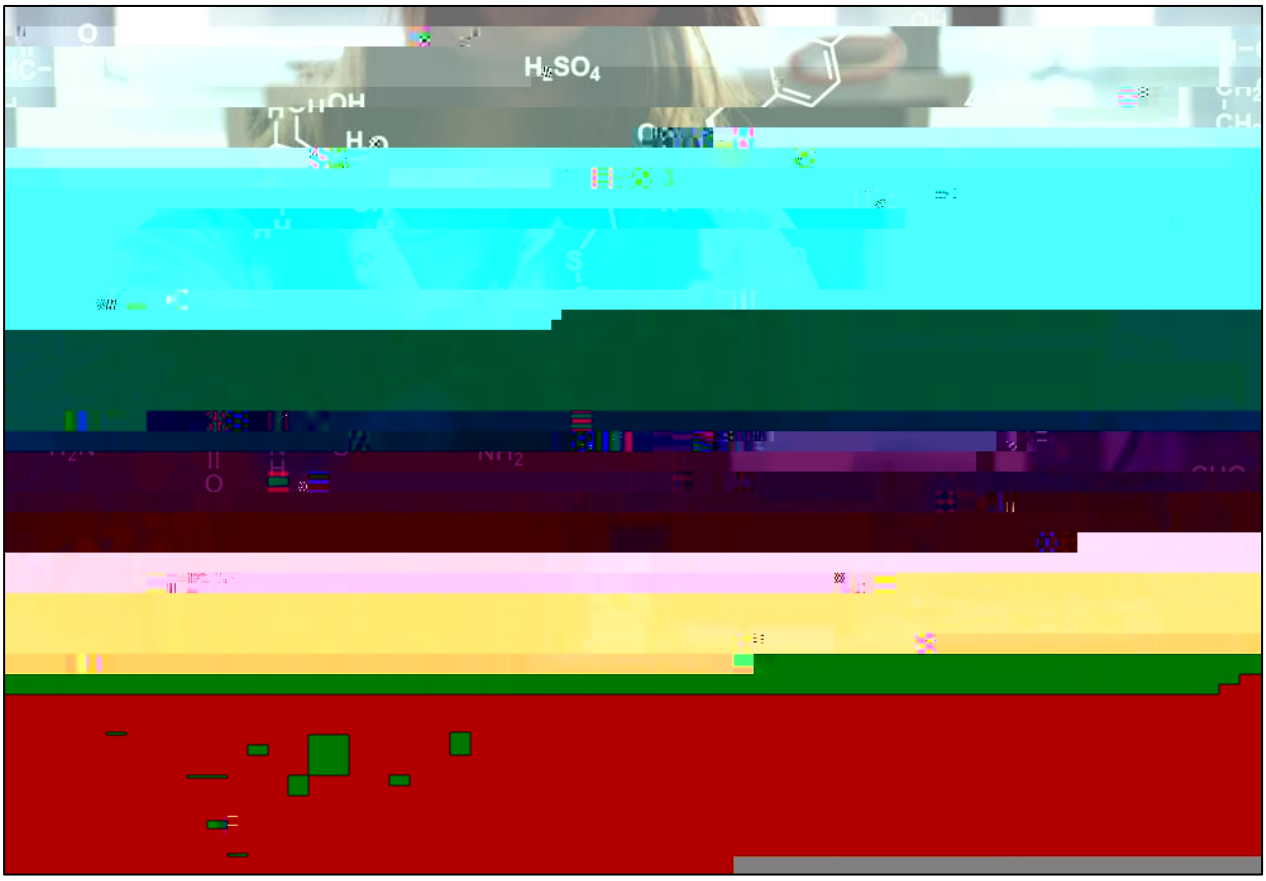
NEWSLETTER

Summer 2024

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.



Introducing the new AIChemy Hub (image credit: Charles Romain). See the full article on p. 51-53.

CICAG Websites and Social Media

http://www.rsc.org/CICAG

http://www.rsccicag.org

https://www.youtube.com/c/RSCCICAG https://www.linkedin.com/groups/1989945/

@rsccicag@science94 Tf1 0 C501 9tpn.r

Contents

The Royal Society of Chemistry's prizes have recognised excellence in the chemical sciences for more than 150 years. This year's winners join a prestigious list of past winners in the RSC's prize portfolio, 60 of whom have gone on to win Nobel Prizes for their work, including 2022 Nobel laureate Carolyn Bertozzi and 2019 Nobel laureate John B Goodenough.

The Volunteer Recognition Prizes celebrate those who give their time freely in numerous ways, from serving

Desperately Seeking Heterocycles

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

An ordinary day in the laboratory in February chasing through citations to find the details of the characterisation of a molecule. This is fundamental to chemistry: how do we know whether a molecule has been made? In this particular case, the trail took me to the journal *Heterocycl*es (published by the Japan Institute of Heterocyclic Chemistry), a good source of analytical information. My institution has a subscription to this journal, so I went to the website, to find a notice: “It has been decided that publication of HETEROCYCLES will be suspended from December 2023 due to various circumstances.” It was disappointing that no new issues would be published and disturbing to discover that on-line access to past editions had also disappeared. My institution switched to on-line only access in 2004, so twenty years of papers are inaccessible to me. These two decades saw the publication of more than five thousand papers, which have been cited more than forty thousand times.

This immediately brought to mind a recent article in *Nature*. The title reports that “Millions of research papers at risk of disappearing from the Internet”. Originally the paper had a different title, as can be gathered from the correction “The headline of this story has been edited to reflect the fact that some of these papers have not entirely k p

Will it be possible to find this article after a couple of decades? As I write this, it is obvious it was written in March and not August, but this will become less clear as time passes. This blog post discusses a paper with a DOI: 10.31274/jlsc.16288. This is a study of archives. Are papers accessible in archives if a publisher goes bankrupt or loses interest in the journal? If you hope that the answer is “yes” you might want to take a deep breath before reading.

Davis and James Rush later collaborated on the 1974 book *Information Retrieval and Documentation in Chemistry*.[3] Another IU PhD thesis was by John Michael Knego.[4]

In 1969/70, I worked as an intern for John Knego, who was then the Head of the IU Chemistry Library. With strong encouragement and support from the Department

programs. He was not convinced. Somewhat discouraged and knowing that I was wearing too many hats and had too many bosses, I decided to give up my involvement with informatics. However, once the School of Informatics was approved, the dean of the new school, J. Michael Dunn, made me an offer I could not refuse. I spent my last

6th International Conference on Chemical Research and Education, Washington, D.C., July 1982.
Above: Gary is pictured alongside Steve Heller, third from the right in the third row up.

100 Years of Markush

Markush structure search (e.g., in CAS REGISTRY) to find them (scenario). This poses a legal problem if it can be argued that the broad Markush in question was deliberately intending to cover the chemical space of interest defined by the query. This is just one illustration and there are other variations to scenario -type circumstances, but broadly speaking we need to be aware of contextual structural similarity as a nuance to Markush searching.

One could argue that ideally, initiation of a secondary broader search (or searches) against the subset of a database of specific compounds corresponding to the content of the Markush search answers could be undertaken immediately following execution of the Markush search, with the marrying of the two answer sets in order to identity potential scenario (and wider, as above) answers. Although such an approach is available manually using today's search systems, this is cost- and time-consuming and is not used as a meth71 0 595.32 8( )-2((wi)( )

3871 0 595.32 841.92 reW* nBT/F3 9.96 Tf12 121 54.024 628.421330526.80.0.13[(-)ETBoq0.000008871 0 595.32 841.92 reW* nBT/F3 9.96 Tf12

the pyrrolopyrimidine. Note that all the conformers have the piperidine methyl group axial. Thus, even if we didn't know anything about the target that tofacitinib binds to, we could still guess the two most likely bioactive conformations of the molecule: One corresponding to the first six structures in Figure 2, and the other corresponding to the last six. With this knowledge, we can imagine two ways to cyclise the molecule and lock the piperidine relative to the pyrrolopyrimidine core, see Figure 3.

Figure 2. Conformational analysis of tofacitinib (Schrödinger MacroModel, OPLS4 in water). Generated conformers are shown in grey with relative energies in kcal/mol. The bioactive conformation from the X-ray crystal structure of tofacitinib in complex with JAK1 (PBD: 3EYG) is shown in green for reference. The closest conformer (green square) is within 1.4 kcal/mol of the global minimum (top left structure). Only the first 12 conformers are shown. A total of 120 were generated within an energy window of 5 kcal/mol.

For more flexible molecules, the conformational space will be too large to effectively inform us on the bioactive conformation. In this situation, two ways forward can be tried:

1. Try to cut off flexible appendages to see if a smaller but more rigid starting point can be found.

Figure 3. Ideas for locking the piperidine ring into one or the other low energy conformation by adding a third ring to the pyrrolopyrimidine core.

6. Convert the energy into a docking score, which is a measure of the strength of the ligand-receptor interaction: The better the docking score, the better the fit of the ligand to the receptor. It is important to know that the docking score is not an estimate of binding affinity! Other methods, e.g., free energy perturbation (FEP), can be used to estimate affinity, but they are best used by compchem experts. This recent perspective provides an overview.

Summary

- Conformational analysis of the ligand(s) can help identify potential bioactive conformations.
- For flexible molecules, cutting off the flexible appendages or rigidifying them can help identify the bioactive conformation.
- New designs should be conformationally analysed to make sure they are biased towards the (assumed) bioactive conformation(s).
- Pharmacophore searches and shape match methods can be used to identify new ligands.

- The ligand-target complex shows us potentially important interactions.
- ViN

that mean? The Hamiltonian is a mathematical rule that describes the laws of physics for the kinetic and

Finally, it is worth noting that with the advent of AI methods like deep learning, it is becoming possible to get QM-quality results at much higher speeds. OrbNet-Equi is one recent example of this approach. Hopefully, such methods will make their way into compchem modelling software in the near future.

Summary

Molecular mechanics (MM) methods

- Describe the system (one or more interacting molecules) as "balls and springs" using the laws of classical, macroscopic mechanics.
- Are very fast (milliseconds to seconds), even for large systems like proteins in complex with small molecules.
- Depend on lookup tables with pre-calculated constants for every atom, bond etc.
- Have limited accuracy and generally give energies within 1-3 kcal/mol of the true value.

Quantum mechanics (QM) methods

- Describe the system (one or more interacting molecules) as a wavefunction.
- Are slow (minutes to hours) even for small molecules with MW < 500.
- Do not depend on lookup tables like MM methods but instead find approximate solutions to the

Careers Support from the Royal Society of Chemistry

Contribution from Dr Robert Bowles MRSC RCDP, Careers and Professional Development Adviser, Royal Society of Chemistry, email: careers@rsc.org

Navigating your career as a chemist can be a challenge for many. At different points in your career you may need help with making a key decision, or developing a vision for your future. Perhaps you would simply like to review your career to date. As a computational chemist, cheminformatician or chemical information specialist you have a wide range of career opportunities open to you. For some,

Currently, the cookbook contains content including:

-

Despite leaving the RSC, over the years through family membership affiliations, I've seen the magazine evolve subject-wise and design-wise. As is the case for any reader, 'my' magazine has sparked interest and occasionally surprise. To stay relevant, *Chemistry World* has to reflect the advancing face of its subject matter and do it with a modern look and feel if it is to maintain its position as a valued benefit to members; Philip Robinson, the current editor, clearly recognises this. Recent changes, to my mind, have been exciting. Feature articles in particular are standout successes: in-depth analyses coupled with innovative layouts unafraid to use white space to enhance the reading experience, with this year's special issue on water a prime example. Cover designs are often inspired, such as last February's story on the chemistry of love. The launch of *Chemistry World*'s new website in 2016 enabled a much wider readership to access the magazine and provided a platform for new and updated content including a members' area.

Looking to the future

What next for *Chemistry World*? As an observer, to me it is to RSC's credit that it has always seen the value of its magazine, without focussing exclusively on its bottom line. However, print publications and their delivery costs are very expensive and impact the environment in ways we have come to understand better over the last 20 years. Many members now consume news through other media, certainly different to when the magazine's first issue appeared. I have no insight into what plans may emerge for the ongoing development of *Chemistry World*. However, to me it seems inevitable that members in time may see yet more of a digital-led publication, perhaps even with less emphasis on content published to a fixed monthly frequency. What I am confident about is that the RSC values *Chemistry World*'s unique ability to connect with its members, and the wider community, whilst its journalists continue to filter, analyse and report on the significant and fascinating contributions

This commentary introduces the general concepts of molecule normalisation in the context of InChI[1] and SMILES,[2,]

InChI and SMILES – key differences in the context of general use-cases

The standard InChI normalises and merges equivalent chemical structures based on resonance, aromaticity, and some mobile hydrogens (e.g., tautomers) into one unique representation. There is no guarantee that the originally drawn structure will be reproduced by either the standard InChI or a SMILES string. Information is

(2,6-dichlorophenyl)methylideneamino]guanidine, PubChem CID 5353646, and this is concordant with the InChI view of this structure. If you are searching a database, the InChI would be a good choice, because the two representations of the one molecule have one InChI. However, if, like the study in reference (10), you are discussing proton localisation, SMILES (or a non-

(12) https://www.nextmovesoftware.com/talks/OBoyle_SmilesBenchmark_ACS_201808.pdf
(13) https:/www.daylight.com/dayhtml/doc/theory/theory.smarts.html
(14) https://www.inchi-trust.org/technical-faq-2/#16.3

- -

Google and DeepMind spun out Isomorphic Labs to focus on AI-driven drug discovery. Given how competitive the field is, DeepMind are obviously keen to show that they are still on top.

That is presumably the main motivator for unveiling AlphaFold3, which was published in Nature last month along with a webserver to run predictions. The controversial nature (pun intended) surrounding the release of the model has also come to symbolise the transformation of DeepMind/Isomorphic from an open-source altruist to a full-blooded biotech.

On the technical side, the headline news is that AlphaFold3 operates on all-atom coordinates and also uses a diffusion model (a kind of generative model I described in my last article) to build 3D coordinates. The former means it can now generalise to more kinds of biomolecular modelling, including predicting ligands, DNAs, RNAs, modified residues and glycans. The latter means that AlphaFold is now a generative model rather than a predictive one, so it does not always output the same structure and it can sometimes confidently hallucinate bogus outputs (more on that later).

Remarkably, many of the ideas and innovations introduced in AF2, such as equivariance and frames, which have become a defining feature for the research community that the model spawned, were unceremoniously thrown out of the window in favour of brutal simplic1gmunity that the model spawn0T/F3 9.96 T63((Q123(8(o)4(d(c.87f)-

The Pairformer architecture. The single representation stores information on a per token basis (e.g. one residue or ligand atom) and the pair representation stores structural information between residues/atoms. Image credit: DeepMind/Isomorphic, taken from Nature paper.

The first major difference is the second module, called the Pairformer, which generalises AF2's Evoformer to all kinds of biomolecules. In AF2, the Evoformer had a two-track setup: one track took a 2D multiple sequence alignment (MSA) as input, and the other stored representations of pairwise relationships between residues. An attention mechanism at each layer extracted co-evolutionary information from the MSA, passing it to the pair representation track for structural reasoning.

In AF3, the Pairformer retains familiar triangle update and attention mechanisms but applies them to all modalities. Amino acids and nucleotides are tokenised by residue or base1 54.0247des ares024r96otides are token

impressive is that, when binding residues are specified (this is an optional input when running AF3), performance on the PoseBusters docking benchmarking set jumps up to 90.2%, up from 76.4%. (Note: I am only discussing the results on PoseBusters V1 for brevity.)

AlphaFold3 performance on PoseBusters set and validity checks. Image credit: DeepMind/Isomorphic, taken from Nature paper.

A subject close to my heart is physical realism of generated poses from ML models, thankfully, the paper also reports the success rate of the

The training datasets used in AlphaFold3. Collecting, cleaning, processing this data and perfecting sampling strategies takes a considerable amount of engineering. DeepMind/Isomorphic, taken from Nature paper.

Reading the supplementary methods, it is obvious that a substantial amount of effort has been put into the data engineering pipeline. This allows them to quickly collect, merge and pre-process all kinds of biological data, and importantly, rapidly try out lots of different training datasets and splits (they do four fine-tuning stages!). Tools like this, along with training code, are rarely published nowadays. The lack of such good open-source tooling for data management here is holding back the research community substantially, so I hope to see a large academic group make one available soon.

A little talked about nugget from the supplementary methods relates to how AF3 was fine-tuned for modelling transcription factors (i.e. protein-DNA complexes). Here, non-structural data was taken from high-throughput SELEX experiments to construct 16k diverse protein-DNA pairs. AF3 was used to make predictions and then fine-tuned on complexes with high predicted confidence. Similarly, the SELEX data was used to construct negative pairs and AF3 was trained to predict idealised monomers of each molecule some distance apart. Although the lack of ablations makes it difficult to measure the importance of this.

This strategy of using large-scale non-structural data, in particular for areas where we are structural data poor such as nucleotides, is quite hot currently. There are a number of startups founded by great ML people moving in this direction, and I predict that we will see Isomporhic try to overcome the data bottleneck with some kind of ultra-high throughput experimental technology such as DELs, deep screening of antibodies or chemotype evolution.

There is a web server that allows you to run AF3, and I'm sure we can expect to see more independent benchmarking studies coming out in the near future. However, a crucial limitation of the web server is that you're not allowed to dock your own ligands; instead, you can only select molecules from a predefined list (e.g., ATP, NAD). These molecules are all well within the training distribution, so if your protein has a binding site for these molecules, AF3 will likely perform well. Notably, this restriction means that others cannot easily measure AF3's generalisation capabilities to novel scaffolds.

The licence for AF3 is much more restrictive than that for AF2. Commercial use of AF3 is understandably prohibited (unless you pay them a lot for partnerships), and even academics cannot use AF3 predicted structures in any software that predicts ligand or peptide binding (e.g. AutoDock). Additionally, users are prohibited from using AF3 outputs to train their own machine learning models. This is notable because AF2 outputs, available from the AlphaFold Database, have 4(ed)-5( )-47(st)3(r)-258(A)-4s63rrocaot ken from

knowledge into cheaper models. Moreover, recent improvements in docking generalisation have been achieved by bootstrapping docking models with high-confidence docked poses.

Initially, I wrote a long, detailed analysis of each sentence in the above paragraph, speculating on what each meant for AF3 performance and what DeepMind might be hiding, both positively and negatively. However, on reflection, it is probably just a case that lawyers in the room wanted to have maximum protection over anything IP generating, allowing them to generate maximum revenue from partnerships down the line.

Much has been said about the AF3 publication and open-sourcing debate (and, to be honest, I will leave it to those not looking for an industry career!). Regardless, I believe AF3 *will* eventually be reproduced and open-sourced due to its comprehensive supplementary methods. However, predicting the exact timeline is challenging.

Training AF3 requires significant resources. Based on my estimates, it took around 5,000 A100 GPU days, costing between $100k and $500k for a single training run if reproduced by an academic group (I'm sure Google also calculated the opportunity cost of not training another LLM!). This is many times more than an average PhD student's total GPU usage during their entire graduate career. The actual compute used in model development was likely much higher.

While few academic groups can reproduce AF3, the OpenFold Consortium, led by Mohammed AlQuraishi at Columbia, is well-positioned to do this first. They successfully retrained and released the AF2 model with a widely permissible licence. The other confounding factor is that Isomorphic have said they will release an open-source version of AF3 in 6 months (for non-commercial use). As happened with OpenFold, it is likely then that the first PyTorch version of the model will initially use the weights released by DeepMind, with them figuring out how to train from scratch later. At which point, researchers can try and build on top of it.

DeepMind's decision to release AlphaFold3 now is likely driven by multiple strategic goals: securing future partnerships, demonstrating their technological superiority over competitors, and attracting top talent. Isomorphic has already established partnerships with pharmaceutical giants like Eli Lilly and Novartis, securing agreements worth an astounding $3 billion in biobucks – a remarkable figure for such a nascent biotech firm. Given Isomorphic's algorithms are incredibly data-intensive, we can expect further collaborations not only with major pharmaceutical companies but potentially also with startups that possess ultra-high throughput technologies capable of generating the vast datasets they need to keep the music going.

Moreover, Isomorphic's strategy extends beyond external partnerships; the company plans to build its own in-house wet labs to support its data acquisition efforts. This approach is partly driven by the challenges in generalising to novel targets, which has prompted many in the field to focus on amassing data for specific targets. For example, Charm Therapeutics, which has its own co-folding model DragonFold, has invested in developing a wet lab to determine numerous crystal structures, thus creating a protective moat of target-specific data which is only accessible with material venture capital investments.

This signifies a broader trend others have noticed in TechBio from a 'first-in-class' to a 'best-in-class' approach. Early startups in the space promised AI could *discover* novel biology, front-loading a large amount of both technological and biological risk by going after novel targets, and have tended to struggle in the clinic. In contrast, newer TechBio companies, riding the hype around AlphaFold and 'GenAI', are collecting large

The technological backbone of NFDI4Chem aims to provide researchers with an easy-to-use and comprehensive research data infrastructure covering the whole research data lifecycle. We have identified the early capture of research data in the lab as the bottleneck of this process.

Figure 1. Capturing data early in the lab enables researchers to overcome typical hurdles in publishing data later in the scientific process.

Only if researchers are empowered to capture and manage all research data in an electronic laboratory notebook with LIMS functions will they be able to easily push well-annotated FAIR datasets into public repositories. One thing that is frequently overlook

establishing the CSD, Dr Kennard made significant contributions as a crystallographer, solving the first structure of ATP and various nucleotide structures. Olga excelled as a successful woman in science during a time when such accomplishments were uncommon. She became a Fellow of the Royal Society and received several awards recognising her scientific contributions throughout her lifetime. Beyond her scientific endeavours, she nurtured a keen interest in art, architecture and design, and served as a Trustee of the British Museum.

I had the privilege of opening the session with a talk on “The

Dirk Trauner was next on the agenda with a talk on “The Chemist and the Architect” which highlighted Olga’s passion for architecture and design. In the talk he used molecular models from the CCDC and PDB databases, to reveal the similarities between synthetic chemistry and architecture, between molecules and buildings. He

during pharmaceutical development. In the presentation, Rajni focussed on case studies that demonstrated how

If we are successful we will eventually have a brand new Heritage Gallery on the first floor to replace the one

planned and the Institute of Chemical Engineers (IChemE) have said that they will provide financial support on that occasion.

Catalyst is determined to become more inclusive in its provision and to bring under-represented groups in to attend its workshops and shows at the weekends and in the holidays. I have recently applied for a Large Grant to the RSC Outreach Fund for money to support a project called IDHHP: Interpreters for Deaf and Hard of Hearing People. If we succeed in getting the funding we will work with the local Deafness Resource Centre to encourage and enable deaf people to join our workshops and shows with a British Sign Language (BSL) interpreter present. The money would also include training of all our visitor services staff in welcoming the deaf community to Catalyst. Keeping fingers crossed for the funding of this much-needed development.

Our new CEO, Mrs Nikki Burton Mallot, is now settled in. She was formerly the Education Manager at Knowsley Safari Park near Liverpool and is a graduate of Liverpool University.

So lots of activities happening at Catalyst. Do come and visit us if you are in the area.

Finally, on Saturday 1 June I was at the Chemistry Department of the University of Edinburgh, my alma mater,

graduates including myself who were felt to be role models for younger women of what women chemists with children can achieve, and who are also deeply involved with the Royal Society of Chemistry.

- - -

RSC Databases Update

Contribution from Richard Kidd, Royal Society of Chemistry, email: KiddR@rsc.org, and Tania Benito Fernandez, Royal Society of Chemistry, email BenitoFernandezT@rsc.org

Following on from the article on ChemSpider published in the last CICAG Newsletter, we are grateful for all the feedback received on our beta site. The team is currently preparing for the full data transfer and switchover

Highlights from the last few months include:

In late 2023 Professor Simon Coles and Dr Samantha Pearman-Kanza were invited to present at the ChemSpider Webinar Series. These webinars were created as a free, three-part series for chemical scientists working with data, learn more about chemistry data today, what the future holds, and the current challenges and opportunities of digital chemistry data.

On 17 November 2023, Simon presented for the second webinar: What does the future hold for digital chemistry? Simon's presentation was entitled "Will an AI win a chemistry Nobel Prize and replace us" which looked at the part AI has played up to now in the physical sciences with a look to the future on how things are going to change moving forward.

On 7 December 2023, Samantha presented for the third and final webinar: Challenges and opportunities for digital chemistry data. Samantha's presentation was entitled "How can we combat heterogeneous, unfair and disparate data in digital chemistry?" and provided a whirlwind tour into a number of different areas that PSDI are working on:

- Understanding the barriers and challenges to digital research.
- Process recording, understanding what researchers actually want and how to choose the right tools for you to record your processes in the first place
- The considerations needed for producing fully fair data, research and code

A more detailed blog post about this event can be found on our website.

In April 2024 Dr Samantha Pearman-Kanza and Dr Nicola Knight gave a presentation on behalf of PSDI for the Keele Open Research Network entitled "Promoting Open & Transparent Research Practices in the Physical Sciences through PSDI". This presentation details the research and service development that PSDI is undertaking with relation to openness and transparency. A more detailed blog post about this event can be found on our website.

In May 2024 Dr Samantha Pearman-Kanza presented at the CLEAR 2024 Global Virtual Symposium on "The role of digital note taking for the 21st Century Scientist". This presentation included topics about Scientific

The author is a Professor of Economics at Pomona College, an expert in statistical analysis and computer systems, and the author of 15 books and more than 100 articles. The examples tend to be US oriented but the resources and effects are global, not parochial.

The Table of Contents (with some examples, including Bitcoin economics) are below.

Introduction: Disinformation, Data Torturing, and Data Mining
The three main topics are defined, discussed, and some examples described. The bad reputation of science and scientists among the public is ironically the result of STEM invented tools.

Part I Disinformation

ESP observances are discussed, the experiments of J.B. Rines and James Randi proving hoaxes are described.

UFOs, alien invasions, crop circles, and moon landing deniers are discussed.

Bill Gates' alleged COVID vaccinations with microchips, the shadowy Illuminati, and conspiracy theories are discussed. The cover-up of the Vietnam Reports by the government, disclosed in the Pentagon Papers, is described.

Evidence-based governmental policies are threatened by the post-fact world, especially COVID treatment, where ideology can trump scientific knowledge. Sources include the *Weekly World News* (a satirical newspaper) – 'fake news', yet believed by many; disinfomedia (e.g. Facebook); cyberwarfare (ISIS, Russia), including by Russia on Ukraine and US elections; Pizzagate (yielding QAnon, a notorious disinformation site).

Part II Data Torturing

Statistical errors, irrelevant and incorrect correlations, often due to deep diving, are described. Post-hoc fallacies (false attribution / "prediction" of deaths but after the fact) are noted. The pitfalls of misinterpretation and design of randomised controlled trials are discussed, as well as over dependence on statistical significance. The hydrochloroquine hoax and validity of Federal funding of vaccine producers (Pfizer no, Moderna yes) are noted. The notorious Wansink Pizza Papers (bad correlations) are discussed, as well as P-Hacking, and false positives in testing. Election models predictability is discussed including statistical significance vs practical importance.

Real and important problems with trials for medical treatments, including real positives vs false positives, are discussed. The failure of Apple watches, leading to a large number of false positives for atrial fibrillation, led to it being withdrawn. Electronic medical records (EMR), instead of providing medical record efficiency, are

The *British Medical Journal* Christmas issue, labelled “goofy”, but often believed, is described. The notorious full m

Contributed to and edited by experts with long-time experience in the field, *Open Access Databases and Datasets for Drug Discovery* includes information on:

- An extensive listing of open access databases and datasets for computer-aided drug design
- PubChem as a chemical database for drug discovery, DrugBank Online, and bioisosteric replacement for drug discovery supported by the SwissBioisostere database
- The Protein Data Bank (PDB) and macromolecular structure data supporting computer-aided drug design, and the SWISS-MODEL repository of 3D protein structures and models
- PDB-REDO in computational aided drug design (CADD), and using Pharos/TCRD for discovering druggable targets

Unmatched in scope and thoroughly reviewing small and large open data sources relevant for rational drug design, *Open Access Databases and Datasets for Drug Discovery* is an essential reference for medicinal and

Computational Phytochemistry

Computational Phytochemistry, 2nd ed., explores how recent advances in computational techniques and methods have been embraced by phytochemical researchers to enhance many of their operations, refocusing and expanding the possibilities of phytochemical studies. By applying computational aids and mathematical models to extraction, isolation, structure determination, and bioactivity testing, researchers can obtain highly detailed information about phytochemicals and optimize working approaches.

This book aims to support and encourage researchers currently working with or looking to incorporate computational methods into their phytochemical work. Topics in this book include computational methods for predicting medicinal properties, optimizing extraction, isolating plant secondary metabolites, and building dereplicated phytochemical libraries. The roles of high-throughput screening, spectral data for structural prediction, plant metabolomics, and biosynthesis are all reviewed before the application of computational aids for assessing bioactivities and virtual screening is discussed. Illustrated with detailed figures and supported by practical examples, this book is an indispensable guide for all those involved with the identification, extraction, and application of active agents from natural products.

This new edition captures remarkable advancements in mathematical modelling and computational methods that have been incorporated in phytochemical research, addressing, e.g., extraction, isolation, structure determination, and bioactivity testing of phytochemicals.

Edited by Satyajit Dey Sarker, Lutfun Nahar. Elsevier, March 2024. Paperback ISBN: 9780443161025. eBook ISBN: 9780443161032.

Exploring Chemical Concepts Through Theory and Computation

Topics discussed include:

- Orbital-based approaches, density-based approaches, chemical bonding, partial charges, atoms in molecules, oxidation states, aromaticity and antiaromaticity, and acidity and basicity
- Electronegativity, hardness, softness, HSAB, sigma-hole interactions, charge transport and energy

UKeiG – Call for Award Nominations

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

The UK e-information Group (UKeiG) is delighted to launch a call for nominations for three international awards in the fields of information retrieval, library and aal

Chor Award

information science courses in 1963 at the precursor to City University, London, where he became Director of the Centre for Information Science in 1966.

Nominations should meet one or more of the following criteria:

- Contributing to the creation, promotion and exploitation of digital resources and services
- Raising the profile of library and information services across the organisation
- Raising awareness of the value and impact of library and information services internally and/or externally
-

Please include testimonials, letters of support, references, a selective bibliography relevant to the nomination,

The emergence of GPT-4 and ChatGPT brought considerable attention to large language models (LLMs) in 2023. In November and December, several large pharmas held "AI Day" presentations featuring LLM applications for clinical trial data analysis. Many of these groups demonstrated the ability of LLMs to ingest large bodies of unstructured clinical data and subsequently generate tables and reports based on natural language queries. Aside from some very brief demos on code generation and literature searches, mentions of

One means of enhancing the ability of LLMs to perform domain-specific tasks is to provide a set of "helper" programs that the LLM can call. A paper by Bran and coworkers from EFPL and the University of Rochester presented ChemCrow, a system for integrating Chemistry capabilities into LLMs. ChemCrow provides software tools for performing domain-specific tasks, including web searches, file format conversions, and similarity searches. Compared with GPT-4, ChemCrow provided superior performance on tasks like synthetic route planning. The authors also point to potential misuse of LLMs and suggest mitigation strategies.

ChemCrow: Augmenting large-language models with chemistry tools https://arxiv.org/abs/2304.05376

Given the complexity of programming multiple instruments and identifying appropriate experimental conditions, laboratory automation is another area that can benefit from applying LLMs. A paper by Bioko and coworkers from Carnegie Mellon University presented Coscientist, a set of LLMs for designing and executing organic syntheses. Coscientist consists of four components designed to search the web, write Python code, extract information from documentation, and program laboratory robotics. The authors test Coscientist using several open and closed-source LLMs and present examples of the system's ability to plan and execute simple organic syntheses.

Autonomous chemical research with large language models
https://www.nature.com/articles/s41586-023-06792-0

Perspective: It's early days for LLMs, and I think it's a stretch to say that GPT-4 or any other LLM understands Chemistry. At this point, LLMs seem to have two general use cases. First, summarization and information retrieval. LLMs can parse vast Qq0.000008871 0 595.32 841.9pes48(m)3(in)-5(u)x(ta)-,95.32 841.92 reW* nBT/F3 9.96 Tf1 0

of a larger dataset. In 2022, several papers, including one of ours, showed the promise of active learning approaches in drug discovery.

In 2023, we saw several papers published describing active learning applications.

A different type of AL approach, called PyRMD2Dock, was reported in a paper by Roggia and colleagues from the University of Campania Luigi Vanvitelli. Their approach is similar to other active learning methods in the initial stage. A set of 1 million molecules is sampled from an ultra-large database, and this subset is docked using AutoDock Vina. A docking score threshold is then used to classify molecules as active or inactive, and this data is used to train PyRMD, a ligand-based virtual screening tool developed by the authors. The trained PyRMD model screens the ultra-large database and selects a subset of molecules to be docked. Final* nckedbasedlarge14(ule

dimensional structure of molecules.

Chemical complexity challenge: Is multi

Explainable AI

A Perspective on Explanations of Molecular Prediction Models
https://pubs.acs.org/doi/10.1021/acs.jctc.2c01235

Explainable AI for Bioinformatics: Methods, Tools and Applications
https://academic.oup.com/bib/article-abstract/24/5/bbad236/7227172?redirectedFrom=fulltext

From Black Boxes to Actionable Insights: A Perspective on Explainable Artificial Intelligence for Scientific Discovery https://pubs.acs.org/doi/10.1021/acs.jcim.3c01642

Generative Models

Integrating structure-based approaches in generative molecular design
https://www.sciencedirect.com/science/article/pii/S0959440X23000337

Deep generative models for 3D molecular structure
https://www.sciencedirect.com/science/article/pii/S0959440X23000404

Generative Models as an Emerging Paradigm in the Chemical Sciences
https://pubs.acs.org/doi/full/10.1021/jacs.2c13467

Open Source

Open-Source Machine Learning in Computational Chemistry
https://pubs.acs.org/doi/10.1021/acs.jcim.3c00643

Multiple Instance Learning

Chemical complexity challenge: Is multi-instance machine learning a solution
https://wires.onlinelibrary.wiley.com/doi/10.1002/wcms.1698

Multi-objective Optimization

Computer-aided multi-objective optimization in small molecule discovery
https://www.cell.com/patterns/fulltext/S2666-3899(23)00001-6

Computational Approaches to Targeted Protein Degradation

Targeted Protein Degradation: Advances, Challenges, and Prospects for Computational Methods
https://pubs.acs.org/doi/10.1021/acs.jcim.3c00603

Data Related Reviews

Data Sharing in Chemistry: Lessons Learned and a Case for Mandating Structured Reaction Data
https://pubs.acs.org/doi/10.1021/acs.jcim.3c00607

- -

Other Chemical Information News

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

5 Tips Every New Information Professional should know

Whether recently graduated or transitioning into the role from a different career, information professionals (IPs) new to th-eir position face unique challenges as they work to help their organisation achieve its goals.
https://www.copyright.com/blog/5-tips-every-new-information-professional-should-know/

Can AI Plan Organic Syntheses alone?

AI can be very useful for different tasks related to chemistry and adjacent sciences. For example, there are successful AI-based approaches to predicting the structure of proteins and AI methods for the optimisation of catalysts. However, retrosynthetic planning in organic chemistry is still challenging. Methods in which AI purely "learns" from the data of example reactions have often been limited to simple target products and can fail for complex molecules. This has sometimes been attributed to insufficient amounts of reaction data.

https://www.chemistryviews.org/can-ai-plan-organic-syntheses-alone/

Source: Chemistry Views

Altmetric 500 data offers wider insight into research's most influential articles

Digital Science has announced an exciting new tranche of data that throws light on how and why research cuts through to society at large – in the shape of the Altmetric 500.

https://www.stm-publishing.com/altmetric-500-data-offers-wider-insight-into-researchs-most-influential-articles/

Source: STM Publishing News

Publicly backed Bioscience spin-outs make big impact on Economy

Hundreds of UK spin-outs, established following taxpayer-funded bioscience research, have contributed £5.2 billion to the economy and created thousands of jobs.

https://www.ukri.org/news/publicly-backed-bioscience-spin-outs-make-big-impact-on-economy/

Source: UKRI

Enhanced AI tool TeXGPT powers up academic writing

Digital Science announces an update to Writefull for Overleaf, which uses AI to help academic authors write better, faster and with more confidence in LaTeX.

https://www.digital-science.com/news/enhanced-ai-tool-texgpt-powers-up-academic-writing/

Source: Digital Science

Is AI's copyright world flat, or will AI flatten the copyright world?

Is offshoring the training of AI a credible and efficient response to minimize copyright compliance risks or is offshoring merely a theoretical argument designed to both influence lawmakers and for government relations purposes?

https://www.copyright.com/blog/is-ais-copyright-world-flat-or-will-ai-flatten-the-copyright-world/

Source: CCC

Clarivate in the age of AI: Innovation rooted in Academia

Clarivate has provided an update on its generative AI strategy and its expanded AI-powered product portfolio. In a virtual presentation, Bar Veinstein, President, Academia & Government, details the company's generative AI strategy for academia, and demonstrates the new AI capabilities developed to drive research excellence and student success. As part of this strategy, Clarivate is both enhancing existing solutions and introducing new solutions built on its specialised Academic AI platform services.

https://www.stm-publishing.com/clarivate-in-the-age-of-ai-innovation-rooted-in-academia/

Source: STM Publishing News

Beyond standard search: Getting the targeted data your organisation needs

To become data-driven effectively, organisations naturally seek more and better data.

https://www.copyright.com/blog/beyond-standard-search-getting-the-targeted-data-your-organization-needs/

Source: CCC

Beyond standard search: solving problems with new datasets from multiple sources

Of the many applications of deep search benefitting organisations, a standout is the ability for an organisation

What the EU's tough AI law means for Research and ChatGPT

The EU AI Act is the world's first major legislation on artificial intelligence and strictly regulates general-purpose models.

https://www.nature.com/articles/d41586-024-00497-8

Source: Nature

Opaque AI tools threaten trustworthiness of scientific findings, says Royal Society report

A new report by the Royal Society raises concerns about the potential downsides of using "opaque" artificial intelligence (AI) tools in scientific research. The report, titled "Science in the Age of AI," explores the opportunities and challenges presented by AI, particularly machine learning and large language models, as transformative tools for 21st-century research.

https://www.knowledgespeak.com/news/opaque-ai-tools-threaten-trustworthiness-of-scientific-findings-says-royal-society-report/

Source: Knowledgespeak

Google's claim of quantum supremacy has been completely smashed

Google's Sycamore quantum computer was the first to demonstrate quantum supremacy – solving calculations that would be unfeasible on a classical computer – but now ordinary machines have pulled ahead again.

https://www.newscientist.com/article/2437886-googles-claim-of-quantum-0 0 1 4BT6(c)-13(ieSwow 841.92 reWq 595.e0 dW

MIT Press's MIT Open Publishing Services (MITops) releases preprints of Generative AI Impact Papers A ground-