ROYAL SOCIETY OF CHEMISTRY | CHEMICAL INFORMATION AND COMPUTER APPLICATIONS GROUP

Chemical Structure Representation: What Would Dalton Do Now?

Abstracts

IUPAC: the role of international organisations in nomenclature and terminology in the internet age

Professor Jeremy Frey, *Department of Chemistry, University of Southampton*

Setting standards is at the heart of international exchange of information and trade and IUPAC's work in establishing Chemical Nomenclature and following this with work on terminology has been critical for international collaboration both within chemistry and to other areas, and has often been driven by developments in chemical structures (novel structures, new characterisation methods). The rise of the World Wide Web is radically changing the way in which chemical concepts are conveyed and the discipline specific international organisation are struggling to maintain their role in setting standards, but it is essential that they find their place in the Web World to ensure that the knowledge that is currently communicated on paper can be readily and accurately communicated in the digital world.

Indescribable structure: finding words for the future

Professor Jonathan Goodman, *Department of Chemistry, University of Cambridge*

Our knowledge of chemistry is limited by the vocabulary we have to describe molecular phenomena. N-heterocyclic carbenes somehow look wrong until you get used to them. Benzyne is not the most plausible of structures. The discovery of enediyne natural products made the Bergman cyclisation become much more interesting.

What are the limits of our current descriptions? Wavefunctions describe molecules well, but do not always make them comprehensible. Structural drawings can make molecules intelligible, but do not always correspond to observations. We could trust our calculations and not worry about their interpretation, but this approach seems to lack ambition. We can try to understand wavefunctions. We can try to make structures more expressive, and adjust our interpretation of them so that they tell us more about structure and reactivity.

Are we close to the limit of what we can describe? A failure to predict, to rationalise or to understand a structure or a reaction might arise because we are using our current tools imperfectly or because our current tools are inadequate. Both areas need to be explored in order to describe the molecules and reactions currently beyond our reach.

Biology: bigger models, bigger confusion

Dr David C Briggs, *Department of Life Sciences, Imperial College*

Despite having a limited repertoire of chemical constituents, biological systems achieve remarkable complexity.

This complexity is primarily due to the sheer size of the individual polypeptide and/or nucleic acids components - a typical enzyme will have a molecular mass of over 30 kDa, and contain several hundred amino acids. The prokaryotic ribosome, a major therapeutic target for novel antibiotics, is around 2.3 MDa in size and is composed of three separate ribosomal RNA subunits, and over 50 separate protein subunits. Such complexity is compounded by the poorer observation-to-parameter ratios (and therefore increased uncertainty) inherent in biological structure determination.

This presentation will give an overview of the different sorts of data obtained by structural biologists, and some of the challenges encountered when representing such complicated molecular machines.

Chemical structure representation challenges encountered when curating the CSD

Dr Matt Lightfoot, *Editor in Chief, Cambridge Structural Database, CDDC*

The Cambridge Structural Database (CSD) provides a comprehensive record of all published organic and metal-organic small-molecule crystal structures. The database has been in operation for over 50 years and contains almost 900,000 entries. Every day an additional 200 structures are added to the database and all these structures are processed both computationally and by expert structural chemists prior to entering the database. A key component of this processing is the reliable association of the chemical identity of the structure studied with the experimental data. This important step helps ensure that data is widely discoverable and readily reusable.

This presentation will focus on the particular challenges that we face when assigning the chemical identity to the structures, look at why consistent and reliable representations of chemical structures are important and look at how improvements to experimental and publishing workflows can help with this challenge in the future.

Describing chemical substances and chemical structures: we have a long way to go

Dr Evan Bolton, *Chemistry Program Head, US National Center for Biotechnology Information (NCBI)*

The description of chemical substances and chemical structures are intertwined. To the casual consumer of chemical information, a chemical substance and a chemical structure are often confused as being the same thing. Why not, as a chemical substance is supposed to be a pure substance and, in many use cases, the chemical structure is referred to as the representation of the chemical substance. There are a number of issues that result from this generalized association. A primary problem is that chemical structures can be a poor conceptualization of 'chemical substances', which can also exist in different phases of matter or interchange between them.

Making matters worse is that a 'chemical substance' can have multiple definitions. IUPAC defines a 'chemical substance' as a matter of constant composition best characterized by the entities it is composed of. There are also legal definitions of 'chemical substance' that typically include mixtures with a defined composition or manufacturing process, enabling for their ready identification even in the absence of known chemical structure information. Real-world use of the term 'chemical substance' often includes anything you can name, identify, isolate, or make.

Chemical structures have their own issues. Different tautomeric forms of a chemical structure may be favored depending on environment. There are no universally adopted standards for chemical structure representation, with each organization or scientist adopting their own approaches independent of the other, giving rise to combinatoric ways to represent the same chemical entity. In addition, the very means to communicate chemical structure information is plagued by 'industry standard' approaches without a unifying standards body, leaving each chemical information vendor or software package to adopt their own (potentially conflicting) extensions or approaches to various file format flavors.

With Prof John Dalton in mind, this talk will explore opportunities to improve chemical information representation, especially when it comes to chemical structures and chemical substances. Using PubChem as a case study, this talk will give an overview of standard pitfalls and misconceptions when handling chemical substance information and best practices to avoid data corruption when communicating this content. In addition, community efforts to improve the plight of chemical structure and chemical substance information everywhere will be provided.

Using data science techniques to put molecules in context

Dr Aileen Day, *Data Scientist, Royal Society of Chemistry*

Most chemical databases, including ChemSpider, rely on finding molecules based on text searches of names, numerical searches on properties and structural searches based on substructure or similarity to another molecule. These can sometimes yield narrow results that are unsatisfactory because of the widespread nature of research questions that are attempting to be answered when someone wants a springboard from an original molecule to another of potential interest. Sometimes the biggest research breakthroughs are due not to non-linear progression, but rather jumps to new pastures.

As a chemical sciences publisher as well as hosts of ChemSpider, the Royal Society of Chemistry has a good amount of data to put molecules in context beyond the structured confines of a chemical database.

As such, we have put together a tool which, given a starting molecule, shows other clusters of molecules related to it, not only by standard cheminformatics fingerprint methods (based on Morgan and Topology similarity from RDKit) but also by user behaviour (scientists' behaviour by mining millions of user interactions from ChemSpider) and text mining (chemistry as mined from the RSC journal corpus and a selection of fingerprints).

Here we present a description of the methodology and comparison of results and we would welcome conference attendees to try out the molecule recommender themselves to feedback their opinions about the methods which yield the potentially most useful results.

We also describe improved chemical Named Entity Recognition based on three different "deep learning" which yielded some of the most successful results in the CEMP task of BioCreative V.5.

Chemical structure representation of inorganic salts and mixtures of gases: A newer system of chemical philosophy

Dr Roger Sayle, *CEO and Founder, NextMove Software Ltd*

In 1808, 209 years ago, John Dalton published his influential work "A New System of Chemical Philosophy". In this work, Dalton reviews the available experimental measurements of the time, and fits them to a proposed "atomic theory", that postulates compounds are made up of atoms of different elements, where each atom of the same element has the same mass, but atoms of different elements have different masses, and that atoms of different elements combine in simple whole number ratios. These insights clearly form the basis of modern molecular chemistry and, with the dawn of the computer age, of modern cheminformatics.

It is interesting to observe that many of the chemical systems that Dalton studied and worked on still continue to present representational challenges (on computer systems) to this day. An entire section of his book is entitled "metallic oxides", yet modern databases such as ChemSpider, PubChem and ChEMBL fail to assign unique identifiers to alternate connection table representations of these compounds. It is a testimonial to the representational complexity of Dalton's case studies that in 2017, the IUPAC InChI project has active working groups on topics of both inorganics and mixtures.

In this presentation, I describe how two of the approaches employed by Dalton, the chemical nomenclature he used, and the molecular formulae implied by his atomic theory, can be used to tackle several of the representational issues still facing cheminformatics today. This holds not only for "sulphuret of potash" and pewter alloys, but also to modern challenges such as atropisomerism, non-standard peptides, glycans, siRNAs and chimeric-antibody drug conjugates.

Extracting medicinal chemistry knowledge by a secured Matched Molecular Pair Analysis platform: standardization of SMIRKS enables knowledge exchange

Dr Al Dossetter, *Managing Director, MedChemica Ltd*

The challenge of extracting pre-competitive knowledge from organizations wishing to gain the benefits of sharing but without compromising their intellectual property is substantial. Matched Molecular Pair Analysis enables this by only extracting fragments of molecules and differences in real measured properties. By standardizing the algorithm, assay units and calibration and running the system within contributing organizations, the output can be merged with a similar analysis from public data to produce a Grand Rule Database (GRD) of medicinal chemistry knowledge. This is now a reality. MedChemica sits at the center of a consortium of large pharmaceutical companies and have produced current three versions of the GRD. We discuss the process of data curation and standarisation (units, species, canonical SMIRKS), algorithm methods and performance on “Big Data” scale. Statistical analysis methods to define medicinal chemistry rules and merging with public data and handling this type of analysis in the ‘Cloud’ using modern secure systems will be shared.

InChis are part of the solution

Mr Richard Kidd, *Publisher, Royal Society of Chemistry & Treasurer, InChi Trust*

The InChI standard identifier was launched over 10 years ago, and its latest update was released earlier this year. IUPAC-sponsored groups are working to extend the standard in several ways, but there are interesting questions arising – how far is it reasonable to extend the standard? How should the standard be maintained end extended in the long term? What other use cases can be built on a structure representation, and should they?

HELM: Setting the standard for biomolecular information exchange

Claire Bellamy, *HELM Project & CSCS Expert Community Manager, Pistoia Alliance*

With the increased focus on biotherapeutics R&D in the biopharmaceutical industry, there has been an increasing need for a robust capability for the electronic representation of an ever-growing variety of complex biologic entities. HELM was implemented to address this need by enabling the representation of many kinds of complex macromolecules such as nucleotides, proteins, antibodies and antibody drug conjugates, including those containing non-natural elements. The latest implementation of the standard, HELM2, further enables scientists to capture structural ambiguity such as unknown conjugation sites or monomers. Supporting the HELM standard is a rich ecosystem of open-source software, public information and an active community, all to be discussed in this presentation.

Chemical Structure Representation: What Would Dalton Do Now?

Speaker Biographies

Professor Jeremy Frey, *Department of Chemistry, University of Southampton*

Jeremy Frey obtained his DPhil on experimental and theoretical aspects of van der Waals complexes at Oxford University, followed by a NATO fellowship at the Lawrence Berkeley Laboratory. In 1984 he took up a lectureship at the University of Southampton, where he is now Professor of Physical Chemistry. His experimental research probes molecular organisation in environments from single molecules to liquid interfaces using laser spectroscopy from the IR to soft X-rays. He investigates how e-Science infrastructure can support scientific research with an emphasis on the way digital infrastructure can support the intelligent access to and analysis of scientific data.

Professor Jonathan Goodman, *Department of Chemistry, University of Cambridge*

Jonathan Goodman studied boron-mediated aldol reactions during his PhD with Professor Ian Paterson FRS at the University of Cambridge. He then did a post-doc with Professor Clark Still at Columbia University, before returning to the chemistry department at Cambridge, where his now Professor of Chemistry and Deputy Director of the Centre for Molecular Informatics. In 2013, he won the RSC’s Bader Award. His research group does experimental chemistry, computational chemistry and data analysis.

Dr David C Briggs, *Department of Life Sciences, Imperial College*

David Briggs is a protein crystallographer who has studied protein-carbohydrate interactions in the mammalian extracellular matrix for the past 10 years, studying the underlying molecular events that govern such diseases as osteoarthritis and muscular dystrophy. More recently, he has turned his attention to the biochemical pathways that control the post-translational modification of proteins with complex carbohydrates. Prior to working at Imperial, David worked at the University of Manchester and Cancer Research UK. He obtained his PhD in Protein Crystallography from Birkbeck College, University of London.

Dr Matt Lightfoot, *Editor in Chief, Cambridge Structural Database, CDDC*

Matthew has a BSc in chemistry from Birmingham University and a PhD in alkali metal coordination chemistry from UMIST. He joined the CCDC after completing his PhD in 2000 as a Scientific Editor in the Database Group where he was responsible for carrying out the curation and validation of crystal structures into the CSD.

During his time at the CCDC, Matthew has been involved in many different aspects of database processing and has developed a deep understanding of how structures are processed into the CSD. More recently, Matthew has been working with the Internal Database Development Team on developing CSD-Xpedite - a completely new platform for the internal processing of data. This went live during 2013 and has provided the CCDC with a much more efficient and flexible system with which to process structures into the CSD. In 2013 Matthew was given the role of Editor in Chief. He now leads the Editorial Team and is responsible for how all data are processed into the CSD.

Dr Evan Bolton, *Chemistry Program Head, National Center for Biotechnology Information (NCBI)*

Dr. Evan Bolton is the Head of the Chemistry Program at the U.S. National Center for Biotechnology Information (NCBI), part of the U.S. National Library of Medicine (NLM), an institute of the U.S. National Institutes of Health (NIH), where he has helped to lead efforts with the PubChem project (https://pubchem.ncbi.nlm.nih.gov) since 2004. Dr. Bolton earned a B.S. in chemistry from Rider College in Lawrenceville, New Jersey in 1991 and a Ph.D. in physical chemistry from the University of Georgia in 1995 (graduate advisor was Prof. Henry F. Schaefer, III). Dr. Bolton received the FDA Commissioner's Special Citation award from the U.S. Food and Drug Administration (FDA) and is co-recipient of the 2016 Herman Skolnik Award from the American Chemical Society (ACS) Division of Chemical Information (CINF).

Dr Aileen Day, *Data Scientist, Royal Society of Chemistry*

Aileen Day (née Gray) has a background in materials science and a PhD in computer modelling at Cambridge University and University College London respectively. She has always worked at the interface between computer programming and science - working in materials and chemistry data management and analysis. Her work aims to use computer technologies to facilitate scientific research, and has covered areas such as ChemSpider, research data and ELNs.

Dr Roger Sayle, ***CEO and Founder, NextMove Software Ltd***

Roger Sayle obtained his Ph.D. in Computer Science from the University of Edinburgh. Roger has worked at GSK in Stevenage and with Daylight Chemical Information System in the US. Between 2001 and 2010, Roger was vice president of software engineering at OpenEye Scientific Software. In 2010, he left OpenEye to found a new cheminformatics software company, NextMove Software, located on the Cambridge Science Park in the UK. One of his early achievements was the molecular graphics program RasMol. For his work on RasMol, he has received the Biochemical Society's Heatley Medal and a Blue Obelisk award.

Dr Al Dossetter, ***Managing Director, MedChemica Ltd***

Al Dossetter joined AstraZeneca (AZ) in 1999 after a PhD from Nottingham University and post-doctoral research at Harvard University. His 13 years of experience in medicinal chemistry has been spread across oncology (hormonal and kinase inhibitors), inflammation (OA and RA, enzyme inhibitors and GPCR targets) and diabetes (obesity, GPCR and enzyme inhibitors). In 2012 Al started MedChemica Limited with Dr Ed Griffen and Dr Andrew Leach centred around the technology of Matched Molecular Pair Analysis (MMPA) as a method of accelerating medicinal chemistry. MedChemica now licenses a suite of software tools for companies to extract and share knowledge from their own data and combine with public data. The software and methodologies have been used by many pharmaceutical companies, universities and biotech to accelerate drug discovery programmes.

Mr Richard Kidd, ***Publisher, Royal Society of Chemistry and Treasurer, InChi Trust***

Richard is responsible for the RSC's initiatives in data publishing and is also Treasurer of the InChI Trust. https://uk.linkedin.com/in/kiddrichard

Claire Bellamy, ***HELM Project & CSCS Expert Community Manager, Pistoia Alliance***

Claire Bellamy manages the HELM project for the Pistoia Alliance. Prior to her current role, Claire was a senior business analyst at AstraZeneca, leading a global team of business analysts implementing R&D IT systems that were rolled out to thousands of users. Claire has an MBA and a BSc in Chemistry from Nottingham University.