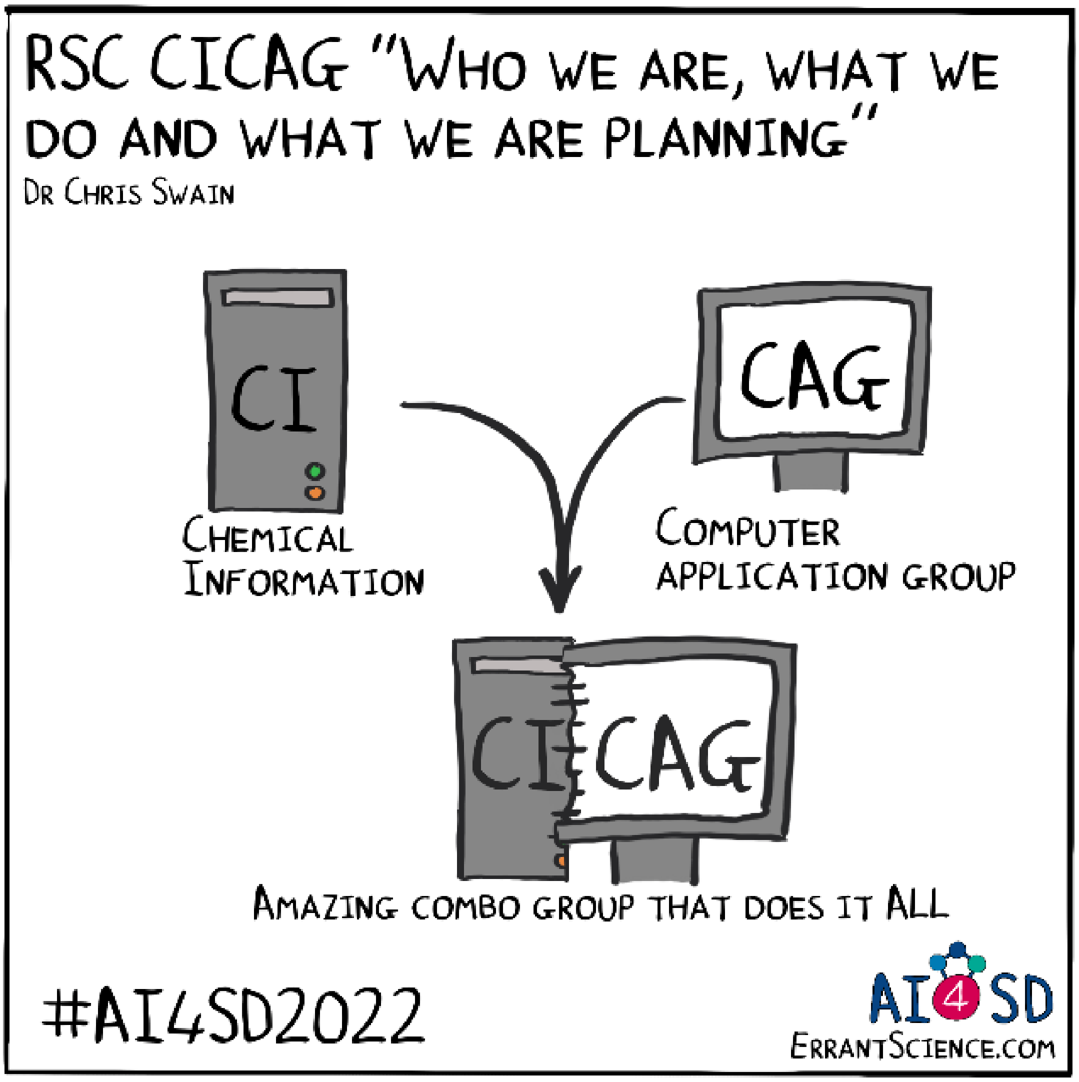
RSC INTEREST GROUP
CHEMICAL INFORMATION AND
COMPUTER APPLICATIONS GROUP

NEWSLETTER

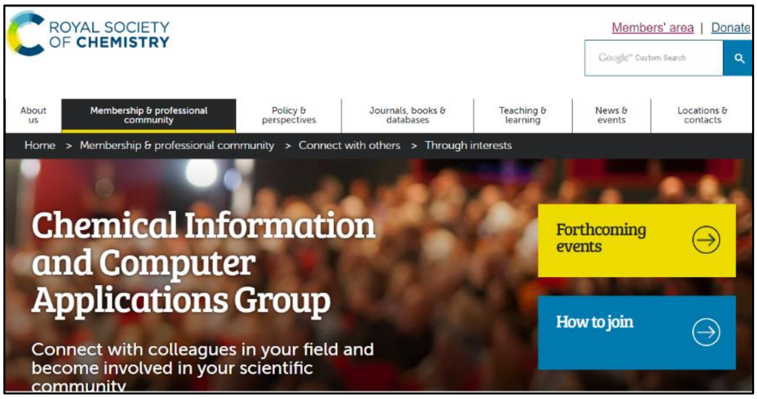
Summer 2022

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.


RSC CICAG "WHO WE ARE, WHAT WE DO AND WHAT WE ARE PLANNING"
DR CHRIS SWAIN
CI
CAG
CHEMICAL INFORMATION
COMPUTER APPLICATION GROUP
CI CAG
AMAZING COMBO GROUP THAT DOES IT ALL
#AI4SD2022
AI4SD
ERRANTSCIENCE.COM

Cartoon accompanying the talk given by Chris Swain, CICAG Chair, at the AI4SD conference 1-3 March 2022. Image credit: ErrantScience.

CICAG Websites and Social Media

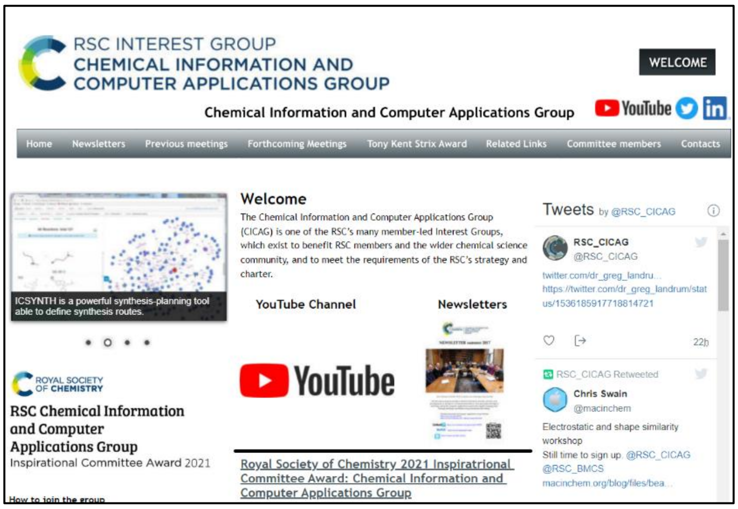


Home > Membership & professional community > Connect with others > Through interests

Chemical Information and Computer Applications Group

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

http://www.rsc.org/CICAG



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

RSC Chemical Information and Computer Applications Group
Inspirational Committee Award 2021

How to join the group

Welcome

The Chemical Information and Computer Applications Group (CICAG) is one of the RSC's many member-led Interest Groups, which exist to benefit RSC members and the wider chemical science community, and to meet the requirements of the RSC's strategy and charter.

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

Tweets by @RSC_CICAG

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

22h

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

http://www.rsccicag.org

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

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

@RSC_CICAG
https://twitter.com/RSC_CICAG

Contents

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

Chemical Information and Computer Applications Group Chair's Report

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

There have been a few changes to roles on the CICAG committee, Neil Berry is taking over from Diana Leitch as Treasurer, Gillian Bell is taking on the role of Secretary from Jeremy Frey and Helen Cooke is the new Editor of the Newsletter, taking over from Stuart Newbold. I'm always conscious that all committee members are volunteers and I am really grateful that they are willing to give up their time. I'm also delighted to be able to introduce three new members to the CICAG Committee, following the ballot organised by the RSC Networks team; the three nominees elected are: Nessa Carson, Hannah Bruce Macdonald and Willem van Hoorn. Also many thanks to all the other people who volunteered to be on the committee, I think it underlines we have a thriving and engaged community.

As events slowly return to normality CICAG are starting to plan several in-person events. The first is a meeting on Ultra Large Chemical Libraries (10 August) at Burlington House followed by the 5th RSC-BMCS/RSC-CICAG Artificial Intelligence in Chemistry (1-2 September) meeting at Churchill College Cambridge. Registration for both meetings is now open.

Social media became an increasingly important way for communicating with members (and non-members) during lockdown and the trend continues, with @RSC_CICAG Twitter now having 1412 followers, and LinkedIn 568 members. The CICAG website is often updated and we would be very interested to receive suggestions for additional content for all communication channels.

CICAG's YouTube channel now has 755 subscribers and contains the 13 video presentations from AI4proteins meetings, in addition to all 18 of the Open-Source Tools for Chemistry workshops. These workshop videos are very popular and have been watched a nearly 17,000 times. At the time of writing this report two more workshops are planned.

CICAG has continued to collaborate with the AI3SD Network+, holding several joint meetings in 2021 and taking part in the AI4SD Network+ Conference in March 2022.

This Newsletter includes contributions from Alpha Lee, describing cheminformatics/compchem work on SAR-Cov-2 and a description of the early days of cheminformatics at Sheffield University from Peter Willett.

CICAG came into existence in 2007 with the merger of the Chemical Information Group and the Computer Applications Group. Currently CICAG membership stands at 667. This figure is increasing at around 60-70 new members per year; the 162 overseas members come from 42 different countries. Many of the UK CICAG members are from the Cambridge-Oxford-London triangle but we are seeing an increase in membership in other areas. Whilst RSC members can join interest groups for free, in practice many members do not take up this opportunity. You can make a request to join a group via email (membership@rsc.org), telephone (01223 432141) or via the RSC's website.

Once again, I'd like to invite contributions to the CICAG Newsletter that would be of interest to the CICAG community. Please contact the Newsletter Editor, Helen Cooke, or me to discuss your ideas.

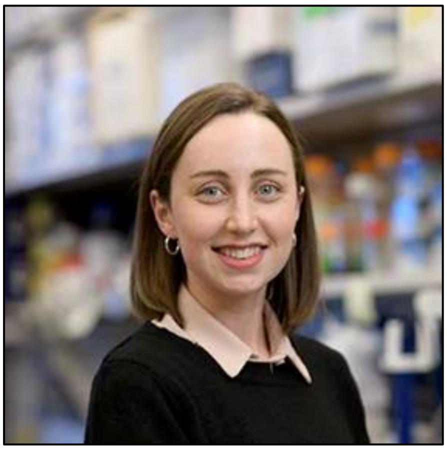
- -

Your CICAG Committee - Introducing Our New Members

CICAG Committee members present at our Zoom meeting on 24 June 2022 (Nathan Brown and Diana Leitch were unable to attend on this occasion).

In June 2022 we were delighted to welcome three new members to the CICAG Committee.

Nessa Carson was born in Warrington, UK. She received her MChem degree from Oxford University, before completing postgraduate studies in catalysis and organic methodology at the University of Illinois at Urbana-Champaign. After her studies, she started in industry at AMRI, initially as a synthetic chemist for AMRI, and then running the high-throughput automation facility on behalf of Eli Lilly in Windlesham, UK, working across both discovery and process chemistry. After this, she moved to process development using high-throughput reaction optimisation and other automation-based experimentation at Pfizer. Nessa started at Syngenta in 2020, where she works with automation, reaction optimisation, and data management, currently holding the title of Automated Data Workflow Specialist. In 2021, she was the recipient of the Salters' Institute Centenary Award for early career chemists with the potential to make a long-term contribution to industrial chemistry. She maintains a website of useful chemistry resources.



Hannah Bruce Macdonald completed her PhD in Jonathan Essex's group at the University of Southampton in 2018, before joining John Chodera's group at the Sloan Kettering Institute, followed by joining MSD in 2020 as a computational chemist supporting drug-discovery projects. Her interests include free-energy methods, molecular-dynamics simulations and medicinal chemistry. Hannah is a co-opted member of the Molecular Graphics and Modelling Society (MGMS) and a co-organiser of the Binding Site group for under-represented groups in computational chemistry.

Willem van Hoorn had a misfire at the beginning of his career when he realised halfway through his chemical engineering degree in the Netherlands that he had chosen the wrong subject, at which time switching was no longer possible. After a PhD in computational chemistry studying organic molecules in organic solvents and a postdoc doing Monte Carlo simulations on similar systems, he joined Pfizer Sandwich in 1999 as a computational chemist not knowing the 20 natural amino acids. At that time data sets generated by combinatorial library design and high-throughput screening became large enough to overwhelm Excel but cheminformatics in the guise of Pipeline Pilot came to the rescue. This finally was the beginning of a career. After ten years at Pfizer and a brief period at Accelrys/Biovia he joined Exscientia in 2013 working on cheminformatics-based tools and active learning. Willem still has a cheat sheet with the 20 amino acids!

- -

CICAG Planned and Proposed Future Meetings

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

Meeting	Date	Location	Further Information
Open-Source Software Workshops	2022	Virtual	An on-going series of workshops
Ultra Large Chemical Databases	10 Aug 2022	London, UK	Organised by CICAG
5th Artificial Intelligence in Chemistry Meeting	1-2 Sep 2022	Cambridge, UK	Joint event from RSC-CICAG and RSC-BMCS division
CICAG/AI3SD Webinars	Autumn 2022	Virtual/TBD	Joint event from RSC-CICAG and AI3SD
Computational Tools for Drug Discovery	23 Nov 2022	The Studio, Birmingham,UK	Joint event from RSC-CICAG and SCI
Python for Chemists	TBD	TBD	Details to follow
Chemical Information during COVID	TBD	TBD	Details to follow
Centenary of Markush Structures	2023	TBD	Details to follow

- -

Free Workshops on Open-Source Tools for Chemistry

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

All workshops are recorded and can be viewed on YouTube.

PDBe Knowledge Base

David Armstrong, EMBL-EBI, Cambridge, UK

This workshop explores the Protein Data Bank in Europe Knowledge Base resource and its tools for the investigation, analysis, and interpretation of biomacromolecular structures. PDBe-KB brings together data from all PDB entries and displays this as aggregated information for individual proteins, including ligand binding sites, macromolecular interactions and more. Furthermore, this community-led resource brings together structural and functional information from a host of other related resources.

In this workshop, you will learn how to use the PDBe-KB aggregated views for proteins to investigate structural and function information for proteins and their associated ligands. We will also demonstrate effective use of novel visualisation components of large-scale structural data on these pages, including 3D visualisation of superposed protein structures with their bound ligands.

KLIFS a kinase database

Albert Jelke Kooistra, Copenhagen University, and Andrea Volkamer, Charité-Universitätsmedizin, Berlin

KLIFS is a kinase database that dissects experimental structures of catalytic kinase domains and the way kinase inhibitors interact with them. The KLIFS structural alignment enables the comparison of all structures and ligands to each other. Moreover, the KLIFS residue numbering scheme capturing the catalytic cleft with 85 residues enables the comparison of the interaction patterns of kinase inhibitors, for example, to identify crucial interactions determining kinase-inhibitor selectivity. The workshop will be in two segments: 1) an introduction to KLIFS and 2) programmatic access and applications of KLIFS.

ESP-Sim shape and electrostatics mapping

Esther Heid, Technical University of Vienna

Electrostatic effects along with volume restrictions play a major role in enzyme and receptor recognition. Evaluating electrostatic and shape similarities of pairs of molecules such as proposed versus known ligands can therefore be valuable indicators of prospective binding affinities. This workshop will demonstrate how to compute electrostatic and shape similarities using the open-source tool ESP-Sim. Available options for comparing electrostatics will be discussed interactively on selected examples of public datasets, along with advice on embedding and aligning molecules prior to computing similarities. All code is available on GitHub.

- -

The COVID Moonshot

Contribution from Alpha A. Lee, PostEra, email: alpha.lee@postera.ai, and Matthew C. Robinson, PostEra, email: matthew.robinson@postera.ai

COVID Moonshot is an open-science drug-discovery campaign targeting SARS-CoV-2 main protease. The SARS-CoV-2 genome encodes two polyproteins and four structural proteins. The polyproteins are cleaved by the cysteine proteases nsp5-Mpro (11 cleavages) and nsp3-PLpro (three cleavages) to liberate shorter viral proteins crucial for viral replication.

With $1M of philanthropic donations and bootstrapped academic grants, and in 18 months, COVID Moonshot built a community resource comprising:

- >500 ligand-bound X-ray structures
- >10,000 assay measurements
- >2,400 synthesised compounds
- Preclinical candidates that are in IND-enabling studies funded by a $11M grant from the Wellcome Trust

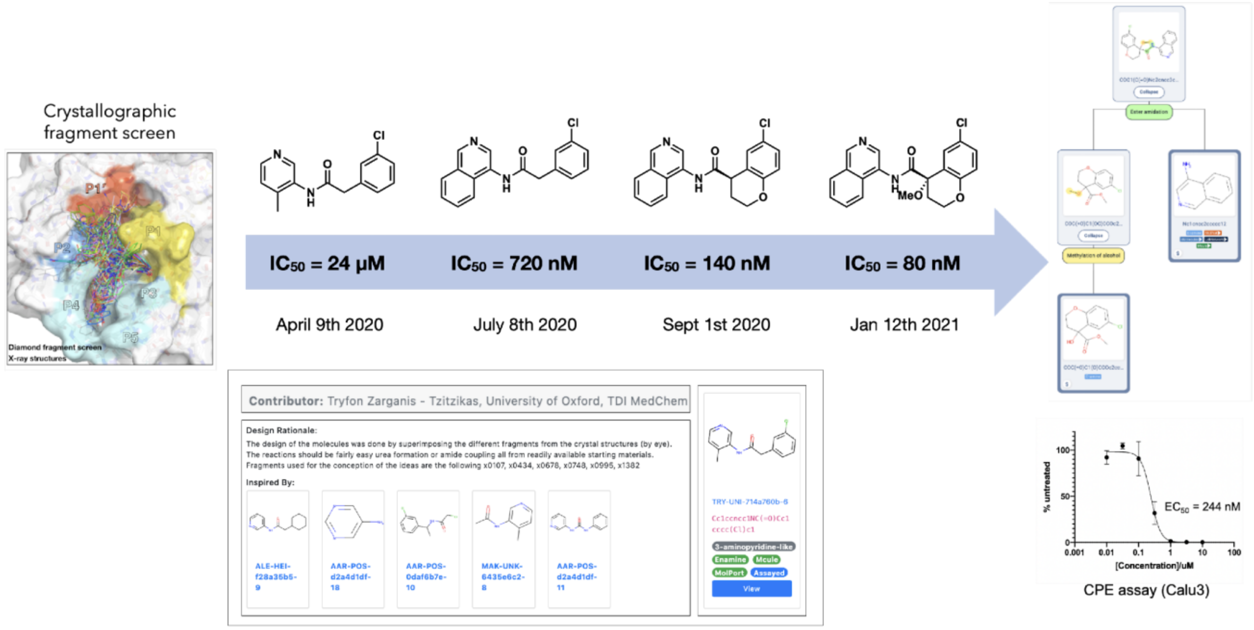
Our starting point was a rapid crystallographic fragment screen that assessed 1,495 fragment-soaked crystals screened within weeks to identify 78 hits that densely populated the active site. This dataset was posted online on 18 March 2020, only days after the screen was completed. The non-covalent fragment hits did not show detectable inhibition. However, they provided a high-resolution map of key interactions that optimised compounds may exploit to inhibit Mpro.

Crowd sourcing drug discovery

We launched an online crowd-sourcing platform on 18 March 2020 inviting participants to submit compounds designed based on the fragment hits. Data from biochemical assays and X-ray crystallography were released rapidly on the same platform, enabling contributing designers to build on all available data, as well as designs contributed by others.

To ensure there would be no delays in ultimately delivering potential drug candidates straight to generics manufacture due to IP licensing issues, all designers were asked to contribute their designs directly into the public domain. Every design and related experimental data were immediately disclosed online and made openly available explicitly free of IP restrictions.

This aggressive open-science policy enabled contributors from multiple fields in both academia and industry to share their ideas freely. Within the first week, we received over 2,000 submissions, representing a diverse set of design strategies. Our initial hit-to-lead strategy focused on compact ligand-efficient designs, heavily triaging based on synthetic complexity forecasted by machine-learning algorithms (detailed below).



Contributor: Tryfon Zarganis - Tzitzikas, University of Oxford, TDI MedChem

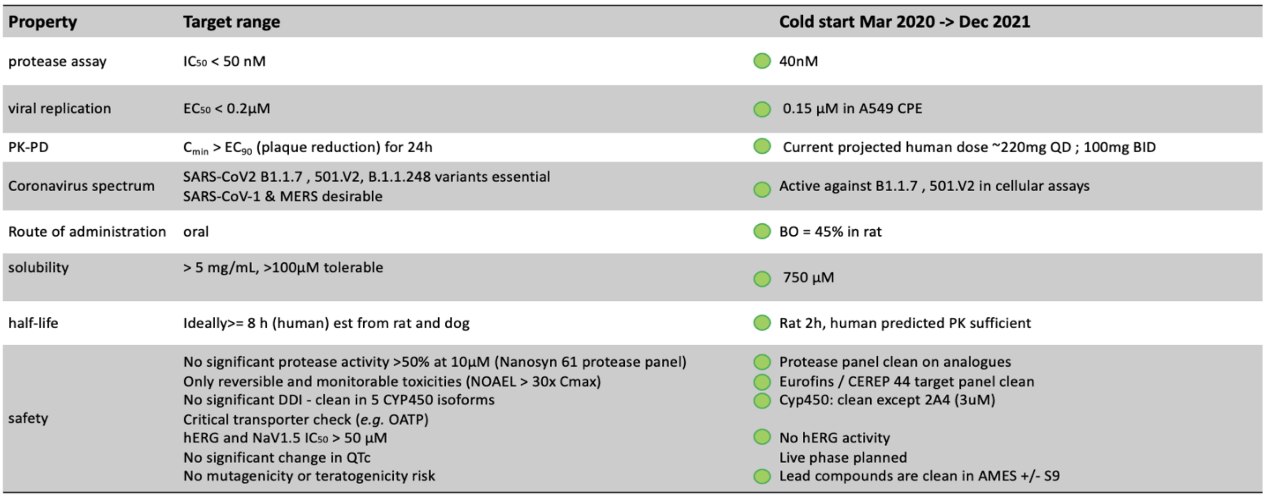
Design Rationale:
The design of the molecules was done by superimposing the different fragments from the crystal structures (by eye). The reactions should be fairly easy urea formation or amide coupling all from readily available starting materials. Fragments used for the conception of the ideas are the following x0107, x0434, x0678, x0748, x0995, x1382

Inspired By:
ALE-HEI-f28a35b5-9
AAR-POS-d2a4d1df-18
AAR-POS-0daf6b7e-10
MAK-UNK-6435e6c2-8
AAR-POS-d2a4d1df-11

TRY-UNI-714a760b-6
Cc1ccncc1NC(=O)Cc1cccc(Cl)c1
3-aminopyridine-like
Enamine Mcule
MolPort Assayed
View

Gratifyingly, many submissions exploited spatially overlapping fragment hits. The submission TRY-UNI-714a760b-6 was inspired by five overlapping fragments, furnishing a non-covalent inhibitor with a SARS-CoV-2 Mpro enzymatic IC_{50} of 24 µM. The next potency jump was a design based on substituting the pyridine with an isoquinoline. Following that, rigidifying the scaffold with a benzopyran further increased potency, without any change in synthetic complexity (one-step synthesis). The final potency jump, introducing a quaternary methoxy, led to a compound with measurable cellular antiviral efficacy.

Our medicinal chemistry journey from lead to candidate nomination is discussed in greater detail in our preprint.[1] Our preclinical candidates satisfy our desired Target Product Profile.



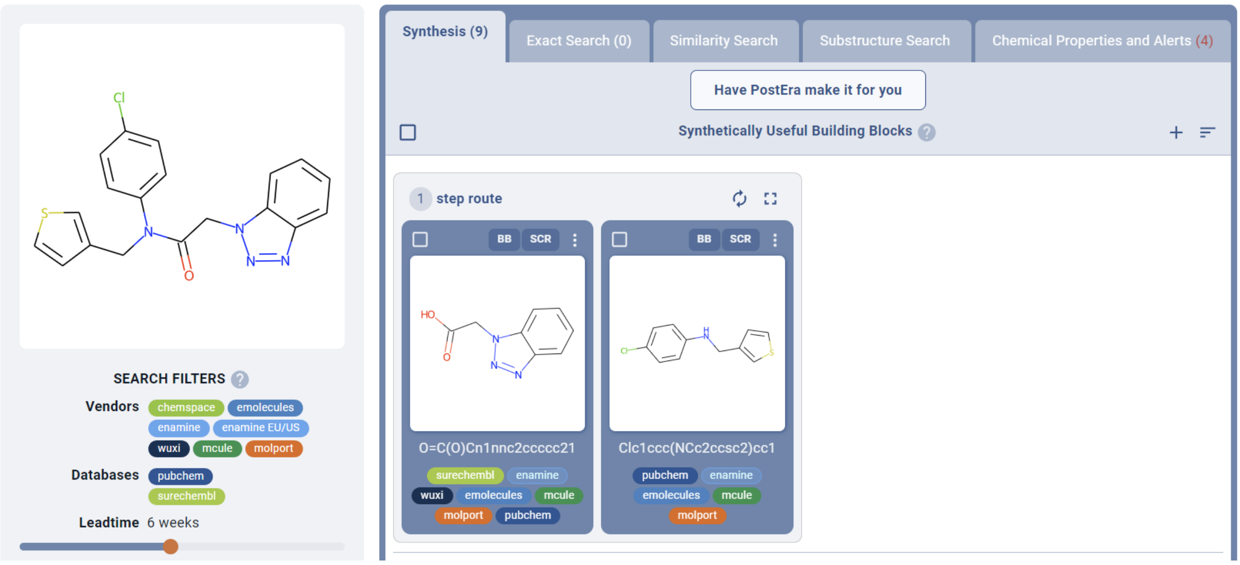
Property	Target range	Cold start Mar 2020 -> Dec 2021
protease assay	IC_{50} < 50 nM	🟢 40nM
viral replication	EC_{50} < 0.2µM	🟢 0.15 µM in A549 CPE
PK-PD	C_{min} > EC_{90} (plaque reduction) for 24h	🟢 Current projected human dose ~220mg QD ; 100mg BID
Coronavirus spectrum	SARS-CoV2 B1.1.7 , 501.V2, B.1.1.248 variants essential SARS-CoV-1 & MERS desirable	🟢 Active against B1.1.7 , 501.V2 in cellular assays
Route of administration	oral	🟢 BO = 45% in rat
solubility	> 5 mg/mL, >100µM tolerable	🟢 750 µM
half-life	Ideally>= 8 h (human) est from rat and dog	🟢 Rat 2h, human predicted PK sufficient
safety	No significant protease activity >50% at 10µM (Nanosyn 61 protease panel) Only reversible and monitorable toxicities (NOAEL > 30x Cmax) No significant DDI - clean in 5 CYP450 isoforms Critical transporter check (*e.g.* OATP) hERG and NaV1.5 IC_{50} > 50 µM No significant change in QTc No mutagenicity or teratogenicity risk	🟢 Protease panel clean on analogues 🟢 Eurofins / CEREP 44 target panel clean 🟢 Cyp450: clean except 2A4 (3uM) 🟢 No hERG activity Live phase planned 🟢 Lead compounds are clean in AMES +/- S9

AI-enabled synthesis driven design

Due to the nature of the project, extensive synthesis prioritisation was needed in order to operate efficiently during a global lockdown, employing multiple contract research organisations across the world. In particular, a detailed understanding of the compound inventories at every synthesis site, as well as the synthetic investment of each design, was crucial to reduce drug-discovery cycle times. To this purpose, we used the PostEra Manifold platform to aid with high-throughput retrosynthetic analysis of prospective designs, as well as ultrafast search across large chemical spaces such as Enamine REAL.

PostEra Manifold, using the underlying Molecule Transformer technology we previously developed[2] augmented by significant engineering efforts to create an extensible cloud-based architecture, was able to score thousands of molecules in a day. All in all, we could find the most accessible crowd-sourced designs out of thousands of submissions in mere hours, when a quote from a CRO would have taken several weeks. Additionally, every submission with molecules found to be available in large (>10 billion) molecule virtual spaces, were automatically annotated with ordering information.

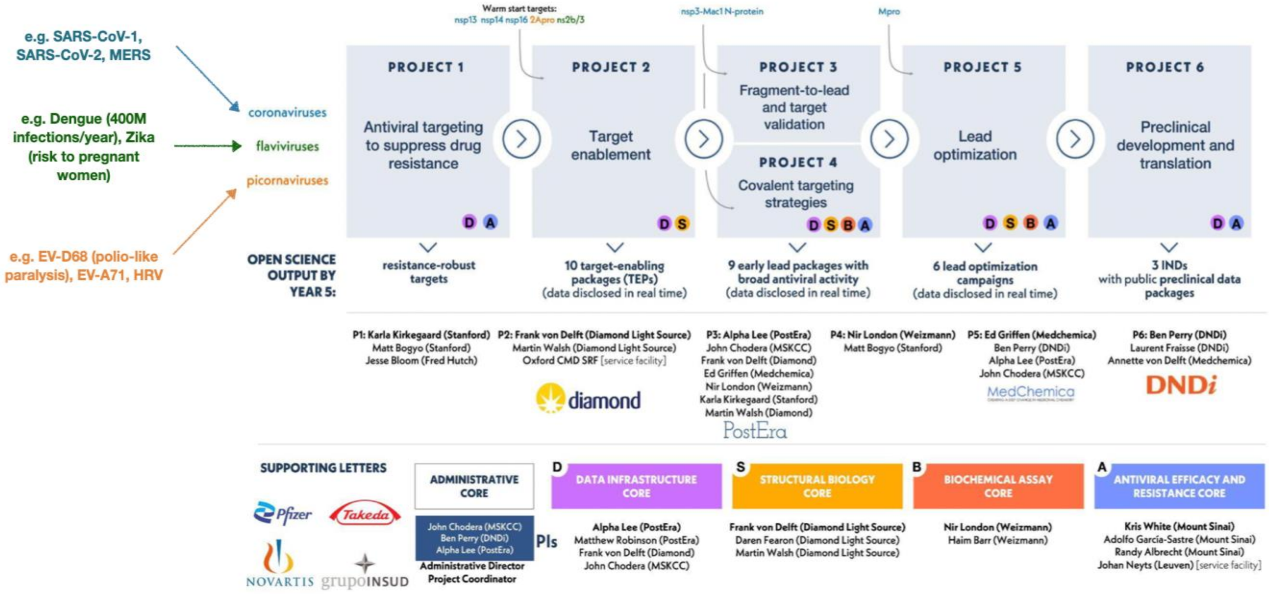
Launched during the COVID Moonshot project, the PostEra Manifold Software-as-a-Service platform has grown into a more general platform used across a wide array of drug-discovery projects. Supporting high-throughput retrosynthesis evaluation, large-scale library enumeration, ultrafast search across essentially every available molecule, and many other novel features, Manifold is currently being used by multiple large biotech and pharma companies.



An example Manifold search of an initial hit on the Moonshot project.

AI-driven Structure-enabled Antiviral Platform (ASAP)

Extending the impact of COVID Moonshot, we are delighted to launch the AI-driven Structure-enabled Antiviral Platform (ASAP). The ASAP pandemic-preparedness platform aims to discover globally and equitably accessible antivirals against coronaviruses, picornaviruses (including potentially debilitating enteroviruses as well other cold-causing viruses), and flaviviruses (responsible for endemic diseases such as dengue and zika). Launched with $68M of initial funding from the US National Institutes of Health over three years, this five-year programme aims to deliver three IND-ready assets, six optimised leads, and chemical probes for at least nine viral targets.


e.g. SARS-CoV-1, SARS-CoV-2, MERS
e.g. Dengue (400M infections/year), Zika (risk to pregnant women)
e.g. EV-D68 (polio-like paralysis), EV-A71, HRV
coronaviruses
flaviviruses
picornaviruses
Warm start targets:
nsp13 nsp14 nsp16 2Apro ns2b/3
nsp3-Mac1 N-protein
Mpro
PROJECT 1
Antiviral targeting to suppress drug resistance
D A
PROJECT 2
Target enablement
D S
PROJECT 3
Fragment-to-lead and target validation
PROJECT 4
Covalent targeting strategies
D S B A
PROJECT 5
Lead optimization
D S B A
PROJECT 6
Preclinical development and translation
D A
OPEN SCIENCE OUTPUT BY YEAR 5:
resistance-robust targets
10 target-enabling packages (TEPs) (data disclosed in real time)
9 early lead packages with broad antiviral activity (data disclosed in real time)
6 lead optimization campaigns (data disclosed in real time)
3 INDs with public preclinical data packages
P1: Karla Kirkegaard (Stanford)
Matt Bogyo (Stanford)
Jesse Bloom (Fred Hutch)
P2: Frank von Delft (Diamond Light Source)
Martin Walsh (Diamond Light Source)
Oxford CMD SRF [service facility]
diamond
P3: Alpha Lee (PostEra)
John Chodera (MSKCC)
Frank von Delft (Diamond)
Ed Griffen (Medchemica)
Nir London (Weizmann)
Karla Kirkegaard (Stanford)
Martin Walsh (Diamond)
PostEra
P4: Nir London (Weizmann)
Matt Bogyo (Stanford)
P5: Ed Griffen (Medchemica)
Ben Perry (DNDi)
Alpha Lee (PostEra)
John Chodera (MSKCC)
MedChemica
P6: Ben Perry (DNDi)
Laurent Fraisse (DNDi)
Annette von Delft (Medchemica)
DNDi
SUPPORTING LETTERS
Pfizer
Takeda
NOVARTIS
grupoINSUD
ADMINISTRATIVE CORE
John Chodera (MSKCC)
Ben Perry (DNDi)
Alpha Lee (PostEra)
Administrative Director
Project Coordinator
PIs
D DATA INFRASTRUCTURE CORE
Alpha Lee (PostEra)
Matthew Robinson (PostEra)
Frank von Delft (Diamond)
John Chodera (MSKCC)
S STRUCTURAL BIOLOGY CORE
Frank von Delft (Diamond Light Source)
Daren Fearon (Diamond Light Source)
Martin Walsh (Diamond Light Source)
B BIOCHEMICAL ASSAY CORE
Nir London (Weizmann)
Haim Barr (Weizmann)
A ANTIVIRAL EFFICACY AND RESISTANCE CORE
Kris White (Mount Sinai)
Adolfo García-Sastre (Mount Sinai)
Randy Albrecht (Mount Sinai)
Johan Neyts (Leuven) [service facility]

ASAP aims to tackle pandemics by targeting two salient processes: mutations, by which viruses mutate into more virulent and pathogenic strains, either in human or animal reservoirs, and transmission, whereby dominant strains spread uncontrollably within human populations. These two forces drive a vicious cycle: increased transmission drives more infections, creating a greater risk of mutations due to intrinsic viral mutation rates, thereby leading to the emergence of more fit and virulent strains.

To address mutations, we identify viral targets that cannot easily evolve resistance to small molecule antivirals generated by our platform, using three complementary strategies: (1) phylogenetic analysis of circulating strains to identify conserved sites across members of a viral family; (2) deep mutational scanning (DMS) to interrogate the fitness cost of mutations in a druggable site; and (3) mechanistic analysis of the viral lifecycle to identify 'dominant targets' where the liganded targets disrupt viral growth and cannot be rescued by mutations arising within the same cell.

Rapid discovery of suitable inhibitors against these targets is enabled by high-throughput structural biology, which illuminates druggable sites with dense fragment maps that can be synthesised into novel potent chemotypes where target engagement is restricted to regions where mutations would compromise fitness. Machine learning is then used to execute structure-based design to drive lead optimisation and rapidly arrive at candidate molecules, whilst simultaneously delivering chemical probes to validate target biology and verify the fitness costs of engaging each target.

To address transmission, we will use an open-science approach to drug discovery, developing new antivirals in a manner that enables them to be brought to market with the goal of immediate, global and equitable access. Equitable access to safe, effective antivirals is the fastest and most effective way to inhibit pandemic transmission. As we have seen with SARS-CoV-2, pandemics are uniquely global threats: an epidemic arising anywhere in the world is a threat everywhere, because it breeds mutations that could be both more transmissible or capable of overcoming any vaccine-induced or natural immunity.

Ultimately, we hope COVID Moonshot, and the successor ASAP platform, will help tackle the current coronavirus pandemic as well as prevent future ones by advancing a pipeline of globally and equitably accessible antivirals. Further, the significant time acceleration that COVID Moonshot has achieved is a milestone for AI/ML-augmented drug design.

References

(1) The COVID Moonshot Consortium, *et al*., Open Science Discovery of Oral Non-Covalent SARS-CoV-2 Main Protease Inhibitor Therapeutics. *Biorxiv*, doi:10.1101/2020.10.29.339317

(2) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A.A. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS Central Science*, **2019**, *5*(9), 1572.

- -

Practical Cheminformatics with Open-Source Software

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

I have collected several (currently 13) of the Jupyter notebooks, which I put together for teaching cheminformatics, in a Git repository. This growing collection of notebooks provides hands-on tutorials on several topics including:

- Fundamentals – working with SMILES and SMARTS
- Clustering
- SAR Analysis – R-group analysis and Positional Analogue Scanning
- Machine learning – classification and regression models

All the notebooks use open-source software components and can be run on Google Colab without the need for local software installation.

- -

The Catalyst Science and Discovery Centre Archives

Contribution from Judith Wilde, Collections Officer, Catalyst Science and Discovery Centre and Museum, email: judith@catalyst.org.uk

Catalyst Science Discovery Centre and Museum in Widnes is housed in the former office building of John Hutchinson's Alkali Works. Gossages Soap Works was established in 1855 on Spike Island and at one stage was the largest soap company in the UK, if not the world.

In October 2021, we were honoured to welcome Mr Peter Gossage along with family members, friends and colleagues to officially open The Gossage Room. The aforementioned building was used for administration of the soap works from 1908 until 1933. This new heritage room, refurbished during the pandemic with funding from the Gossage family, now proudly displays the portraits of the founder of the soap works William Gossage (1799-1877) and his son Frederick Herbert Gossage (1831-1907), artefacts and photographs relating to the soap industry and a huge, wall-filling bookcase containing much of the Catalyst library archive.

This new reading room enables some of our collection to be available without the need to make an appointment. The room itself can been booked for meetings and is regularly used by local groups.

The Gossage Room.

The Gossage Room houses around three-quarters of our reference library. Our book and journal collection comprises several sets of bound volumes including the *Journal of the Society of Chemical Industry* and many chemical company in-house magazines and newspapers. We also have a rare set of minutes of the Alkali Inspectorate's annual meetings. There are also many individual books relating to aspects of chemistry and industry, and to local matters. This is an extensive collection which compliments that held by the British Library. Please note that some of the rarer books can only be viewed by appointment.

So, what is in our wider archive?

Special areas of interest for local and family historians alike are the collections which answer the question "what did my ancestor do?" and "what is a chemical / alkali worker?". Our collection consists of a large ICI archive for the Mond and Alkali Divisions with records pertaining to Widnes, Runcorn, mid-Cheshire and elsewhere, plus factories in Cornwall and Stoke Prior, Worcestershire. In January 2022, we published a list of Workers' Records mainly relating to the ICI and Brunner Mond mid-Cheshire sites. Similar to the release of the 1921 Census, the records are for people who would have been more than 100 years of age during the current year. The oldest recorded worker was born in the 1840s. A full list of over 13,000 named individuals is available and people can order a copy of their ancestor's record. Full details and some background information can be found on our Brunner Mond Workers' Records website.

The employment records are only part of the story though. Another absolute gem is the ICI Magazines and their regional supplements. Starting in Jan 1928, the series runs into the 1990s and we have all except for a few elusive editions from the 1970s. The early magazines – 1928 to Aug 1939 (none between Sep 1939 and Dec 1946 due to WW2) – show the entire ICI worldwide family. ICI began in 1926 from a merger between Brunner Mond, Nobel Industries, United Alkali Co. (which itself was made up of lots of small chemical firms – mostly in the north-west of England) and British Dyestuffs Corporation. Any individuals who were long servers could expect to be represented in the magazines with a few lines on retirement etc. and even a photograph. There is often a potted history of their employment and the different areas they may have worked in, along with references to their family members, who may also have worked for the company.

Items from the workers' archive.

In addition to this, we hold the Weston Point Studios Archive of negatives. Over 10,000 images taken from around the 1950s-1990s of retirements, sports presentations etc. were taken mainly for the magazines or local / national newspapers – you name it – they photographed it!

So did your ancestor work in Research for ICI? Well maybe they will crop up in the 25,000 General Chemicals Reports that we hold (of which only 14% are currently digitised – a huge undertaking!). Did they work on ground-breaking discoveries or do important 'secret' work for the Government during the war?

Maybe they made a suggestion! The ICI Suggestion Scheme paid out some figures that many of us would find very generous nowadays. Who better than the worker to spot where savings and efficiencies could be made? A quick submission of an idea could easily gain him a few shillings for a good suggestion, even if it wasn't implemented. Others could expect to see a handsome reward on the basis that ICI would pay a percentage of the amount of money the company could expect to save by carrying out the amendments. The ledgers and several files of the original handwritten submissions are yet another part of the jigsaw that we hold.

Just because ICI is well represented, doesn't mean that other firms are in the shadows. Gossages Soap Works in its heyday had a magazine entitled *The Wheel*. Although only a fairly short run by comparison (1921-1930) they are similar to the ICI set where they focus on the people rather than corporate issues. Weddings, retirements, sports, Scouts, evening class / examination prizes are all covered, along with a rather charming section called "Long Servers With The Old Firm". Here you will find usually one or two pages of portrait photos of men who had, in some cases, over 60 years' service. Some of them were born in the 1840s so again, a rare chance to see a photograph of your ancestor. These magazines are quite fragile, so photocopies and digital versions are available to view in the Gossage Room.

Also, Vine Chemicals, Laportes, Croda Bowmans, to name but a few – if there's an in-house magazine then we want it! Many of these are digitised and available to view, fully searchable, in our library.

Other collections include the Hutchinson Estate Papers which when said quickly can seem just a minor detail but is in fact boxes upon boxes of legal agreements, letters, maps etc. from around 1841 onwards.

We hold in excess of 10,000 accessioned items. These range from large chemical industry and general laboratory equipment; a large solid brass calculator "The Millionaire" which is a Swiss-made calculator manufactured in the 1890s and could, for the first time calculate multiplications of large numbers; and some very rare and precious items such as Ludwig Mond's private papers. A few examples from the latter include his indentures from his apprenticeship in Germany and his Dutch passport as well as family correspondence which has been translated into English (a copy of which is in our library).

Way back in 1982 when the collection started as the Halton Chemical Industry Museum Project – a joint enterprise between the Manpower Services Commission and Halton Borough Council – could we have envisaged how far we would come? From the early days of the research project a team of people dug deep into the history and methods of the chemical industry of our area and further afield (and yes, we do still have the research notes – they are indexed and available on the bottom shelves!). We have continued to collect and preserve our local and industrial history and the donations of artefacts, books and papers still come, and continue to surprise us as well!

We are open six days a week (seven days during the school holidays) from 10.00-17.00. An entry ticket (standard or concession) entitles you to visit Catalyst as often as you like during the following 12 months. Pop along and see for yourself. If you wish to view anything specific which may be in our vault please email heritage@catalyst.org.uk or phone ahead so we can make sure someone is available to assist you.

- -

Chemical Data Recovery 3: Legacy Chemical Data Recovery

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

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

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

Figure 1. A chem-archaeologist has discovered an ancient library filled with long-lost molecular secrets. She transcribes some of the information using a chemist's triangle.

Introduction

Cheminformatics solutions can be incredibly challenging to implement. What this really means is cheminformatics problems are incredibly rewarding to solve. While cheminformatics work is difficult, such solutions are very important to our scientists and our society. As we discover more about the universe we exist in, the already impressive work created by those in the cheminformatics field only grows in usefulness.

Of core importance in cheminformatics is the actual storage and communication of chemical information. The protocols we implement to describe chemistry and molecules make it possible to quickly load and share data for use in the algorithms we create. Imagine if the PDB format never existed and we would need to hard-code the data for every atom in a protein structure. Today, we have access to a multitude of chemistry file formats for handling chemistry information. The most popular include the MDL connection tables, CDX files, SMILES, ChemDoodle JSON and CML.

You have probably heard of all of these formats, but have you ever heard of Wiswesser Line Notation (WLN)? At one time, WLN was the most popular protocol for describing molecular information. Books were published about it, conferences focused on it, high schoolers were taught to transcribe it (such skills would get you a decent pay check), and the largest databases of the time primarily stored WLN codes. But no more. How did the most popular chemistry protocol disappear from our collective knowledge?

Almost a decade ago, I was fortunate to meet and befriend William L. Todsen, a WLN enthusiast. He taught me a lot about WLN and I decided it would be interesting to implement a WLN reader, allowing us to recover legacy chemical data and bring back to life an important chemical protocol. Along the way, we learned much about the history of cheminformatics, the ways in which WLN excels and the reasons why it is no longer used. Even if it is an abandoned technology, we want to make sure WLN is not forgotten and guarantee those in the future always have the tools to observe and learn from the ingenuity of the cheminformaticists that came before us.

A (brief) history of Wiswesser Line Notation

WLN was invented by William J. Wiswesser in 1949. His goal was to devise a unique identifier for a molecule to be used by both humans and machines, in a simpler manner than an IUPAC name, adopting common chemical notations chemists were already familiar with. Keep in mind WLN originated before computers were widely accessible to chemists. Wiswesser describes in a 1952 Chemical and Engineering News (C&EN) article, "More than a decade ago, the author recognised the need for a truly universal and fully systematic chemical structure notation... Having learned the penetrating value of molecular models and logical symbolism, the author has favored a pictorially obvious symbolism." The system Wiswesser devised was very understandable and professionally presented. WLN quickly gained in popularity and application over competing solutions.

Three official editions of the WLN specification were eventually published. The first edition was written by Wiswesser and published in 1954. In the forward to the first edition, Elbert G. Smith, states "[WLN] is a new chemical notation by which even complicated chemical structures may be expressed concisely and without ambiguity in a single line of letters, numbers and punctuation marks. It has been designed to provide a straightforward way of indexing chemical compounds and so to bypass the present growing confusions and frustrations in chemical nomenclature. Chemists generally seem aware that future progress in communication and utilization of chemical knowledge will require some sort of new chemical notation." A second edition of the WLN specification was published in 1968 under Smith's leadership and the newly created Chemical Notation Association (CNA). In the forward to the second edition, I. Moyer Hunsberger states "Since a computer can transform a Wiswesser notation to an atom connection table (which completely represents the structure of a chemical compound on magnetic tape), the notation may find favor as an economical input device for computer-based information retrieval systems." In 1976, Chemical Information Management, Inc. (CIMI) published the third and final official edition of the WLN specification, edited by Smith and Peter A. Baker, under the governance of the CNA. Figures 2a-2c are images of the three editions of *The Wiswesser Line-Formula Chemical Notation*.

WLN was at the height of its popularity in the 1960s and 1970s. Cheminformatics was becoming an essential discipline and protocols like WLN were necessary to organise the growing chemical data available. One only needs to look at the vast amount of publication material related to WLN during this period to understand its impact. WLN was mainly used as an indexing solution. Innovators were also discovering ways to use WLN for molecular structure matching. A program called Pathfinder was developed by Carlos M. Bowman's group at Dow Chemical and its development continued by Tommy Ebe and Antonio Zamora at the Chemical Abstracts Service for elucidating canonical WLN paths in ring graphs. In 1968, Charles E. Granito (who later founded CIMI) at the Diamond Alkali Company documented WLN for registration systems and later at the Institute for Scientific Information (ISI) with Murray D. Rosenberg created the Chemical Substructure Index allowing for

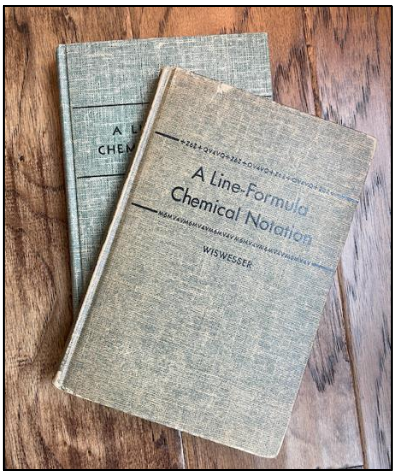


Figure 2a. First edition.

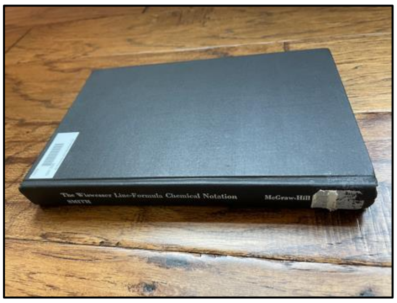


Figure 2b. Second edition.

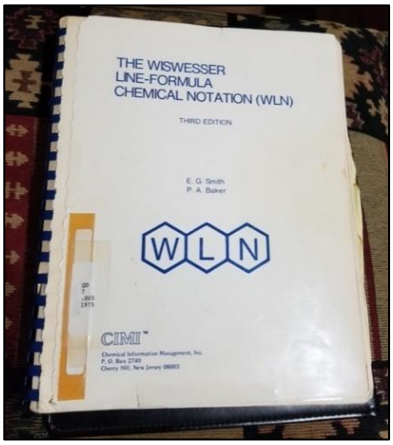


Figure 2c. Third edition.

novel substructure searching based on WLN strings. The ISI was heavily invested in the WLN protocol and provided several programs and searching solutions, including the Index Chemicus Registry System for recording chemical data in the Index Chemicus as WLN codes and the popular CROSSBOW program for handling WLN strings as computer data structures.

The International Union of Pure and Applied Chemistry (IUPAC) even considered WLN as their standard line notation, before choosing the competing Dyson notation instead. The decision was very controversial and resulted in a lot of protest. Bonnie Lawlor from the Chemical Structure Association (CSA) Trust (the CSA Trust evolved from the CNA) summarises the importance of WLN and IUPAC's choice in a 2016 article. Neither notation is maintained by IUPAC today.

In a 1982 publication, Wiswesser postulated what might become of WLN in the future, "Soon online computer and word-processing terminals will be as commonplace as IBM's Selectric typewriters are today; by 1999 high schools and alert grade schools will have color-coded chemistry in educational entertainment that goes far beyond today's pinball games of skill: computer-weaned grade-school science students might well be able to tap an online 'Chemical Picture Book for Children' with advanced WLN descriptions." In retrospect, we now know that did not happen. The world began to change, as it always does. As Wiswesser predicted, computers became more commonplace and many new programmers were entering the industry. New solutions were necessary to solve more difficult problems, but WLN did not keep up.

In his 1952 C&EN explanation, Wiswesser was clairvoyant in reasoning a concise and easy to use chemical protocol would be necessary, "Simplicity of usage is the prime requirement of a good structure notation... Conciseness is intimately related to the [ideal usefulness], and particularly to the one specifying ease of manipulation by machine methods. It should be obvious that conciseness is desirable for the efficient use of any tabulating machinery. Even with new machines, card punching and verifying will remain the most expensive of the numerous ingenious tabulating operations, and it is almost directly proportional to the number of symbols required." While WLN did achieve the goal of being compact and the basics were relatively easy to learn, the protocol was hardly simple to program (as we will find out below), requiring individuals to manually generate and decipher WLN strings. As Wendy A. Warr states bluntly in a 1982 review of WLN applications, "The principle disadvantage of WLN is that it is not user friendly...no one has yet produced a cost-effective program to [derive a canonical WLN] for over 90% of compounds...one has to balance the cost of hardware plus software against the costs of extra WLN-skilled personnel." As computers gained adoption in scientific laboratories and academic settings, simpler solutions overcame the conciseness of the WLN protocol. Today, the most popular chemical protocols are the easiest to program, regardless of their verbosity.

It may seem logical that the introduction of the much more readily implementable Daylight SMILES line notation in 1988 led to the replacement of WLN, but in reality, WLN had already fallen out of favour as more practical methods for storing chemical information were developed for the many computer systems introduced. The MDL connection tables, *circa* 1979, are ASCII formats many programmers could easily use, and Mike Elder produced the DARING software to help aid in the conversion from WLN to MDL connection tables. New developments in computer algorithms were also complicit in the demise of WLN. J. R. Ullman published an algorithm for graph isomorphism in 1976, enabling cheminformatics applications to directly and efficiently match parts of chemical structures based on the constituent atoms and bonds. Granito's Chemical Substructure Index was no longer the optimal solution and WLN was losing popularity by the time CIMI published the 3rd edition; Granito would soon change business directions.

Wiswesser would pass away in 1989, leaving one of the most impactful and impressive chemistry protocols ever created as his legacy. To date, WLN is still the most concise, lossless, string representation of chemical information. The WLN protocol is a passion project of a talented group of cheminformatics experts, and a work I hold in very high regard.

A breakdown of Wiswesser Line Notation

WLN is a substructure-based, canonical, line notation for molecular structure(s). The characters in a WLN string define the atoms and bonds in the molecular structure(s). The entirety of the periodic table of elements is supported, and single, double, triple and dative bond types are available. Any type of complex ring system is compatible, including polycyclic fused, perifused, spiros, bridged and pseudo-bridged structures. There is no explicit aromaticity model, but unsaturations in rings are fully defined. Charges, radicals and isotopes are included. Stereochemistry is supported, but not using CIP. Beyond basic molecule structures, there are special rules for handling chelate compounds, metallocenes and catenanes, polymers, inorganic formulas, uncertainties and MANTRA (Mixture, Alternative possibility, Not assigned, Tautomer, Reactant, Addition) suffixes. One should be aware the WLN specification evolved through the editions, and the changes are not all backwards compatible. For instance, WLN version 2 removes lower-case locant symbols and version 3 removes methyl and ring contractions.

WLN is a line notation, which means the chemical information a WLN string contains is defined in a single line of text. The characters in the WLN string are limited to those found on a standard typewriter. Contemporary line notation protocols include SMILES (Daylight and OpenSMILES), IUPAC InChI and technically IUPAC naming.

WLN is substructure-based, differentiating it from other line notations. SMILES is an all-atom representation and InChI is based on information layers. IUPAC names are also substructure-based, but are meant for written and spoken language between chemists. The substructure types in WLN are predetermined, for instance, a *W* symbol defines a dioxo group and an *R* symbol represents a phenyl group (when not preceded by a space, after which they would be locants instead). Chains are defined by numbers and rings are defined by symbol sequences beginning with a *L* (carbocyclic), *T* (heterocyclic) or *D* (chelate). WLN is therefore an intrinsically compact representation of a molecular connection table. Take a look at the following table for a comparison of WLN strings to SMILES, InChI and IUPAC names, and notice how much shorter the WLN strings are.

To finish the description, WLN is a canonical protocol, and therefore one and only one WLN string is theoretically acceptable for any chemical entity. Canonicalisation is typically used for indexing and exact matching in databases. SMILES is not a canonical protocol, but Daylight Informatics did publish a vague and incomplete CANGEN algorithm for canonicalising SMILES strings. Canonical SMILES algorithms today are unique to the developer and are not cross-compatible. InChI is canonical by definition, and its implementation is so complex, only one official codebase exists. Traditional IUPAC names are not canonical, as several correct names are possible given the various IUPAC rules, but the latest 2013 IUPAC naming specification defines Preferred IUPAC Names (PINs), which are meant to be canonical.

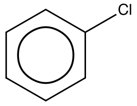
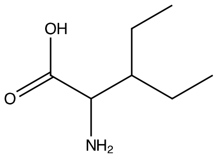
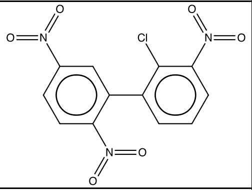
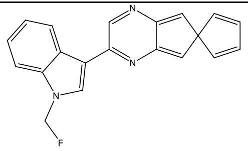
Molecular Structure	Line Notation	Value
CH_4	*IUPAC Name*	methane
	WLN	1H
	SMILES	C
	InChI	InChI=1S/CH4/h1H4
	*IUPAC Name*	chlorobenzene
	WLN	GR
	SMILES	c1ccccc1Cl
	InChI	InChI=1S/C6H5Cl/c7-6-4-2-1-3-5-6/h1-5H
	*IUPAC Name*	2-Amino-3-ethylvaleric acid
	WLN	QVYZY2&2
	SMILES	OC(=O)C(C(CC)CC)N
	InChI	InChI=1S/C7H15NO2/c1-3-5(4-2)6(8)7(9)10/h5-6H,3-4,8H2,1-2H3,(H,9,10)
	*IUPAC Name*	2-Chloro-2',3,5'-trinitrobiphenyl
	WLN	WNR DNW BR BG CNW
	SMILES	c1ccc(N(=O)=O)c(Cl)c1c1c(N(=O)=O)ccc(N(=O)=O)c1
	InChI	InChI=1S/C12H6ClN3O6/c13-12-8(2-1-3-11(12)16(21)22)9-6-7(14(17)18)4-5-10(9)15(19)20/h1-6H
	*IUPAC Name*	5'-[1-(Fluoromethyl)-1*H*-indol-3-yl]-4',7'-diazaspiro[2,4-cyclopentadiene-1,2'-indene]
	WLN	T56 CX FN INJ C-& AL5XJ& G- DT56 BNJ B1F
	SMILES	C1(=C2)C(=CC2(C=C2)C=C2)N=C(C=N1)C1=CN(C(=CC=C2)C1=C2)CF
	InChI	InChI=1S/C20H14FN3/c21-13-24-12-15(14-5-1-2-6-19(14)24)18-11-22-16-9-20(7-3-4-8-20)10-17(16)23-18/h1-12H,13H2

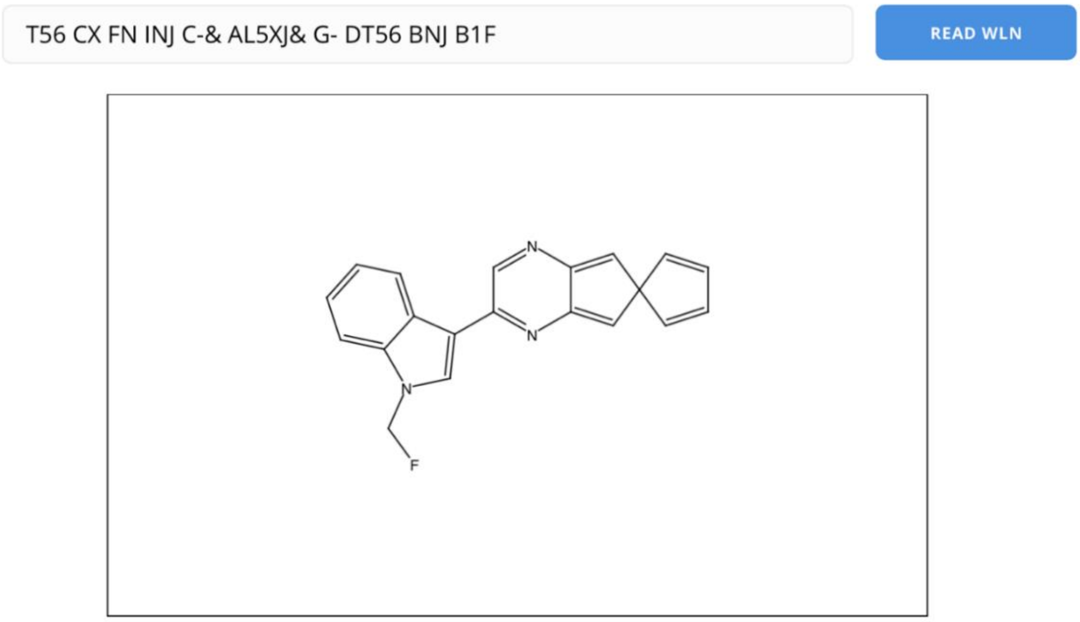
Table 1. Comparison of line notation values.

Implementation

Today, Todsen has been unofficially maintaining the WLN specification, incrementing the minor version number. Todsen is currently polishing version 3.2 of the WLN specification and he states, “This project has been a labor of love, consuming a lot of my off-duty time for the better part of a decade. By and large (and by design!), my refinements resulted in almost no differences in the WLN codes. Indeed even going from the 1968 version of the rules to the 1975 version, the bulk of the codes stayed the same or could be easily updated.” Todsen has provided a full description of his experience with WLN, which is well worth a read.

I was intrigued after learning more about WLN, and I am always interested in implementing unique chemistry protocols. There is no commercial application for WLN, so I pursued this project as an intellectual curiosity and wrote a WLN reader. Here is a picture of the reader in action. Just type in your WLN string (v3.2) and press the **Read WLN** button to see the related chemical drawing.

I have to admit, implementing WLN is much more difficult than I anticipated. There is a significant level of detail concerning each aspect of chemical structures. One difficult concept is the connectivity of each symbol type. Care must be taken when completing valences by WLN rules to correctly place implied hydrogens and charges. The chain terminator symbol, &, is more complicated than expected; it is not trivial to understand where the previous connection may be, and mixed with terminal structure symbols, caused quite some confusion. But while annoying, everything, from symbols with multiple meanings to ring delocalisation to dative bonds, is consistent once implemented properly. The worst procedures to implement are the complex ring rules. I had to stop at perifused systems, as the suggested algorithms required an unintuitive guess-and-check method for the correct answer. Altogether there are nine sets of rules for handling different types of ring systems. By comparison, any ring system may be defined in a SMILES string using a simple molecular spanning tree and ring closure indexes. I now understand why there is no fully compliant program for handling the WLN protocol.



T56 CX FN INJ C-& AL5XJ& G- DT56 BNJ B1F

Figure 3. A ChemDoodle Web Components demo allows you to recover the chemical structures defined in a WLN string.

As for application, I want to briefly discuss canonicalisation, which is a requirement for WLN strings. Canonicalisation is an attempt to simulate graph-theoretical functionality without making use of graph-theoretical algorithms. One using a canonical WLN string for a structure would be confident in matching another structure converted into a WLN string if the two strings are identical. However, specifications and software implementations continuously change, invalidating previously generated canonical output. So the integrity of your dataset will eventually be dependent on using old and obsolete software. The correct way to handle molecule structure comparisons is through graph isomorphism implementations, perhaps after a fingerprint pruning, which neither existed nor were practical when WLN was conceived. Canonical string comparisons are also only applicable to exact matching, and not substructure, superstructure, query or maximum common substructure. That being said, canonical WLN strings do provide limited ability for substructure matching without graph-isomorphism algorithms due to the notation symbols, for instance, you will know a structure contains a benzene ring from the inclusion of an *R* symbol with no space before it.

WLN is a very interesting format. I feel it fits right in between SMILES and IUPAC naming in terms of theory and implementation. The following image shows the print specifications for SMILES, WLN and IUPAC naming, where you can observe a direct correlation between page count and implementation difficulty. SMILES is relatively simple to implement, although Noel O'Boyle will provide you with an endless number of reasons why SMILES is not as simple as it appears. InChI, while very complex, is an open standard and the software to handle it is funded and open source. IUPAC naming is the most massive undertaking, by far, but that is a whole other complex discussion.

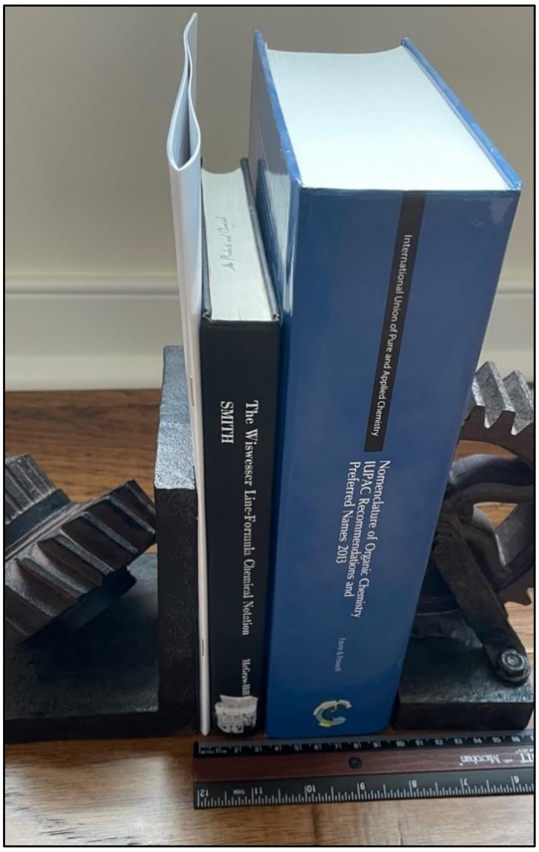

International Union of Pure and Applied Chemistry
Nomenclature of Organic Chemistry
IUPAC Recommendations and
Preferred Names 2013
The Wiswesser Line-Formula Chemical Notation
SMITH
McGraw-Hill

Figure 4. Comparing the specification sizes for Daylight SMILES (left), WLN (centre) and IUPAC naming (right). Page count correlates well with implementation difficulty.

Regardless, I enjoyed my time with WLN. It is a unique perspective on chemistry data and I hope to return to find a better algorithm for the complex WLN ring systems. If you have the time, please learn about it, try to implement it, and teach it to your students and colleagues. If you are interested in cheminformatics, writing a WLN parser or writer is an incredibly challenging project providing you with very thorough experience into the concepts of chemical structure and graph theory. It will certainly be an impressive statement in your portfolio.

Preservation

We may now understand why such an important chemistry protocol, which was widely adopted in the mid-1900s, is no longer known today. The importance of preserving the innovation of those in cheminformatics is significant. When we decide to investigate new solutions or new problems, we may find past work provides inspiration or even the answer. Due to WLN losing popularity 50 years ago, those with WLN experience are now retiring or have unfortunately passed away. It is very possible we will lose all WLN expertise in my lifetime.

To preserve WLN, Todsen is continuing to maintain the latest WLN specification. As a companion, iChemLabs has developed a WLN parser based on Todsen's work. Our goal is to continue to develop this parser until we can handle any WLN code created to spec. Other groups are also helping to preserve WLN. Both the PubChem and ChemSpider databases store WLN strings for many entries, although no defining version is provided in either. The Pistoia Alliance updated their UDM format specification in 2018 to allow WLN as an acceptable chemical protocol. In 2019, Roger Sayle developed a WLN parser based on the second edition of the specification and contributed it to the open source OpenBabel project.

The most difficult problem in preserving WLN is not in the complex theory or the difficult algorithm, but in the specification itself. The three editions of the specification are protected by copyright. This does not prevent us from discussing or implementing WLN, but precludes the redistribution of the WLN specification. There is no online copy and all three editions are out of print. You may occasionally find the first or second edition in a chemistry library or on Amazon or eBay, but the third edition is very rare. Without the ability to access the specification, the usefulness of the protocol is limited.

We searched for the copyright holders, but they cannot be found or are deceased. Efforts to locate Granito have been unsuccessful. Smith and Baker, editors of the third edition of the specification, are deceased, and Chemical Information Management, Inc., which holds the copyright on the third edition, no longer is in operation.

Our own legal team investigated this matter and has concluded that the copyrights have not expired and the specification is not currently in the public domain. Accordingly, we cannot redistribute the specification at this time. It is unlikely Wiswesser wanted copyright issues to cause his work to be lost to history, so please reach out if you have any information.

Altogether, WLN does not provide much benefit over the line notations widely used today, even if it is more compact. The size of databases can be reduced, but it is unlikely such an improvement would be significant given continued advances in computer systems. The difficult work in implementing the protocol would make any practical usage a chore. Rather, WLN is a part of the history of cheminformatics, a part of our culture. It needs to be preserved.

Final word

I want to thank Todsen for working with me on WLN theory, helping me write this article, and for his continued friendship. Todsen continues to pursue his interest in WLN, and if you would like to know more, he may be reached by email.

If you enjoyed the artwork leading these articles, they were commissioned to Etienne Delalande, an artist in France who I have known for several years now. He is incredibly talented and I highly recommend him, but his art speaks for itself. You may reach out to him here if you are interested in working with him.

I was very excited to put these articles together as they place a coda on three massive projects I and my team at iChemLabs have been working on for a decade. But now that everything is discussed, I have to admit I feel sad it is all over. The work, however, will continue, as we will be looking to make our chemical data recovery tools even better. And, of course, we have some more projects in the pipeline, which I look forward to telling you about in the future.

I hope this series of articles has provided insight into what iChemLabs does and the problems we enjoy solving. I, myself, have been programming chemistry solutions since 2006, back when I was in college, and I don't ever plan to stop. If you have read this far, I want to say thank you. Please do reach out. I want to know about you and your interests in the cheminformatics industry. Feel free to connect to or contact me on LinkedIn, or address an email to me using our contact form. Until next time!

- -

Svante Wold, 1941-2022

Contribution from Professor Johan Trygg, Sartorius & Umeå University, email: jonke3@gmail.com

Original version posted on LinkedIn.

Svante Wold left us on 4 January 2022, at the age of 81. His legacy includes many people from his professional career who he touched in different ways and impacted both professionally and personally. I post this message on behalf of Svante's family and also colleagues and friends at Umeå University and Sartorius Data Analytics.

Professionally, Svante had a very impressive career, being a great scientist, a successful entrepreneur with a clear vision and mission that the combination of data, domain expertise and data science would be key to finding solutions to real, complex and societal problems. This is what he named chemometrics, in the year I was born, and he never stopped his focus on spreading that vision and focusing on that mission. Myself together with colleagues continue to drive his mission and vision at both Umeå University and Sartorius.

Personally, Svante was a great people person. For him it was all about people first and his genuine, generous and deep commitment to help and support the people around him was everlasting. In this element, Svante really shone through and we all enjoyed listening to his stories, laughing at his Norwegian jokes and Swedish limericks, typically while enjoying food and drinks, e.g., Swedish potato dumplings along with a beer or two.

Here's a personal statement from Nouna, his wife, his partner and soulmate: "I feel a tremendous loss as Svante has had an enormous impact in his field. He was a pioneer deeply committed in changing the way people process and extract information from large chemical and other data sets. Change is always difficult but Svante persevered, coined the word chemometrics, and with Bruce Kowalski, founded the Chemometrics Society, teaching and preaching Design of Experiments and use of Multivariate Analysis. Svante was the creative partner, developing algorithms to solve specific real problems. I was his sounding board, helping sort through ideas and implementing into user-friendly software. By any metrics, his scientific and societal impact achievements will continue to impress. I owe him a wonderful life, exciting, always thriving to learn and improve and full of love. The world owes him to have opened doors and started a new field. Johan Trygg, his PhD student, in many ways part of our family, continues his work as professor of chemometrics and carries his mantel. May Svante rest in peace and his soul continues to guide us."

Svante Wold, Nouna Kettaneh-Wold, and Johan Trygg.

On behalf of Svante's family, we have setup a digital condolence book. Please share your own stories, photos, express your thoughts and prayers or other ways you'd like to commemorate Svante.

Johan & Nouna

- -

Cheminformatics: a Digital History - Part 1. Early days at Sheffield: a Personal Perspective

Contribution from Professor Peter Willett, Information School, University of Sheffield, email: p.willett@sheffield.ac.uk

I have been asked to provide what is intended to be the first of a series of articles in the *CICAG Newsletter* that aims to describe the origins and early days of cheminformatics from the viewpoints of some of those who were involved at the time. This aim immediately raises the question of when cheminformatics first started and hence when were these early days. The term, in its alternative form of chemoinformatics, appears to have been first formally defined in an article by Brown in 1998, who stated "The use of information technology and management has become a critical part of the drug-discovery process. Chemoinformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and optimization."[1] That said, many of the discipline's basic techniques had already been established well before 1998:[2] for example, the first reports of substructure searching and of QSAR appeared in 1957[3] and 1962[4] respectively; and the leading journal in the field, now the *Journal of Chemical Information and Modeling* but founded as the *Journal of Chemical*

Documentation, published its first articles in 1961.[5] For the purposes of this article I shall take the turn of the century as the cut-off point since by then the discipline had become sufficiently well established for the introduction in 2000 of the first specialist masters course in cheminformatics at the University of Sheffield, with others at the University of Manchester Institute of Science and Technology and at Indiana University starting in the following year.[6]

UKeiG
The Chemoinformatics Research Group
Information School, University of Sheffield

Peter Willett receiving the 2012 Jason Farradane Award, made in recognition of a major contribution to the theory and practice of information science, from Martin White on behalf of the Sheffield chemoinformatics research group.

Research into cheminformatics started in Sheffield in 1965 with the appointment of Mike Lynch to a position in what was then called the Postgraduate School of Librarianship. Most academic research in cheminformatics over the years has, not surprisingly, been carried out in chemistry (or, to a lesser extent, computing or life science) departments and it is primarily due to Mike's pioneering work that a library science department came to play an important role in the subject's development. He came to Sheffield following four years working in, and finally directing, the research department of Chemical Abstracts Service in Columbus Ohio. While there he had overseen some of the earliest studies anywhere on the use of computers for the storage and retrieval of both textual and chemical structural information, work that played an essential part in the development and introduction of the CAS Registry System. On coming to Sheffield he rapidly established research into computational techniques for generating printed subject indexes and for identifying the structural changes involved in chemical reactions, and then followed this with an extended programme of research into the design of fragment-based screening systems for 2D substructure searching.[7,8] The statistical criteria that he had developed for identifying highly discriminating fragment screens subsequently formed the basis for a range of textual analyses.[9] Following this, he directed an extensive programme of research to develop the connection-table, screening and atom-by-atom components of conventional 2D substructure searching systems so that they could be used to represent and to search databases of the generic Markush structures that are found in many chemical patents. The programme extended over some 15 years in collaboration with a range of industrial partners, and resulted in techniques that contributed to operational systems for structure-based access to generic chemical structures.[10]

After studying Chemistry at Oxford, I came to Sheffield in 1975 to take the MSc Information Science program, and it was here that I first encountered computers, with an initial compulsory programming course in assembler and then an optional one in COBOL. I was absolutely fascinated by this and decided to do my dissertation on an extension of some of Mike's previous work on indexing chemical reactions using Wiswesser Line Notations. It was at this point I encountered the second formative influence on me. George Vladutz was a Romanian chemist who, after working for many years in senior positions in VINITI in the USSR, came to Sheffield on a one-year visiting fellowship.[11] He was a remarkable man – he was fluent in seven languages and could get by in several others – and in the early 1960s had been the first to suggest that computers could be used for the automatic indexing of chemical reactions and for suggesting novel synthetic pathways. During the year he was in Sheffield (after which he went to work for the Institute for Scientific Information in the USA) he returned to this early work and devised an approach to reaction indexing based on a maximal common subgraph isomorphism algorithm. The implementation of George's ideas formed an important part of the PhD that I undertook under Mike's supervision on the completion of my MSc, and of subsequent operational systems for searching reaction databases.

During the PhD I began to read the extensive literature that already existed on the use of similarity and clustering techniques by academic researchers in textual information retrieval (IR). Although elegant in concept, the methods then available were restricted to the processing of small datasets, and my postdoc sought to extend the available methods to large databases of documents, seeking methods that were not only sufficiently efficient for large-scale use but that were also effective in operation, i.e., methods that could retrieve documents that were relevant to a user query. It was at this stage that I began to recognise the significant analogies that exist between searching chemical and textual databases.[12] For example, textual documents can be characterised for retrieval by their constituent words, and molecules can be characterised by their constituent substructural fragments, which suggested that similarity and clustering techniques based on the identification of matching words in pairs of documents might also be based on the matching of common fragments in pairs of molecules. Also, a document can be relevant or irrelevant to a query and a molecule can be active or inactive in a biological screen, which suggests that analogous performance criteria can be used to evaluate the effectiveness of the two types of database system.

These analogies provided the starting point for a programme of research that occupied much of my time until our turn-of-the-century cut-off point (and indeed for many years after that). The work commenced with a collaboration with Pfizer's research centre in Sandwich, the first of so many of our projects that have been supported by pharmaceutical and agrochemical companies both in the UK and elsewhere. The research demonstrated that fragment-based measures of intermolecular structural similarity were able to identify molecules with similar biological properties, an early example of what is often referred to as the Similar Property Principle. This finding provided the basis for one of the first systems for chemical similarity searching (which is still one of the most common forms of ligand-based virtual screening) and for a non-random approach to biological screening (where a database was clustered and then the clusters tested to identify those containing active molecules).[13] Other successful applications of IR techniques to the cheminformatics context included relevance feedback and data fusion *inter alia*. Relevance feedback involves using user judgements of the relevance or otherwise of the results of an initial search of a document database to weight the query terms for a second, and hopefully more effective, search.[14] The weighting schemes used in IR (many of which had a strong theoretical basis) were found to enhance the performance of substructural analysis: this was one of the first chemical applications of machine learning and involves calculating weights that reflect the extent to which the presence of individual fragments contribute positively or negatively to the bioactivity of molecules.[15] Data fusion was originally developed for signal processing in defence applications, but was adapted for use in IR by combining the rankings produced by different retrieval mechanisms when applied to the search of a document database.[16] The cheminformatics analogy is to conduct multiple similarity searches using, e.g., several different types of fingerprint or several different target structures, procedures that have been shown to result in significant increases in the effectiveness of similarity-based virtual screening.[17]

It was from reading the IR literature that I became familiar with the work of Karen Sparck Jones at the University of Cambridge's Computer Laboratory.[18] Karen was one of the pioneers of IR research, helping to develop many of the techniques that underlie modern search engines, and demonstrating the value of extremely detailed, systematic comparisons of different approaches to identify those that were the most effective in operation. We adopted her approach in the Pfizer project mentioned previously, comparing a number of different fragment-based similarity coefficients and of clustering methods, and hence identifying the Tanimoto coefficient and the Jarvis-Patrick method as the most suitable for chemical applications (although later hardware and software developments meant that other clustering methods are now used). This comparative approach was subsequently used to compare the effectiveness of several other important procedures in cheminformatics, e.g., algorithms for subgraph and maximum common subgraph isomorphism and for diversity-based compound selection. While perhaps not as striking as research that develops new algorithms and methods, such studies serve to identify tools of demonstrable effectiveness for what was by then a rapidly growing scientific community (as evidenced by the appearance of the first issues of the *Journal of Molecular Graphics and Modelling* and of the *Journal of Computer-Aided Molecular Design* in 1983 and 1987 respectively). Later work in Sheffield made increasing use of another IR approach with which Karen had been heavily involved, that of using specially curated document test collections, i.e., databases where it was known which individual documents were relevant to each of a set of associated user queries. In the chemical context, the test collections that we, and subsequently many others, used, were sets of molecules with associated bioactivity data (the first of these being the MDL Drug Data Report database) that could be used as community-wide benchmarks for studies in virtual screening.

A retirement get-together in 2019 with collaborators and current and previous members of Peter's research group. Photographer: Peter Bath.

The increasing availability of 3D structure generation programmes such as CONCORD and CORINA in the late 1980s meant that it became possible to create databases of 3D structures from existing 2D databases, and it was hence natural to consider how the methods that had been developed for 2D substructure searching might be applied to searches for 3D pharmacophoric patterns. Mike Lynch's work had shown that fragments describing patterns of atoms and bonds provided an efficient screening mechanism in 2D substructure searching, followed by a subgraph isomorphism check for an exact match with the atoms and bonds comprising the query substructure. Focussing on the distances separating atoms, rather than the bonds linking them together, provided a natural way of adapting these 2D techniques to permit the introduction of operational systems for 3D searching, with subsequent developments based on smoothed bounded distances permitting the extension of these ideas to encompass the conformational flexibility that characterises many small molecules of biological interest. Other research in the 1980s and 1990s involved work on molecular diversity, genetic algorithms for molecular docking and pharmacophore mapping, and the application of small-molecule graph matching algorithms to the analysis of biological macromolecules *inter alia*.[19,20]

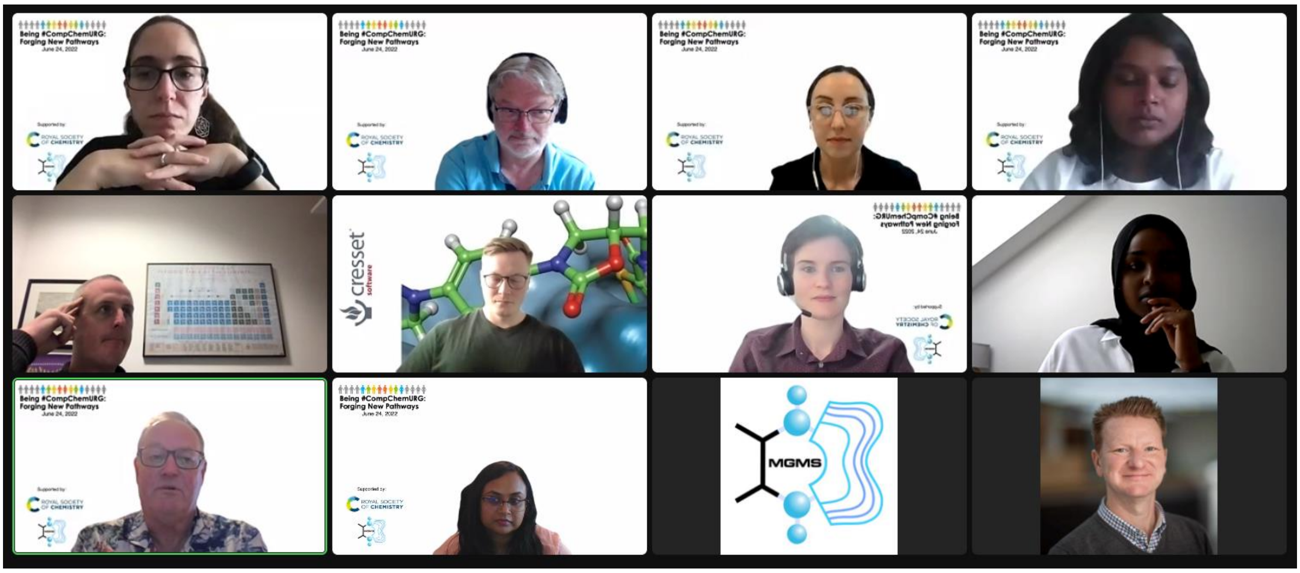
In conclusion, it is perhaps not unreasonable to suggest that the research in Sheffield has played a significant role in the development of cheminformatics – most obviously in ways of processing the databases that increasingly underlie so much research in modern chemistry – and these studies have continued[21] since the early days that have been reviewed here.

(1) Brown, F.K. Chemoinformatics: what is it and how does it impact drug discovery? *Annual Reports in Medicinal Chemistry*. **1998**, *33*, 375-384.
(2) Hann, M.; Green, R. Chemoinformatics: a new name for an old problem? *Current Opinion in Chemical Biology*. **1999**, *3*, 379-383.
(3) Ray, L.C.; Kirsch, R.A. Finding chemical records by digital computers. *Science*. **1957**, *126*, 814-819.
(4) Hansch, C. *et al*. Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature*. **1962**, *194*, 178-180.
(5) Willett, P. From chemical documentation to chemoinformatics: 50 years of chemical information science. *Journal of Information Science*. **2008**, *34*, 477-499.
(6) Schofield, H. *et al.* Recent developments in chemoinformatics education. *Drug Discovery Today*. **2001**, *6*, 931-934.
(7) Lynch, M.F. *et al. Computer Handling of Chemical Structure Information*. Macdonald, 1971.
(8) Lynch, M.F.; Willett, P. Information retrieval research in the Department of Information Studies, University of Sheffield: 1965-1985. *Journal of Information Science*. **1987,** *13*, 221-234.
(9) Lynch, M.F. Variety generation—a reinterpretation of Shannon's mathematical theory of communication, and its implications for information science. *Journal of the American Society for Information Science*. **1977**, *28* , 19-25.
(10) Lynch, M.F.; Holliday, J.D. The Sheffield generic structures project: a retrospective review. *Journal of Chemical Information and Computer Sciences*. **1996**, *36*, 930-936.
(11) Lynch, M.F.; Willett, P. George Vladutz, 1928-1990. *Journal of Chemical Information and Computer Sciences*. **1990**, *30*, 349.
(12) Willett, P. Textual and chemical information processing: different domains but similar algorithms. *Information Research*. **2000**, *5*, at http://informationr.net/ir/5-2/paper69.html.
(13) Willett, P. *Similarity and Clustering in Chemical Information Systems.* Research Studies Press, 1987.
(14) Robertson, S.E.; Sparck Jones, K. Relevance weighting of search terms. *Journal of the American Society for Information Science and Technology*. **1976**, **27**, 129-146.
(15) Ormerod, A. *et al.* Comparison of fragment weighting schemes for substructural analysis. *Quantitative Structure-Activity Relationships*. **1989**, *8*, 115-129.
(16) Belkin, N.J. *et al.* Combining the evidence of multiple query representations for information retrieval. *Information Processing and Management*. **1995**, *31*, 431-448.
(17) Ginn, C.M.R. *et al.* Combination of molecular similarity measures using data fusion. *Perspectives in Drug Discovery and Design*. **2000**, *20*, 1-16.
(18) Robertson, S.E.; Tait, J. Karen Sparck Jones. *Journal of the American Society for Information Science and Technology*. **2008**, *59*, 852-854.

(19) Bishop, N. *et al.* Chemoinformatics research at the University of Sheffield: a history and citation analysis. *Journal of Information Science.* **2003**, *29*, 249-267.
(20) Willett, P. Searching techniques for databases of two- and three-dimensional chemical structures. *Journal of Medicinal Chemistry*. **2005**, *48*, 4183-4199.
(21) Gillet, V.J. *et al.* Chemoinformatics at the University of Sheffield 2002–2014. *Molecular Informatics.* **2015**, *34*, 598-607.

Being #CompChemURG: Forging New pathways

Contribution from Hannah Bruce Macdonald, Senior Scientist I, Computational Chemistry, MSD, London



The Binding Site held its second virtual conference *Being #CompChemURG: Forging New Pathways* on 27 June 2022, supported by the Molecular Graphics and Modelling Society (MGMS) and the RSC's Inclusion and Diversity Fund. Talks from Chris Swain, Scott Midgely, Colin Edge, Anita Nivedha, Idil Ismail, Andrew Leach, and Caroline Lynn Kamerlin covered a range of personal career stories, focussing on career transitions. The talks were followed by a panel discussion with the speakers and joined by Nessa Carson. The meeting was well-attended, with over 80 online attendees over the afternoon.

The Binding Site took the opportunity to introduce their mentorship scheme, to give an opportunity for under-represented groups in computational chemistry going through career transitions. They are looking for applicants for both mentors and mentees, with more details available on the website.

UKeiG Call for Nominations for the Prestigious Tony Kent Strix Award 2022

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

The UK e-information Group (UKeiG) is delighted to announce the call for nominations for the prestigious Tony Kent Strix Award 2022. Nominations should be received by 6 pm GMT on Friday 30 September 2022.

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

The Award is given in recognition of an outstanding practical innovation or achievement in the field of information retrieval in its widest sense. This could take the form of an application or service, or an overall appreciation of past achievements that have led to significant advances. The award is open to individuals or groups from anywhere in the world and submissions would be welcomed from CICAG members.

Nominations must be for a major, sustained or influential achievement that meets one or more of the following criteria:

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

Service delivery
The development and management of systems, networks or services:

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

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

Submissions must include:

- The name, institutional address and qualifications of the nominee
- A brief biography (maximum one A4 page)
- A selective bibliography (key publications relevant to the nomination)
- A justification for the nomination (maximum one A4 page) showing clearly which of the award criteria the nominee meets and how they are met
- Additional material – letters of support, for example (letters from previous winners would be especially valuable)

It is possible that the Award Committee will request additional information from the nominators for those nominees considered suitable candidates for the award. An unsuccessful nomination from previous years may be reconsidered provided the nominator updates it, if necessary, to reflect the current extent of the candidate's achievement.

Nominations for the 2022 award must reach the judges by 6 pm GMT on Friday 30 September 2022. Please email to: John Wickenden secretary.ukeig@cilip.org.uk (Hon. Secretary UkeiG), and copy in Gary Horrocks info.ukeig@cilip.org.uk (UKeiG administrator) and Sue Silcocks treasurer.ukeig@cilip.org.uk (Hon. Treasurer UKeiG).

For more information about UKeiG, the Tony Kent Strix Award and previous winners of this prestigious international award is available on the CILIP website.

A video of UKeiG's 7th Tony Kent Strix Annual Memorial Lecture 2021 – delivered by the 2020 Strix award winner Ian Ruthven, Professor of Information Seeking and Retrieval at the Department of Computer and Information Sciences, University of Strathclyde – is available. The Award was presented in recognition Professor Ruthven's outstanding practical innovation and achievements in the field of information retrieval.

Professor Ruthven's lecture was entitled: Google's what you use when Alexa doesn't know the answer, Uncle Ian.

> Abstract: Search is now a pervasive online activity. The ability to successfully interact with the information tools we have available to us is a key life skill, one that forms part of what is often seen as essential information literacy. However, even though we may not be able to imagine everyday life without these tools, they are a staggeringly recent phenomenon. Key to their success has been the interfaces and interaction models that underpin these information tools. Interactive search has been the site of rich study and experimentation and this research has taught us much about how we work with information and how information systems can support our interactions with information. In this presentation, I shall look at some of the history of interactive searching, discuss why some of the tools we use now are more successful than others, and look forward to how we might be interacting with information in the future.

A PDF of the slides is also available.

There was no Strix Prize winner in 2021, but the UKeiG is still planning to hold a lecture in December in 2022. More information will be available closer to the time.

- -

Welcome to the New Era of Scientific Publishing

Contribution from Tobias Kuhn, Assistant Professor, Department of Computer Science, VU University of Amsterdam, email: kuhntobias@gmail.com

Originally posted on Github, reproduced with permission from the author. Reporting on work with Cristina Bucur, Davide Ceolin and Jacco van Ossenbruggen[1,2]

I believe we have made the first steps venturing into a new era of scientific publishing. Let me explain. [Update: At what point exactly a new era begins and what counts as first steps are of course subject to debate. I therefore added a section on related initiatives and further reading.]

Science is nowadays communicated in a digital manner through the internet. We essentially have a kind of 'scientific knowledge cloud', where researchers with the help of publishers upload their latest findings in the form of scientific articles, and where everybody who is interested can access and retrieve these findings. (This is in fact only true for articles that are published as open access, but that is not the point here.)

scientific knowledge
?

In a sense, this scientific knowledge cloud has been a big success, but it also has serious limitations. Imagine somebody wanting to build an application that feeds on all these findings, for example to help researchers learn about interesting new developments or to summarise scientific consensus for laypeople. Concretely, a developer might face a task like this one:

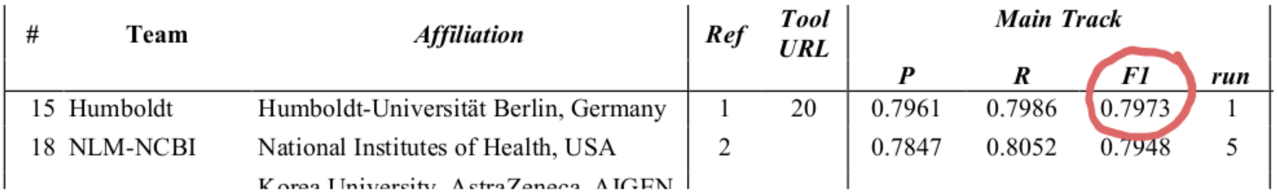
> Retrieve all genes that have been found to play a role in a part of the respiratory system in Covid-19 patients. Only include results from randomised controlled trials published in the last seven days.

To a naive developer without experience in how scientific knowledge is communicated, this might sound quite easy. One would just have to find the right API, translate the task description into a query, and possibly do some post-processing and filtering. But everybody who knows a bit how science is communicated immediately realises that this will take a much bigger effort.

Why text mining is not the solution

The problem is that scientific results are represented and published in plain text and not in a structured format that software could understand. So, in order to access the scientific findings and their contexts, one has to apply text mining first. Unfortunately, despite all the impressive progress with deep learning and related techniques in the past few years, text mining is not good enough, and probably will never be.

To illustrate the point, we can look at the results of the seventh BioCreative workshop held in November 2021, where world-leading research teams competed in extracting entities and relations from scientific texts. Just to detect the type of a relation between a given drug and a given gene out of 13 given relation types, the best system achieved an F-score of 0.7973.



#	Team	Affiliation	Ref	Tool URL	Main Track			
					P	R	F1	run
15	Humboldt	Humboldt-Universität Berlin, Germany	1	20	0.7961	0.7986	0.7973	1
18	NLM-NCBI	National Institutes of Health, USA	2		0.7847	0.8052	0.7948	5
		Korea University, AstraZeneca, AIGEN						

But that is just the relation type. To get the full relation out, we first have to know the entity on the left-hand side (subject) and right-hand side (object) of the relation. We can look at a different task of the BioCreative workshop to get a feeling of how well extracting these subjects and objects work. The task focused on extracting chemicals, and this is done in a two-stage process. First, the entities are recognised in the text, with an F-score of the best system of 0.8672, and then the recognised chemicals are linked to the corresponding formal identifiers, with the best F-score being 0.8136:

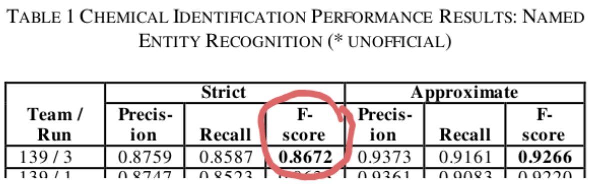


TABLE 1 CHEMICAL IDENTIFICATION PERFORMANCE RESULTS: NAMED ENTITY RECOGNITION (* UNOFFICIAL)

Team / Run	Strict			Approximate		
	Precis-ion	Recall	F-score	Precis-ion	Recall	F-score
139 / 3	0.8759	0.8587	**0.8672**	0.9373	0.9161	**0.9266**
139 / 1	0.8747	0.8523	[illegible]	0.9361	0.9083	0.9220

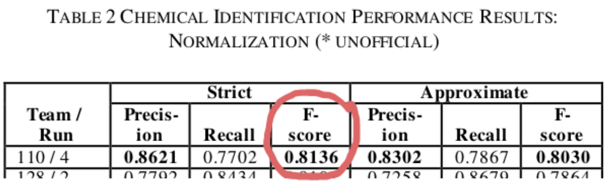
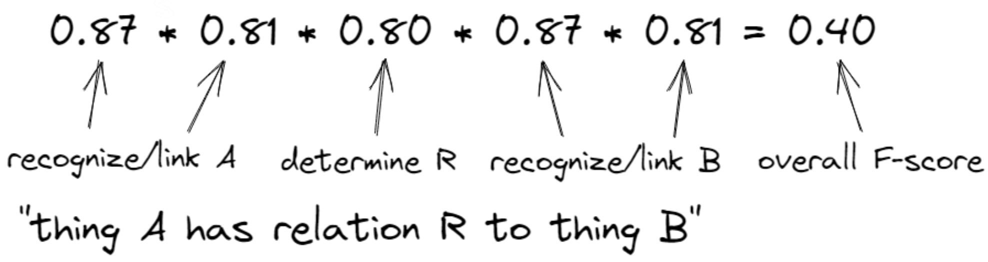


TABLE 2 CHEMICAL IDENTIFICATION PERFORMANCE RESULTS: NORMALIZATION (* UNOFFICIAL)

Team / Run	Strict			Approximate		
	Precis-ion	Recall	F-score	Precis-ion	Recall	F-score
110 / 4	**0.8621**	0.7702	**0.8136**	**0.8302**	0.7867	**0.8030**
128 / 2	0.7792	0.8434	[illegible]	0.7258	0.8679	0.7864

A very rough back-of-the-envelope calculation can give us an estimate of the quality of mining such entire relations:



An overall F-score of 0.40, as resulting from this calculation, roughly means that around 60% of retrieved relations are wrong and 60% of existing relations are not retrieved. This is clearly not even close to good enough for most types of possible applications. And mind you, these numbers come from the best performing systems when world-leading research groups compete and probably put months of effort into optimising their systems for the specific problem. Moreover, we are talking here only about the simplest possible kind of relations of the form subject-relation-object.

This seems to point to a deeper problem. Text mining is just a work-around, and the real problem is elsewhere. As Barend Mons rhetorically asked "Why bury it first and then mine it again?". Instead of seeing text mining as the ultimate solution, we should just stop burying our knowledge in the first place.

In the work I will explain below, we wanted to find out how we can practically publish findings without burying them.

Representing scientific findings in logic

As a first step to experiment with such a new way of publishing, we needed to find a general way of how to represent high-level scientific findings in some sort of formal logic. Even though such findings are arguably the most important outcome of science, there was no prior work on practically mapping (most of) these findings to formal logic across domains. To better understand the logical structure of such high-level scientific findings (e.g. that a gene tends to have a certain effect on the course of a given disease), we selected a random sample of 75 research articles from Semantic Scholar.

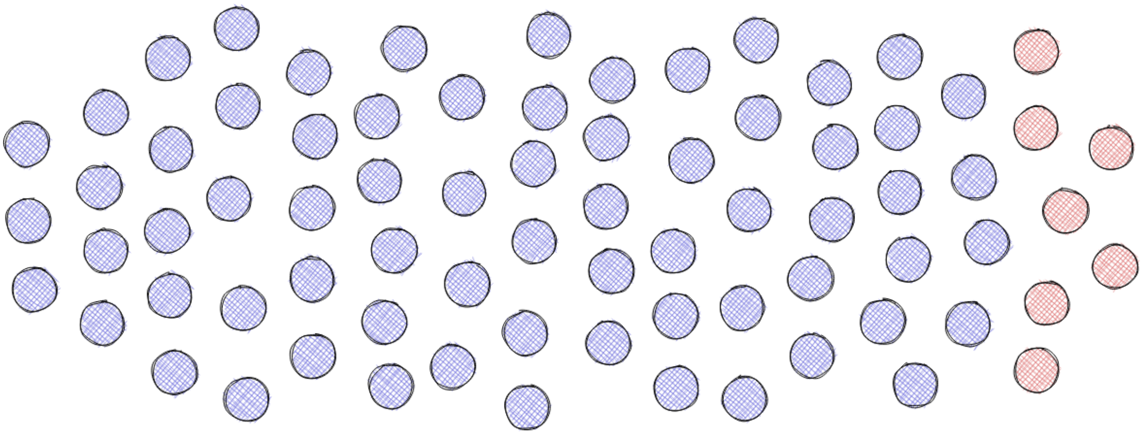
Studying the high-level findings from these random articles, we managed to elicit a widespread logical pattern, which we then turned in the super-pattern model. It consists of five slots to be filled, and translates directly to a logical formula. Our paper[1] explains the details, and I give here just one example. The finding of the article entitled P-glycoprotein expression in rat brain endothelial cells: evidence for regulation by transient oxidative stress can be expressed with the super-pattern as follows:

- **Context class:** rat brain endothelial cell
- **Subject class:** transient oxidative stress
- **Qualifier:** generally
- **Relation:** affects
- **Object class:** Pgp expression

The three class slots take any class identifier (here shown by class name for simplicity), whereas the qualifier and the relation come from closed lists of a few possible values. Informally, the example above means that whenever there is an instance of 'transient oxidative stress' in the context of an instance of 'rat brain endothelial cell' then generally (being defined as in at least 90% of the cases) it has a relation of type 'affects' to an instance of 'Pgp expression' that is in the same context. Formally, it corresponds to this logical formula (in a bit of a non-standard notation using conditional probability):

$$P(\ \exists z(\text{pgp-expression}(z) \wedge \text{in-context}(z,x) \wedge \text{affects}(y,z))\ |\ \text{transient-oxidative-stress}(y) \wedge \text{rat-brain-endothelial-cell}(x) \wedge \text{in-context}(y,x)\ ) \geq 0.9$$

Once we discovered this pattern we tried to formalise the 75 high-level findings, and this was the result:



68 of the 75 findings, shown in blue, could be represented with the super-pattern. And the remaining seven findings, shown in red, are in fact *easier* to represent, as they can be captured with a simple subject-relation-object structure. (So these latter simpler structures are the ones that existing text-mining approaches struggle with, and it is safe to assume that they would perform even worse on the more complicated statements here shown in blue.)

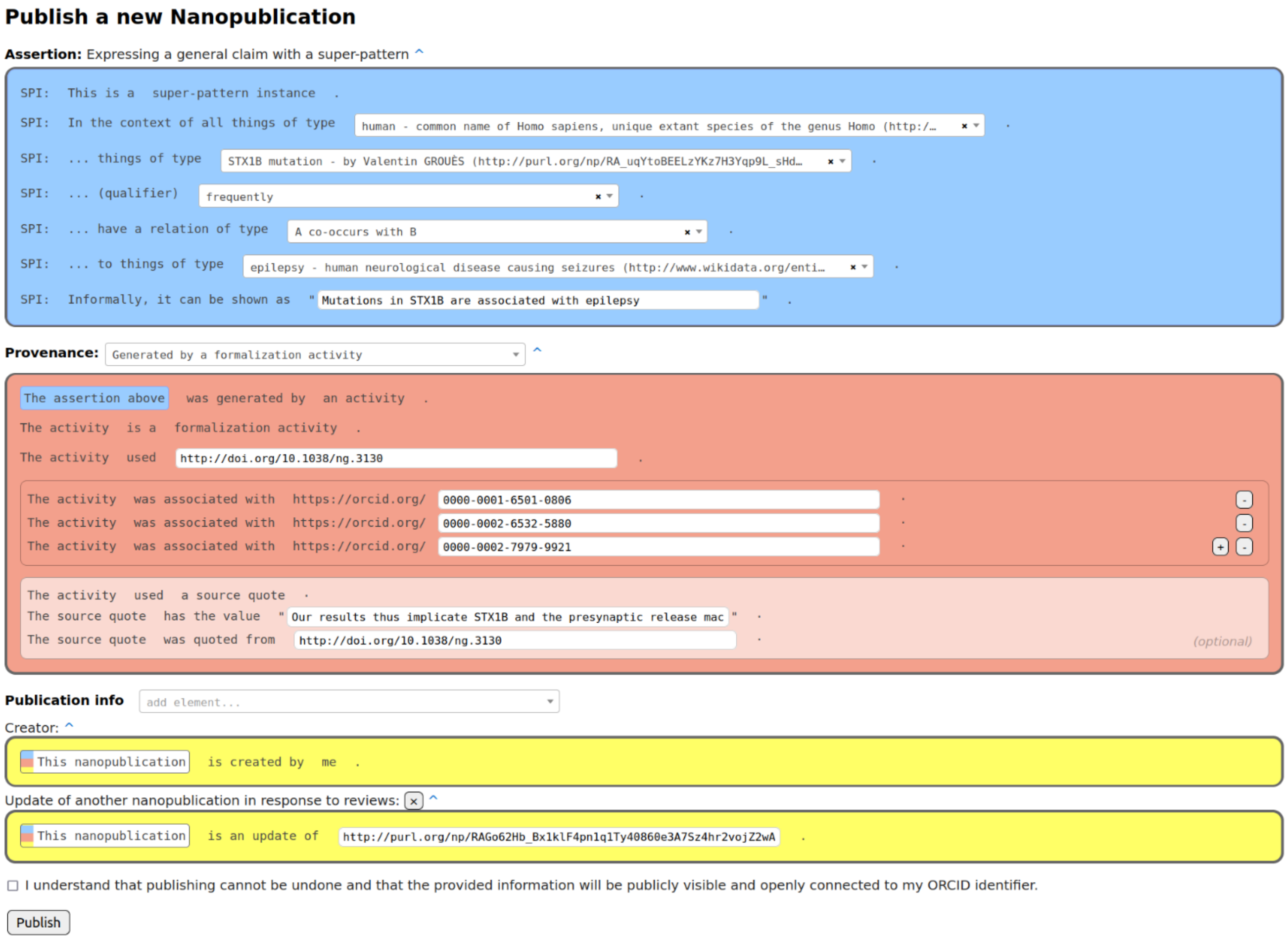
So, it seems that we have found a logical pattern that allows us to represent most high-level scientific findings from different disciplines.

Stepping into the new era of scientific publishing

Next, we wanted to make a serious practical step into the new era where findings are machine-interpretable from the start. We designed this as a kind of field study with a special issue of machine-interpretable papers at an existing journal. We wanted these papers to look like regular papers to those who want to look at them in that way, but they should also come with representations in formal logic for anyone or anything that knows how to deal with that. For that special issue, we chose the journal Data Science, of which I am an editor-in-chief.

We also had to make a practical concession though: while the whole setup *could* be used to publish novel findings, we restricted ourselves to findings from existing publications. For that, we introduced the concept of a 'formalisation paper' whose novel contribution is the formalisation of an existing finding. So, authors of a formalisation paper take credit for the formalisation of the finding, but not for the finding itself.

To represent these findings and thereby the formalisation papers, we used the nanopublication format and the Nanobench tool. Researchers who contributed to this special issue filled in a form to express and submit their formalisation that looked like this:



Publish a new Nanopublication

Assertion: Expressing a general claim with a super-pattern ^

SPI: This is a super-pattern instance .
SPI: In the context of all things of type human - common name of Homo sapiens, unique extant species of the genus Homo (http:/… × ▾ .
SPI: ... things of type STX1B mutation - by Valentin GROUÈS (http://purl.org/np/RA_uqYtoBEELzYKz7H3Yqp9L_sHd… × ▾ .
SPI: ... (qualifier) frequently × ▾ .
SPI: ... have a relation of type A co-occurs with B × ▾ .
SPI: ... to things of type epilepsy - human neurological disease causing seizures (http://www.wikidata.org/enti… × ▾ .
SPI: Informally, it can be shown as " Mutations in STX1B are associated with epilepsy " .

Provenance: Generated by a formalization activity ▾ ^

The assertion above was generated by an activity .
The activity is a formalization activity .
The activity used http://doi.org/10.1038/ng.3130 .

The activity was associated with https://orcid.org/ 0000-0001-6501-0806 . -
The activity was associated with https://orcid.org/ 0000-0002-6532-5880 . -
The activity was associated with https://orcid.org/ 0000-0002-7979-9921 . + -

The activity used a source quote .
The source quote has the value " Our results thus implicate STX1B and the presynaptic release mac " .
The source quote was quoted from http://doi.org/10.1038/ng.3130 . *(optional)*

Publication info add element... ▾

Creator: ^

This nanopublication is created by me .

Update of another nanopublication in response to reviews: ☒ ^

This nanopublication is an update of http://purl.org/np/RAGo62Hb_Bx1klF4pn1q1Ty40860e3A7Sz4hr2vojZ2wA .

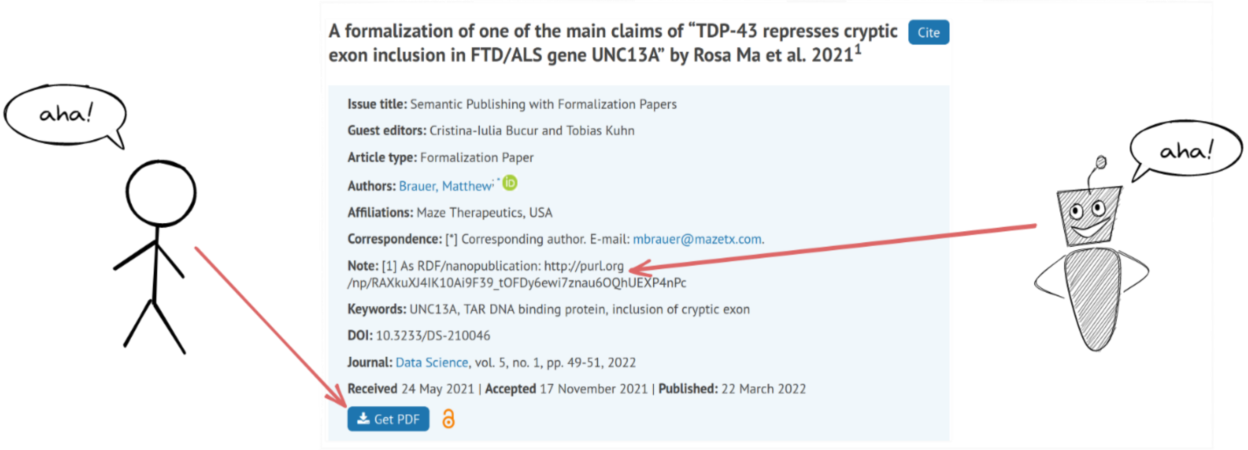
☐ I understand that publishing cannot be undone and that the provided information will be publicly visible and openly connected to my ORCID identifier.

Apart from the actual formalisation in the assertion part of the nanopublication (blue), this also involved specifying the provenance of that formalisation (red) and further metadata (yellow). The provenance in this case states that the assertion was derived by a formalisation activity taking an existing paper as input. This interface provides auto-completing drop-down menus and maps the result to the logic language RDF in the back. I won't go into the details of defining new classes and reviewing (both done with nanopublications and Nanobench too), but you can have a look at the preprint of our paper[2] on that.

We ended up with 15 formalisation papers in our special issue, as summarised by this table:

	Authors	CONTEXT	SUBJECT	QUALIFIER	RELATION	OBJECT
1	Joslin	early human adipogenesis	regulatory element within the first intron of FTO	generally	affects	expression of genes IRX3 and IRX5
2	Nichols	human motor neuron	TAR DNA binding protein	can generally	contributes to	transcription of stmn2
3	Mietchen	dejellied fertilizable stage VI Xenopus laevis oocyte	strong static magnetic field	generally	affects	cell cortex
4	Ehrhart, Evelo	(no context class)	genes associated with CAKUT	sometimes	is same as	targets of vitamin A
5	Patrinos	patient undergoing PCI	pharmacogenomics guided clopidogrel therapy	generally	enables	cost-effective treatment
6	Martorana	human	smoothened signaling pathway	mostly	affects	astrocyte development
7	Mietchen, Penev, Dimitrova	biodiversity data	license with non-commercial clause	generally	inhibits	data reuse
8	Dimitrova	release of OpenBiodiv knowledge graph	triple in OpenBiodiv knowledge graph	generally	is same as	semantic triple extracted from biodiversity literature
9	Brauer	UNC13A	TAR DNA binding protein	generally	inhibits	inclusion of cryptic exon
10	Dumontier	data set	adherence to the FAIR guiding principles	can generally	enables	automated discovery
11	Queralt Rosinach	human	NGLY1 deficiency	always	is caused by	dysfunction of ERAD pathway
12	Usbeck	social group	relative neocortex size	never	affects	social group size
13	Bainer	ecm bound cancer cell	glycocayx bulk	generally	increases	integrin clustering
14	Grouès, Vega Moreno, Satagopam	human	STX1B mutation	frequently	co-occurs with	epilepsy
15	de Boer	digital humanities research	usage of Linked Data Scopes	can generally	contributes to	transparency

Each row of this table corresponds to a formalisation, and can be written out as a logic formula. So, these papers look here quite different from what we are used to. But if you go to the official page for the special issue on the publisher's website, they look like regular papers. Besides the regular button to download the paper as a PDF, there is also a link that points to the nanopublication representation, thereby connecting the two ways of looking at this paper:



A formalization of one of the main claims of "TDP-43 represses cryptic exon inclusion in FTD/ALS gene UNC13A" by Rosa Ma et al. 2021[1]

Issue title: Semantic Publishing with Formalization Papers

Guest editors: Cristina-Iulia Bucur and Tobias Kuhn

Article type: Formalization Paper

Authors: Brauer, Matthew[*]

Affiliations: Maze Therapeutics, USA

Correspondence: [*] Corresponding author. E-mail: mbrauer@mazetx.com.

Note: [1] As RDF/nanopublication: http://purl.org/np/RAXkuXJ4IK10Ai9F39_tOFDy6ewi7znau6OQhUEXP4nPc

Keywords: UNC13A, TAR DNA binding protein, inclusion of cryptic exon

DOI: 10.3233/DS-210046

Journal: Data Science, vol. 5, no. 1, pp. 49-51, 2022

Received 24 May 2021 | **Accepted** 17 November 2021 | **Published:** 22 March 2022

So, for the first time, software can reliably interpret the main high-level findings of scientific publications. This special issue is a just a small first step, but it could prove to be the first step into a new era of scientific publishing. The practical immediate consequences of this might seem limited, but I think the longer-term potential of making scientific knowledge accessible to the interpretation by machines is hard to overstate.

References

1. Bucur, C.-J.; Kuhn, T.; Ceolin, D.; van Ossenbruggen, J. Expressing high-level scientific claims with formal semantics. In *Proceedings of K-CAP '21*. ACM, **2021**.
2. Bucur, C.-J.; Kuhn, T.; Ceolin, D.; van Ossenbruggen, J. Nanopublication-based semantic publishing and reviewing: a field study with formalization papers. *arXiv 2203.01608*, **2022**.

[Update:] Related initiatives and further reading

For completeness, I list here some related initiatives and further reading. I only include references where machine-interpretable representations of findings are part of or treated as scientific publications.

- **Initiative:** GloBI with its integrations to Plazi/Zenodo and Pensoft
- **Journal:** Biodiversity Data Journal
- **Viewpoint paper:** Upham, N.S. *et al.* Liberating host–virus knowledge from biological dark data. *The Lancet Planetary Health*. **2021**, *5*(10), E746-E750.
- **Implementation:** Penev, L. *et al*. Implementation of TaxPub, an NLM DTD extension for domain-specific markup in taxonomy, from the experience of a biodiversity publisher, In *Journal Article Tag Suite Conference (JATS-Con) Proceedings* **2012.**
- **Editorial:** Penev, L. *et al.* Taxonomy shifts up a gear: new publishing tools to accelerate biodiversity research. *ZooKeys*. **2010**, *50*, 1-4.
- **Position paper:** Kuhn, T.; Dumontier, M. Genuine semantic publishing. *Data Science*. **2017**, 1(1-2), 139-154.

- -

Greg Landrum Receives the Mike Lynch Award

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

The Mike Lynch Award recognises and encourages outstanding accomplishments in education, research, and development activities that are related to the systems and methods used to store, process, and retrieve information about chemical structures, reactions, and properties.

The Trustees of the CSA Trust awarded the 2022 Award to Greg Landrum in recognition of his work on the development of RDKit and his fostering of the community around it, a transformative software resource for cheminformatics and machine learning. The Award was presented at the 12th International Conference on Chemical Structures (ICCS), Noordwijkerhout, in June 2022.

Greg is a senior scientist in Sereina Riniker's group at the ETH Zurich, Founder and Managing Director of T5 Informatics GmbH, a Senior Advisor to Knime, and the primary developer for the RDKit.

Greg Landrum receiving the Mike Lynch Award from Noel O'Boyle and Wendy Warr. Image credit: Wendy Warr.

Jonathan Goodman, chair of the CSA Trust, commented: "I am delighted that Greg Landrum has accepted this award. His work on RDKit has made chemical informatics techniques more accessible to scientists worldwide both in industry and academia. When introducing students to cheminformatics, becoming familiar with RDKit is a key part of the learning process, and makes it possible to explore new ideas in chemical information rapidly and reliably."

Greg Landrum, who gave a keynote address at the ICCS, said: "I am really honoured to have been selected for this award; it's especially meaningful to me because of the foundational importance of Mike Lynch and the 'Sheffield school' to our field. I would particularly like to thank the Trust for recognising the importance of the RDKit community."

- -

Bioinformatics in the Post-AlphaFold 2 Era

Contribution from Carlos Outeiral, Postdoctoral Fellow at the Oxford Protein Informatics Group, Stipendiary Lecturer in Biochemistry at St Peter's College, University of Oxford, email: carlos.outeiral@gtc.ox.ac.uk

Scientific revolutions are historical events, much like those we have been living through recently, that many read about yet few expect to experience. Truly ground-breaking ideas are rare enough that you never expect to be amongst the first to witness them. But that is exactly what happened in November 2020, when, as myself and a couple of hundred protein scientists around the world tuned in to the CASP14 conference, a scientific team at Google DeepMind announced AlphaFold 2, a deep-learning system able to accurately predict the structure of individual protein chains. The discovery, solving a central problem in biochemistry that had remained unresolved for six decades, rippled waves throughout the life sciences and started a scientific revolution in its own right.

The announcement filled the computational biology community with awe and worry in equal parts. Awe, because we quickly understood that structure prediction would lead to enormous advances. Easy access to protein structures would sidestep classical roadblocks in drug discovery, aid to unravel the function of many unknown genes, and supercharge protein engineering. In essence, it would better our world, which is why many of us are in science in the first place. But it also brought forward a deep sense of worry – because we were struggling to grasp how a central problem to computational biology, object of significant research effort for decades, had been solved out of the blue by a relative outsider. This pushed the community to intense self-reflection, eventually changing the way we conduct our research.

One and a half years later, this existential dread seems to have given way to one of the most productive periods in computational biology, ever. We have new tools, we know new tricks, and we work on new problems. It truly has been a wonder to be part of the field. In reflection of this growth, I would like to use this brief essay examine how the field has reacted in the past year and a half, and what the current research trends look like – and, of course, I would like to make a few bets about the future so I can be embarrassed about my naivete not too long after this article gets published.

1. What happened last year (if you haven't been paying attention to science news)

Proteins are molecular machines that mediate almost any chemistry that occurs in the cell. Crucial to their functional diversity is their ability to adopt complicated, highly conserved three-dimensional structures. Indeed, the central tenet of structural biology is that structure determines function. Enzymes, for example, have an intricate scaffold that places the right chemical groups in adequate positions of space so as to stabilise a transition state, transfer or abstract protons, start nucleophilic attacks, or other operations, to speed up chemical reactions. Unfortunately, these structures have been notoriously difficult to determine experimentally: individual crystal structures of proteins have often been enough to get a paper published in a top journal like *Cell*, *Nature* or *Science*. Similarly, many biochemical projects, for example in drug discovery or in the characterisation of molecular processes, have been hurdled – or even abandoned – by the lack of structural information about certain proteins.

Sidestepping costly experimental determination has therefore been a chief objective of computational biology for most of the past five decades. Progress in the field has been driven by the Critical Assessment of protein Structure Prediction (CASP), a biennial blind test where computational biologists try to predict the structure of several proteins whose structure has been determined experimentally – yet not publicly released. This assessment has witnessed significant leaps over the years. Early approaches were often focused on physical modelling, but soon became overtaken by methods incorporating knowledge-based heuristics. In recent years, it has become commonplace to use statistical approaches exploiting large databases of evolutionary information to predict which pairs of amino acids were in contact. However, one thing was clear: except for some special cases, we were not very good at predicting the structure of proteins.

Then, out of the blue, the algorithm presented by DeepMind at the fourteenth edition of CASP beat every other group, and then some. Their methodology introduced a large number of novel ideas, working in unison to produce predictions of unprecedented accuracy. For example, rather than using a costly exploratory algorithm, like simulated annealing, AlphaFold 2 predicts a structure 'end-to-end' on a single shot. The model takes a protein sequence (well, a multiple sequence alignment) on one end, and outputs a list of coordinates on the other. Borrowing from the well-established ideas in the field, they also used their machine-learning expertise to extract as much information as possible from evolutionary databases. Finally, and what perhaps has had most influence in the field afterwards, they used a novel type of deep neural network, namely SE(3)-equivariant transformers. In combination, these advances brought together a step change that marked for the first time an artificial intelligence system solving a large open scientific problem.

There have been some concerns along the way. In the interlude between CASP14 and the publication, there was much discussion in the community about whether DeepMind would publish their findings in a way that was 'runnable' by the community, or if (as they did with the first version of the system) it could only be run on a small number of examples. In the end, they have been a good example of open science. After six long months of waiting, the paper describing AlphaFold 2 was published in *Nature* in July 2021, alongside fully working code and a 30+ page Supplementary Information describing the methods in depth. Barely a week later, the team announced the AlphaFold Database, an online database of predicted structures aiming to have a prediction of every known protein in a couple of years. In the past year, the paper has accrued well over 3,000 citations – and it is well-poised to become one of the influential papers in history.

The magnitude of the breakthrough is hard to put into words. The average prediction of the AlphaFold 2 team was over a standard deviation better than the second-best team, which was already very much world class. One of the experimental groups, upon examining the prediction, realised that they had made a mistake assigning the isomerisation state of a proline (trans, instead of cis). Another, who had struggled to refine the experimental data for months, was able to completely solve the structure within a few hours after witnessing the AlphaFold 2 prediction. Many authorities of the field cite AlphaFold 2 as the greatest advance they have witnessed in their career, and respectable observers expect that a Nobel Prize may be awarded for this work in the near future.

And yet, while there is talk that the protein structure problem has been 'solved', there is no doubt that it is far from closed. There exist several reports of proteins that were wrongly predicted by the system, and there is still much of the human proteome that cannot be characterised structurally – arguably because these proteins' structure depends on the interaction with partners. In particular, AlphaFold 2 is not, as many news outlets claimed, a solution of the 'protein folding problem', or a physicochemical explanation to why (and how) proteins adopt the complicated structures they do. Put another way, this system has yet to teach us any physics or chemistry about the structure of proteins. Or, as I like to think about it: the enormous breakthrough has shone light not only on the problem itself, but on the interesting scientific questions that we are now free to explore after nearly six decades focused almost exclusively on structure prediction.

2. What is happening right now

Few areas of the life sciences have been impartial to the release of AlphaFold 2. Protein structure becoming much more accessible has been the immediate consequence of AlphaFold 2. In the past, a biochemist working with a protein could only hope that their system would, someday, be characterised by a structural group.

Today, if you have an internet connection, you can pretty much type the name of your protein on the AlphaFold Database published by DeepMind and get a decent hit. Even if the protein has not yet been added to the database, community-led tools like ColabFold have made predicting structures as easy as typing a sequence in a text box.

Easily accessible structural predictions have not replaced experimental approaches, and in fact they may well have contributed to revitalise them. In crystallography, for example, it has become commonplace to use predicted protein structures to refine difficult datasets, a technique known as molecular replacement. In CryoEM, where atomic resolution is still very much a challenge, AlphaFold predictions have been combined with low-resolution maps to reconstruct complex systems, such as the widely discussed structure of the nuclear pore complex structure recently published in *Science*. Even in protein NMR there has been some suggestion to combine predicted structures with recorded data. In what I think has been a trend throughout the sciences, rather than bring decay and obsolescence, the strength of AlphaFold has brought widespread enthusiasm and readied a field to tackle new, more interesting problems.

What about computational biology, specifically? In my opinion, the effect has been felt in roughly two ways. On the one hand, since the algorithm was released as an easy-to-use tool, researchers have incorporated it into a variety of pipelines, as well as extended and modified it to solve related problems. On the other hand, the ideas presented in the paper, mainly in terms of the deep-learning architectures used, have provided impetus for many related problems in structural bioinformatics.

The original version of AlphaFold could only predict the structure of individual chains. However, an updated version capable of predicting the structure of protein complexes, that is, proteins that consist of two or more chains that are not tethered by peptide bonds, was released under the name of AlphaFold-Multimer shortly after the original paper had been published. The system is still very much experimental, though. One of its earlier versions, recently patched, had a tendency to produce unresolvable steric clashes, for example building several intertwining alpha-helices in the same region of space. The performance is also much less impressive than the original AlphaFold, and it is known to struggle in many crucial protein-protein interaction problems, such as the binding between an antibody and its antigen. And yet it outperforms most known protein-protein docking computational tools and has been explored in a range of downstream applications. Think of any problem that involves understanding how a protein latches onto another – I bet you someone has thought of applying AlphaFold-Multimer to that problem.

The community has also produced several ‘hacks’ of AlphaFold 2. Some of these have involved modifying the architecture of the model to reduce the runtime, such as faster multiple sequence alignment (MSA) construction by using pre-clustered sequence databases, or altering the parameters of the model (e.g. reducing the recycling iterations). Many of these improvements have been led by the ColabFold project, which has very much opened AlphaFold 2 to non-computational scientists. Other ideas have concerned the prediction of specific target classes, for example by enriching the AlphaFold MSA with special sequence databases. Finally, several papers have been attempting to make AlphaFold 2 produce not a single crystal structure, but an ensemble of conformations, ideally representing different functional states of a protein. This has included a variety of tricks, from running the model multiple times with different seeds, to performing intricate surgery on the MSA. However, while there certainly is progress, the approaches are fairly limited in terms of dynamics.

The ability to produce accurate structure predictions at scale has also enabled new tasks. Several groups have been exploring ‘inverse folding’ models: neural networks that are trained on the backbone positions of a protein, and used to predict protein sequences that would reliably adopt these structures. The idea being, of course, that if you are trying to design a protein, you could just draw up the scaffold in some kind of 3D editor, and then use this model to find potential target sequences. While several papers have explored this idea, one of the most widely discussed used an impressive 12 million predicted protein structures.

Beyond protein structure prediction problems *per se*, the ideas presented by the AlphaFold team have had a wider impact on the field by providing novel tricks to tackle existing problems. Perhaps the most important is the concept of 'SE(3)-equivariant neural networks', deep-learning layers that yield the same result regardless of arbitrary geometric changes, such as rotations and translations. This may be understood as a form of data augmentation: if the model knows that any possible rotation or translation of the data will lead to the same answer, it will need a lot less data to pull it away from wrong models and will therefore be able to learn much more. Informally, this seems to work quite well in a variety of problems involving atoms and coordinates. These networks have been used in areas such as molecular property prediction, pose scoring, representation of molecular graphs, and even molecular force fields.

Docking and scoring are two areas where these new architectures have been broadly used. For example, a widely discussed paper presented EquiBind, an end-to-end approach to protein-ligand docking. The idea is that, instead of using a complicated search function to find ligand-protein poses, a neural network can be trained to take a ligand and a protein and output a potential structure. The method produces structures that are modestly superior to the predictions of other common packages, but at a rate several orders of magnitude faster. Similar methodologies have been applied to develop scoring functions that reflect how strongly a small molecule (or a protein) will bind to (another) protein.

In summary, the impact of AlphaFold 2 on computational biology certainly goes beyond just predicting protein structures and has outlined novel ways to do research.

3. What will probably happen (that is, if you are naive enough to believe me)
Computational biology is an exciting place to be these days, but the million dollar question is, where will the field be going next? This is something I have been thinking deeply about since CASP14, and I think I am ready to make a few bold bets. There is, of course, no guarantee that any of this will happen – having chosen protein structure prediction as my PhD topic, my track record of predicting the future is not precisely without fault. And yet, I think we are starting to see some powerful trends that are bound to shape the field.

As a disclaimer, I will focus on topics that I expect to be widely pursued, but not necessarily those that are scientifically most interesting. For example, one of the problems with AlphaFold 2 is that it is, in essence, a smart way to extract information from multiple sequence alignments, with some extra steps. And yet this is completely contrary to the way protein folding occurs *in vivo*: a protein, as a biophysical entity, knows nothing about its evolutionary history. Some people have been working on sidestepping the multiple sequence alignment (typically by using a large protein language model), but while I believe furthering our understanding of the biophysics of protein folding would pay enormous dividends in multiple fronts, I am also deeply aware that the interests of researchers (and funding bodies) will likely focus on more immediate applications.

Now that this is clear, my first bet is that the next big problem is predicting protein-ligand interactions. Whereas AlphaFold is quite successful at predicting protein structure, and to a limited extent, the interaction between proteins, the problem of proteins binding ligands introduces a new set of complications. Whereas there are only 20 possible amino acids (and, perhaps, about an order of magnitude more after chemical modifications that occur in the cell), the space of chemicals resembling typical small molecules is typically estimated to contain about 10^{60} molecules with enormous diversity. This chemical space is also very variable – think, for example, of the many known cases of pairs of compounds, differing in as little as a single functional group, that display vast differences in binding affinity, a phenomenon known as 'activity cliff'. In contrast, AlphaFold 2 is not even very good at predicting the effect of a single amino acid mutation.

There are important industrial players forming in this quest. Unless you have lived blind to the world for the past years, you must be aware that Demis Hassabis, the CEO of DeepMind, has embarked on a new venture, Isomorphic Labs. The project – "reimagining the entire drug discovery process from first principles with an AI-first approach" – is fazed with the standard secretism, but few doubt that predicting protein-ligand figures amongst their objectives. Another competitor, Charm Therapeutics, featuring David Baker, recently raised a $50M series A round to take Isomorphic Labs in what may be one of the most interesting technological competitions of this decade.

My second bet is that the battle over protein-ligand interactions will not be fought over computational methods, but the availability of data. The AlphaFold breakthrough was possible because over 170k protein structures had been painstakingly solved by structural biologists, at an estimated aggregated cost of ~$10 billion. Unfortunately, data on protein-ligand interactions is much more scarce – and, for the increased complexity discussed above, probably several more orders of magnitude would be required to train a model that rivals AlphaFold's complexity. Furthermore, much of this data is held in the internal databases of pharmaceutical companies which, for good reasons, will protect them fiercely to guarantee their intellectual property.

There is a deeper problem: drug-discovery datasets are plagued by biases. Think about it: most of our best data comes from the hands of incredibly talented medicinal chemists who have to work hard for every datapoint, and thus invest significant thought to design molecules with a serious chance of binding. These chemists, after all, know a thing or two about chemistry. But deep neural networks do not. In order to learn these chemical principles, they need to have a good portion of 'dumb' negative examples to teach them all the chemistry that these medicinal chemists already know. Thus, producing large amounts of 'machine-learning-grade data', free of biases and containing positive as well as negative examples, will (should) become a priority in the drug development industry.

My third bet is that physics will become the endgame of structural bioinformatics. Although we are living in an era of increasingly high-throughput measurements, the level of detail to which we can measure the structural choreography of biomolecules is still far behind what we would like. Fortunately, or not, this could be palliated if our models were given some meaningful intuition about the physics of our system – for example, that no, two carbon atoms cannot be less than 2Å apart from each other, and that yes, it is very favourable to have that amide forming a hydrogen bond with the neighbouring residue. In a way, physics is a natural informative prior that may supercharge our models.

Some central players have been considering how to use machine-learning methods to produce better physical models – for example, by training point cloud models on high-accuracy quantum calculations to produce rapid and accurate force fields. While this is a very interesting problem, we need better ways to encode the physics of biomolecules into our models. Unfortunately, the problem here is almost tautological: we want our models to learn from physics (so as to benefit from our prior knowledge), but at the same time we don't want them to learn from 'our' physics (the vastly incorrect, highly biased models that we successfully use to understand the behaviour of matter). Finding intelligent ways to introduce physics into deep learning may well be one of the most interesting questions for computational biophysics in the next few years.

My final bet is that, as our methods gain in prowess, we will move away from the concept of protein structure towards the more interesting questions of function and systems-wide biology. As I tell my biochemistry students, "no-one cares about protein structure, it's just more convenient than the abstract concept of protein function.” Easy structure prediction has been a powerful step in this direction: by predicting the structure of a protein and comparing against known structures, it is possible to elucidate function in some cases. In a way, searching structure databases using predictions may well become a substitute to sequence search methods like the BLAST family.

What is certain is that we are likely to witness an era of flourishing computational tools that enable deeper insight into interesting problems, as well as exciting applications throughout the life sciences.

4. What needs to happen (the rant section)

Many good things have come from AlphaFold 2 beyond scientific progress and new ideas. Perhaps the most important one is openness. In computational biology, a field, where advances were sometimes kept private until the following CASP exercise, it now seems like the floodgates have truly been opened. Ideas have flown back and forth, often through preprints, but in many cases also through Twitter and other social media. In fact, I recently reviewed a couple of papers that included a tweet or two among their references (dear authors, I pity you dealing with the journal’s proof software).

There are, however, things that we need to fix if we want to build a community that thrives on discovery. In particular, I would like to discuss two that keep coming up: the availability of data, and the access to computational resources.

Reliable data has always been central to scientific discovery, but the progress in artificial intelligence has made it more crucial than it has ever been. I have already discussed why I think many central problems in our field will likely be solved by mass producing machine-learning-grade protein-ligand binding data, rather than by grinding new techniques and better models. But there are many, many other problems, some of enormous practical importance, which we could solve with simple machine-learning algorithms if only we had good databases.

Think of any problem that takes some time to work out in a laboratory, say, the right conditions to crystallise a protein. This is a problem that depends on a number of variables, including temperature, pH, degree of saturation, and up to a couple dozen other parameters. This can be casted as an optimisation algorithm in a relatively small optimisation space, and it should not be difficult to convince yourself that you can build a decent model in a couple of weeks’ time. Perhaps not a ‘will crystallise any protein’ kind of model, but more robust than ‘the wise postdoc of the lab says so’. So, why is there no such model? Because there is a lack of high-quality data, and because we mostly have data about the times when it worked, but very little data about the failures.

In all frankness, I think the problem runs much deeper. Since I am already ranting, consider a pet peeve of mine: the state of datasets in protein folding. In a paper I published last year, we tried to understand if the advances in protein structure prediction could somehow translate into improved understanding of the mechanisms of folding. A simple litmus test is to check whether the predicted mechanism agrees with the experimental kinetics of protein. If you look into the literature, however, you will find that there are, perhaps a couple of hundred proteins for which such data is available. These are simple experiments. Automatable protocols that could be run in a couple of weeks for a few thousand pounds in consumables – why is there no such dataset for a problem so central to biophysics? Could we not use some of the resources that we put into computational studies to generate this dataset?

The next problem is access to computational resources. Many of the *en vogue* computational techniques are reliant on massive computational scales. Much has been said about DeepMind’s success being linked to their endless computer budget (spoiler: they are very, very clever, but would not have solved this problem with pen and paper only), and clearly even reproducing their results, without any of the intermediate models that did not work, would be impossible for most academic labs. Just as this piece is sent to print, a team led by Mohammed AlQuraishi at Columbia has published the first public reproduction of AlphaFold 2 – taking over 100,000 hours on a Nvidia A100 40GB GPU (roughly ~$2M if you were using AWS, probably cheaper if you manage the GPUs locally).

The capabilities of current supercomputers are quite limited, in many ways. Most of them are based on CPUs, rather than GPUs or other types of processing units that are specialised in machine learning. Successful applications typically provide a few hundreds or thousands of hours, which is great when you have a defined large simulation that you want to run – not so much if you want to try a number of tricks and hope some of them work. The only alternative is managing a local server, which requires a large initial capital expenditure (and which, depending on where you are located, can be difficult to fund via traditional methods). If we want computational science to thrive, we need to think of better ways to provide access to computational resources.

5. Final remarks

A year after the publication of the AlphaFold 2 paper, computational biology looks like a new field – a shinier, more open discipline that is buzzing with ideas. We have rapidly moved on to solve new problems with new tools: leveraging rapid and accurate predictions, extending the original code to produce different results, and exporting the deep-learning advances to new problems. These ideas have led to steady progress in areas ranging from protein-ligand docking to protein dynamics, and the best is yet to come.

The solution of the structure prediction problem has opened new avenues of research. Some of the big questions on the horizon include predicting how a ligand interacts with a protein; meaningfully introducing physics into machine learning models; and, eventually, moving the focus from structure alone into function. These prospects will require new ideas, but above all, a focus in producing high-quality ('machine-learning-grade') data that is free of biases.

There are challenges on the horizon, and many things that we need to work out about the future of research in our field, but one thing is clear – it is a great time to be working in computational biology.

I would like to thank the members of the Oxford Protein Informatics Group for insightful discussions about many of the ideas discussed in this essay, and in particular to Professor Charlotte M. Deane for, among others, suggesting the term 'machine-learning-grade data'.

References

Outeiral, C. CASP14: What Google DeepMind's AlphaFold 2 really achieved, and what it means for protein folding, biology and bioinformatics. *RSC-CICAG Newsletter, 2021*, 5-12; Oxford Protein Informatics Blog post.

Jumper, J., *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature*, **2021**, *596*(7873), 583-589.

Zhu, X., *et al.* Structure of the cytoplasmic ring of the Xenopus laevis nuclear pore complex. *Science*, **2022**, *376*(6598): eabl8280.

Fontana, P., *et al.* Structure of cytoplasmic ring of nuclear pore complex by integrative cryo-EM and AlphaFold. *Science*, **2022**, *376*(6598): eabm9326.

Mosalaganti, S., *et al.* AI-based structure prediction empowers integrative structural analysis of human nuclear pores. *Science*, **2022**, *376*(6598): eabm9506.

Hsu, C., *et al.* Learning inverse folding from millions of predicted structures, **2022**, bioRxiv.

Ganea, O.-E., *et al.* Independent SE(3)-equivariant models for end-to-end rigid protein docking. **2021**. arXiv preprint arXiv:2111.07786.

- -

Diana Leitch – Reflections on her Life in Chemistry, Chemical Information and Librarianship

Contribution from Dr Diana M Leitch, MBE, BSc, PhD, FRSC, CICAG Committee Member and Chair of Trustees at the Catalyst Science Discovery Centre, email: diana.leitch@googlemail.com

I was born in Liverpool just after the Second World War ended and brought up in Runcorn – the heart of the UK chemical industry. I come from a long line of those involved with the chemical industry in north-west England. My mother's ancestors brought the salt from Winsford and Northwich on Weaver flats to Runcorn from the 18th century onwards. She worked for ICI at Castner Kellner Works in the engineering drawing department. My father's grandfather came to St Helens from Oxfordshire to work in the industry in 1870 and my father, Clarence Bridges, started work at Winnington Works of ICI at Northwich before being transferred to Widnes Research Lab where he was part of the team working on the separation of uranium isotopes in WWII. Dad had been taught at Widnes Technical College in the early 1930s by Harold Wilson's father, as part of day-release studies, and was awarded a chemistry degree from UMIST. He later went on to head several ICI analytical laboratories in Runcorn at Rocksavage Works and to market Arctons. My Aunt, Alice Marsh, worked in the accountants' department of Gossages Soap Works in Widnes in the building that is now the Catalyst Science Discovery Centre and Museum where I am the Chair of Trustees.

Diana (second from the right) in the school chemistry lab in 1964.

My early schooling began at a state primary school in Runcorn but I was quickly picked out from a class of 46 as a bright child who wasn't progressing, as I was used by the teacher to dress other children as I could already read and write! I passed the exam to go to The Queen's School, Chester, in 1954, which had a huge formative influence upon me. It was an all-girls' school and the girls were expected to do everything boys did. There was also a strong emphasis on science, charitable work and community service. Every day from the age of 8 to 18 I travelled 16 miles each way from Runcorn to Chester by train to school, fortunately just before Dr Beeching cut the line.

At the age of 16, my love of the world of chemical information started. I spent my summer vacation helping out in the Information Department at ICI's R&D Division at the Runcorn Heath complex and whilst there came under the influence of the work of John Wales, the Information Manager, and Angela Haygarth-Jackson, at ICI Pharmaceuticals Division, who became a lifelong inspiration to me of what women could achieve in a very male-dominated world. Chemical information was almost the only way that women could become senior managers in ICI.

When I was 18 in 1965, the magic of academic chemistry beckoned and I went to study chemistry at Edinburgh University. There were only 10 girls in my year but we did have two female academic staff, both crystallographers, which was very unusual. In my first year I also studied maths, physics and meteorology (in which I won the Class Medal) and, in my second year, geology (in which I jointly won the Class Medal). But the highlight of my time in Edinburgh was undertaking my fourth-year project with Evelyn Algernon Valentine Ebsworth, Crum Brown Professor of Inorganic Chemistry and recently arrived from Cambridge, as my supervisor. A transformative and interesting experience if a slightly eccentric one. I was awarded a BSc First Class Honours in 1969.

Like the majority of us, post an undergraduate university course was a time of uncertainty for me. What to do? A PhD? A chemical information masters course? Employment in industry? A geochemistry course at Bristol with Professor Geoffrey Eglington? After much thought, I chose the PhD route and returned to work with the Ebsworth group which was a truly rewarding experience. I vividly recall the frustration during my PhD at not being able to find information easily. Of course there were the print versions of *Chemical Abstracts* and *Beilstein* but the Chemistry Library had a fierce librarian and books and journals were kept behind locked grills! You had to ask her to get them out for you and put them back. My research was on silyl and germyl complexes of iridium (III). I was told to find the original paper by Vaska who had made Vaska's Compound. What, where, how did I start? It took me many weeks but taught me a lot and mostly that there must be a different way forward in future to find information.

Learning technical German or Russian was a necessity and mandatory to get a PhD and I passed the German exam with a dictionary's assistance.

In 1971, having completed my PhD research, I started looking for a job. However, at the time, jobs were thin on the ground as the chemical industry was in a downturn. I was offered a post with Shell International in London in technical marketing but it entailed flying to Marseille twice a week and I was not a good plane traveller so had to turn it down even though they were going to pay me £1,800 per annum. Finally, I found employment in Didsbury, Manchester at the Cotton Silk and Man-made Fibres Research Association (the Shirley Institute), working on *World Textile Abstracts* as an abstractor, indexer and translator for £1,200 per annum. The publication had just been computerised and one of my daily tasks was checking the ticker-tape. However, the job was very boring and I only lasted a year there. My second job was at multinational CPC International (now Cerestar) in Trafford Park, Manchester (on the Manchester Ship Canal), where I was appointed as Assistant Technical Information Officer. This was a dream job, doing patent searching, preparing scientific information reports for management and being on daily watch for cornflour explosions (Brown and Polson factory) and 'sniff testing' to stop complaints from local residents! However, after just 10 months CPC (US) decided to close the Information Centre and create a joint European Centre in Brussels and I was made redundant. This was not a good time to be made redundant as I had just got married and my husband and I had recently taken out our first mortgage. Luckily, I saw an advert for a job at The University of Manchester; a chemist was required to work in the University Library. I was interviewed by the University Librarian, Dr Fred Ratcliffe (and what seemed like 20 men in grey suits) and was offered the job. You had to have a First Class Honours degree or a PhD to pass muster. I accepted the post despite the fact that I had no idea what the job entailed – the job description did not make that clear. Recruitment was different in those days!

On 6 July 1973 I turned up for work in the Christie Science Library. I was given a manual typewriter and told to catalogue and classify books. Not quite what I expected as I had never classified a book in my life or done a library training course but it was another formative experience! I took over from another chemist, Dr Alan Neville (who sadly died in 2020), who had been moved to the Medical Library to work on the first experiments on the use of digitised *Index Medicus* (the Medlars system). I found myself in good company. Also working in the John Rylands University Library (JRUL) of Manchester at the time were Bill Simpson, Chris Hunt, Reg Carr, John Hall, John Henshall, John Lancaster, John Tuck and Ian Lovecy, who all went on to become University Librarians and senior ambassadors for the library profession in the UK and abroad. 'Information' was not a word that was allowed to be used in the overall Library at that time and neither was training in information usage for students and staff. I spent the next four years as a cataloguer and then in 1977 I found that I was pregnant. I asked to take maternity leave but was told this would not be possible. I needed a job as my husband, also a research chemist from Edinburgh, had been made redundant from ICI Dyestuffs Division at Blackley in Manchester. Our younger readers will probably be astonished to know that maternity-leave legislation was not in place in 1977. In fact it took a petition to the Vice-Chancellor to allow me to return to work part-time in late 1978 – I was the first woman to return to work in the Library having had a baby. No-one had tackled any of my work as maternity cover was not practised at that time. Following the birth of my second child in 1981 all the staff from the Christie Science Library were moved to a new extension of the Main Library building. This coincided with the appointment of a new University Librarian, Dr Mike Pegg, and I was finally released from the cataloguing department. Alleluia, but it had paid the bills. My new job was working in a team of three people, including Alan Neville, in the newly-created Science and Medical Information Office – finally I was working in an outward-facing team, as I had always wanted to do, and had contact with users. A whole programme of training of users was set up making the best use of all resources at our fingertips. These were still mainly print based.

These were the early days of online information and the team used dumb terminals initially to search Medline and other online databases through ESA, Questel, DataStar and STN International. But technology moved on quickly. By the mid-1980s CD-ROMs were beginning to proliferate; indeed Alan Neville in my team was the first person in the UK to network CD-ROMs in a LAN. Working with colleagues from UMIST he then worked out how to make a wide-area network (WAN) of CDROMs.

The late 1980s and early 1990s proved highly significant as far as my information career was concerned. I began my long-standing involvement with the Joint Information Systems Committee (JISC) in June 1995 by becoming a member of the JISC ISSC (Information Services Sub Committee). Also, work was beginning in Germany on computerising *Beilstein* and *Gmelin*, and there was discussion of a new service called CrossFire. I still had involvement with the Beilstein Institut as Manchester was the only UK library still buying the printed copy. I got involved in the early trials of CrossFire at the Daresbury Laboratory which proved that it was a service that chemists really wanted. Nobel Prize-winner, Sir John Cornforth, still researching in his late 70s, wrote to express his delight at this new service. From then began a long association of working with Helen Cooke of CICAG on various projects. Furthermore, my involvement with the Chemical Abstracts Service (CAS) led to my being involved in brokering a deal with CHEST for the UK scholarly community for SciFinder Scholar.

In 1991 Chris Hunt was appointed University Librarian at Manchester and one of his early actions was to appoint me and my colleague, John Tuck (recently retired from Royal Holloway), as his deputies. My focus as Head of Information Services was the e-environment and I threw myself wholeheartedly into the development of e-resource services for all our users. My huge enthusiasm and knowledge of the e-environment has been valued by the users, both in the library and the publishing world, and I served on the library advisory boards of Springer, Wiley-Blackwell, Elsevier, the Royal Society of Chemistry, the American Chemical Society, CAS and OUP as a UK academic representative. Managing the transition from print to electronic proved a tremendous challenge at Manchester (as in many libraries) and one of my major tasks from 2002 was sorting out the space problems in the library. The purchase, which I managed, of digital back-files of journals meant that many runs of printed journals were hardly, if ever, consulted, and it was agreed by the University that these could be moved to storage in a cotton mill in Stockport. I oversaw the whole operation. Over 50 miles of materials were moved around the whole JRUL system in the two-and-a-half years to provide space to maintain services to users. I did get nasty emails from some academics in different disciplines about removing their right to have access to printed copies of journals but in general the benefits of having remote access to information was understood and finally welcomed. The Head of the Chemistry Department told me he had disposed of all his offprints in his filing cabinet as it was easier to go online than open the drawer to retrieve an article. A great accolade.

My contribution to the information profession at large and to The University of Manchester specifically has been prolific. It has been recognised by library colleagues and publishers as well as professional organisations such as UKSG (UK Serials Group), JISC and the RSC. In the late 1990s I became the first woman at The University of Manchester to become a University Presenter – introducing Honorary Graduands at their Graduation Ceremony. But perhaps the greatest highlight of my career came in 2005 when Professor Sir Harry Kroto, with whom I had worked on the Dalton Bicentenary celebrations, invited me to seek election as a Fellow of the Royal Society of Chemistry – the first ever academic librarian to become a Fellow of the Society, following, once again, in the footsteps of Angela Haygarth-Jackson.

The amalgamation of the Victoria University of Manchester and UMIST was challenging but because of earlier cooperative schemes between the library and information staff of those two institutions it went well in our areas and I was deeply involved in that process.

Diana's retirement from the John Rylands Library at the University of Manchester, 2008.

In July 2008 I retired from my post as Deputy University Librarian and Associate Director of the John Rylands Library at the University of Manchester and started on another episode of my involvement with the chemistry world and chemical information. Initially I did some consultancy work and went to various parts of Europe to promote electronic databases for a variety of publishing companies. However, I soon found that the chemical information world and its computer applications were moving on apace and my knowledge was becoming rusty. I wrote scientific articles about the local history of the area where I live in Manchester for a local newspaper and started to give scientific and historical talks to local groups in north-west England – U3As (University of the Third Age), local history societies, men's Probus groups, WIs (Women's Institute) and these continued being very popular until COVID abruptly stopped them in February 2020. In 2019 I did 95 talks. I have been able to give a few by Zoom in the subsequent period to now.

In 2012 I went back to my original roots and became a Trustee at the Catalyst Science Discovery Centre and Museum in Widnes. It is the only museum of the chemical industry in the UK and has vast numbers of industrial archives from ICI, Brunner Mond and other local companies. In November 2018 I was elected the Chair of Trustees. The period since has been challenging in the extreme and not what was expected, but we have survived financially, reopened after 18 months closure and the public have come back; school visits started slowly as did sleepovers but are building up. The Centre and Museum has a great future to continue the education of prospective chemists and maintain the great chemical industry tradition of north-west England.

Diana receiving her MBE at Buckingham Palace in 2014.

I am very proud to have been part of the development of the chemical information world and was delighted to be awarded an MBE "for services to chemistry" in 2014.

I have two children – a daughter, Fiona, who went to Glasgow University to study medicine and is now a Consultant Surgeon at Glasgow Royal Infirmary with two daughters, and a son, Andrew, who also went to Glasgow to study Systems Engineering and is now the Chief Technical Officer of a bank in Manchester with one daughter. My husband, David, after being made redundant by ICI, helped to set up a drug-trials organisation called Medeval at The University of Manchester. He then went back to the family tradition of pharmacy, became a mature student, took a pharmacy degree and became the Regional Drug Information Officer and Medicines Information Manager for the NW Region and Central Manchester Healthcare Trust until retirement.

I have had a wonderful career and seen many changes particularly for women in my time. With one or two exceptions I have thoroughly enjoyed my career in the information world, particularly the people I have met and worked with, and I hope I can continue to keep in touch in the future. My involvement with RSC CICAG, latterly as Treasurer, has enabled me to do that.

- -

Meeting Report: AI in Drug Discovery

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

Introduction

Drug discovery, the process of discovering new medicines or therapies, is witnessing a revolution. Machine-learning (ML) techniques are replacing traditional methods in hypothesis generation, target identification and molecular design in the hope of reducing costs and development times. Billions of dollars in venture capital are being poured into new start-ups, large pharmaceutical companies are building in-house AI capabilities and conventional tech companies are now investing heavily in biology and drug discovery. However, like any new and exciting technology, it remains an open question whether AI will live up to its promises as it passes through the hype cycle and one of the main challenges remains its integration into real-world experimental techniques and commercial discovery pipelines.

At this turning point for drug discovery, the 1st Cambridge University AI in Drug Discovery Conference (CamAIDD) took place on 26 February 2022 to help students, academia and industry gain a sense of community and direction in this rapidly changing field. We aimed to give a broad yet comprehensive overview of both the frontier of scientific research transforming drug discovery as well as glean insights from some of the most innovative entrepreneurs in the field.

The conference was co-organised by the Cambridge University Artificial Intelligence Society (CuAI) and Cambridge University Artificial Intelligence in Medicine Society (CUAIM). The conference was designed to be highly interdisciplinary in nature, with a 50/50 split between academics and founders giving talks, and panel discussions on business and AI policy.

Petar Veličković

First up to speak was Petar Veličković, who is a Staff Research Scientist at DeepMind and an affiliated lecturer at the University of Cambridge. Petar is not only a leading figure in the field of Graph Neural Networks (GNNs), which have exploded into the chemistry and biology scene in recent years, but he is also one of the best science communicators in the space. We were incredibly lucky to have him start off the day. It was fitting having someone from DeepMind explain GNNs given the company's record of producing one of the most impactful pieces of GNN (and possibly ML) research to date, AlphaFold2. DeepMind have also recently spun out Isomorphic Labs, a new company focused on AI-first drug discovery.

When I was entering the field of GNNs, I found Petar's categorisation of different types of GNNs into three different 'flavours' (convolutional, attentional and message-passing) particularly useful. While many newer GNN architectures within a certain flavour are usually more advanced than their progenitor architecture, and it is common for models to compose principles from multiple flavours, they nevertheless provide an excellent theoretical basis for someone entering the space.

Andreas Bender

Next was Andreas Bender, Professor of Molecular Informatics at the University of Cambridge, co-founder of HealX and PharmEnable and previously holder of numerous industry positions including directorships at Novartis and AstraZeneca. Outside of his work in cheminformatics, Andreas is probably most (in)famous for his soberingly realistic views on where we are with applying AI to drug discovery which he formulated brilliantly in his talk titled *Aspects of Translation, Platform Validation, and Where We are on the Hype Cycle*.

Andreas does not let his realistic views cloud his ambition, however. He has recently co-founded and now serves as Chief Technology Officer of Terra Lumina which aims to use AI to harness the power of small molecules found in nature, doing something similar to what the London company Basecamp Research does for proteins but with small molecules.

Michael Bronstein

Michael Bronstein is the newly appointed DeepMind Professor for AI at the University of Oxford, Head of Graph ML at Twitter and a world-famous researcher in the field of GNNs. He is widely regarded one of the founders of Geometric Deep Learning (a term which he coined) and has published a highly regarded textbook of the same name, which provides an excellent bridge between abstract theory and thoughtful application. A serial entrepreneur, one of Michael's companies, Fabula AI, was the first known commercial success of GNNs when it was acquired by Twitter in 2019.

Michael's presentation focused on GNNs through the lens of Differential Geometry and Algebraic Topology. Michael explains that many common discrete geometric domains (e.g. grids and meshes) have analogous continuous objects (e.g. planes and manifolds respectively). There is, however, no obvious continuous analogy for graphs, which Michael characteristically describes as "concerning". However, he went on to describe how this can be achieved by a process called Graph Neural Diffusion expressed in the form of a Partial Differential Equation (PDE).

Rabia Khan, Neel Madukar and Daniel Jamieson

We were very fortunate to have the talks end with a succession of excellent presentations from three Founders and CEOs. These were Rabia Khan of Ladder Therapeutics, Neel Madhukar of OneThree Biotech and Daniel Jamieson of Biorelate.

Ladder Therapeutics is a Y-Combinator-backed company using machine learning to design drugs for a relatively new class of target, RNAs. Rabia's talk focused on how she is currently building an *in-silico* and *in-vitro* platform to design small-molecule modulators to target RNAs in the often overlooked 'dark transcriptome'.

OneThree Biotech is a start-up spun out of Cornell by Neel Madhukar and three other co-founders in 2017. Neel explained how his company is using AI in their pipeline to de-risk the earliest stage of the drug-discovery pipe that often hampers clinical success, the discovery of novel disease biology and target identification.

Our last talk was by Daniel of Biorelate, a company which he spun out of the University of Manchester in 2013. Daniel's talk titled *Using Cause-and-Effect to Empower Drug Discovery* described Galactic AITM, a supercomputing platform that automatically curates biomedical research literature to distinguish simple correlation from casual therapeutic pathways. By connecting obfuscated evidence, Biorelate's goal is to accelerate development of important new therapies.

Start-up panel
I knew this section of the conference would be the one of the best the second we all joined the private Zoom lobby before going live; I hardly needed to encourage the Founders to bounce ideas off of each other and the mood, even though virtual, was electric. In attendance were Rabia, Neel and Dan from before as well as another panelist, Laksh Aithani. Laksh is the CEO of CHARM Therapeutics, a start-up he co-founded along with the world-famous protein designer David Baker.

During the panel, we touched on all the usual pillars of entrepreneurship, such as finding the right co-founder, raising funds and recruiting the right people into your team early on. Of particular interest was the discussion about how there has been a recent shift of AI in drug-discovery start-ups away from simply being the AI partner of a traditional pharma company and instead developing their own 'end-to-end' drug-discovery capabilities. Indeed, many of our AI co-founders were actively building out their own wet lab capabilities. This trend is exemplified by companies like Exscientia and BenevolentAI going public with very successful IPOs based on their AI-driven discovery platform and current in-house drug-development portfolio.

Since the event, the Founders have been leading their companies from success to success. A recent example is with CHARM Therapeutics, which officially came out of stealth mode and announced a $50M Series A round co-led by F-Prime Capital and OrbiMed at the time of writing. Finally, all our panellists were keen to point out that they are all actively hiring!

Science panel
The day ended with a final panel discussion on the *Scientific Frontiers of AI and Drug Discovery*. The discussion was aimed at being highly interdisciplinary and our panellists reflected this. In attendance was Michael Bronstein and two excellent complementary guests. The first was Nathan Benaich who is the Founder and General Partner of Air Street Capital, a London-based VC firm specialising in making investments into AI and Life Science companies (including LabGenius, Exscientia and Valence Discovery) and is also the Founder of London.AI, Spinout.fyi and the RAAIS Foundation. He is not just a businessman however, Nathan holds a PhD in computational biology from the University of Cambridge, is co-author in the acclaimed *State of AI Report* and moves fluently between discussions on biology, AI, business and geopolitics.

Our final panellist for the evening was Sir Tom Blundell. Sir Tom is a highly interdisciplinary biochemist, structural biologist, computational biologist and science administrator. He is currently Professor Emeritus of Biochemistry at Cambridge and holds an impressive list of achievements. These included: being part of the team that solved the first structure of insulin, founding Astex pharmaceuticals (which successfully commercialised fragment-based drug design) and serving as a Scientific Advisor to the British Prime Minister in the 80s.

The discussion started with each of the guests describing what they felt was the main challenge for making AI have a real-world impact on drug discovery and biology more generally. Answers given ranged from data availability and quality, current model architecture and slow organisational embracement of AI technology. Tom was also keen to point out that there is currently a parallel revolution going on within drug discovery alongside AI and that is Cryo Electron Microscopy (CryoEM). He went on to describe his (at the time) radical adoption of CryoEM into Astex and that better integration of ML and CyroEM is an increasingly important area of research.

Conclusion

The majority of talks are available online via the CuAI YouTube channel, however, the panel discussions were not recorded to encourage openness in the discussions. If you were not able to make it, however, there is no need to worry! We are already planning a second conference of the same name next year which will likely be in person and have an even greater emphasis on interdisciplinary discussions between science, business and governmental experts. More information will be on the CamAIDD website.

I would like to end this piece with a number of overdue thank yous. First and foremost, of course, would be our fantastic set of speakers for their time and inspiring thoughts. There were also many amazing volunteers working hard behind the scenes from CuAI and CUAIM Societies who made the event the success that it was. The event would also have not been possible without the fantastic support of my two PhD supervisors Pietro Lio and Sir Tom Blundell, who always tirelessly throw their hat in the ring for me behind the scenes. Finally, I would like to thank everyone who attended for their overwhelming turnout and support. I hope you all enjoyed the event as much as we did putting it together.

This article was written with generous assistance from my colleague Simon Mathis.

- -

AI4SD News

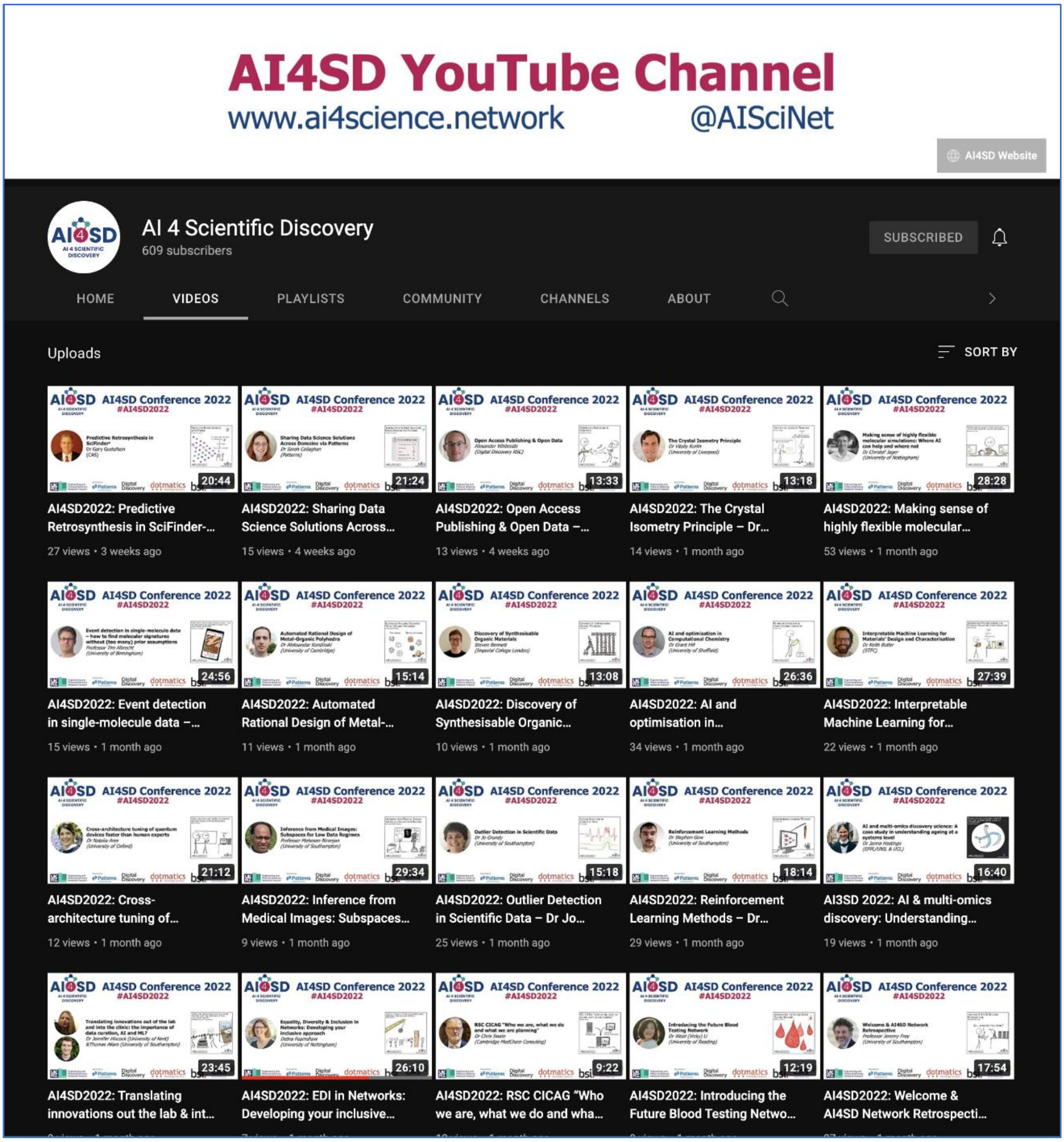
Contribution from Dr Samantha Kanza, AI4SD Network+ Coordinator, University of Southampton, email: s.kanza@soton.ac.uk

AI4SD Conference

On 1-3 March 2022 at Chilworth Manor in Southampton, it was fantastic to finally run our first in-person event for two years! The event was a mixture of keynote talks, discussion sessions and networking opportunites, as well as some fantastic musical entertainment from the University of Southampton Music Department.

Wendy Warr is in the process of preparing a detailed report on this event, so watch this space!

Almost all of our videos are now online on our YouTube Channel. These can all be found in the AI4SD Conference 2022 Playlist. If you haven't already done so, you can subscribe to our YouTube Channel here: AI4SD YouTube Channel.



AI 4 Scientific Discovery
609 subscribers

Uploads

AI4SD2022: Predictive Retrosynthesis in SciFinder-... 20:44
27 views • 3 weeks ago

AI4SD2022: Sharing Data Science Solutions Across... 21:24
15 views • 4 weeks ago

AI4SD2022: Open Access Publishing & Open Data –... 13:33
13 views • 4 weeks ago

AI4SD2022: The Crystal Isometry Principle – Dr... 13:18
14 views • 1 month ago

AI4SD2022: Making sense of highly flexible molecular... 28:28
53 views • 1 month ago

AI4SD2022: Event detection in single-molecule data –... 24:56
15 views • 1 month ago

AI4SD2022: Automated Rational Design of Metal-... 15:14
11 views • 1 month ago

AI4SD2022: Discovery of Synthesisable Organic... 13:08
10 views • 1 month ago

AI4SD2022: AI and optimisation in... 26:36
34 views • 1 month ago

AI4SD2022: Interpretable Machine Learning for... 27:39
22 views • 1 month ago

AI4SD2022: Cross-architecture tuning of... 21:12
12 views • 1 month ago

AI4SD2022: Inference from Medical Images: Subspaces... 29:34
9 views • 1 month ago

AI4SD2022: Outlier Detection in Scientific Data – Dr Jo... 15:18
25 views • 1 month ago

AI4SD2022: Reinforcement Learning Methods – Dr... 18:14
29 views • 1 month ago

AI3SD 2022: AI & multi-omics discovery: Understanding... 16:40
19 views • 1 month ago

AI4SD2022: Translating innovations out the lab & int... 23:45

AI4SD2022: EDI in Networks: Developing your inclusive... 26:10

AI4SD2022: RSC CICAG "Who we are, what we do and wha... 9:22

AI4SD2022: Introducing the Future Blood Testing Netwo... 12:19

AI4SD2022: Welcome & AI4SD Network Retrospecti... 17:54

Upcoming events

We are running two hybrid events in July 2022.

***Failed it to Nailed it*: Nailing your data visualisation: a hands-on training workshop – 13-14 July 2022**

- **Where:** This is a hybrid meeting, the physical sessions will take place at the Chilworth Manor Hotel in Southampton. Virtual attendees can join via Zoom.
- **What:** This event forms part of the *Failed it to Nailed it* series run by the Artificial Intelligence for Scientific Discovery Network+ (AI3SD), the *Cell Press Patterns Journal* and the Physical Sciences Data-Science Service (PSDS). This event is the second in our 2022 edition of the *Failed it to Nailed it* series and is designed to teach how to nail data visualisation. There will be talks on the importance of data visualisation and how to tell your story through your data. There will some technical training sessions

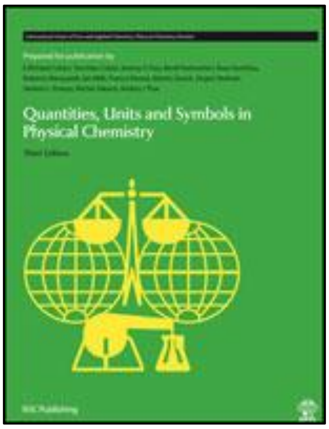
that delve into some of the tools and techniques that can be used for this and a number of challenges to choose from. Feel free to come in a team or come solo and we will match you up. There will be helpers on hand if anyone needs advice.
- **Practical requirements:** This is a hands-on workshop so after the talks there will be some team-based practical activity. You will be expected to bring your own device. If this is not possible please let us know when you register.

AI4SD & Directed Assembly Early Career Researchers Event – 25-26 July 2022

- **Where:** This is a hybrid meeting, the physical sessions will take place at the Chilworth Manor Hotel in Southampton. Virtual attendees can attend via Zoom.
- **What:** This is a joint event between the AI4SD and Directed Assembly networks. It is specifically designed to inform, upskill and facilitate networking between early career researchers. The event will contain talks on scientific publishing, ED&I (equality, diversity & inclusion), grant and fellowship applications, networking and much more. There will also be dedicated time for networking.
- **About AI4SD:** The AI4SD Network+ is funded by EPSRC and hosted by the University of Southampton. It aims to bring together researchers looking to show how cutting-edge artificial and augmented intelligence technologies can be used to push the boundaries of scientific discovery.
- **About Directed Assembly:** We aim to gain unprecedented control of the assembly of molecules and structures that are the building blocks of many functional materials, consumer and industrial products. The methods our members develop will enable the production of new materials at length scales from the nano to the everyday. These materials will have major impact in areas of societal importance such as personalised healthcare and food production, transport systems and fuel production, housing construction and consumer electronics.

Equality, diversity & inclusion survey

In Summer 2021 the Network of Networks sent out an equality, diversity and inclusion survey to look at how diverse and inclusive our Network are and to obtain thoughts from our members about how to move forward with running blended events and returning to physical-based events. We are now running the 2022 version of this survey, which is for all members of all the Networks and is entirely voluntary and anonymous. It takes approximately 5-10 minutes to complete. This survey, developed by the Network of Networks, will help us understand more about the inclusivity of our Network, so the activities offered are inclusive and representative of their communities. We will use the results to benchmark and continually track our progress in terms of improving the inclusivity of our activities. The 2021 survey results influenced the way that activities were planned and delivered to improve what we do for all our members. We have received full ethics approval from the University of Southampton Ethics and Research Governance Team to run this survey under ERGO No. 66119.



The IUPAC Green Book

Contribution from Professor Jeremy Frey, IUPAC 5th Edition Project, AI4SD Network, and School of Chemistry, University of Southampton, email: J.G.Frey@soton.ac.uk

The IUPAC Green Book, *Quantities, Units, and Symbols in Physical Chemistry*, provides a readable compilation of widely used terms and symbols from many sources together with brief understandable definitions.

We are currently working on a potentially major revision of the Green Book for the 5th edition IUPAC Project and we would like to consult as widely as possible about the content of the Green Book. The 4th edition which has been updated with the new SI will be published soon along with an abridged edition but we are considering more major changes for the 5th edition.

Please fill in our survey to help us gather community views on what should be included in the 5th edition. The survey will remain open until the end of the summer but early replies will be very useful in planning the 5th edition project.

We have received full ethics approval from the University of Southampton Ethics and Research Governance Team to run this survey under ERGO No. 72139. Please read the following participant information to make sure that you understand and agree to the terms of this study.

To contribute to this discussion please go to the Survey.

RSC Historical Group – Women in Chemistry Symposium

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

200 years ago there were few opportunities for women to study or practise any form of science. By the late 19th century, the door was beginning to open for women to study for a university degree. Since the beginning of the 20th century there has been a huge change in the career opportunities available to women.

This one-day symposium organised by the RSC's Historical Group will take place on 13 October 2022 at Burlington House. The focus is the 'hidden' women of chemistry and explores the barriers they faced, their roles, contributions to chemistry, and how information about their pioneering efforts can be uncovered. Attendance is free of charge.

Speakers

- Anne Barrett, Imperial College, London: *How archives can reveal hidden women in chemistry*.
- Sally Horrocks, University of Leicester: *Women in industrial chemistry in inter-war Britain* (provisional title).
- Patricia Fara, Clare College, Cambridge: *Listening to the canaries: munitions workers in World War One*.
- Annette Lykknes, NTNU, Trondheim, Norway: *A seat at the table: women and the periodic system*.
- Marelene Rayner-Canham and Geoff Rayner-Canham, Grenfell Campus, Memorial University, Newfoundland, Canada: *"Let us in!" – the opposition to the admission of women to the professional societies.*
- Professor Gill Reid, University of Southampton, President RSC: *My journey with chemistry*.

ACS CINF Report for July 2022

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

Greetings to RSC CICAG members from ACS CINF.

American Chemical Society (ACS) Chemical Information Division (CINF) members recently participated in the ACS Spring 2022 conference in March. The hybrid conference was attended by over 9,000 in-person attendees in San Diego, CA. Most of the CINF sessions were virtual, with the rest hybrid with some in-person talks in San Diego. A listing of the symposia and presentations was published in the spring 2022 *Chemical Information Bulletin* (CIB), with summaries of some of the symposia to appear in a future issue. You can also view some of the presentations from the spring 2021 meeting.

Our next meeting will be from 21-25 August with mainly in-person and some virtual/hybrid sessions in Chicago, Il. Most notably, we will finally celebrate our Herman Skolnik Award winner Wendy Warr. You can see the list of our planned symposia.

This year, our focus is to improve our outreach using social media and other means, looking to organisations like yours to see what we can learn.

CINF members are a collaborative bunch and we'd like to work together with you. We can start by sharing science and information, as well as participating in each other's events, especially the virtual ones. After that if you like, there are opportunities to join our committees, to get involved in program planning for the national meetings, and to write for the CIB. We are also looking for people who wish to develop their leadership skills by volunteering on our executive committee. If you'd like to discuss opportunities, reach out to me or any of our executive committee members.

If you have ideas for projects that CICAG and CINF can work on together, such as developing educational programs or hosting themed events, please reach out to us. We may be able to obtain some funding also.

Thank you for reading.

News from CAS

Contribution from Yvette Lawal, STN IP Suite Solutions Marketing Lead, email: YLawal@cas.org, and Dr Anne Jones, Senior Customer Success Specialist, email: ajones2@acs-i.org

Everyday CAS scientists collect and analyse published scientific literature from around the globe, building the highest quality and most up-to-date collection of scientific information in the world. CAS is proud to cover advances in chemistry and related sciences, and at the heart of the CAS Content Collection is human intelligence.

CAS Common Chemistry

CAS Common Chemistry is an open community resource for accessing reliable chemical information, including nearly 500,000 substances from the CAS Content Collection™. Our *recent publication in the Journal of Chemical Information and Modeling (JCIM)* overviews enhancements to CAS Common Chemistry in 2021 – including an expanded data set, integrated API access, and reusable licensing of content through a Creative Commons Attribution-Non-Commercial (CC-BY-NC 4.0) licence.

Scientific insights

CAS scientists are passionate about advancing scientific breakthroughs. Our experts publish in peer-reviewed journals on a variety of emerging topics (RNA therapeutics, Li-ion recycling, lipid nanoparticles, and beyond) that leverage our unique landscape view of the world's science. These recent peer-reviewed articles are being recognised and cited by WHO, high (top 5%) altimetric attention scores, and top-read article lists. Visit our resources page to learn more.

Webinars and virtual events

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

Blog

The CAS blog features deep insights into various issues that impact the scientific community. Recent topics included:

- Monkeypox: scientifically, how worried should we be?
- Green hydrogen economy: game-changing technologies to transform the world's energy supply.

Visit and follow the CAS blog to stay updated on new insights and resources.

CAS SciFinder Discovery Platform

CAS SciFinder Discovery Platform, an enterprise-wide platform solution with workflow tools and capabilities designed to support multiple scientific research requirements, was launched in 2021. The Platform includes CAS SciFinder^n, CAS Formulus, CAS Analytical Methods, as well as all the new enhancements to Retrosynthetic Planning, and our newest capabilities in Biosequences. Additional enhancements have been made throughout 2022.

CAS SciFinder^n

Expanding our core capabilities to meet the growing needs of the scientific community has taken a significant step with continued development of biosequences functionality within CAS SciFinder^n. In addition, our continued focus on improving the research efficiency and effectiveness of our traditional users remains of utmost importance to our team of technology and scientific experts.

Recent notable enhancements include:

- Workflow enhancements to further improve the experience of biologists using our biosequences functionality.
- Integration of GetFTR full-text links to give users direct access to the full text of entitled and open access articles.
- Inclusion of ORCID iDs as a search option for improved author searching.
- The introduction of the CAS Lexicon Query Builder to enable users to quickly build more thorough search queries.

More detailed information on recent enhancements can be found by viewing the monthly "What's New" release notes within CAS SciFinder^n, or feel free to get in touch and we'll be happy to provide you with more information about items that are most meaningful to you.

CAS Formulus
2022 is off to a great start with CAS Formulus. The year started by focusing on improving the search experience for our users. This included making enhancements to autosuggest, improving controlled vocabulary handling, and incorporating supplier trade names as searchable ingredient identifiers.

In addition to improving the search experience, CAS is investing in expanding our formulations content. CAS Formulus now includes expanded surfactant formulations, covering products such as detergents, hard surface cleaners, and consumer products. Throughout this year, we will also include formulations from coatings, inks, and paints.

Looking ahead for the rest of 2022, CAS Formulus will offer the ability to save items of interest as well as setting alerts on search queries. These key features will offer better workflow support to our formulators. Another upcoming aspect of workflow support is expanded exporting capabilities, including the ability to export information such as *Commonly Formulated With…*, *Commonly Used As…*, and formulation-centric regulatory data.

STN IP Protection Suite
CAS provides IP solutions and services while seeking innovative ways to identify and navigate the IP landscape. Our CAS STNext has AI which effectively improves patent office efficiency and application timeliness. Our STN IP Protection Suite is comprised of FIZ Patmon (end-user tool for worldwide monitoring of IP rights), CAS STNext®, CAS Scientific Patent Explorer™ and CAS Search Guard™.

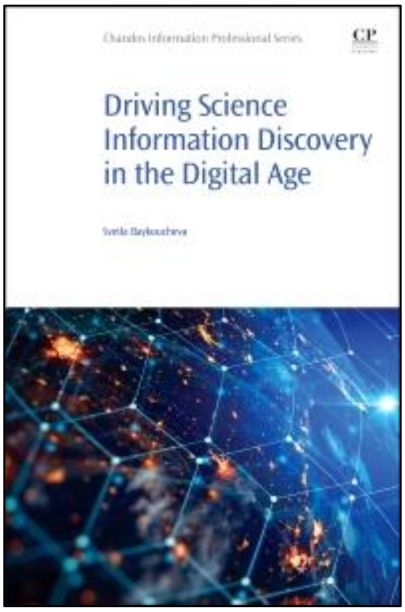
- -

Chemical Information / Cheminformatics and Related Books

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

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

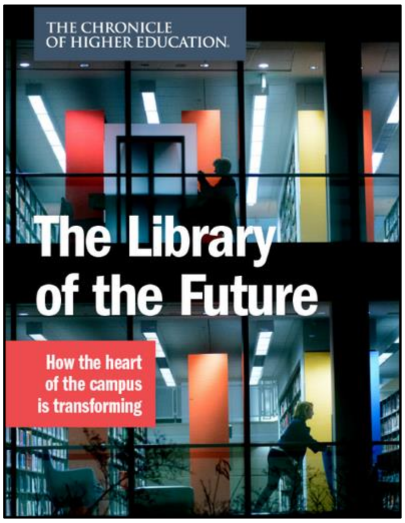
Driving Science Information Discovery in the Digital Age


Chandos Information Professional Series
CP
Driving Science Information Discovery in the Digital Age
Svetla Baykoucheva

New digital technologies have transformed how scientific information is created, disseminated – and discovered. The emergence of new forms of scientific publishing based on open science and open access have caused a major shift in scientific communication and a restructuring of the flow of information. Specialised indexing services and search engines are trying to get into information seekers' minds to understand what users are actually looking for when typing all these keywords or drawing chemical structures. Using artificial intelligence (AI), machine learning, and semantic indexing, these 'discovery agents' are trying to anticipate users' information needs. In this highly competitive environment, authors should not sit and rely only on publishers, search engines, and indexing services to make their works visible. They need to communicate about their research and reach out to a larger audience. *Driving Science Information Discovery in the Digital Age* looks through the "eyes" of the main "players" in this "game" and examines the discovery of scientific information from three different, but intertwined, perspectives: discovering, managing, and using information (information seeker perspective); publishing, disseminating, and making information discoverable (publisher perspective); creating, spreading, and promoting information (author perspective).

Svetla Baykoucheva. Elsevier/Chandos, 2021, ISBN 9780128237236, e-book ISBN 9780128237243

The Library of the Future

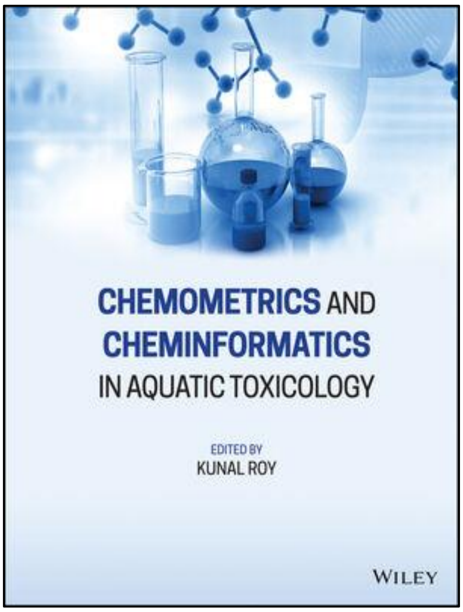

THE CHRONICLE OF HIGHER EDUCATION
The Library of the Future
How the heart of the campus is transforming

Academic libraries occupy a central position on campus, literally and figuratively. For scholars and students, they serve as an essential gateway to knowledge. For publishers, they have been both a partner and a shaper of debates over copyright, censorship, and free and open access to information. For the campus and community, they are a place to connect. Recent years have seen enormous upheaval for libraries and librarians, and the ways they've adapted to such changes point to larger lessons about how colleges and universities can transform, too.

Report, The Chronicle of Higher Education, 2022

Chemometrics and Cheminformatics in Aquatic Chemistry

Chemometrics and Cheminformatics in Aquatic Toxicology delivers an exploration of the existing and emerging problems of contamination of the aquatic environment through various metal and organic pollutants, including industrial chemicals, pharmaceuticals, cosmetics, biocides, nanomaterials, pesticides, surfactants, dyes, and more. The book discusses different chemometric and cheminformatic tools for non-experts and their application to the analysis and modelling of toxicity data of chemicals to various aquatic organisms.

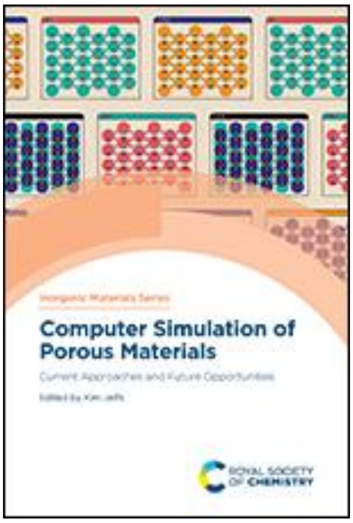


You'll learn about a variety of aquatic toxicity databases and chemometric software tools and webservers as well as practical examples of model development, including illustrations. You'll also find case studies and literature reports to round out your understanding of the subject. Finally, you'll learn about tools and protocols including machine learning, data mining, and QSAR and ligand-based chemical design methods.

Kunal Roy. Wiley, 2021, ISBN: 9781119681601

Computer Simulation of Porous Materials: Current Approaches and Future Opportunities

Computer Simulation of Porous Materials covers the key approaches in the modelling of porous materials, with a focus on how these can be used for structure prediction and to either rationalise or predict a range of properties including sorption, diffusion, mechanical, spectroscopic and catalytic. The book covers the full breadth of



(micro)porous materials, from inorganic (zeolites), to organic including porous polymers and porous molecular materials, and hybrid materials (metal-organic frameworks). Through chapters focusing on techniques for specific types of applications and properties, the book outlines the challenges and opportunities in applying approaches and methods to different classes of systems, including a discussion of high-throughput screening. There is a strong forward-looking focus, to identify where increased computer power or artificial intelligence techniques such as machine learning have the potential to open up new avenues of research. Edited by a world leader in the field, this title provides a valuable resource for not only computational researchers, but also gives an overview for experimental researchers. It is presented at a level accessible to advanced undergraduates, postgraduates and researchers wishing to learn more about the topic.

Kim Jeffs (ed.). RSC, 2022, ISBN 9781788019002

- -

Other Chemical Information News

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

Jisc and the RSC sign new Transitional Open Access Agreement
Jisc and the RSC have extended and revised their transformative agreement until the end of 2024. Now utilising all previous expenditure to support open access (OA) publications, the deal covers all expected publishing output in the RSC's hybrid journals portfolio.
https://www.stm-publishing.com/jisc-and-the-royal-society-of-chemistry-sign-new-transitional-open-access-agreement/
Source: STM Publishing News

Unique Insights Afforded to us by Computational Chemistry
Though experimentation is still king in most chemists' minds, computational chemistry has the potential to transform the field.
https://www.advancedsciencenews.com/unique-insights-afforded-to-us-by-computational-chemistry/
Source: Advanced Science News

Stabilising Pharmaceuticals with Deuterium
An epoxide ring opening reaction could help stabilise biomolecules by replacing hydrogen with deuterium with a high degree of selectivity.
https://www.advancedsciencenews.com/stabilizing-pharmaceuticals-with-deuterium/
Source: Advanced Science News

Chemists Debate Machine Learning's Future in Synthesis Planning and ask for Open Data
A controversial paper ignites discussion around how to use data-based algorithms and report results.
https://cen.acs.org/physical-chemistry/computational-chemistry/Chemists-debate-machine-learnings-future/100/i18
Source: CEN

BioWorld by Clarivate Identifies the Top Six Global Biopharmaceutical Stories and Trends of 2021
Advances in Alzheimer's, innovation in DNA vaccines, increased regulatory collaboration and the evolving impact of Artificial Intelligence, join Covid-19 as key stories and trends of last year.
https://clarivate.com/news/bioworld-by-clarivate-identifies-the-top-six-global-biopharmaceutical-stories-and-trends-of-2021/
Source: Clarivate

Benefits of Semantic Enrichment Across the Drug Development Pipeline
Discussing practical use cases using semantic enrichment technology in various stages of drug development to show the benefits that leveraging semantic search can bring to the table.
https://www.copyright.com/wp-content/uploads/2021/09/CCC-SciBite-White-Paper-Benefits-of-Semantic-Enrichment-Across-the-Drug-Development-Pipeline.pdf
Source: CCC

Nature Masterclasses Online Training made free to access for Researchers in Lower Income Countries
Nature Masterclasses offers highly targeted courses designed to enhance the skills and boost the confidence of early career researchers, with over 60,000 researchers globally having benefited from the platform to date. The world-class training is delivered by editors from across the Nature Portfolio of journals as well as experienced researchers, funders and professionals via a subscription.
https://www.stm-publishing.com/nature-masterclasses-online-training-made-free-to-access-for-researchers-in-lower-income-countries/
Source: STM Publishing News

Newly launched SearchRxiv builds Search Community to foster easier, quicker Research
CABI launched SearchRxiv, its new open access platform. The website is designed to let researchers report, store and share their searches, thus helping with the review and re-use of existing searches to make research quicker and easier.
https://www.infotoday.eu/Articles/News/Featured-News/Newly-launched-searchRxiv-builds-search-community-to-foster-easier-quicker-research-150706.aspx
Source: Information Today

Advancing Open Science in Africa – three Organisations Collaborate to implement Open Science Principles in Seven Partner States
The East African Science and Technology Commission (EASTECO), the Public Library of Science (PLOS), and the Training Centre in Communication (TCC Africa) have announced that they will collaborate in the implementation of Open Science and Open Access principles for EAC Partner States, which include Burundi, Kenya, Rwanda, South Sudan, United Republic of Tanzania, Democratic Republic of Congo and Uganda.
https://www.stm-publishing.com/advancing-open-science-in-africa-three-organizations-collaborate-to-implement-open-science-principles-in-seven-partner-states%ef%bf%bc/
Source: STM Publishing News

Digital History of Science Collection, ready to Launch with nearly one Million Pages
For the first time researchers, teachers and students can access digitally more than 90% of the British Association for the Advancement of Science – Collections on the History of Science (1830s-1970s). Free to Jisc members and affiliates, the move to digitise this collection, much of which was previously unpublished, began in 2020,when leading UK university libraries and archives were invited to put forward their archives. This was the first time they had an opportunity to influence a commercial publisher's decisions about what to digitise. The collection includes maps, photographs, slides and documents from the core years of the British Empire, documenting the

efforts of the British scientific community to establish science as a profession and establish Britain as a world leader in science. It connects the works, thoughts and interactions of some of the most influential scientists of the time, from Charles Darwin to Sir William Ramsay.
https://www.stm-publishing.com/digital-history-of-science-collection-ready-to-launch-with-nearly-one-million-pages/
Source: STM Publishing News

Half of UK Librarians 'won't meet users' needs'
Nearly half of UK university librarians do not believe they will be able to meet student needs in the 2022/23 academic year due to inadequate funding, according to new research.
https://www.researchinformation.info/news/half-uk-librarians-wont-meet-users-needs
Source: Research Information

MRC prizes 2022: Nominations are Now Open
The MRC is now inviting nominations for their 2022 prizes, including the new MRC Impact Prize 2022 and the MRC Millennium Medal 2022.
https://www.ukri.org/news/mrc-prizes-2022-nominations-are-now-open/
Source: UKRI

Dimensions Life Sciences & Chemistry Expands Patent Coverage
Dimensions Life Sciences & Chemistry (Dimensions L&C) has extended its patent data to include more than 140 million patents from around the world. Dimensions L&C – part of Digital Science – now includes 140+ million patents from over 100 patent offices including English translations of Asian and other non-English patent offices. Geographic coverage includes China, Japan, United States, Germany, European Patent Office (EPO), South Korea, World Intellectual Property Organisation (WIPO), United Kingdom, France and Australia, in addition to others. The expansion of patent data has been made possible due to Digital Science's acquisition in 2021 of IFI CLAIMS.
https://www.knowledgespeak.com/news/dimensions-life-sciences-chemistry-expands-patent-coverage/
Source: Knowledgespeak

ACS and Jisc partner to enable Open Access Publishing for Researchers across the UK
The Publications Division of ACS and the Jisc consortium have announced a landmark transitional agreement that will serve researchers in the UK across all fields of chemistry. The three-year agreement, which will last through 2024, provides the ability for all scientific articles published by researchers at UK universities and research institutes in ACS journals to be open access (OA) at no cost to the researcher.
https://www.stm-publishing.com/american-chemical-society-and-jisc-partner-to-enable-open-access-publishing-for-researchers-across-the-uk/
Source: STM Publishing News

ISI Proposes new Citation Credit Indicator
Clarivate has released a report from the company's Institute for Scientific Information, proposing a new method for analysing the credit authors of academic papers receive via citations. In an increasingly global and collaborative world, where the number of articles naming dozens or even hundreds of researchers as authors is rapidly increasing, the need for informed, data-driven analysis on credit that works across research disciplines and regions is essential, Clarivate says. However, existing methods for analysing credit can become distorted by exceptionally high author counts.
https://www.researchinformation.info/news/isi-proposes-new-citation-credit-indicator
Source: Research Information

Improving International Research Collaboration
A consortium led by ARMA UK will explore how recommendations made by an earlier report could support the aims of Trusted Research.
https://www.ukri.org/news/improving-international-research-collaboration/
Source: UKRI

Springer Nature and Figshare launch Pilot Program to enhance Data Sharing
Springer Nature and Figshare have launched a free pilot to better support authors in making their data openly available. Authors submitting to a number of Nature research journals and Academic Journals will now be able to easily opt into data sharing, via Figshare, as part of one integrated submission process.
https://www.knowledgespeak.com/news/springer-nature-and-figshare-launch-pilot-program-to-enhance-data-sharing/
Source: Knowledgespeak

A Smarter Way to Develop New Drugs
MIT researchers have developed a machine learning model that proposes new molecules for the drug discovery process, while ensuring the molecules it suggests can actually be synthesised in a laboratory.
https://news.mit.edu/2022/ai-molecules-new-drugs-0426
Source: MIT News

OA Content up 40% across Springer Nature's Transformative Journals
Data released IN June showed that in 2021 Springer Nature's Transformative Journals (TJs) published 40% more gold open access (OA) research articles than in 2020. 730 Springer Nature journals also met cOAlition S's challenging TJ requirement targets, meaning that more Springer Nature titles achieved the required metrics than those from all other TJ publishers put together. The articles were used on average 2.8 times more than subscription articles in the same journals, demonstrating the value to authors of publishing via the OA route.
https://www.stm-publishing.com/oa-content-up-40-across-springer-natures-transformative-journals-2/
Source: STM Publishing News

RSC Launches three new Sustainability-focused Journals
The RSC has launched three new journals themed around sustainability, as part of its ongoing commitment to support the chemical sciences in facing up to global sustainability challenges. The new journals, *RSC Sustainability*, *Sustainable Food Technology*, and *EES Catalysis* are gold open access, and the RSC will be covering all article processing charges (APCs) until mid-2025 - enabling scientists and institutions from around the world to share research at no charge.
https://www.knowledgespeak.com/news/royal-society-of-chemistry-launches-three-new-sustainability-focused-journals/
Source: Knowledgespeak

Make Research Integrity Training Mandatory say 73% of Australian Researchers in First National Survey
The results of the first national survey to investigate research integrity in Australia, a collaboration between the Australian Academy of Science and publisher Springer Nature, indicate broad support for mandatory research integrity training. The survey found that whilst 68% of respondents stated that their institution offered research integrity related training and 50% stated it was mandatory, 73% felt that such training should be mandatory for all those holding a research position.
https://www.stm-publishing.com/make-research-integrity-training-mandatory-say-73-of-australian-researchers-in-first-national-survey%ef%bf%bc/
Source: STM Publishing News

Science Europe welcomes new DFG position paper on Academic Publishing and Research Assessment

Science Europe has welcomed the German Research Foundation (DFG) position paper "*Academic Publishing as a Foundation and Area of Leverage for Research Assessment*" on two important and timely topics. The paper reflects broad discussions that are taking place across the research landscape and neatly ties together the topics of academic publishing and research assessment, rightly highlighting their interconnections. Research assessment often focuses on quantifiable elements of research outputs: most often simple journal- and author-based publication metrics – these do not capture substance or quality of the research being assessed.
https://www.knowledgespeak.com/news/science-europe-welcomes-new-dfg-position-paper-on-academic-publishing-and-research-assessment/
Source: Knowledgespeak

Karger Publishers and Molecular Connections Collaborate to Enrich Metadata and Build Decision Support System

Molecular Connections Pvt. Ltd. and Karger Publishers have entered into a technical collaboration to enhance search and retrieval of Karger content, build an analytics dashboard to support business decisions, and improve editorial efficiency by machine learning applications. The platform will help Karger make better and faster decisions in terms of content recommendation, identify content gaps, improve editorial workflows, and boost sales. Several technologies will be embedded by Molecular Connections in developing the platform: Molecular Connections' award-winning AI and ML technology that builds custom ontologies, disambiguates institutions, and recommends a manuscript to the right journal. The platform will be able to accommodate updates and content changes both in terms of type and volume. In addition, it will include services that are inbuilt with due diligence, data security, usage policies, and more.
https://www.stm-publishing.com/karger-publishers-and-molecular-connections-collaborate-to-enrich-metadata-and-build-decision-support-system/
Source: STM Publishing News

Frontiers is the first Publisher to sign 'Stick to Science' Initiative

Initiated by Universities UK, Swiss Federal Institute of Technology Lausanne (EPFL), public research university ETH Zurich, the ETH Board, Wellcome and The Royal Society, the '*Stick to Science*' campaign calls for an open, inclusive, and collaborative research and innovation landscape in Europe that is free from political barriers.
https://www.stm-publishing.com/frontiers-is-the-first-publisher-to-sign-stick-to-science-initiative-%ef%bf%bc/
Source: STM Publishing News

ACS CEO Announces his Retirement

The ACS has announced that Thomas Connelly, Jr., Ph.D., will retire from his position as ACS chief executive officer at the end of the year after more than seven years of serving in this role. In addition to effectively navigating the organisation through the unprecedented challenges of a global pandemic, he has led the ACS as it strengthened its core values, transformed its membership and launched new initiatives.
https://www.stm-publishing.com/american-chemical-society-ceo-announces-his-retirement/
Source: STM Publishing News

The Future of Research Revealed

Researchers predict increased use of AI, greater collaboration and more open knowledge.
https://www.elsevier.com/about/press-releases/corporate/the-future-of-research-revealed
Source: Elsevier

ACS partners with Jisc to enable Open Access Publishing for Researchers across the UK

The Publications Division of the ACS has announced a landmark transitional agreement with the Jisc consortium that will serve researchers in the UK across all fields of chemistry. The three-year agreement, which will last through 2024, provides the ability for all scientific articles published by researchers at UK universities and research institutes in ACS journals to be open access (OA) at no cost to the researcher.

https://www.knowledgespeak.com/news/acs-partners-with-jisc-to-enable-open-access-publishing-for-researchers-across-the-uk/

Source: Knowledgespeak

GOBI® Library Solutions Partners with Knowledge Unlatched to Support Open Access Initiatives

GOBI® Library Solutions from EBSCO (GOBI Library Solutions) now supports the Knowledge Unlatched (KU) Open Access (OA) e-books funding model, providing the opportunity for academic libraries to support OA funding initiatives within their GOBI workflow. The addition of the Knowledge Unlatched Open Research Library E-Book platform will make the complete collection of Knowledge Unlatched.

https://www.stm-publishing.com/gobi-library-solutions-from-ebsco-partners-with-knowledge-unlatched-to-support-open-access-initiatives-in-academic-libraries/

Source: STM Publishing News

Research Square and Aries Systems Support Authors, Publishers through AI-based Digital Editing

Research Square, an innovative quality service provider for the global research community, is pleased to announce its partnership with Aries Systems, a leading technology provider of workflow management solutions for scholarly publishers, to improve the quality of submitted research.

https://www.stm-publishing.com/research-square-and-aries-systems-support-authors-publishers-through-ai-based-digital-editing/

Source: STM Publishing News

AIP Publishing launches new OA Scientific Journal, APL Machine Learning

AIP Publishing has announced the launch of a new open access scientific journal — *APL Machine Learning*. The journal will feature research addressing how machine learning (ML) and artificial intelligence (AI) can aid physicists, material scientists, engineers, chemists, and biologists in advancing scientific discovery, and how insights from these disciplines can pave the way for the development of better AI systems.

https://www.knowledgespeak.com/news/aip-publishing-launches-new-oa-scientific-journal-apl-machine-learning-dr-adnan-mehonic-named-founding-editor-in-chief/

Source: Knowledgespeak

BioSpace Introduces Diversity in Life Sciences Content Series

https://www.biospace.com/article/biospace-introduces-diversity-in-life-sciences-content-series-/

Source: BioSpace

CCC Acquires Ringgold

CCC has acquired Ringgold, a longstanding provider of persistent organisation identifiers widely used by the scholarly communications community. With offices in the US and UK, Ringgold is now a wholly owned subsidiary of CCC.

https://www.copyright.com/ccc-announces-acquisition-of-ringgold-leading-provider-of-organization-identifiers-in-scholarly-communications/

Source: CCC

Elsevier releases Research Futures 2.0 Report, Researchers predict increased use of AI, greater Collaboration, and more Open Knowledge

Research launched by Elsevier reveals the challenges and opportunities facing researchers in a post-COVID-19 world. The findings are published in Elsevier's *Research Futures 2.0 report*, which is free to download.
https://www.knowledgespeak.com/news/elsevier-releases-research-futures-2-0-report-researchers-predict-increased-use-of-ai-greater-collaboration-and-more-open-knowledge/
Source: Knowledgespeak

Publishers Collaborate to Launch 'COVID and Beyond: Living with Pandemics

A group of scholarly publishers is launching a major new website to explain research about pandemics and infectious diseases to the public, policy makers, media, and other audiences within and beyond academia. *COVID and Beyond: Living with Pandemics* has been sponsored by Kudos, Impact Science, the ACS, AIP Publishing, De Gruyter, Hindawi, SAGE Publishing, the University of Toronto Press, and Wolters Kluwer Health.
https://www.knowledgespeak.com/news/publishers-collaborate-to-launch-covid-and-beyond-living-with-pandemics/
Source: Knowledgespeak

IntechOpen Launches Portfolio of Open Science Journals

OA publisher IntechOpen has launched a portfolio of Open Science journals covering rapidly expanding areas of interdisciplinary research. IntechOpen was founded by scientists, for scientists, in order to make book publishing accessible around the globe.
https://www.knowledgespeak.com/news/intechopen-launches-a-portfolio-of-open-science-journals/
Source: Knowledgespeak

PLOS and Kudos partner to Break Boundaries and Empower Researchers

Kudos, the award-winning service for helping more people find, understand, and use research, has announced a new partnership with the Public Library of Science (PLOS), the non-profit Open Access publisher.
https://www.knowledgespeak.com/news/plos-and-kudos-partner-to-break-boundaries-and-empower-researchers/
Source: Knowledgespeak

Core Facilities are Central Hubs of Discovery

Core facilities can boost an institution's capacity for research collaboration, but they present challenges for those who run them.
https://www.natureindex.com/news-blog/core-facilities-are-central-hubs-of-discovery
Source: Nature

New Science Journals site named among best Science Websites in 26th Annual Webby Awards

The new Science journals website, which launched last September, has been honoured among the *Best Science Websites* in the 26th Annual Webby Awards.

Being honoured as a Webby Award finalist makes the site for the Science journals, which are published by AAAS, as one of the five best websites in the world in its category, and among the top 12% of the over 14,000 sites entered this year.
https://www.knowledgespeak.com/news/new-science-journals-site-named-among-best-science-websites-in26th-annual-webby-awards/
Source: Knowledgespeak

Partnership aims to use AI to Identify novel cMET Inhibitors for Lung Cancer

Arctoris, a tech-enabled biopharma company, and *Evariste Technologies*, an AI-drug discovery company, have announced a joint venture to identify novel small molecule kinase inhibitors for the treatment of patients with NSCLC. The partnership brings together two highly synergistic approaches for AI-guided and robotics-powered molecule design, to significantly accelerate the DMTA (Design, Make, Test, Analyse) cycle.
https://www.scientific-computing.com/news/partnership-aims-use-ai-identify-novel-cmet-inhibitors-lung-cancer
Source: Scientific Computing World

cOAlition S and ALPSP release Open Access Toolkit

Smaller independent publishers, libraries, and consortia can now more easily enter into Open Access agreements because of a set of new tools published by cOAlition S and the Association of Learned and Professional Society Publishers (ALPSP). In order to foster a diverse, open scholarly publishing landscape, libraries and consortia need to broaden the scope of their negotiation strategies to embrace smaller independent publishers, but tailoring each agreement can take considerable time and resources. Shared standards and greater automation are required, and these tools provide a sound foundation from which to build, notes Colleen Campbell, coordinator of the OA2020 Initiative.
https://www.knowledgespeak.com/news/coalition-s-and-alpsp-release-open-access-toolkit/
Source: Knowledgespeak

Google aiming to Unpick Protein 'Dark Matter' with AI

Google has announced the development of a technique that uses AI and machine learning to reliably predict the function of proteins, helping us decrease the size of the 'dark matter' in the protein universe. Dark matter in this instance refers to the unknowns in the protein universe that this research aims to unlock. This approach, developed in collaboration with the EBI, uses deep learning and allows for the prediction of protein function, functional effects of mutations, and protein design. The research, published in Nature Biotechnology, can be applied to drug discovery, enzyme design, and even understanding the origins of life.
https://www.scientific-computing.com/news/google-aiming-unpick-protein-dark-matter-ai
Source: Scientific Computing World

Writefull's AI-based Language Services integrated into ACS Publications

Writefull's world-leading AI-based language services have been integrated into the ACS Publications workflow. In a partnership that began almost two years ago, ACS has now progressed to a full integration of Writefull's application programming interfaces (APIs) for three key uses. Writefull's proprietary AI technology is trained on millions of scientific papers using Deep Learning. It identifies potential language issues with written texts, offers solutions to those issues, and automatically assesses texts' language quality. Thanks to Writefull's APIs, its tech can be applied at all key points in the editorial workflows.
https://www.digital-science.com/press-release/writefull-ai-language-services-integrated-into-acs-publications/
Source: Digital Science

CAS applies AI-driven Approach in Collaboration with INPI Brazil to Transform Patent Examinations

CAS has completed a major project with The National Institute of Industrial Property (INPI) of Brazil to implement an enhanced examination workflow solution for chemistry patent applications using a unique blend of technology, artificial intelligence, data, and expertise. The new solution reduced application examination times by up to 50%, helping the office achieve its goal of clearing 80% of its multi-year application backlog.
https://www.knowledgespeak.com/news/cas-applies-ai-driven-approach-in-collaboration-with-inpi-brazil-to-transform-patent-examinations/
Source: Knowledgespeak

Gmelin-Beilstein Memorial Medal 2022

The Gmelin-Beilstein Memorial Medal 2022 has been awarded to Professor Gisbert Schneider, Swiss Federal Institute of Technology (ETH) Zurich, Switzerland, by the Gesellschaft Deutscher Chemiker (GDCh, German Chemical Society). The prize was presented at the 17th German Conference on Cheminformatics, Garmisch-Partenkirchen, on May 9, 2022.

https://www.chemistryviews.org/gmelin-beilstein-memorial-medal-2022/

Source: ChemistryViews

ACS and Universities of The Netherlands renew Partnership to Advance OA Publishing

https://www.knowledgespeak.com/news/acs-and-universities-of-the-netherlands-renew-partnership-to-advance-oa-publishing/

Source: Knowledgespeak

CCC Collaborates with Springer Nature to enable direct access to AdisInsight within RightFind Navigate

CCC has announced a collaboration with Springer Nature to provide mutual customers access to *AdisInsight* from within CCC's *RightFind Navigate*. Springer Nature's *AdisInsight* is an integrated database of reports authored by Springer editors on drugs in development, clinical trials, drug safety, company deals, and patents.

https://www.knowledgespeak.com/news/ccc-collaborates-with-springer-nature-to-enable-direct-access-to-adisinsight-within-rightfind-navigate/

Source: Knowledgespeak

AAAS and scite partner to develop Smart Citations

scite, an award-winning tool that helps students and researchers discover and understand research findings better, has announced a partnership with AAAS , the world's largest general scientific society and publisher of the six Science family journals. Using advanced machine learning techniques, scite has developed a novel system that provides enriched citation information, showing not only how many times an article has been cited but how it was cited by a paper by displaying the surrounding textual context, which section it was referenced in, and a classification indicating whether the claims from the cited paper were supported, mentioned, or contrasted. These next-generation citations, termed *Smart Citations*, help readers better understand and contextualise findings in the literature.

https://www.knowledgespeak.com/news/the-american-association-for-the-advancement-of-science-and-scite-partner-to-develop-smart-citations/

Source: Knowledgespeak

GetFTR now supports half of Global Research Output

Get Full Text Research (GetFTR), a free service that enables faster access for researchers to published journal articles, now supports access to more than half of global research output.

https://www.researchinformation.info/news/getftr-now-supports-half-global-research-output

Source: Research Information

Frontiers and ChemRxiv integration now live

Frontiers has announced that authors can now easily and quickly submit manuscripts to their OA publishing platform directly from the preprint repository ChemRxiv. *ChemRxiv* is the fourth preprint platform partnering with Frontiers. Other successful integrations include *Chronos*, *BioRxiv*, and *MedRxiv*. The integration with *ChemRxiv* is the latest development in Frontiers' mission to make all science open by improving the services offered to authors, and streamlining the whole publishing process.

https://www.knowledgespeak.com/news/frontiers-and-chemrxiv-integration-is-now-live/

Source: Knowledgespeak

CHORUS and ChemRxiv sign MOU to pilot Preprint Dashboard Service

CHORUS and ChemRxiv have signed a one-year Memorandum of Understanding (MOU) to pilot a preprint dashboard service. By using persistent identifiers, CHORUS will create a dashboard for *ChemRxiv* that connects preprints to funders and datasets as well as information related to public accessibility and other key metadata to be added later.
https://www.knowledgespeak.com/news/chorus-and-chemrxiv-sign-mou-to-pilot-preprint-dashboard-service/
Source: Knowledgespeak

Small-Molecule Antiviral could be used in Anti-COVID-19 Nasal Spray

N-0385 is a peptidomimetic compound with the sequence Ms-Gln-Phe-Arg-kbt (Ms = mesyl, Gln = glutamine, Phe = phenylalanine, Arg = arginine, kbt = ketobenzothiazol). The drug targets the enzyme TMPRSS2 (transmembrane protease serine 2), which is important for cell entry.
https://www.chemistryviews.org/details/news/11347216/Small-Molecule_Antiviral_Could_Be_Used_in_Anti-COVID-19_Nasal_Spray/
Source: ChemistryViews

Prototype for Game-Changing Ammonia Plant

Engineers are designing a green ammonia plant that can reliably and efficiently generate ammonia using only intermittent renewable energy as the source of power.
https://www.ukri.org/news/prototype-for-game-changing-ammonia-plant/
Source: UKRI

Elsevier's Reaxys becomes the leading source of Curated Chemistry Patents in Collaboration with LexisNexis' PatentSight

Elsevier has announced its market-leading position in chemistry patent coverage and the extension of its collaboration with *LexisNexis® PatentSight®*. In March 2021, Elsevier launched its initiative to strengthen the existing patent coverage in Reaxys®. The content expansion resulted in a 15-fold increase in patent coverage and ensures pharma and chemical companies and their researchers do not miss key competitive intelligence insights.
https://www.knowledgespeak.com/news/elseviers-reaxys-becomes-the-leading-source-of-curated-chemistry-patents-in-collaboration-with-lexisnexis-patentsight/
Source: Knowledgespeak

How Institutions can better Support their Postdoc Fellows

Competition for fellowships can be fierce, but experiences will vary.
https://www.natureindex.com/news-blog/how-institutions-can-better-support-their-postdoc-fellows
Source: Nature

CAS SciFinder includes unique Biosequence data Collection and Capabilities

CAS has announced the launch of a major expansion of the CAS SciFinder Discovery Platform into life sciences. The enhanced platform includes over 2 million modified biosequences, 60 years of patent literature, and one of the largest collections of journal information including PubMed's biomedical and life science data.
https://www.knowledgespeak.com/news/cas-scifinder-discovery-platform-includes-unique-biosequence-data-collection-and-capabilities/
Source: Knowledgespeak

Institutions partner with ACS to advance first California-wide Transformative OA Agreement
Three California consortia, representing nearly 60 academic and research institutions, and the ACS have announced the first-ever California-wide transformative open access agreement. It is also ACS' first "read and publish" agreement in the U.S. composed of multiple consortia. Through a partnership with the 10-campus University of California (UC) system, the 23-campus California State University (CSU) system, and 25 subscribing institutions represented by the Statewide California Electronic Library Consortium (SCELC), readers and researchers at dozens of California research institutions will be able to benefit from full access to subscription content.
https://www.knowledgespeak.com/news/institutions-partner-with-acs-to-advance-first-california-wide-transformative-oa-agreement/
Source: Knowledgespeak

All UKRI Councils now on one, Integrated Website
UKRI has launched a new integrated website delivering a simpler, more efficient user experience. They have brought all seven research councils, *Innovate UK*, and *Research England* websites into one site, providing a strong unified voice for UKRI and its councils.
https://www.ukri.org/news/all-ukri-councils-now-on-one-integrated-website/
Source: UKRI

Second call for Global AI Leaders to Launch
Up to £20 million funding is available through a further two rounds of the Turing AI World-Leading Researcher Fellowships. This will support exceptional researchers to advance AI through world-leading programmes of research.
https://www.ukri.org/news/second-call-for-global-ai-leaders-to-launch/
Source: UKRI

Clarivate Identifies Seven Medtech Trends to Watch in 2022
Report finds innovation, healthcare decentralisation and mergers & acquisitions among key factors set to impact medtech markets in 2022.
https://clarivate.com/news/clarivate-identifies-seven-medtech-trends-to-watch-in-2022/
Source: Clarivate

Elsevier announces intention to acquire Interfolio
Elsevier has announced that it has entered into an agreement to acquire Interfolio, a provider of advanced faculty information solutions for higher education, headquartered in Washington DC, US. Founded in 1999, Interfolio supports over 400 higher education institutions, research funders and academic organisations in 25 countries, and over 1.7 million academic professionals and scholars.
https://www.elsevier.com/about/press-releases/corporate/elsevier-announces-its-intention-to-acquire-interfolio
Source: Elsevier

Google announced new fact checking features
Recognising that information these days comes to us from an enormous number of sources and different directions, Google's new fact checking features helps searchers sort out what information is credible and what isn't.
https://www.infotoday.eu/Articles/News/Featured-News/Google-announced-new-fact-checking-features-152209.aspx
Source: Information Today Europe

SCOAP^3 reaches 50,000 articles milestone

The Sponsoring Consortium for Open Access Publishing in Particle Physics (SCOAP^3), the world's largest disciplinary open access initiative, has reached the milestone of over 50,000 research articles published.
https://www.knowledgespeak.com/news/scoap3-reaches-50000-articles-milestone/
Source: Knowledgespeak

EBSCO Information Services Releases FSTA with Full Text

For the foodies and food researchers among us, EBSCO now has extensive full-text coverage of food and nutrition resources available.
https://www.infotoday.eu/Articles/News/Featured-News/EBSCO-Information-Services-Releases-FSTA-with-Full-Text-151611.aspx
Source: Information Today Europe

Elsevier began a Pilot Project to make Articles from Four Major Publishers available via ScienceDirect

The pilot, a collaboration with the publishers in an attempt to help researchers find and access content more easily, is intended to last between four and six months, with articles appearing in search and browse lists. The GetFTR button will allow authorised users to access the full text of articles.
https://www.infotoday.eu/Articles/News/Featured-News/Elsevier-began-a-pilot-project-to-make-articles-from-four-major-publishers-available-via-ScienceDirect--151159.aspx
Source: Information Today Europe

Two new journals to support ACS' growing Applied Materials Portfolio

The ACS has announced the launch of two new journals this spring, *ACS Applied Engineering Materials*, and *ACS Applied Optical Materials*. The journals will be guided by ACS Applied Materials & Interfaces Editor in Chief Kirk Schanze, Ph.D., and each will be under the leadership of a deputy editor.
https://www.knowledgespeak.com/news/two-new-journals-to-support-acs-growing-applied-materials-portfolio/
Source: Knowledgespeak

Institutions partner with ACS to advance first California-wide Transformative OA Agreement

Three California consortia, representing nearly 60 academic and research institutions, and the ACS have announced the first-ever California-wide transformative open access agreement. It is also ACS' first "read and publish" agreement in the U.S. composed of multiple consortia. Through a partnership with the 10-campus University of California (UC) system, the 23-campus California State University (CSU) system, and 25 subscribing institutions represented by the Statewide California Electronic Library Consortium (SCELC), readers and researchers at dozens of California research institutions will be able to benefit from full access to subscription content.
https://www.knowledgespeak.com/news/institutions-partner-with-acs-to-advance-first-california-wide-transformative-oa-agreement/
Source: Knowledgespeak