hellauer, T.; Görögh, E. Guidelines for open peer review implementation. *Research Integrity and Peer Review* **2019,** *4* (1), 4.
(13)Huang, C.-K.; Neylon, C.; Hosking, R.; Montgomery, L.; Wilson, K. S.; Ozaygen, A.; Brookes-Kenworthy, C. Evaluating the impact of open access policies on research institutions. *eLife* **2020,** *9*, e57067.
(14)Guha, R.; Willighagen, E. Helping to improve the practice of cheminformatics. *J. Cheminf.* **2017,** *9* (1), 40.
(15)Guha, R. Implementing cheminformatics. *J. Cheminf.* **2019,** *11* (1), 12.
(16)Willighagen, E.; Jeliazkova, N.; Guha, R. Journal of Cheminformatics, ORCID, and GitHub. *J. Cheminf.* **2019,** *11* (1), 44.
(17)Guha, R.; Willighagen, E. Learning cheminformatics. *J. Cheminf.* **2020,** *12* (1), 4.
(18)Willighagen, E. Adoption of the Citation Typing Ontology by the Journal of Cheminformatics. *J. Cheminf.* **2020,** *12* (1), 47.
(19)Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W., *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016,** *3*, 160018.
(20)Hanspers, K.; Riutta, A.; Summer-Kutmon, M.; Pico, A. R. Pathway information extracted from 25 years of pathway figures. *Genome Biol.* **2020,** *21* (1), 273.
(21)Haak, L. L.; Fenner, M.; Paglione, L.; Pentz, E.; Ratner, H. ORCID: a system to uniquely identify researchers. *Learned Publishing* **2012,** *25* (4), 259-264.
(22)Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992,** *32* (3), 244-255.
(23)Samwald, M.; Jentzsch, A.; Bouton, C.; Kallesøe, C. S.; Willighagen, E.; Hajagos, J.; Marshall, M. S.; Prud'hommeaux, E., *et al.* Linked open drug data for pharmaceutical research and development. *J. Cheminf.* **2011,** *3* (1), 19.

Speaker Biographies

Dr Martin Hicks
Scientific Director, Beilstein Institut

Martin G Hicks is a member of the board of management of the Beilstein-Institut where he serves as scientific director. He received an honours degree and PhD in chemistry from Keele University and completed a postdoctoral position at the University of Wuppertal, where he carried out research on semi-empirical quantum chemical methods.

In 1985, Martin joined the computer department of the Beilstein-Institut where he worked on the Beilstein Database project. His subsequent activities involved the development of cheminformatics tools and products in the areas of substructure searching and reaction databases. He organized the first

Dr Egon Willighagen
Team Leader, Departmernt of Bioinformatics, Maastricht University & Editor-in-Chief, Journal of Cheminformatics (OA)

Egon Willighagen studied chemistry and cheminformatics at the Radboud University in Nijmegen, The Netherlands, with a minor in organic chemistry (lipid aggregation) and a major in data analysis of chemical data. He obtained his PhD at the same university in the field of chemometrics in the Analytical Chemistry group of Prof Lutgarde Buydens. This was followed by various research positions at Cologne University, Wageningen University, Uppsala University, University of Cambridge, Karolinska Institute and

Beilstein Symposium in 1988 and after joining the board of management in 2002 he launched the first Beilstein Open Access journal, Beilstein Journal of Organic Chemistry in 2005.

The Beilstein-Institut is a non-profit foundation based in Frankfurt and fulfils its mission primarily through its own projects involving the dissemination of high-quality scientific information. Open Science is at the heart of the Beilstein-Institut's philosophy. The current core projects revolve around Open Access and Open Data involving developing new publishing paradigms and infrastructures in chemistry and together with the Beilstein Symposia form an important part of the Beilstein-Institut's forward-looking strategy.

Dr Tony Ross-Hellauer

Leader, Open and Reproducible Research Group, Graz University of Technology

Tony Ross-Hellauer is leader of the Open and Reproducible Research Group (ORRG) at TU Graz and Senior Researcher at Know-Center. His research focuses on a range of issues related to open science evaluation, skills, policy, governance, monitoring and infrastructure. Tony has a PhD in Information Studies (University of Glasgow, 2012), as well as degrees in Information and Library Studies and Philosophy. He is coordinator of the EC H2020 project ON-MERRIT researching issues of equity in Open Science and FAIR Data Austria, a major project to advance FAIR data practices at the national level. He is formerly is OpenAIRE Scientific Manager and Editor-in-Chief of the MDPI open access journal 'Publications', is co-author of the Open Science Training Handbook, and co-leads Transpose, a grassroots initiative to build a crowdsourced database of journal policies for preprints and peer review.

Professor Cameron Neylon

Maastricht University, with research in the fields of drug discovery, metabolomics and toxicology. He is now a team leader at the Department of Bioinformatics at Maastricht University and participates as PI in various international projects, with a total funding of over two million euros in the past six years.

His research basically focuses on the question of how to capture chemical and biological knowledge to enable machine learning. He takes advantage of his education in the fields of analytical and organic chemistry. A secondary objective is to enable open science approaches and learning requires data and knowledge to learn from.

Dr Willighagen is recognised for his open science and cheminformatics work, advises in various projects, has been involved in many Open Science projects, such as the Chemistry Development Kit, Jmol, and WikiPathways, is Editor-in-Chief of the Journal of Cheminformatics (one of two), is chair of the Data Management Working Group of the EU NanoSafety Cluster and founding member of the GO FAIR Chemistry Implementation Network.

Rachel Bruce

Head of Open Science, UK Research and Innovation (UKRI)

Rachel Bruce is Head of Open Research at UK research and innovation, she is responsible for the breadth of open research across UKRI but this primarily focuses on open access, open research data and responsible research assessment. She has worked in the area of digital technologies and the support of research for a number of years, responsible for developing a network of repositories for research in the UK, in partnership with UK HEIs and technical innovators, and also the introduction of more open research and learning practices taking advantage of digital technologies. She has advised the European Commission on the European Open Science Cloud and been the UK representative in its Governance, sat on the UK forum for responsible research assessment and is currently a member of G7 Open Science Working Group focused on data infrastructure and incentives. Currently she is leading the UKRI open access policy review.

Clair Castle MCILIP

Professor of Research Communication, Centre for Culture and Technology, Curtin University

Cameron Neylon is Professor of Research Communication at the Centre for Culture and Technology at Curtin University and well-known agitator for opening up the process of research. His current work focusses on how the cultures of research affect and effect change in research communications. He speaks regularly on issues of Open Science including Open Access publication, Open Data, and Open Source as well as the wider technical and social issues of applying the opportunities the internet brings to the practice of science. He was named as a SPARC Innovator in July 2010 for work on the Panton Principles and was a co-author of the Altmetrics manifesto and the Principles for Open Scholarly Infrastructures. He is a proud recipient of the Blue Obelisk for contributions to open data. He writes regularly at his blog, Science in the Open.

Helen Malone

Senior Director, External Data Lead, Data & Computational Science, GlaxoSmithKline

Helen Malone is Senior Director, External Data Lead in the Data & Computational Science (DCS) department at GSK. Helen has driven multiple initiatives to evolve the quality and delivery of scientific information to GSK R&D scientists. In her current role, Helen has developed and is leading an innovative External Data Strategy for DCS. Helen is passionate about leveraging relationships between the external information / data industry and information consumers to generate insights and knowledge for our organisations.

Helen is also the President of the Pharma-Documentation-Ring (P-D-R). The P-D-R is a community of experienced knowledge managers

Librarian, Department of Chemistry, University of Cambridge

© Gabriella Bocchetti, University of Cambridge

Clair graduated from Loughborough University in 1996 with a Joint Honours degree in Information & Library Studies and German and is a Chartered member of CILIP (Chartered Institute of Library and Information Professionals). Clair has worked at the University of Cambridge for over 20 years in various libraries. She has been managing the Department of Chemistry Library service since October 2013, supporting all research and teaching activities that take place at the Department. She has developed a special interest in research data management - which she teaches to all new Chemistry graduate students - and in scholarly communication and open research.

Dr Neil Hammond

Publisher for Open Access Journals, Royal Society of Chemistry

Neil holds a PhD in Nuclear Physics from Liverpool University, and conducted research at Argonne National Lab in the USA, co-authoring more than 50 articles in peer-reviewed journals, before leaving scientific research to begin a career in academic publishing. He has served in various roles for a number of publishers over the last 15+ years and has worked with a diverse range of academic communities and learned societies. He is currently the Publisher for Open Access Journals at the Royal Society of Chemistry, carrying responsibility for the development and growth of the RSC's open access journals programme.

and directors from 26 leading pharmaceutical companies in the world.

Dr Marshall Brennan

Manager of ChemRxiv, American Chemical Society

Marshall received his Bachelor's in Chemistry with honors from Northeastern University in Boston, MA, during which time he carried out research at Harvard University under the supervision of Professor Tobias Ritter. He completed his graduate work in 2015 at the University of Illinois at Urbana-Champaign in the laboratory of Prof. Alison Fout, studying low-valent cobalt complexes featuring strongly-donating bis(silyl)amide ligands and their catalytic reactivity toward C—N bond formation. He then carried postdoctoral research on rhodium-mediated C—C activation methodology before joining *Nature Chemistry* as an editor for nearly two years. In 2017, Marshall was recruited to lead the launch and development of the American Chemical Society's new preprint server, ChemRxiv, where he now oversees day-to-day operations and business development, leading the service to more than 10,000 daily active users and 2,000,000 total downloads in less than two years. In 2018, Marshall was a recipient of the ACS Catalyst Award and a nominee for Forbes Magazine's "30 Under 30 in Science & Technology". Marshall is based out of the Washington, D.C. office.

Open Data for Chemistry – Meeting Reports

Meeting reports provided by Dr Wendy Warr, email: wendy@warr.com

Open Chemical Science: Reports from the online meetings and workshops, 9-13 November 2020

Large-scale bioactivity data for Drug Discovery: Some History and the Future

John Overington, Chief Informatics Officer, Medicines Discovery Catapult

From 2000 until 2008, John worked for Inpharmatica where his team built a large SAR database "Structure Activity Relationships from the Literature" (StARLITe), outsourcing the data entry and performing extensive manual curation and automated indexing in-house.[1] StARLITe was an attempt to "codify" the rules for lead optimisation, and also support target identification and validation. Chemical structures and biosequences were linked and treated with equal emphasis in curation and search capabilities. In 2005 there were 429,000 chemical compounds, 1.4 million biological activities and 3,440 functional molecular targets. There were links to synthetic routes and assay protocols and other resources at Inpharmatica: DrugStore (a database of about 1,500 known drugs, their indications and molecular targets),[2] BioPendium (a proteome annotation resource), and SARfari (integration portals around gene families such as GPCRs and kinases, and also ADME properties).[3]

The concept of compound druggability was popular at that time, with the Rule of Five being ubiquitous in compound design, and Inpharmatica attempted to extend this thinking to a corresponding set of rules for targets. Matched molecular pairs could also be calculated and analysed from the data stored in StARLITe. John showed a screenshot of an early MDL ISIS Base interface to StARLITe. BLAST search of sequences could be carried out avoiding the issues of confusing synonyms in name-based search (e.g., KDR or FLT1). Close and more distant homologues could be located and ranked. Compound and target spaces could be searched bidirectionally. Exact, substructure, and similarity search of chemical structures could also be carried out.

The underlying data structure allowed selectivity profiles to be plotted. Other applications included chemical starting points, SAR datasets, alternative targets for compounds, scaffold hopping, and focused compound sets. John presented a Bemis-Murcko-like analysis finding privileged scaffolds. There was a trade-off against high molecular weight and specificity for a particular target bioactivity and so an "elegance parameter" was devised.[4] StARLITe was also used to suggest bioisosteric replacements and the impact on bioactivity and molecular properties. This work also led to advances in the support for, and scope of, the druggable genome.[5,6] Nowadays, all of these use cases can be attempted using ChEMBL[7-9] with, currently, 2.4 million compound records and 13,377 targets. Moving StARLITe into the public domain was a big step.[1] Inpharmatica was sold to Galapagos and became part of the BioFocus CRO division. In July 2008, BioFocus DPI announced the transfer of its databases to the European Bioinformatics Institute-European Molecular Biology Laboratory (EMBL-EBI), a transaction funded by a £4.7 million award from the Wellcome Trust. In ChEMBL, the databases are now freely and publicly available online to drug discovery researchers worldwide.

Semantic Web for Chemistry: How to use these Technologies Effectively, and what not to do
Samantha Kanza, Enterprise Fellow, University of Southampton

The Semantic Web is the web of linked data. It is a way to bring context and meaning to data, and a set of standards for data representation, integration, and search. The Resource Description Framework (RDF) is its machine-readable linked data format. The atomic data entity in the RDF model is a semantic triple (Figure 1). The subject, predicate and object are often Uniform Resource Identifiers (URIs).

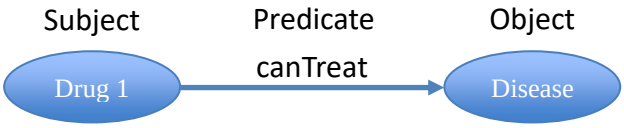


Figure 1. Semantic triple.

An ontology defines different concepts within a domain including hierarchies, relationships to other concepts, and terms used to refer to them. Ontologies are written in Web Ontology Language (OWL) or RDF Schema. Formats for semantic annotation of documents and web pages with metadata are Resource Description Framework in Attributes (RDFa), Javascript Object Notation for Linked Data (JSON-LD) and Microdata (an HTML extension specification). SPARQL is a query language to retrieve data stored in RDF format; it facilitates search on concepts rather than text. Knowledge graphs are graph network structures to describe real world entities and their relationships through the combination of linked data and ontologies.

Samantha discussed affordances of the Semantic Web. Common shared vocabularies, making data machine readable, and producing rich interoperable metadata, make it easier to search web pages and documents and to mark-up text. This can be very useful in electronic lab notebooks (ELNs).[10-12] Scientific research can be enhanced by linking datasets together to find undiscovered links and answer questions that cannot be addressed with a single data source (e.g., in drug discovery).[13] Description logic can be embedded into ontologies, enabling machines to "infer" additional information that is not explicitly defined in the data. For example, if Samantha is allergic to juniper, and gin has botanical juniper, then we can infer that Samantha is allergic to gin. Semantics unlocks the potential of AI and machine learning with high quality data.[14]

Ontologies include RSC Ontologies (the Name Reaction Ontology (RXNO), the Chemical Methods Ontology (CMO) and the Molecular Process Ontology (MPO); the Chemical Information Ontology; Chemical Entities of Biological Interest (ChEBI); the Gene Ontology; and the BioAssay Ontology (BAO). Ontology repositories include OBO Foundry, BioPortal, and EMBL-EBI. Knowledge bases of semantically annotated chemical and biological data include ChemSpider RDF, ChEMBL RDF, Open PHACTS, and DrugBank. There are also many general semantic resources. Dublin Core describes datasets. Simple Knowledge Organization System (SKOS) organises hierarchies. DBPedia is a crowd-sourced effort to extract structured data from Wikipedia. Schema.org is an ontology community.

Samantha summarised some best practices. Reuse ontologies, design patterns, taxonomies, and schemas where possible. Do not forget standards; even standards have standards (e.g., OBO Foundry). Think why you are making ontologies, and what they are for; think through the entire life cycle. Break your ontologies down into smaller, related modules. Data should be interoperable but that is only one of the "findable, accessible, interoperable and reusable" (FAIR) criteria.[15] Play to OWL's strengths. For example, OWL does not work well with numerical-based classifications: it might be best to handle mathematics in a separate algorithm. Consider your data formats. Converting your data into a linked data format can be quite time consuming and tricky and you should think about the final appearance of your linked datasets before you design an ontology, not afterwards. Remember the "garbage in, garbage out" rule: it still applies in data conversion. Note also that the Semantic Web is a sociotechnical phenomenon: it requires human effort just as much as a technological one. Samantha concluded by saying: "To the well-organised, linked dataset, AI is but the next great adventure".

Digital Detective Work: Connecting Cheminformatics, Mass Spectrometry and our Environment *via* Open Data

Emma Schymanski, FNR ATTRACT Fellow and Head of Environmental Cheminformatics, Luxembourg Centre for Systems Biomedicine, University of Luxembourg

The environment of chemicals to which we are exposed is incredibly complex, with around 100 million chemicals in open databases such as PubChem.[16] Detectable molecules in complex samples can now be captured using high resolution mass spectrometry (HRMS), which provides a "snapshot" of all chemicals present in a sample and allows for retrospective data analysis through digital archiving[17,18], but scientists cannot yet identify most of the tens of thousands of features in each sample, leading to bottlenecks in identification and data interpretation. The "exposome" concept[19] strives to capture the diversity and range of exposures to synthetic chemicals, dietary constituents, psychosocial stressors, and physical factors, and their corresponding biological responses.[18]

Emma and her co-workers have developed a toolkit including Shinyscreen to extract and automatically quality-control HRMS data and the NORMAN Suspect List Exchange (NORMAN-SLE) to capture expert knowledge. More than 73 lists of nearly 145,000 substances are now included in the NORMAN SLE, which is also integrated in PubChem. PubChem is used as a large knowledge base to find "known unknowns". Also included is MetFrag for computer-assisted identification of small molecules from mass spectra. MetFrag has recently been enhanced[20] and a connection to MassBank Europe has been added. MetFrag, MassBank Europe, NORMAN SLE, and PubChem are all used to connect various lines of evidence for identification in HRMS experiments.

Wide coverage and high efficiency are needed to address the challenge of the growing number of candidates. Rather than screen over 100 million compounds in PubChem, only the 371,663 relevant, annotated ones are used in the new PubChemLite for Exposomics open dataset. Calculations show that using PubChemLite plus other lines of evidence correctly ranks ~80 % of small molecules in the test set correctly in first place in a fraction of the time, compared with 70 % if full PubChem had been used instead. The searchable NORMAN Suspect List Exchange is being added to PubChem, and gaps in the dataset have been used to add further annotations, to ensure that high quality missing compounds are added.

A flow chart from Jessy Krier's master's thesis in 2020 is shown in Figure 2. Loop (1) uses the S69 LUXPEST Pesticide Screening List from the NORMAN-SLE and in loop (2) are agrochemical transformation products and metabolites that were extracted from the Transformations section and Hazardous Substances Data Bank information (to yield the HSDBTPS List) on PubChem.

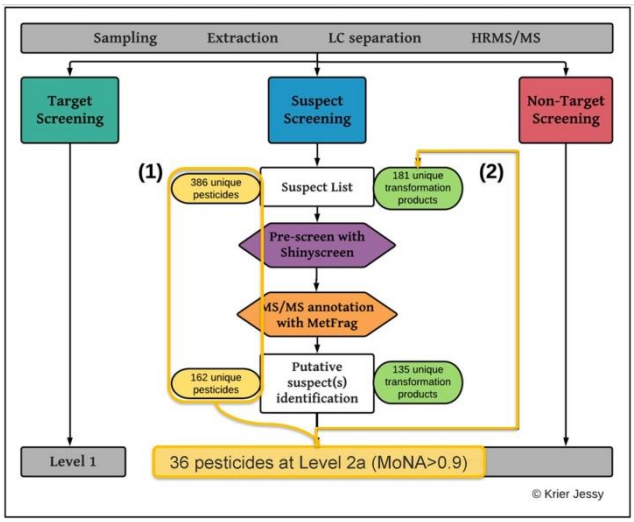


Figure 2. Data-driven transformation product and metabolite search. Reproduced with permission.

Another example is third hand smoke in dust. A case study demonstrates the pros and cons of various environmental cheminformatics approaches to interpret non-target HRMS data, and highlight challenges still facing the field, such as limited coverage of mass spectral libraries.[21] Emma has published a recent perspective on these challenges in the context of neurotoxicity.[22] Anjana Elapavalore is currently working on rapid classification and interpretation of large collections as an extension of her master's thesis together with collaborators, in connection with chemicals from the Environmental Protection Agency's (EPA's) Non-targeted Analytical Collaborative Trial (ENTACT).[23]

Trying to increase the number of compounds identified and to improve the interpretation is challenging. "Digital detective work" involves capturing expert knowledge in both machine- and human- readable forms; connecting this to environmental observations; identifying and closing knowledge gaps; and supporting interpretation of complex data. Information in the public domain helps everybody; you never know when it will help *you*.

Sharing Chemical Data Through a Structural Database

Suzanna Ward, Head of Data, Cambridge Crystallographic Data Centre (CCDC)

CCDC is a not-for-profit charity which curates and disseminates 3D structural data, delivers knowledge-based solutions, and promotes research collaborations, outreach, and education. It produces the Cambridge Structural Database (CSD), a repository of 1.09 million small-molecule organic and metal-organic crystal structures from X-ray and neutron diffraction analyses. Olga Kennard, the founder, and John Bernal "had a passionate belief that the collective use of data would lead to the discovery of new knowledge which transcends the results of individual experiments".[24,25]

Quality of data was a CCDC maxim from the beginning. In AI, the consequences of using poor quality data are incorrect conclusions, wasted effort, loss of trust, and poor business decisions. Underpinning the CSD are deposited pre-publication datasets that capture the results of structure determination experiments. CCDC validates the data during the guided deposition process and adds value. First, a validated Open Researcher and Contributor ID (ORCID) is used as the unique identifier for a researcher.[26] CCDC also benefits from the fact that the Crystallographic Information File (CIF)[27] is a standard for information interchange in crystallography. The scientific validation program checkCIF checks the consistency and integrity of the CIF data and generates alerts to faults that should either be corrected or explained. Automated enrichment steps also include resolving disorder, identifying the monomeric unit for polymers, deducing a chemical structure

from the coordinates[28] (making the data FAIR),[15] assigning a reliability score, and generating a chemical diagram and name. Finally, each deposition is reviewed and validated by CCDC's expert editorial team.

Challenges in data curation include poor geometry, missing hydrogen atoms, ambiguous oxidation states, disordered structures, and greater diversity and complexity. The data are revisited and improved periodically. Research integrity is much more than misconduct:[29] it also concerns data completeness, consistency and trustworthiness. CCDC has taken steps to address this issue. The Protein Data Bank (PDB), CSD, the Inorganic Crystal Structure Database (ICSD) and the database of the International Centre for Diffraction Data (ICDD) are linked. The use of the InChI standard for chemical structure representation[30] enables greater discovery across multiple resources. When data are deposited at the CCDC, the depositor is sent an accession number that can be used by a publisher; publishers notify CCDC when an article associated with a dataset has been published. The exchange of IDs between CCDC and publisher enables reciprocal links between dataset and article. Since not every dataset is published, *CSD Communications* was established to allow a dataset to be published through the CSD.

CSD data are used through several solutions. CSD-Community offers a range of free-of-charge tools. The components in CSD-System provide search, visualisation and analysis features for structural chemists. CSD-Discovery provides pharmaceutical and agrochemical researchers with tools for discovering new molecules. CSD-Materials provides solid form informatics capabilities. CSD-Enterprise gives access to the CSD and all CCDC application software. The CSD can be searched with web-based interfaces, desktop tools, and programmatic interfaces.

Generating income through value-added services and software is challenging. It can lead to misconceptions about public availability of data. The restrictions imposed to ensure long-term preservation of the data can hamper reusability and could ultimately limit innovation. CCDC has built a recognised, trusted repository. How might it conform to FAIR data[15] principles? Every individual dataset is free to view and download; DOIs are assigned to deposited data; there are links to and from other resources; and there is a free teaching subset. Country-wide and campus-wide licences are offered. For individual researchers, the cost of CSD can be included in a grant application. Low-, and middle-income countries, receive discounts. The Frank Allen International Research and Education Programme supports the use of CSD in developing countries. CCDC is also developing partnerships with industry to help shape the future, aid innovation, and ensure the sustainability of the CSD. By sharing data CCDC has learned general research data management expertise, has developed an understanding of users' needs, and has delivered services tailored to support domain requirements. Nevertheless, balancing sustainability with openness and FAIRness is hard.

Chemotion: Infrastructure to Provide Open Data
Nicole Jung, Group Leader, Compound Platform, Karlsruhe Institute for Technology

The Chemotion projects consist of several software tools that facilitate digital work processes in experimental chemistry. They support two open systems: an electronic laboratory notebook (ELN) and a repository for research data. Data from the ELN can be selected and published in the repository.

Currently, most synthetic organic chemistry data are not FAIR,[15] a lot of information is lost, the correlation of experiment to data is not clear and the availability of data depends heavily on the goodwill of a scientist. Although many journals require data availability, the data are not open in terms of the DFG Code of Conduct. In the Chemotion ELN you can save your reactions and samples, make calculations, and generate reports. Data can be imported with ChemScanner which extracts chemical information from ChemDraw files (.cdx, .cdxml, or cdx(ml) files containing .doc and .docx files). ChemSpectra allows you to view, edit and export spectra from Joint Committee on Atomic and Molecular Physical Data (JCAMP-DX) files. The spectral viewer function does not require any other software to be installed. You can publish your chemical structures, attach characterisation data, and make them citable by DOI using the Chemotion repository for molecules, reactions and research data. Registration with a few scientific data providers is automated. Data from the repository are

checked for input to PubChem. Data can be displayed from Chemical Abstracts Service databases (*via* SciFinder) or PubChem. Whole Chemotion data collections can be exported as SDfiles, .xlsx, .docx, and JSON. Videos of Chemotion functionality can be seen on the Chemotion website.

Chemotion concepts are molecules (with analyses and properties); reactions (descriptions and calculations); wellplates (lists and compound assignment); screens (basic plans and attachments); and a flexible module for text, tables, images, and files. Generic settings are currently being prepared for devices, metal organic frameworks, materials, and mixtures. Soon, molecules and materials will be linked, and it will be possible to order samples of reference materials.

Germany is funding its National Research Data ("Forschungsdaten") Infrastructure (NFDI), providing over 85 million euros over the next 10 years. This covers research data management for all areas of science, represented by 30 consortia. The NFDI recognises that digital data storage is an indispensable prerequisite for treating new research issues, generating findings, and making innovations. NFDI4Chem is the chemistry consortium in the NFDI. It is an initiative to build an open and FAIR infrastructure for research data management in chemistry. Free facilities are now available. There are no excuses for not sharing data.

Learning from Massive Scale Absorption, Distribution, Metabolism, Excretion and Toxicity (ADMET) and Potency Data Analysis
Ed Griffen, Co-founder, MedChemica

MedChemica researchers have developed an automated approach to analysing large ADMET and potency datasets based on advanced matched molecular pair analysis (MMPA).[31-33] The method has a strict requirement for normalised structures and data, otherwise inferences that could be gained from discrete structural changes will be masked in the noise. MMPA recognises molecules that differ only by a particular, well-defined, structural transformation, and it captures the change and environment of that transformation. Statistical analysis defines "medicinal chemistry rules": transformations with a high probability of improving the properties of molecules. The rules are stored in a high-performance database with an intuitive user interface. Standardisation of units, species, assays and the MMPA environment enabled knowledge sharing for Roche, Genentech and AstraZeneca.[34] About 500 million molecular pairs were aggregated by MedChemica and converted into a "grand rule database" which the three companies independently exploited.

Sets of transformations can be calibrated against each other, since the differences between two values, not the absolute values in assays, are compared. Assays are usually linearly displaced against each other. The database contains large numbers of statistically valid transforms (e.g., 153,349 rules for log*D*). Ed illustrated some trends and exceptions in solubility *versus* log*D*, and clearance *versus* log*D*, and related them to transformations.

If chemical structures are described inconsistently that inconsistency will be captured rather than a genuine structural change. Thus, it is important to correct inconsistent tautomeric forms and charge states. Ed displayed an example where three tautomers of one compound have only 60% similarity. Some tautomers are simple to fix using SMARTS as a stopping condition and SMIRKS to edit the compounds. SMIRKS with explicit hydrogen is used for precision. Nevertheless, some structures (e.g., chlorhexidine) can get stuck in loops where two tautomerisable groups are linked and rules can fight each other. MedChemica has also had to address (with SMIRKS) the intricacies of charge normalisation. Large companies need to code weakly acidic heterocycles and moderate bases, and need to be aware of permanently charged states. As for assays, experimental protocols may be available, and agonists *versus* antagonists for receptors, and binding *versus* activation for ion channels must be recognised: the BioAssay Ontology (BAO) is used wherever possible. The metadata must make clear which units are used (e.g., μM, nM, or pIC_{50}), what the species is, and how protein binding and clearance are measured. Both the original units and a transformed canonical unit are stored. For MMPA, log units are used so that the matched pair difference value makes sense. Rules for increasing potency are gathered. Individual assay rules are grouped as a "broad" goal so that historical rules can be applied to new hits for a particular target (e.g., dopamine). Managing metadata is not easy but is vital. Ed presented some

screen shots (e.g., a table of transformations with changes in property values) to show how using clear data makes for easier analysis. Structure clean-up makes automated compound comparison easier; aggregation of multiple measurements is possible, and anomalous data are flagged.

Ed concluded with some comments on the impact of automated heuristic generation to drug discovery programs[31] related, in particular, to explainable AI and transparent QSAR. Visualisable descriptors can be checked and precise ones can be distinguished from simple ones. Topological distances are precisely specified and can be exactly visualised on the molecules, so that there can be no doubt which features are correlated with activity. MedChemica offers transparent algorithms: MMPA, and *k* nearest neighbour (*k*NN) and random forest (RF) models built on understandable descriptors. An analysis of errors from the models identifies the exact pharmacophore descriptors that are less accurately predicted, suggesting areas for new compound acquisition and testing.

Good data are at the heart of AI. Medicinal chemists should ask themselves who cleaned their data and how, how the data were validated, and whether it is possible to check for outliers, whether the units are understandable, whether the descriptors can be validated and whether the under-represented data can be described.

Mining Data from the Open Domain for Modelling Transporter-Mediated Toxicity
Barbara Zdrazil, University of Vienna

Understanding factors that drive the onset of liver disease and toxicity is an important field of research because the human liver is the prime organ for drug metabolism; 80% of drug safety events are mediated by some degree of off-target pharmacology. In this talk Barbara focused on ATP binding cassette (ABC) and solute carrier (SLC) transporters which transport nutrients across the membrane to sustain normal cellular and organ function.[35,36]

Genetic variants in the Organic Cation Transporter 1 (OCT1) are known to affect pharmacokinetics and efficacy of tramadol, morphine, and codeine. Barbara's team took an in-house dataset of nine synthetic opioids and by scaffold clustering[37] and chemical similarity search in ChEMBL (for six different substructures) they enriched the dataset by 45 active and 97 inactive compounds. They derived trends in physicochemical properties and established a ligand-based shared pharmacophore model for synthetic opioids and morphinans binding to OCT1.[38]

Barbara's team has studied transporters of the SLC family, and in particular hepatic Organic Anion Transporting Polypeptides (OATP1B1, OATP1B3, OATP2B1)) and Organic Cation Transporter 1 (OCT1). Through mining public domain data, cheminformatics analysis, machine learning and virtual screening approaches, they aim to understand the basis of ligand recognition and selectivity, and determine critical chemical substructures enabling ligand-protein interactions.[39]

In silico tools to predict inhibition and substrate profiles towards the human ABC transporters Breast Cancer Resistance Protein (BCRP) and Multidrug Resistance Protein 1 (P-gp) might serve as early filters in the drug discovery and development process. Barbara and her co-workers retrieved BCRP and P-gp inhibitors from open data and from data manually curated from the literature, and they used machine learning approaches to establish multi-label classification models.[40] They compared label-powerset, binary relevance, and classifier chain[41] models. Label-powerset revealed important molecular features for selective or polyspecific inhibitory activity. Just two descriptors (the numbers of hydrophobic and aromatic atoms) were sufficient to separate selective BCRP inhibitors from selective P-gp inhibitors. A KNIME workflow proved to be a useful tool to merge data from diverse sources.

Barbara's student Alzbeta Türkova has integrated ligand bioactivity data for three hepatic OATPs from five open data sources in a KNIME workflow.[42] Highly curated datasets were analysed with respect to enriched

scaffolds. Sixty-five per cent of compounds in the dataset were measured more than once. Activity profiles and interesting scaffold series could be extracted, providing an indication for selective, dual-, or pan-inhibitory activity toward hepatic OATPs. In addition, a sequential binary modelling approach revealed common and distinctive ligand features for inhibitory activity towards the individual transporters. The workflows for data curation and analysis are freely available.

Together with Rajarshi Guha, Barbara mined ChEMBL20 to examine the evolution of scaffold features, such as enumerated compounds, biological activity, and liabilities, over 17 years.[43] They attempted to explain why a scaffold receives more attention over time and highlight that obvious aspects such as synthetic feasibility do not explicitly drive attention. In later work, target data from ChEMBL were linked with curated gene-disease associations in DisGeNET and biological process annotations in the Gene Ontology to monitor target classes over time and capture different aspects of target innovation. Workflows, scripts, data and a web application are publicly available.[44]

Linking knowledge can lead us towards a mechanistic understanding. For example, Barbara showed a schematic depiction of adverse outcome pathways (AOPs) leading to hepatic steatosis,[45] derived from QSAR modelling of assays provided by the ToxCast programme. Barbara and her colleagues have combined *in vitro* and *in vivo* information in QSAR work on steatosis soon to be published.

In order to model toxicity better, we need more compounds measured *in vitro* on multiple related targets; better accessibility to more *in vivo* data such as ToxRefDB;[46] better understanding of *in vitro* to *in vivo* extrapolation and animal to human extrapolation; integration of chemical data with toxicogenomics data and pathway information; and automated aggregation of AOP information.

A Frosted Window: Sharing Chemical Data in Open Drug Discovery Networks
Matthew Todd, Professor of Drug Discovery, University College London, School of Pharmacy

An open source approach to drug discovery and development has been tested and Matthew has distilled the lessons learned into six laws of operation[47] that help to clarify working practices:

- all data are open and all ideas are shared
- anyone can take part at any level
- there will be no patents
- suggestions are the best form of criticism
- public discussion is much more valuable than private email
- an open project is bigger than, and is not owned by any given lab

The approach has been used for antimalarials,[48,49] tuberculosis,[50] and antifungals.[51] Matthew's team has developed platforms to promote collaborative working and to facilitate publication.[52,53] After Martin Shkreli hiked the price of daraprim from $13.50 to $750 a tablet, schoolboys in Sydney used the Open Source Malaria platform to make the same drug in their school lab for about $2 a dose.

Open science has been shown to be an accelerator of biomedical research, and the COVID-19 crisis has made openness seem the logical choice,[54] but can the approach be scaled up to more programmes for more diseases where pharma struggles to compete? For example, the rapidly emerging antiviral drug discovery initiative (READDI) aims to give humanity an early advantage against future diseases but who will fund drug discovery for a disease that does not yet exist? Another initiative is Target 2035, an international federation of scientists from the public and private sectors developing new technologies to create, by year 2035, chemogenomic libraries and chemical and biological probes for the entire proteome.[55]

Patents are not necessary for funding drug discovery: public-private partnerships involved in the Drugs for Neglected Diseases initiative (DNDi) developed the anti-malarial combination, ASAQ Winthrop[56] and

fexinidazole[57] for treating sleeping sickness. As an alternative to patents, a form of data exclusivity[58] can be used to protect medicines from competition. M4K Pharma was incorporated to launch an open science drug discovery programme that relies on regulatory exclusivity as its primary intellectual property and commercial asset, in lieu of patents,[59] and this has led to the founding of M4ID pharma which aims to do the same for drugs for infectious diseases.

Another challenge to scaling up open science is that molecules in projects may be understandable to humans but not by machines (like looking through a "frosted window"). To address this issue, Matthew has been using InChIs[30] in his LabArchives ELN. Google can also find molecules by InChIKey but it is still too easy to miss relevant work. We need better ways to "mention" molecules such as the Science Introduction Robot (SCINDR)[60] added to ELNs, although "molecule of the day" on Twitter could be done with a Twitter robot from open science projects.

Attempts have been made at "telling stories". The Open Source Malaria repository contains a review of medicines for malaria, their past, present, and future, starting from a publication by Matthew's team.[61] It is a *living* review which can be edited and expanded by anyone. Another repository houses the living review of open resources for drug discovery and development for SARS-CoV-2.[54] Traditional outputs remain important but supporting information needs to be handled carefully.[48]

Matthew concluded with some comments on FAIR data. It is difficult to ensure data are well folded in and the cells of a project are not well interlinked. It is difficult to have interoperable procedures when there are many legacy format preferences. It is easy to work with a consortium of people, but accessibility still relies on human gatekeepers and automatic metadata generation is not great. Reusability is possible if the licence of the project is CC-BY-4.0 but there are lots of gaps.

Some challenges go beyond FAIR. Mistakes must be corrected, so an immediately-online project must ensure the good data percolate through; but what if the data have already been indexed elsewhere? Economic viability is another issue: openness makes research better and faster; it can lead to big-scale, manufactured outcomes that impact society, but will people invest?

Exploring Chemical Space for Molecular Material Discovery
Kim Jelfs, Imperial College London

Kim's team develop software to assist in the discovery of porous molecular materials, polymers, and organic electronics. Initially they focused on porous molecular materials which are typically synthesised from organic precursors through dynamic covalent chemistry (DCC). Kim showed an octahedral structure the vertices of which are replaced with organic molecules to make a symmetrical molecule with a cavity. From 800,000 aldehyde and amine precursors, more than 830 million porous organic cages could be made. It is not possible to make and screen all these.

To provide support for high-throughput screening of large batches, Kim's team developed stk, a tool for the automated assembly, molecular optimisation, and property calculation of supramolecular materials.[62] An evolutionary algorithm assembles hypothetical molecules from a library of precursors. A Python library provides an API and integrates with third party codes, allowing users to explore the potential energy landscape of supramolecules, and then calculate structural features (e.g., pore size and chemical similarity) and properties (e.g., energy and ionisation potential).

The team has addressed the computational challenges encountered when trying to predict the most likely topological outcomes from DCC reactions of organic building blocks[63] and has fused computation with robotic synthesis.[64] They have predicted solvent effects on the structure of porous organic molecules[65] and used machine learning to predict shape persistence and cavity size.[66] The computational predictability of organic cage crystal packing has been explored in conjunction with Graeme Day's team.[67,68]

All this research work has been brought together and extended to synthesis using robotic screening.[64,69] Becky Greenaway's work on exploring precursor space and making complex molecules from the precursors using robotic screening found that 40% of the synthesis attempts were successful. Computational prediction of topology was found to be successful in the cases where there was a strong thermodynamic preference, but this is not always the case. Property prediction has been reported separately.[70,71] The pyWINDOW Python package[72] for structural analysis of discrete molecules with voids and windows is openly available.

The majority of hypothetical organic cages suffer from lack of shape persistence and thus lack intrinsic porosity. Kim's team have used machine learning to predict shape persistence and cavity size and have created a database of 63,472 cages, formed through a range of reaction chemistries and in multiple topologies. Using this database they developed machine learning models capable of predicting shape persistence[66] with an accuracy of up to 93%. The data and models are openly available, and so is an app to predict if a cage will be shape persistent or collapsed. It was found that the imine condensation of trialdehydes and diamines in a [4 + 6] reaction is the most likely to result in shape persistent cages, whereas thiol reactions are most likely to give collapsed cages.

The stk tool has also been applied to polymer screening, using neural networks, in collaboration with Martijn Zwijnenburg's group at University College London.[73-75] Kim's team has also worked on polymer membranes for separations.[76,77] Such membranes can be used for energy efficient gas separations. Qi Yuan found a database of experimentally measured polymer gas permeabilities but it was incomplete, so he filled in missing values using the multivariate imputation by a chained equations (MICE) algorithm to fill in the missing data through an iterative procedure of predictive models.[78] The researchers re-analysed historical polymers, looked for potential missed candidates with promising gas selectivity, and identified polymer membranes worthy of further investigation. The code is available on GitHub.

Kim's team has also demonstrated the application of deep recurrent neural networks (RNN) and transfer learning for the exploration of the chemical space of building blocks for certain donor-acceptor oligomers with specific electronic properties.[79] They generated about 1,700 new oligomers with an RNN network tuned to target oligomers with a HOMO–LUMO gap <2 eV and a dipole moment <2 Debye. These could have potential application in organic photovoltaics. Thus, the team is now generalising the research that was initially focused on porous molecular materials to other molecular materials and their applications such as organic semiconductors and photocatalysis.

References

(1)Overington, J. P. ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). Interview by Wendy A. Warr. *J. Comput-Aided Mol. Des.* **2009,** *23* (4), 195-8.
(2)Overington, J. P.; Al-Lazikani, B.; Hopkins, A. L. How many drug targets are there? *Nat. Rev. Drug Discovery* **2006,** *5* (12), 993-996.
(3)Davies, M.; Dedman, N.; Hersey, A.; Papadatos, G.; Hall, M. D.; Cucurull-Sanchez, L.; Jeffrey, P.; Hasan, S., *et al.* ADME SARfari: comparative genomics of drug metabolizing systems. *Bioinformatics* **2015,** *31* (10), 1695-1697.
(4)Gleeson, M. P.; Hersey, A.; Montanari, D.; Overington, J. Probing the links between in vitro potency, ADMET and physicochemical parameters. *Nat. Rev. Drug Discovery* **2011,** *10* (3), 197-208.
(5)Nguyen, D.-T.; Mathias, S.; Bologa, C.; Brunak, S.; Fernandez, N.; Gaulton, A.; Hersey, A.; Holmes, J., *et al.* Pharos: Collating protein information to shed light on the druggable genome. *Nucleic Acids Res.* **2017,** *45* (D1), D995.
(6)Santos, R.; Ursu, O.; Gaulton, A.; Bento, A. P.; Donadi, R. S.; Bologa, C. G.; Karlsson, A.; Al-Lazikani, B., *et al.* A comprehensive map of molecular drug targets. *Nat. Rev. Drug Discovery* **2017,** *16* (1), 19-34.
(7)Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S., *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012,** *40* (D1), D1100-D1107.

(8)Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krueger, F. A.; Light, Y., *et al.* The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* **2014,** *42* (D1), D1083-D1090.
(9)Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F., *et al.* The ChEMBL database in 2017. *Nucleic Acids Res.* **2016,** *45* (D1), D945-D954.
(10)Kanza, S.; Willoughby, C.; Gibbins, N.; Whitby, R.; Frey, J. G.; Erjavec, J.; Zupancic, K.; Hren, M., *et al.* Electronic lab notebooks: can they replace paper? *J. Cheminf.* **2017,** *9*, 31.
(11)Kanza, S. What influence would a cloud based semantic laboratory notebook have on the digitisation and management of scientific research? Ph.D. Thesis, University of Southampton, Southampton, U.K., April 2018 http://eprints.soton.ac.uk/421045/ (accessed April 15, 2020).
(12)Kanza, S.; Gibbins, N.; Frey, J. G. Too many tags spoil the metadata: investigating the knowledge management of scientific research with semantic web technologies. *J. Cheminf.* **2019,** *11*, 23.
(13)Kanza, S.; Graham Frey, J. Semantic technologies in drug discovery. In *Systems Medicine;* Wolkenhauer, O., Ed.; Academic Press: Oxford, 2021; pp 129-144.
(14)Kanza, S.; Frey, J. G. A new wave of innovation in semantic web tools for drug discovery. *Expert Opin. Drug Discovery* **2019,** *14* (5), 433-444.
(15)Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W., *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016,** *3*, 160018.
(16)Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A., *et al.* PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* **2019,** *47* (D1), D1102-D1109.
(17)Schymanski, E. L.; Singer, H. P.; Longrée, P.; Loos, M.; Ruff, M.; Stravs, M. A.; Ripollés Vidal, C.; Hollender, J. Strategies to Characterize Polar Organic Contamination in Wastewater: Exploring the Capability of High Resolution Mass Spectrometry. *Environ. Sci. Technol.* **2014,** *48* (3), 1811-1818.
(18)Vermeulen, R.; Schymanski, E. L.; Barabási, A.-L.; Miller, G. W. The exposome and health: Where chemistry meets biology. *Science* **2020,** *367* (6476), 392.
(19)Wild, C. P.; Scalbert, A.; Herceg, Z. Measuring the exposome: A powerful basis for evaluating environmental exposures and cancer risk. *Environ. Mol. Mutagen.* **2013,** *54* (7), 480-499.
(20)Ruttkies, C.; Schymanski, E. L.; Wolf, S.; Hollender, J.; Neumann, S. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J. Cheminf.* **2016,** *8* (1), 3.
(21)Oberacher, H.; Sasse, M.; Antignac, J.-P.; Guitton, Y.; Debrauwer, L.; Jamin, E. L.; Schulze, T.; Krauss, M., *et al.* A European proposal for quality control and quality assurance of tandem mass spectral libraries. *Environ. Sci. Eur.* **2020,** *32* (1), 43.
(22)Schymanski, E. L.; Baker, N. C.; Williams, A. J.; Singh, R. R.; Trezzi, J.-P.; Wilmes, P.; Kolber, P. L.; Kruger, R., *et al.* Connecting environmental exposure and neurodegeneration using cheminformatics and high resolution mass spectrometry: potential and challenges. *Environ. Sci.: Processes Impacts* **2019,** *21* (9), 1426-1445.
(23)Ulrich, E. M.; Sobus, J. R.; Grulke, C. M.; Richard, A. M.; Newton, S. R.; Strynar, M. J.; Mansouri, K.; Williams, A. J. EPA's non-targeted analysis collaborative trial (ENTACT): genesis, design, and initial findings. *Anal. Bioanal. Chem.* **2019,** *411* (4), 853-866.
(24)Kennard, O. *Bernal's Vision: from Data to Insight. J. D. Bernal Lecture 1995*; Birkbeck College: London, 1996.
(25)Kennard, O. From Private Data to Public Knowledge. In *The Impact of Electronic Publishing on the Academic Community;* Butterworth, I., Ed.; Portland Press: London, 1997; pp 159-166.
(26)Haak, L. L.; Fenner, M.; Paglione, L.; Pentz, E.; Ratner, H. ORCID: a system to uniquely identify researchers. *Learned Publishing* **2012,** *25* (4), 259-264.
(27)Hall, S. R.; Allen, F. H.; Brown, I. D. The crystallographic information file (CIF): a new standard archive file for crystallography. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **1991,** *47* (6), 655-685.
(28)Bruno, I. J.; Shields, G. P.; Taylor, R. Deducing chemical structure from crystallographically determined atomic coordinates. *Acta Crystallogr., Sect. B: Struct. Sci.* **2011,** *67* (4), 333-349.
(29)Anon. Research integrity is much more than misconduct. *Nature* **2019,** *570* (7759), 5.
(30)Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J. Cheminf.* **2015,** *7*, 1-63.
(31)Griffen, E. The rise of the intelligent machines in drug hunting? *Future Med. Chem.* **2009,** *1* (3), 405-408.
(32)Griffen, E.; Leach, A. G.; Robb, G. R.; Warner, D. J. Matched Molecular Pairs as a Medicinal Chemistry Tool. *J. Med. Chem.* **2011,** *54* (22), 7739-7750.

(33)Lukac, I.; Zarnecka, J.; Griffen, E. J.; Dossetter, A. G.; St-Gallay, S. A.; Enoch, S. J.; Madden, J. C.; Leach, A. G. Turbocharging Matched Molecular Pair Analysis: Optimizing the Identification and Analysis of Pairs. *J. Chem. Inf. Model.* **2017,** *57* (10), 2424-2436.
(34)Kramer, C.; Ting, A.; Zheng, H.; Hert, J.; Schindler, T.; Stahl, M.; Robb, G.; Crawford, J. J., *et al.* Learning Medicinal Chemistry Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADMET) Rules from Cross-Company Matched Molecular Pairs Analysis (MMPA). *J. Med. Chem.* **2018,** *61* (8), 3277-3292.
(35)Niemi, M.; Pasanen, M. K.; Neuvonen, P. J. Organic anion transporting polypeptide 1B1: a genetically polymorphic transporter of major importance for hepatic drug uptake. *Pharmacol. Rev.* **2011,** *63* (1), 157-181.
(36)Lai, Y. Transporter Role in Drug Efficacy and Organ Toxicity: Four Pillars of Clinical Trial Survival. *J. Drug Metab. Toxicol.* **2016,** *7*, 4.
(37)Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996,** *39* (15), 2887-2893.
(38)Meyer, M. J.; Neumann, V. E.; Friesacher, H. R.; Zdrazil, B.; Brockmoeller, J.; Tzvetkov, M. V. Opioids as Substrates and Inhibitors of the Genetically Highly Variable Organic Cation Transporter OCT1. *J. Med. Chem.* **2019,** *62* (21), 9890-9905.
(39)Tuerkova, A.; Zdrazil, B. Current Advances in Studying Clinically Relevant Transporters of the Solute Carrier (SLC) Family by Connecting Computational Modeling and Data Science. *Comput. Struct. Biotechnol. J.* **2019,** *17*, 390-405.
(40)Montanari, F.; Zdrazil, B.; Digles, D.; Ecker, G. F. Selectivity profiling of BCRP versus P-gp inhibition: from automated collection of polypharmacology data to multi-label learning. *J. Cheminf.* **2016,** *8*, 7.
(41)Read, J.; Pfahringer, B.; Holmes, G.; Frank, E. Classifier chains for multi-label classification. In *Lecture Notes in Computer Science, Volume 5782;* Buntine, W., Grobelnik, M., Mladenić, D., Shawe-Taylor, J., Eds.; Springer: Berlin, Germany, 2009.
(42)Tuerkova, A.; Jain, S.; Zdrazil, B. Integrative Data Mining, Scaffold Analysis, and Sequential Binary Classification Models for Exploring Ligand Profiles of Hepatic Organic Anion Transporting Polypeptides. *J. Chem. Inf. Model.* **2019,** *59* (5), 1811-1825.
(43)Zdrazil, B.; Guha, R. The Rise and Fall of a Scaffold: A Trend Analysis of Scaffolds in the Medicinal Chemistry Literature. *J. Med. Chem.* **2018,** *61* (11), 4688-4703.
(44)Zdrazil, B.; Richter, L.; Brown, N.; Guha, R. Moving targets in drug discovery. *Sci. Rep.* **2020,** *10* (1), 20213.
(45)Gadaleta, D.; Manganelli, S.; Roncaglioni, A.; Toma, C.; Benfenati, E.; Mombelli, E. QSAR Modeling of ToxCast Assays Relevant to the Molecular Initiating Events of AOPs Leading to Hepatic Steatosis. *J. Chem. Inf. Model.* **2018,** *58* (8), 1501-1517.
(46)Watford, S.; Ly Pham, L.; Wignall, J.; Shin, R.; Martin, M. T.; Friedman, K. P. ToxRefDB version 2.0: Improved utility for predictive and retrospective toxicology analyses. *Reprod. Toxicol.* **2019,** *89*, 145-158.
(47)Todd, M. H. Six Laws of Open Source Drug Discovery. *ChemMedChem* **2019,** *14* (21), 1804-1809.
(48)Williamson, A. E.; Ylioja, P. M.; Robertson, M. N.; Antonova-Koch, Y.; Avery, V.; Baell, J. B.; Batchu, H.; Batra, S., *et al.* Open Source Drug Discovery: Highly Potent Antimalarial Compounds Derived from the Tres Cantos Arylpyrroles. *ACS Cent. Sci.* **2016,** *2* (10), 687-701.
(49)Tse, E.; Aithani, L.; Anderson, M.; Cardoso-Silva, J.; Cincilla, G.; Conduit, G.; Galushka, M.; Guan, D., *et al.* An Open Drug Discovery Competition: Experimental Validation of Predictive Models in a Series of Novel Antimalarials. http://doi.org/10.26434/chemrxiv.13194755.v1 (accessed December 8, 2020).
(50)Guardia, A.; Baiget, J.; Cacho, M.; Perez, A.; Ortega-Guerra, M.; Nxumalo, W.; Khanye, S. D.; Rullas, J., *et al.* Easy-To-Synthesize Spirocyclic Compounds Possess Remarkable in Vivo Activity against Mycobacterium tuberculosis. *J. Med. Chem.* **2018,** *61* (24), 11327-11340.
(51)Lim, W.; Melse, Y.; Konings, M.; Duong, H. P.; Eadie, K.; Laleu, B.; Perry, B.; Todd, M. H., *et al.* Addressing the most neglected diseases through an open research model: The discovery of fenarimols as novel drug candidates for eumycetoma. *PLoS Neglected Trop. Dis.* **2018,** *12* (4), e0006437.
(52)Badiola, K. A.; Bird, C.; Brocklesby, W. S.; Casson, J.; Chapman, R. T.; Coles, S. J.; Cronshaw, J. R.; Fisher, A., *et al.* Experiences with a researcher-centric ELN. *Chem. Sci.* **2015,** *6* (3), 1614-1629.
(53)Robertson, M. N.; Ylioja, P. M.; Williamson, A. E.; Woelfle, M.; Robins, M.; Badiola, K. A.; Willis, P.; Olliaro, P., *et al.* Open source drug discovery - a limited tutorial. *Parasitology* **2014,** *141* (1), 148-157.
(54)Tse, E. G.; Klug, D. M.; Todd, M. H. Open science approaches to COVID-19. *F1000Research* **2020,** *9*, 1043.

(55)Carter, A. J.; Kraemer, O.; Zwick, M.; Mueller-Fahrnow, A.; Arrowsmith, C. H.; Edwards, A. M. Target 2035: probing the human proteome. *Drug Discovery Today* **2019,** *24* (11), 2111-2115.
(56)Bompart, F.; Kiechel, J.-R.; Sebbag, R.; Pecoul, B. Innovative public-private partnerships to maximize the delivery of anti-malarial medicines: lessons learned from the ASAQ Winthrop experience. *Malar. J.* **2011,** *10*, 143.
(57)Torreele, E.; Trunz, B. B.; Tweats, D.; Kaiser, M.; Brun, R.; Mazue, G.; Bray, M. A.; Pecoul, B. Fexinidazole - a new oral nitroimidazole drug candidate entering clinical development for the treatment of sleeping sickness. *PLoS Neglected Trop. Dis.* **2010,** *4* (12), e923.
(58)Marden, E. Open Source Drug Development: a Path to More Accessible Drugs and Diagnostics? *Minnesota Journal of Law, Science and Technology* **2010,** *11* (1), 10.
(59)Morgan, M. R.; Roberts, O. G.; Edwards, A. M. Ideation and implementation of an open science drug discovery business model - M4K Pharma. *Wellcome Open Res.* **2018,** *3*, 154.
(60)Smith, C. C.; Todd, M.; Patiny, L.; Swain, C.; Southan, C.; Williamson, A. E.; Clark, A. M. SCINDR - The SCience INtroDuction Robot that will Connect Open Scientists. *Research Ideas and Outcomes* **2016,** *2*, e9995.
(61)Tse, E. G.; Korsik, M.; Todd, M. H. The past, present and future of anti-malarial medicines. *Malar. J.* **2019,** *18* (1), 93.
(62)Turcani, L.; Berardo, E.; Jelfs, K. E. stk: A python toolkit for supramolecular assembly. *J. Comput. Chem.* **2018,** *39* (23), 1931-1942.
(63)Santolini, V.; Miklitz, M.; Berardo, E.; Jelfs, K. E. Topological landscapes of porous organic cages. *Nanoscale* **2017,** *9* (16), 5280-5298.
(64)Greenaway, R. L.; Santolini, V.; Bennison, M. J.; Alston, B. M.; Pugh, C. J.; Little, M. A.; Miklitz, M.; Eden-Rump, E. G. B., *et al.* High-throughput discovery of organic cages and catenanes using computational screening fused with robotic synthesis. *Nat. Commun.* **2018,** *9* (1), 2849.
(65)Santolini, V.; Tribello, G. A.; Jelfs, K. E. Predicting solvent effects on the structure of porous organic molecules. *Chem. Commun.* **2015,** *51* (85), 15542-15545.
(66)Turcani, L.; Greenaway, R. L.; Jelfs, K. E. Machine Learning for Organic Cage Property Prediction. *Chem. Mater.* **2019,** *31* (3), 714-727.
(67)Jones, J. T. A.; Hasell, T.; Wu, X.; Bacsa, J.; Jelfs, K. E.; Schmidtmann, M.; Chong, S. Y.; Adams, D. J., *et al.* Modular and predictable assembly of porous organic molecular crystals. *Nature* **2011,** *474* (7351), 367-371.
(68)Pyzer-Knapp, E. O.; Thompson, H. P. G.; Schiffmann, F.; Jelfs, K. E.; Chong, S. Y.; Little, M. A.; Cooper, A. I.; Day, G. M. Predicted crystal energy landscapes of porous organic cages. *Chem. Sci.* **2014,** *5* (6), 2235-2245.
(69)Greenaway, R. L.; Santolini, V.; Pulido, A.; Little, M. A.; Alston, B. M.; Briggs, M. E.; Day, G. M.; Cooper, A. I., *et al.* From Concept to Crystals via Prediction: Multi-Component Organic Cage Pots by Social Self-Sorting. *Angew. Chem., Int. Ed.* **2019,** *58* (45), 16275-16281.
(70)Miklitz, M.; Jiang, S.; Clowes, R.; Briggs, M. E.; Cooper, A. I.; Jelfs, K. E. Computational Screening of Porous Organic Molecules for Xenon/Krypton Separation. *J. Phys. Chem. C* **2017,** *121* (28), 15211-15222.
(71)Jackson, E.; Miklitz, M.; Song, Q.; Tribello, G. A.; Jelfs, K. E. Computational Evaluation of the Diffusion Mechanisms for C8 Aromatics in Porous Organic Cages. *J. Phys. Chem. C* **2019,** *123* (34), 21011-21021.
(72)Miklitz, M.; Jelfs, K. E. pywindow: Automated Structural Analysis of Molecular Pores. *J. Chem. Inf. Model.* **2018,** *58* (12), 2387-2391.
(73)Sprick, R. S.; Aitchison, C. M.; Berardo, E.; Turcani, L.; Wilbraham, L.; Alston, B. M.; Jelfs, K. E.; Zwijnenburg, M. A., *et al.* Maximising the hydrogen evolution activity in organic photocatalysts by co-polymerisation. *J. Mater. Chem. A* **2018,** *6* (25), 11994-12003.
(74)Wilbraham, L.; Berardo, E.; Turcani, L.; Jelfs, K. E.; Zwijnenburg, M. A. High-Throughput Screening Approach for the Optoelectronic Properties of Conjugated Polymers. *J. Chem. Inf. Model.* **2018,** *58* (12), 2450-2459.
(75)Wilbraham, L.; Sprick, R. S.; Jelfs, K. E.; Zwijnenburg, M. A. Mapping binary copolymer property space with neural networks. *Chem. Sci.* **2019,** *10* (19), 4973-4984.
(76)Thompson, K. A.; Mathias, R.; Kim, D.; Kim, J.; Rangnekar, N.; Johnson, J. R.; Hoy, S. J.; Bechis, I., *et al.* N-Aryl-linked spirocyclic polymers for membrane separations of complex hydrocarbon mixtures. *Science* **2020,** *369* (6501), 310-315.

(77)Tan, R.; Wang, A.; Malpass-Evans, R.; Zhao, E. W.; Liu, T.; Ye, C.; Zhou, X.; Darwich, B. P., *et al.* Hydrophilic microporous membranes for selective ion separation and flow-battery energy storage. *Nat. Mater.* **2020,** *19* (2), 195-202.
(78)Yuan, Q.; Longo, M.; Thornton, A.; McKeown, N. B.; Comesana-Gandara, B.; Jansen, J. C.; Jelfs, K. Imputation of Missing Gas Permeability Data for Polymer Membranes Using Machine Learning. http://doi.org/10.26434/chemrxiv.13124993.v1 (accessed December 10, 2020).
(79)Yuan, Q.; Santana-Bonilla, A.; Zwijnenburg, M. A.; Jelfs, K. E. Molecular generation targeting desired electronic properties via deep generative models. *Nanoscale* **2020,** *12* (12), 6744-6758.

Workshops on Open Source Tools for Chemistry – Workshop Reports

Workshop reports provided by CICAG Chair Dr Chris Swain, email: swain@mac.com

Open Chemical Science: Reports from the online meetings and workshops, 9-13 November 2020

One of the highlights of the Open Chemical Sciences meeting were the workshops, where we were fortunate to have outstanding contributions demonstrating six critical open-source software packages, PyMOL, KNIME, Fragalysis, Google CoLab, ChEMBL and DataWarrior. These workshops were all recorded and are all on the YouTube channel, and have already been viewed over 1000 times in total.

DataWarrior workshop by Isabelle Giraud
DataWarrior combines dynamic graphical views and interactive row filtering with chemical intelligence. Scatter plots, box plots, bar charts and pie charts not only visualize numerical or category data, but also show trends of multiple scaffolds or compound substitution patterns. This workshop was an introductory tutorial, and DataWarrior can be downloaded here.

PyMOL workshop by Garrett Morris
PyMOL is a comprehensive software package for rendering and animating 3D structures, in particular biomolecules. You can install PyMOL via Conda, Miniconda, Anaconda Cloud, OmicX, or from GitHub.

UsingGoogleCoLab workshop by Jan Jensen
Colaboratory, or "Colab" for short, allows you to write and execute Python in your browser, with Zero configuration required. The notebooks used are available online – Initial, and Final.

ChEMBL workshop by Anna Gaulton
ChEMBL is a manually curated database of bioactive molecules with drug-like properties. It brings together chemical, bioactivity, and genomic data, to aid the translation of genomic information into effective new drugs. This workshop was an introductory tutorial.

Fragalysis workshop by Rachel Skyner
Fragalysis (fragment analysis) is a web-based platform for fragment-based drug discovery). Its initial use case is focussed on the fragment screening experiments at Diamond. This workshop was an introductory tutorial.

KNIME workshop by Greg Landrum
KNIME Analytics Platform is the open-source software for creating data science. Intuitive, open, and continuously integrating new developments, KNIME makes understanding data and designing data science workflows and reusable components accessible to everyone. This workshop was an introductory tutorial. Knime can be downloaded here. Data used in tutorial are here.

RSC Open Access Journals and Future Plans

Contribution from Dr Neil Hammond, Publisher for Open Access Journals, RSC, Email: hammondn@rsc.org

It was a great pleasure to participate as a speaker and panellist at the recent CICAG Open Chemical Science workshop. The aim of my presentation, within the Open Access Publishing session, was to provide a perspective from the RSC's publishing division with regards to the challenges of supporting a transition towards greater openness in the publication of scientific research. I take this opportunity to expand on that theme in a little more detail.

"Dissemination of chemical knowledge" is one of the core objectives specified in the charter of the Royal Society of Chemistry. As a mission-led organisation we are thus mindful that a culture of openness in scientific research and maximum access to the published outputs of that research are principles that align strongly with our purpose. We have therefore been committed for some time to advancing the Open Access (OA) agenda within the chemical sciences, whilst seeking a sustainable approach which enables us to continue reinvesting a surplus back into our charitable activities in the chemistry community (events, grants, bursaries, awards, sponsorships, etc). We are proud of the significant steps we have already taken in that regard. In 2012 we introduced our 'Gold for Gold' initiative, which provided gold OA vouchers for subscription customers to our largest subscription content package at no extra charge. From January 2015 we transitioned our flagship journal Chemical Science to a gold OA model whilst waiving any publication charge, such that the journal presented, and still does to this day, a 'free to read' and 'free to publish' option to the community. The transition of our largest journal, *RSC Advances*, to a gold OA model in late 2016 followed, with a commitment to maintaining a competitive article processing charge (APC, currently £750, with a full waiver for researchers in over 100 countries in the developing world and a reduced rate for researchers in India and other selected countries). More recently, in 2018 we launched our first born-OA journal, *Nanoscale Advances*, which has now published more than 1000 articles, with APCs waived until later this year.

As we enter 2021, the OA landscape continues to evolve. The share of research in chemical sciences published under a gold OA model is growing at a rate of close to 30% year on year and gold OA articles accounted for 23% of all chemistry articles in 2020 ; funder mandates in support of OA are being refined and strengthened, notably with the efforts of cOAlition S in Europe; and publishers, both commercial and non-profit, are expanding options to authors through dedicated OA journals as well as transformative agreements which strive to transition traditional subscription journals to an OA model.

Substantial challenges continue to exist, however, to a more complete transition from a subscription model to OA. There are large global asymmetries in the provision of APC funds and policies supporting OA, often reflecting existing embedded asymmetries in the balance of research expenditure and subscription costs. Perceptions also still abound in some communities that OA is synonymous with lower editorial standards – which, whilst not true *a priori*, does in many cases reflect certain structural factors. Furthermore, whilst in principle most researchers are supportive of OA publishing, in practise the choice to publish under an OA model too often places additional work and administrative burden on researchers.

At this time the RSC has a renewed and strengthened commitment to accelerate the transition to open access within the chemical sciences and indeed to support the broader open science movement.

- We are continuing to expand our number of gold OA journals. In doing so we will provide options for researchers across all subject areas to publish their work under an OA model with the same levels of customer service and peer review fairness and rigour that we have established across our publishing portfolio.

- We are maintaining our commitment to highly competitive APCs, alongside discounted and waived APCs for researchers in developing countries. Our new OA journals will waive all APCs until they have become established, ensuring that authors can build trust in our new products before committing limited funds towards publication costs.

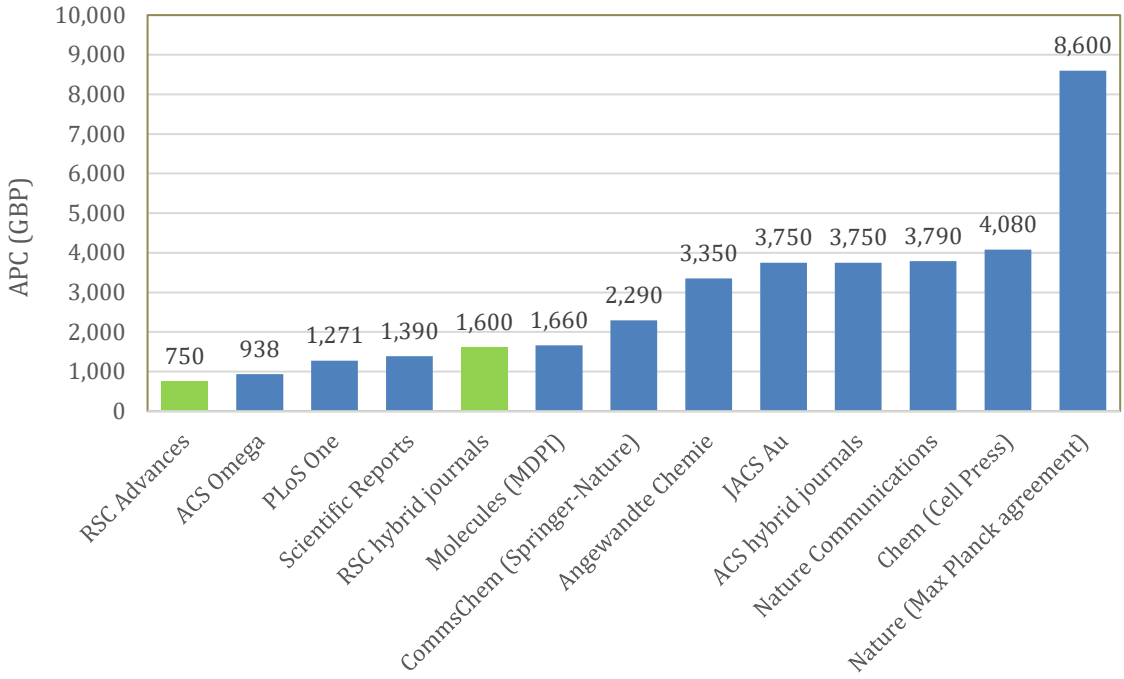


Figure 1. Comparison of article processing charges from leading publishers in chemical sciences and prominent multidisciplinary OA journals. Values shown are full rates (i.e. excluding discounted categories) and converted into GBP where set in another currency. RSC rates are shown in green.

- Building on the recent launches of *RSC Chemical Biology* and *Environmental Science: Atmospheres*, as well as the example set by our flagship *Chemical Science*, we will continue to build a strong correlation between OA and high editorial standards through highly selective OA journals. We will also balance that approach with the launch of OA journals that support early-career researchers and communities that are still developing their science infrastructures (e.g. *Materials Advances*).

- Within our portfolio of subscription journals (which all offer a gold OA option, and are thus identified as 'hybrid' titles), we will continue to promote the OA option and support the potential transition of those journals to a fully OA model. Our Read-and-Publish (R&P) deals, which provide institutions with unlimited OA publication within hybrid journals as part of a journal package subscription, have proven highly effective in this regard. Such deals have facilitated high levels of OA publishing in RSC journals from corresponding countries – for example, as the result of a transformative deal with JISC in 2020 more than half of UK articles in RSC journals in the last year were published under a fully gold OA model. At present we have 100 R&P agreements in place, primarily in Europe.

- We have built, and continue to develop, author services that remove the burden on authors opting for OA publication. In 2020 we released across all of our journals a streamlined author interface for making the article submission and APC payment process simpler. We are also continuing to refine and improve the necessary 'behind the scenes' processes that authors should be able to take for granted, such as automated deposition of articles in relevant repositories, as often mandated by funders.

- We continue to experiment with the offering of different peer review models, in support of greater transparency and fairness. Many of our journals now offer authors the option of double-blind peer review to help counter the implicit bias that we know exists in science and society as a whole. We have more recently initiated 'transparent peer review' on two of our newest OA journals (*RSC Chemical Biology* and *Environmental Science: Atmospheres*). Under this initiative, authors can opt-in to have the peer review history of their article (referee reports, decision letters) published openly alongside the final article.

- We will continue to provide, participate in and support open forums that provide a means for the communities we represent to express their diverse views on OA and Open Science (see, for example, our synthesis of contrasting opinions on Plan S here).

It has become something of a cliché within the open access movement to stress the limitations of any single actor within the broader publishing ecosystem to bring about change, and to emphasise the need for holistic solutions. It remains true nonetheless. The RSC is no more able than any other publisher or learned society to single-handedly effect a transformation which is equally dependent upon the actions and motivations of funders, institutions and individual researchers. What we can, and will, do is provide choice, remove barriers and respond to needs of the communities we serve.

[1] Figure calculated from Web of Science data across a representative list of 620 chemistry journals (over 250,000 articles).
[2] For example, whilst it is true that business models associated with OA publishing tend to incentivise 'scale' (i.e. a large number of articles), there are many examples of highly selective OA journals, such as Chemical Science. Indeed, RSC Advances, though a much larger and relatively less selective journal than Chemical Science, still rejects the majority of the manuscripts it receives.

RSC's Journal Archives now Available to Companies for Text and Data Mining

Contribution from Nathan Price, RSC Marketing Executive, Corporate, Americas, email: pricen@rsc.org

The Royal Society of Chemistry has now launched our new Text and Data Mining (TDM) solution, making our full collection of research journals available to companies for AI and machine learning applications. With the volume of published research available growing hourly, access to the RSC's collection in machine-readable format offers companies the opportunity to extract, pinpoint, and apply insights from research that stretches across 160 years, covering the full breadth of chemical science.

Companies are increasingly looking to AI and machine learning to accelerate their R&D projects and need to be able to integrate published research with public and corporate resources. Richard Kidd, Head of Chemistry Data at the Royal Society of Chemistry said: "*It's near impossible for researchers to be sure they've read all the relevant material, let alone set it in context with their companies' knowledge. An important but hidden piece of information, or new connection, could enable new leaps of progress to a programme. TDM can reach into the literature and, like pulling ingredients out of a cake, infer structure and results to integrate with other resources to build knowledge. Being able to do this with our high-quality scientific research opens new possibilities and can significantly enhance large-scale research projects*."

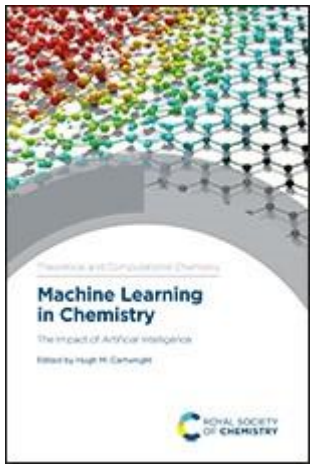
Our research collection delivers this as highly-structured XML, tables, and images – allowing for integration with cross-discipline research projects. Practical licensing tackles one of the key issues raised by users of existing TDM services.
This also allows for integration to machine learning applications, supporting the growth of chemical science knowledge.
Richard continued: "*The Royal Society of Chemistry's Digital Futures report highlighted the significant opportunity presented by smart technology, with AI and machine learning absolutely key to accelerating research and innovation. We are working towards a future in which science can be easily interrogated by machine applications as soon as it is published – making our publications available as XML to our industry customers is the first step to achieving this*."

More information on our Text and Data Mining service can be found at: rsc.li/tdm

Chemical Information / Cheminformatics and Related Books

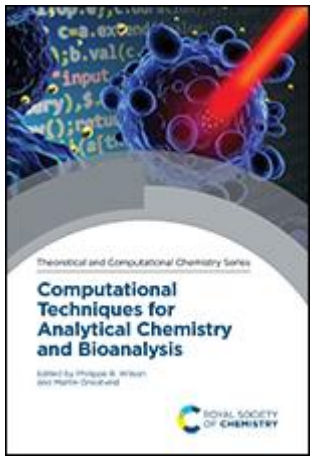
Contributed by RSC CICAG Newsletter Editor Stuart Newbold, email: stuart@psandim.com

Machine Learning in Chemistry: The Impact of Artificial Intelligence



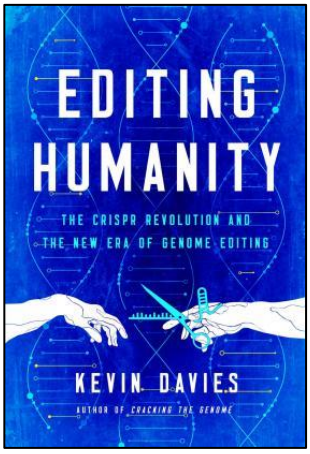
Progress in the application of machine learning (ML) to the physical and life sciences has been rapid. A decade ago, the method was mainly of interest to those in computer science departments, but more recently ML tools have been developed that show significant potential across wide areas of science. There is a growing consensus that ML software, and related areas of artificial intelligence, may, in due course, become as fundamental to scientific research as computers themselves.
Yet a perception remains that ML is obscure or esoteric, that only computer scientists can really understand it, and that few meaningful applications in scientific research exist. This book challenges that view. With contributions from leading research groups, it presents in-depth examples to illustrate how ML can be applied to real chemical problems. Through these examples, the reader can both gain a feel for what ML can and cannot (so far) achieve, and also identify characteristics that might make a problem in physical science amenable to a ML approach.
This text is a valuable resource for scientists who are intrigued by the power of machine learning and want to learn more about how it can be applied in their own field.
RSC Publishing
Editor: Hugh M Cartwright

Computational Techniques for Analytical Chemistry and Bioanalysis



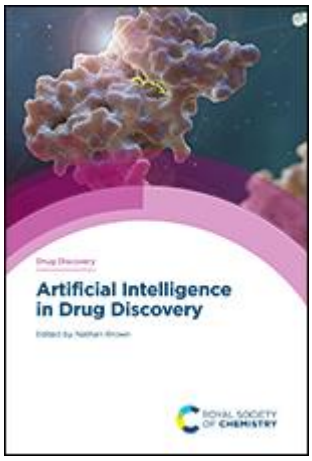
As analysis, in terms of detection limits and technological innovation, in chemical and biological fields has developed so computational techniques have advanced enabling greater understanding of the data. Indeed, it is now possible to simulate spectral data to an excellent level of accuracy, allowing chemists and biologists access to robust and reliable analytical methodologies both experimentally and theoretically.
This work will serve as a definitive overview of the field of computational simulation as applied to analytical chemistry and biology, drawing on recent advances as well as describing essential, established theory. Computational approaches provide additional depth to biochemical problems, as well as offering alternative explanations to atomic scale phenomena. Highlighting the innovative and wide-ranging breakthroughs made by leaders in computational spectrum prediction and the application of computational methodologies to analytical science, this book is for graduates and postgraduate researchers showing how computational analytical methods have become accessible across disciplines. Contributed chapters originate from a group of internationally-recognised leaders in the field, each applying computational techniques to develop our understanding of and supplement the data obtained from experimental analytical science.
RSC Publishing
Editors: Philippe B Wilson, Martin Grootveld

Editing Humanity: The CRISPR Revolution and the New Era of Genome Editing



One of the world's leading experts on genetics unravels one of the most important breakthroughs in modern science and medicine.
If our genes are, to a great extent, our destiny, then what would happen if mankind could engineer and alter the very essence of our DNA coding? Millions might be spared the devastating effects of hereditary disease or the challenges of disability, whether it was the pain of sickle-cell anaemia to the ravages of Huntington's disease.
But this power to "play God" also raises major ethical questions and poses threats for potential misuse. For decades, these questions have lived exclusively in the realm of science fiction, but as Kevin Davies powerfully reveals in his new book, this is all about to change.
Engrossing and page turning, Editing Humanity takes readers inside the fascinating world of a new gene editing technology called CRISPR, a high-powered genetic toolkit that enables scientists to not only engineer but to edit the DNA of any organism down to the individual building blocks of the genetic code.
Davies introduces readers to arguably the most profound scientific breakthrough of our time. He tracks the scientists on the front lines of its research to the patients whose powerful stories bring the narrative movingly to human scale.
Pegasus Books
Author: Kevin Davies

Artificial Intelligence in Drug Discovery



Following significant advances in deep learning and related areas interest in artificial intelligence (AI) has rapidly grown. In particular, the application of AI in drug discovery provides an opportunity to tackle challenges that previously have been difficult to solve, such as predicting properties, designing molecules and optimising synthetic routes. Artificial Intelligence in Drug Discovery aims to introduce the reader to AI and machine learning tools and techniques, and to outline specific challenges including designing new molecular structures, synthesis planning and simulation. Providing a wealth of information from leading experts in the field this book is ideal for students, postgraduates and established researchers in both industry and academia.
RSC Publishing
Editor: Nathan Brown

News from AI3SD

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

AI3SD is on YouTube

AI3SD has established our own YouTube Channel, and would be pleased to have CICAG newsletter readers amongst our rapidly growing subscriber community. Like most organisations, the network has moved online, and the majority of all online events are available on YouTube. There are already over 25 videos and this content will be growing rapidly over the coming months. There two more playlists on our channel in addition to the existing AI3SD Summer Seminar Series:

Failed it to Nailed it: How to get Data Sharing Right Seminar Series

This 'Failed it to Nailed it – Getting Data Sharing Right' series is a series of four events run by the AI3SD, Cell Press Patterns Journal, and the Physical Sciences Data-Science Service (PSDS). These events are a product of the data sharing survey we ran earlier this year. Each event in the series handles a different aspect of dealing with data aiming to educate and inform researchers about how to work well with their data, as well as encouraging discussion along the way. Following on from these events the organisers hope to be able to organise more face-to-face events in 2021 which will expand this event series.

The table below details the event pages, blog posts, video playlists and reports that have been made available so far as part of this event series.

Title	Event Blog Post	Video Playlist	Report
Dealing with Data: Tips & Tricks	Blog post	Playlist	Report
Data Standards	Blog post	Playlist	Report
Responsible Data Management: Legal & Ethical Aspects	Blog post	Playlist	Available soon
Data Citations & Publishing	Blog post	Playlist	Available soon

AI3SD Winter Seminar Series

After the success of our Summer Seminar Series, we have decided to run a Winter Seminar Series starting in November 2020 and running into April 2021. The Series will continue to run on Wednesday afternoons, although rather than running individual talks, we will be running themed sessions of 2-3 talks. Each seminar session will commence at 14:00 and typically finish around 16:45. The full timetable of events can be found here.

Other new Event Blogs & Reports available –

1. AI3SD Network+ Conference Blog Post & Report
2. AIReact2020 Blog Post & Report
3. AI 4 Good @ WebSci'20 Blog Post & Report
4. New Reports from our Funded Projects

Other Chemical Information Related News

Contributed by RSC CICAG Member Dr Keith White and RSC CICAG Newsletter Editor Stuart Newbold

[*All hyperlinks correct & working as of 11 Jan 2021*]

Digital Futures: A new Frontier for Science Exploration and Discovery

Recent advances in AI, robotics, data analysis, modelling and simulation have allowed scientists to augment their research, advancing discovery more quickly, reducing the time it takes to do some tasks in the labs from weeks or months to just hours and identifying patterns and possibilities that humans alone would not see.

The RSC's Digital futures report is a follow-up to Science Horizons, which engaged over 700 academic researchers globally to seek views on key trends and emerging research areas in the chemical sciences and its interfaces. Data and digital technology emerged as one of the main themes. The report sets out to gain a more in-depth understanding of the long-term promise of and concerns about the use of data and digital technologies for scientific discovery by inviting 14 experts from different scientific fields and sectors to our first Strategic Advisory Forum, held in September 2019.

https://www.rsc.org/new-perspectives/discovery/digital-futures

Source: RSC Policy & Perspectives

Elsevier Acquires SciBite

Elsevier has announced that it has acquired SciBite, a semantic AI company headquartered in Cambridge, UK, that helps customers discover insights in life science data through the use of text and data intelligence software. SciBite's software solutions are used to identify and extract scientific insight from structured and unstructured text and content, to identify key concepts such as drugs, proteins, companies, targets, and outcomes. This semantically-enriched, machine-readable data, helps SciBite's customers around the world make streamlined, more efficient decisions.

https://www.scientific-computing.com/news/elsevier-acquires-scibite

Source: Scientific Computing World

Does Tweeting about Research attract more Citations?

Tweeting about your research increases the number of citations it goes on to attract, according to a new analysis. The study analysed the citation patterns of 112 papers in the field of thoracic surgery published between 2017 and 2018 in the Annals of Thoracic Surgery and The Journal of Thoracic and Cardiovascular Surgery.

https://www.chemistryworld.com/news/does-tweeting-about-research-attract-more-citations/4012066.articleSource:/

Source: Chemistry World

CAS Collaborates with MIT on Research to enhance Predictive Chemical Synthesis Planning

CAS, a division of the American Chemical Society, has collaborated with the Massachusetts Institute of Technology (MIT) on research to enhance predictive chemical synthesis planning.

https://coley.mit.edu/news/

Source: MIT Coley Research Group

cOAlition S releases the Journal Checker Tool

cOAlition S has announced the release of the Journal Checker Tool (JCT) in beta. The JCT is a web-based tool that provides clear advice to researchers on how they can comply with their funder's Plan S-aligned Open Access policy when seeking to publish in their chosen journal. During the open testing phase, the community

will have the opportunity to be acquainted with the JCT and provide feedback, so that the tool increases its readiness ahead of the implementation of Plan S in January 2021.
https://www.knowledgespeak.com/news/coalition-s-releases-the-journal-checker-tool/
Source: Knowledgespeak

SciBite releases SciBiteAI Relationship Extraction models
SciBite, the award-winning semantic technology company, has announced the release of its SciBiteAI Relationship Extraction models, which provide the enhanced ability to identify complex relationships within text to further unlock insights from Life Sciences data. Deployed within the recently launched SciBiteAI framework, the deep learning models identify context between terms, such as protein-protein interactions or reporting of drug adverse events. The use of advanced search technologies, such as TERMite, SciBite's Named Entity Recognition engine, have empowered researchers to identify relevant hits from biomedical text. Identifying the relationship between two entities is the key to establishing an additional level of semantic understanding.
https://www.knowledgespeak.com/news/scibite-releases-scibiteai-relationship-extraction-models/
Source: Knowledgespeak

Clarivate Completes Acquisition of CPA Global to Form a Global Intellectual Property Powerhouse
The acquisition of CPA Global creates an end-to-end solution covering the entire intellectual property, science and innovation lifecycle. Clarivate will now offer thousands of law firms and corporate customers world class IP solutions from leading brands covering patent and trademark research solutions, expanded IP services capabilities, IP management and renewal solutions and domain management, all underpinned by extensive human expertise, unparalleled data and powerful technology.
https://clarivate.com/news/clarivate-completes-acquisition-of-cpa-global-to-form-a-global-ip-powerhouse/
Source: Clarivate

Oxford University Press's flagship What Everyone Needs to Know® series available in one digital space for the first time
OUP has released over 100 titles from the What Everyone Needs to Know® series online, giving readers access to them in one digital space for the very first time. The new online version of What Everyone Needs to Know® is designed with our users in mind: easy-to-use search and browse tools allow researchers, lecturers, and students to find the content they need quickly. Integrated technology also makes it easy for readers to share precise content with colleagues and students, facilitating seminar discussions and sparking essay ideas. There is no shelf life for the digital product so users can return to online pages again and again, year after year.
Oxford University Press's flagship What Everyone Needs to Know ® series available in one digital space for the first time | STM Publishing News (stm-publishing.com)
Source: STM Publishing News

ACS Announces Editor-in-Chief for Open Access Journal JACS Au
The Society who plans to add 9 additional open access journals to its suite, announced the appointment of Christopher W. Jones to take the helm of the American Chemical Society's newest fully open access journal, JACS Au. JACS Au (pronounced "JACS Gold") will complement the ACS's flagship Journal of the American Chemical Society by publishing high-impact research in chemistry and related fields through a pay-to-publish, free-to-read model.
https://cen.acs.org/acs-news/publishing/ACS-announces-editorchief-open-access/98/web/2020/08
Source: CEN

Environmental Science: Atmospheres
A new RSC interdisciplinary open access journal advancing the understanding of atmospheric science and related challenges.
Source: RSC Publishing

CAS and MoA Technology Embed Chemical Substance Data into MoA's R&D Workflow Through API
CAS and MoA Technology, an innovator in herbicide R&D and related technology, have announced an agreement to embed chemical substance data from the CAS Registry®, as well as related reference and property information, directly into MoA's R&D workflow via an API. The agreement ensures MoA researchers have access to the most relevant, actionable chemical substance data within their daily workflow to quickly assess activity, enhancing productivity and expediting innovation.
https://www.knowledgespeak.com/news/cas-and-moa-technology-embed-chemical-substance-data-into-moas-rd-workflow-through-api/
Source: Knowledgespeak

Clarivate Releases Web of Science Journal Citation Reports to Identify the World's Leading Journals
the 2020 update to its annual Web of Science Journal Citation Reports™ (JCR). The annual JCR release enables the research community to evaluate the world's high-quality academic journals using a range of indicators, descriptive data and visualisations. The reports are used by academic publishers across the globe to evaluate the impact of their journals relative to their field and promote them to the research community.
https://clarivate.com/news/clarivate-releases-web-of-science-journal-citation-reports-to-identify-the-worlds-leading-journals/
Source: Clarivate

The State of Open Data 2020 – Global Attitudes towards Open Data
New report outlines the impact of the Covid-19 pandemic on research and how this has influenced researchers' data sharing behaviour.
https://www.digital-science.com/press-releases/the-state-of-open-data-2020-global-attitudes-towards-open-data/
Source: Digital Science

Chemistry's Reproducibility Crisis that you've probably never heard of
Computational chemistry faces a coding crisis.
https://www.chemistryworld.com/news/chemistrys-reproducibility-crisis-that-youve-probably-never-heard-of/4011693.article
Source: Chemistry World

Clarivate Introduces Next Generation Insights Platform with Cortellis Generics Intelligence
Cortellis Generics Intelligence powered by Clarivate Research Intelligence Cloud replaces and enhances the well-known Newport platform from Clarivate. It includes new intuitive search capabilities together with expertly curated data on market performance, patents and litigation, and manufacturers to assist finished dose manufacturers, API manufacturers, and pharmaceutical marketers in their pursuit to successfully launch generic drugs and maintain a steady supply chain.
https://clarivate.com/news/clarivate-introduces-next-generation-insights-platform-with-cortellis-generics-intelligence/
Source: Clarivate

Can Computational Chemistry benefit from Blockchain?
Is blockchain a new paradigm in safeguarding science, or does it just tie our hands?
https://www.chemistryworld.com/opinion/can-computational-chemistry-benefit-from-blockchain/4012030.article
Source: Chemistry World

UKeiG's eLucidate Journal Publishes on new Platform
The UK electronic information Group (UKeiG) is now publishing its eLucidate journal on a new web platform, using the Public Knowledge Project's Open Journal Systems (OJS) hosted by the University of Alberta.
https://www.infotoday.eu/Articles/News/Featured-News/UKeiGs-eLucidate-journal-publishes-on-new-platform-141674.aspx

Source: Information Today

Digital Future can be Grasped across Chemistry-Related Industries
We've had a glimpse of a digitally enhanced future through the coronavirus crisis.
https://www.icis.com/explore/resources/news/2020/07/08/10527971/insight-digital-future-can-be-grasped-across-chemistry-related-industries
Source: ICIS

Frost & Sullivan Radar ranks Wolters Kluwer as a top 20 AI Innovation Leader in Healthcare IT
The independent analysis evaluated a field of more than 200 healthcare IT companies and ranked Wolters Kluwer among the top 20 for continuous innovation and growth focussing on areas where AI solutions are most relevant for hospitals, physicians and payers.
https://www.wolterskluwer.com/en/news/frost-sullivan-radar-wolters-kluwer-top-20-ai-innovation-leader
Source: Wolters Kluwer

SCIP Database Tracks Chemicals of Concern in Products
The European Chemicals Agency (ECHA) launched the SCIP database. Companies can now start to submit data on substances of very high concern (SVHCs) in their products to SCIP. The Waste Framework Directive requires companies to submit their data from January 5, 2021. Consumers and waste operators can access and use the data from February 2021.
https://www.chemistryviews.org/details/news/11274764/SCIP_Database_Tracks_Chemicals_of_Concern_in_Products.html
Source: ChemistryViews

Machine Learning Advances Materials for Separations, Adsorption, and Catalysis
An artificial intelligence technique, machine learning, is helping accelerate the development of highly tunable materials known as metal-organic frameworks (MOFs) that have important applications in chemical separations, adsorption, catalysis, and sensing.
https://www.sciencedaily.com/releases/2020/11/201110102536.htm
Source: ScienceDaily

Atos partners with University of Oxford on Largest AI Supercomputer in the UK
Atos, a global leader in digital transformation, has signed a four-year contract worth £5 million with the University of Oxford to deliver a new, state of the art, deep learning supercomputer built on the NVIDIA DGX SuperPOD™ architecture, which will enable UK academics and industry to drive forward scientific discoveries and innovation in machine learning and artificial intelligence, as part of the JADE2 project.
http://www.globenewswire.com/news-release/2020/08/04/2072168/0/en/Atos-partners-with-University-of-Oxford-on-largest-AI-supercomputer-in-the-UK.html
Source: Intrado Global Newswire

Computers Excel in Chemistry Class
Creating computers that can teach themselves how chemical structure dictates the fundamental properties of molecules and then using that knowledge to predict the properties of novel molecules could help to design cleaner energy and industrial systems.
https://www.eurekalert.org/pub_releases/2020-08/kauo-cei082420.php
Source: EurekAlert! Science News

RoboRXN: Automating Chemical Synthesis
Synthetic chemistry, or the art of making materials, remains a most traditional discipline in terms of digitization and acquisition of new technologies. Chemists still rely on many of the same protocols and little progress has been made to modernize the ancient practices of trial and error to enable a new era of accelerated discovery. A dynamic group of scientists at IBM Research Europe set out to change this using modern tools such as artificial intelligence (AI), cloud technology and robotics.

https://www.ibm.com/blogs/research/2020/08/roborxn-automating-chemical-synthesis/
Source: IBM Research Blog

Materials Informatics: Choosing a Successful Strategic Approach
Materials informatics is quickly becoming one of the most significant areas of interest for chemical and material companies. There have been significant early adopters, notable news stories, internationally renowned consortia established, large funding rounds, and still many yet to get started. Navigating a successful path is challenging, a misstep could have a large implication on cost and competitive advantage.
https://www.idtechex.com/en/research-article/materials-informatics-choosing-a-successful-strategic-approach/21576
Source: IDTechEx Research Article

Nature Index Artificial Intelligence supplement investigates emerging trends in AI research output
Escalating computing power, expanding data sets, and algorithms of unprecedented sophistication have led to a massive increase in the number of journal and conference papers referring to Artificial Intelligence (AI) in recent years. The Nature Index AI supplement draws on Nature Index data and the larger Dimensions from Digital Science database to analyse the rapidly advancing and controversial topic. For the first time, the supplement also includes summaries of research articles created using AI, and it looks more broadly at how AI is being used in scholarly publishing.
https://www.knowledgespeak.com/news/nature-index-artificial-intelligence-supplement-investigates-emerging-trends-in-ai-research-output/
Source: Knowledgespeak

Society of Young Publishers (SYP) Expand Mentorship Schemes
The Society of Young Publishers (SYP) is proud to announce the expansion of its mentorship schemes to now cover all areas of the UK and Ireland. As a result, more of those looking to get into or move ahead in the industry than ever before will have the opportunity to receive personalised, structured career guidance from publishing experts.
https://www.alpsp.org/news/syp-expand-mentorship-schemes-sep-2020
Source: ALPSP

Springer Nature Collaborates with the OAPEN Foundation on Open Access (OA) Toolkit for Researchers
Springer Nature has announced it is one of the founding members of the new OAPEN toolkit for researchers and academic book authors. The toolkit is a free-to-access, stakeholder-agnostic resource that aims to help authors better understand open access (OA) for books, to increase trust in OA book publishing, to provide reliable and easy-to-find answers to questions from authors, and to provide guidance in the process of publishing an OA book.
https://www.scientific-computing.com/news/springer-nature-collaborates-oapen-foundation-open-access-oa-toolkit-researchers
Source: Scientific Computing World

Chemists Develop a New Drug Discovery Strategy for 'undruggable' Drug Targets
A research team has developed a new drug discovery method targeting membrane proteins on live cells.
https://www.sciencedaily.com/releases/2020/12/201228095422.htm
Source: ScienceDaily

Physics Societies Unite in Support of OA
Major physics societies, which support physical science researchers with the publication of more than 75,000 peer-reviewed journal articles each year, have joined forces to show their commitment to open access (OA) for physics research.
https://www.researchinformation.info/news/physics-societies-unite-support-oa
Source: Research Information

GII 2020: COVID-19 Pandemic's Expected Impact on Global Innovation; Annual Ranking Topped by Switzerland, Sweden, U.S., U.K. and Netherlands

The COVID-19 pandemic is severely pressuring a long-building rise in worldwide innovation, likely hindering some innovative activities while catalyzing ingenuity elsewhere, notably in the health sector, according to the Global Innovation Index (GII) 2020.

https://www.wipo.int/pressroom/en/articles/2020/article_0017.html

Source: WIPO

IOP Publishing Commits to Adopting Double-Blind Peer Review for all Journals

IOP Publishing (IOPP) is proposing to move all its owned journals to double-blind peer review, making it the first physics publisher to adopt the approach portfolio-wide. The move is part of IOPP's dedication to tackle the significant gender, racial and geographical under-representation in the scholarly publishing process. Double-blind peer review – where the reviewer and author identities are concealed – has the potential to reduce bias with respect to gender, race, country of origin or affiliation which should lead to a more equitable system.

https://ioppublishing.org/news/iop-publishing-commits-to-adopting-double-blind-peer-review-for-all-journals/

Source: IOP Publishing

Clarivate Launches Online Innovation Exchange to Advance Research, Development and Commercialization Activities

New platform features curated ecosystem of partners, enabling life science, healthcare and research professionals across enterprises to access innovative software and data offerings.

https://clarivate.com/news/clarivate-launches-online-innovation-exchange/

Source: Clarivate

Wiley Announces the Acquisition of Hindawi

John Wiley & Sons has announced the acquisition of Hindawi Limited, an innovator in open access (OA) publishing and one of the world's fastest growing scientific research publishers, for a total purchase price of $298 million. The acquisition of Hindawi significantly increases Wiley's position as a global leader in research by adding quality, scale, and growth to the company's open access publishing program.

https://newsroom.wiley.com/press-releases/press-release-details/2021/Wiley-Announces-the-Acquisition-of-Hindawi/default.aspx

Source: Wiley News

OA books 'have Greater Usage and Higher Citations'

Open access (OA) books are reaching more countries and have greater usage and higher citation numbers than non-OA books. A new analysis collaboratively produced by Springer Nature and COARD (Collaborative Open Access Research & Development) presents these and other key findings in a new white paper that explores how OA affects the geographical diversity of readers.

https://www.researchinformation.info/news/oa-books-have-greater-usage-and-higher-citations

Source: Research Information

Special Issue on Insights Gained While Teaching Chemistry in the Time of COVID-19

The global COVID-19 pandemic has greatly challenged teachers' abilities to interact with students in a classroom or laboratory settings. The response of chemistry educators to rapid changes in teaching modalities that resulted from the disruption associated with the global COVID-19 pandemic is chronicled in a wide range of papers in this special issue from the Journal of Chemical Education. This special issue provides a nearly real-time snapshot of ideas and approaches to the rapid changes effecting chemistry teaching and their outcomes. Learn from the experiences of educators worldwide who are persevering through the disruptions caused by COVID-19.

https://pubs.acs.org/toc/jceda8/97/9

Source: ACS

Artificial Chemist 2.0: Quantum Dot R&D in less than an Hour
A new technology, called Artificial Chemist 2.0, allows users to go from requesting a custom quantum dot to completing the relevant R&D and beginning manufacturing in less than an hour. The tech is completely autonomous, and uses artificial intelligence and automated robotic systems to perform multi-step chemical synthesis and analysis.
https://www.sciencedaily.com/releases/2020/12/201211083041.htm
Source: ScienceDaily

Wiley becomes First Global Publisher to Join AAPA
John Wiley & Sons Inc., a global leader in research and education, has been elected by members of the Audiovisual Anti-Piracy Alliance (AAPA) to join its alliance. The first global publisher to join the AAPA, Wiley shares the organisation's commitment to strengthen the fight against online piracy and the protection of intellectual property rights in Europe, the Middle East and around the globe. AAPA's mission is to lead the fight against audiovisual piracy through effective lobbying, supporting law enforcement and building partnerships to tackle piracy. In partnership with Wiley, they aim to continue raising awareness to discourage piracy; building private and public partnerships; lobbying for better legislation and enforcement; sharing content protection strategies; and providing a unique perspective from the publishing industry on today's piracy challenges.
https://www.knowledgespeak.com/news/wiley-becomes-first-global-publisher-to-join-aapa/
Source: Knowledgespeak

Elsevier and Pending.AI collaborate on AI-driven Chemistry Retrosynthesis Tool
New AI leverages high quality, experimental reactions featured in Reaxys' chemistry database, will allow chemists to increase the success rate of synthetic chemistry.
https://www.prnewswire.co.uk/news-releases/elsevier-and-pending-ai-collaborate-on-ai-driven-chemistry-retrosynthesis-tool-893138713.html
Source: PR Newswire

Sliced, Diced and Digested: AI-generated Science Ready in Minutes
AI can decide which papers are worth reading, and condenses them to make the literature more accessible.
https://www.natureindex.com/news-blog/sliced-diced-and-digested-ai-generated-science-ready-in-minutes
https://www.nature.com/articles/d41586-020-03415-w
Source: Nature News

Knowledge Resources
This collection of documents provides access to e-learning and knowledge resources produced by the Government Chemist team alone or in collaboration with others.
https://www.gov.uk/government/collections/knowledge-resources
Source: UKRI

Automatic Database Creation for Materials Discovery: Innovation from Frustration
A collaboration between the University of Cambridge and Argonne has developed a technique that generates automatic databases to support specific fields of science using AI and high-performance computing.
https://www.anl.gov/article/automatic-database-creation-for-materials-discovery-innovation-from-frustration
Source: Argonne National Laboratory

BenchSci Expands Scientists' Access to Critical Experimental Insights
BenchSci has announced an agreement to analyse world-class research from Taylor & Francis Group's top scientific journals using advanced biomedical artificial intelligence. This will facilitate scientists' access to critical experimental insights, empowering them to run more successful experiments and bring new medicine

to patients faster. Over the last decade, pharmaceutical research & development costs have been steadily increasing, while the rate of drug approval has remained roughly constant. In addition to costing more, drug development is taking longer, increasing from about 9.7 years in the 1990s to 10-15 years now. This means it takes longer to bring new treatments to patients. One underlying cause that is receiving increasing attention is Avoidable Experiment Expenditure (AEE). This refers to inefficiencies and productivity challenges in designing and carrying-out preclinical experiments. Experiments are the foundation of preclinical research and development, but irreproducibility rates in preclinical experiments exceed 50%. To address the issue, BenchSci uses advanced biomedical artificial intelligence to analyse published scientific papers and related data sources, understand methodology and results, and use this to help scientists design more successful experiments. For example, BenchSci's AI-Assisted Reagent Selection helps scientists select appropriate reagents, avoiding a cause of more than 36% of irreproducibility.
https://www.knowledgespeak.com/news/benchsci-expands-scientists-access-to-critical-experimental-insights/
Source: Knowledgespeak

How to Digitise your Lab Notebooks
Converting paper records to digital formats provides secure back-ups that researchers can access from anywhere.
https://www.nature.com/articles/d41586-020-02728-0
Source: Nature

Research Square Platform Celebrates Publication of Its 50,000th Preprint
Research Square, a multidisciplinary publishing platform for preprints, has surpassed 50,000 submissions since its debut in October 2018. Preprints, which are scholarly manuscripts published before they are formally peer reviewed, exploded at the start of the COVID-19 pandemic, when the science and publishing community began using preprint servers to share large volumes of research as quickly and broadly as possible.
https://www.stm-publishing.com/research-square-platform-celebrates-publication-of-its-50000th-preprint/
Source: STM Publishing News

How Trump damaged Science — and why it could take Decades to Recover
The US president's actions have exacerbated the pandemic that has killed more than 200,000 people in the United States, rolled back environmental and public-health regulations and undermined science and scientific institutions. Some of the harm could be permanent.
https://www.nature.com/articles/d41586-020-02800-9
Source: Nature

Machine-generated Content: Boon to Information Professionals or the end of the World as we know it?
Advances in machine-generated content are changing scholarly publishing and Springer Nature is in the forefront of adopting its usage. What does this mean for researchers, scholarly publishing and librarians?
https://www.infotoday.eu/Articles/News/Featured-News/Machine-generated-content-Boon-to-Information-Professionals-or-the-end-of-the-world-as-we-know-it-144522.aspx
Source: Information Today

Nobel Prizes have a Diversity Problem even worse than the Scientific Fields they Honor
This is a problem much larger than simply bias on the part of the Nobel selection committees – it's systemic.
https://www.natureindex.com/news-blog/nobel-prizes-have-a-diversity-problem-even-worse-than-the-scientific-fields-they-honor
Source: Nature News

Cerner acquires Kantar Health
Cerner Corporation has agreed to acquire Kantar Health, a division of Kantar Group, for $375 million in cash, subject to adjustment. Kantar Health is a leading data, analytics and real-world evidence and commercial research consultancy serving the life science industry. With the acquisition, Cerner plans to harness data to improve the safety, efficiency and efficacy of clinical research across life sciences, pharmaceuticals and health

care at large. The acquisition is expected to allow Cerner's Learning Health Network SM client consortium to more directly engage with life sciences for funded research studies.
https://www.knowledgespeak.com/news/cerner-acquires-kantar-health/
Source: Knowledgespeak

Cochrane Library Adopts Editorial Manager® and ProduXion Manager® to Manage Publishing Workflows
Aries Systems Corporation, a leading technology workflow solutions provider to the scholarly publishing community, is pleased to announce the adoption of Editorial Manager® (EM) and ProduXion Manager® (PM) by the Cochrane Library. The Cochrane Library is a prestigious collection of databases comprised of high-quality, independent research evidence to promote informed healthcare knowledge and decision-making.
https://www.stm-publishing.com/cochrane-library-adopts-editorial-manager-and-produxion-manager-to-manage-publishing-workflows/
Source: STM Publishing News

Google Spam Report
Google's latest annual spam report said its preventative spam efforts resulted in over 99% of search results being spam-free and that it found 25 billion pages discovered every day are spammy.
https://www.infotoday.eu/Articles/News/Featured-News/Google-Spam-Report-141652.aspx
Source: Information Today

Springer Nature unveils Alternative OA Route
All authors submitting to Nature and the Nature research journals will have the option to publish open access from January.
https://www.researchinformation.info/news/springer-nature-unveils-alternative-oa-route
Source: Research Information

Robot Swarms Follow Instructions to Create Art
Controlling a swarm of robots to paint a picture sounds like a difficult task. However, a new technique allows an artist to do just that, without worrying about providing instructions for each robot. Using this method, the artist can assign different colors to specific areas of a canvas, and the robots will work together to paint the canvas. The technique could open up new possibilities in art and other fields.
https://blog.frontiersin.org/2020/10/14/robotics-ai-collaborative-robot-swarm-creates-art-painting/
Source: Frontiers Science News

Machine Learning with Limited Data
A guide to machine learning techniques for limited data problems, including approaches for small amounts of data and for large amounts of unlabelled data.
https://www.gov.uk/government/publications/machine-learning-with-limited-data
Source: UKRI

PLOS Announces Peer-Reviewed Article Types for PLOS ONE
The Public Library of Science (PLOS) and protocols.io have announced an extension to their partnership to support increased sharing of open research methodologies. In early 2021 PLOS is launching two new peer-reviewed article types in PLOS ONE: Lab Protocols and Study Protocols. The new article types are intended to address three issues familiar to researchers: the rigor and reproducibility of research, efficiency in getting feedback, and recognition for developing and sharing diverse research contributions.
https://www.knowledgespeak.com/news/plos-announces-peer-reviewed-article-types-for-plos-one/
Source: Knowledgespeak

When is a Scientific Collaboration Unfair?
A study enlists the h-index to try to find out why some research partnerships fizzle.
https://www.natureindex.com/news-blog/scientific-research-collaboration-unfair-h-index
Source: Nature News

£20 Million Boost for World Class AI Research could Transform Cancer Treatment and Save Lives
Turing AI Acceleration Fellowships will give 15 of the UK's top AI innovators the resources to drive forward their ground-breaking research.
https://www.gov.uk/government/news/20-million-boost-for-world-class-ai-research-could-transform-cancer-treatment-and-save-lives
Source: UKRI

Elsevier Expands Open Access Options for Cell Press Journals from January 2021
Elsevier has announced that the Cell Press portfolio of journals will be expanding open access publishing options for authors from 1 January 2021. Open access is an integral part of Elsevier's commitment to a more collaborative, inclusive, and transparent world of research where authors, researchers, and academic institutions can share knowledge and build on each other's work to advance outcomes. As one of the fastest-growing open access publishers in the world, nearly all of Elsevier's 2,600 journals now enable open access publishing, including 500 fully open access journals.
Elsevier Expands Open Access Options for Cell Press Journals from January 2021 | STM Publishing News (stm-publishing.com)
Source: STM Publishing News

The National Academy of Sciences to move PNAS to Atypon's Online Platform Literatum
The NAS will move PNAS, the official journal of the National Academy of Sciences, to Atypon's online publishing platform, Literatum, in 2021. The move will encompass the entire PNAS archive, with articles dating back to 1915, as well as PNAS Front Matter, the Science Sessions podcast, the Journal Club blog, Special Features and Colloquia, and the journal's wealth of Profiles, Commentaries, Perspectives, and other nonresearch content.
https://www.knowledgespeak.com/news/the-national-academy-of-sciences-to-move-proceedings-of-the-national-academy-of-sciences-to-atypons-online-platform-literatum/
Source: Knowledgespeak

Artificial-Intelligence Research Escalates amid calls for Caution
This supplement explores artificial intelligence (AI), one of the most rapidly advancing and controversial topics in scientific research.
https://www.natureindex.com/news-blog/artificial-intelligence-research-escalates-amid-calls-for-caution
Source: Nature News

EMBO Launches Early Evidence Base Site
EMBO is launching the Early Evidence Base site, an experimental platform that blends human scientific expertise with artificial intelligence (AI) to highlight scientific findings posted in preprints. Preprints are manuscripts that are publicly shared online by researchers before formal publication in a peer-reviewed journal.
https://www.knowledgespeak.com/news/embo-launches-early-evidence-base-site/
Source: Knowledgespeak

Four AI Technologies that could Transform the way we Live and Work
From facial recognition to drug discovery, these emerging technologies are the ones to watch.
https://www.natureindex.com/news-blog/four-ai-technologies-that-could-transform-the-way-we-live-and-work
Source: Nature News

Cambridge University Press and Jisc reach UK-wide Flexible Open Access Read and Publish Agreement
- Institutions with single subscription receive upgrade to access full collection
- Menu of Read and Publish options to meet different publishing profiles, requirements and budgets.
- Flexible, inclusive agreement means institutions can join and transition at their own pace, but none are left behind in transition to open access

Cambridge University Press (CUP) and Jisc, the education and research not for profit, have reached a UK-wide open access (OA) agreement that offers a range of flexible Read and Publish options to all UK institutions. Institutions who did not previously take the complete collection, will receive a full upgrade to the complete collection for no additional fee.
Cambridge University Press and Jisc reach UK-wide flexible open access read and publish agreement | STM Publishing News (stm-publishing.com)
Source: STM Publishing News

Paul Jagodzinski selected as chair of American Chemical Society's board of directors
Dr. Paul Jagodzinski has been selected as chair of the board of directors of the ACS, a scientific membership organisation with more than 152,000 members, from January 1, 2021. As chair of the Society's 16-member chief governing body, Jagodzinski will preside over board and executive committee meetings, appoint chairs and members of board committees and task forces, oversee the performance of the chief executive officer, and support strategic planning and evaluation of progress toward the goals, among other duties. Jagodzinski is a professor of chemistry at Northern Arizona University in Flagstaff. He earned a bachelor's degree from the Polytechnic Institute of Brooklyn in 1973. He has been a member of ACS since 1976.
https://www.knowledgespeak.com/news/paul-jagodzinski-selected-as-chair-of-american-chemical-societys-board-of-directors/
Source: Knowledgespeak