How do we shape the future? What are the critical issues to be addressed by IUPAC?

by Jeremy G. Frey

IUPAC first 100 years and the World Chemistry Leadership Meeting

Following the 100 Year Anniversary of IUPAC celebrated in Paris in 2019, Artificial Intelligence was made the focus of the following global 2021 World Chemistry Leadership Meeting (WCLM).[1] Once it was clear the prevailing COVID-19 conditions would mean that the 2021 WCLM would be virtual, the planning committee decided to take advantage of an online setting to have a truly global meeting and to have the event follow the sun round the world using three time zones and "visiting" the next two IUPAC Congress sites, *i.e.* The Netherlands 2023 and Malaysia 2025.

As the form of the meeting took shape, I was asked to give the closing talk bringing together the ideas raised by the speakers in their recorded talks, the highlights of the discussions in Malaysia and The Netherlands, the panel led from Montréal, and perhaps even suggest directions for IUPAC to take the lead in the transformation of chemistry as a science in the digital age. Which is why the title of my closing keynote was, "*How do we shape the future?*".

Pure and Applied Chemistry

It is very important to remember that IUPAC stands for and represents both *Pure* and *Applied* Chemistry; perhaps this is not a distinction that should be made in an age where we all need to justify the relevance of our work and with such critical global issues around sustainability (UN Sustainability Goals [1]) abounding, in which Chemistry and Chemists can clearly play such an important role.

It is also an interesting or significant fact that many of the fundamental developments in the field of digital chemistry, the use of Artificial Intelligence, and Machine Learning in chemistry are being driven by industry. In some cases, industry could certainly be considered to be in the forefront (Deep Mind AlphaFold2) of these areas. So, as we move forward into IUPAC's second century, into the digital age, what should we be doing to address these challenges and maximise the effectiveness of new technologies in Chemistry?

Digital *i*-UPAC

As is probably clear from what I have said above and for those who may have read my earlier article on the Digital IUPAC (*i*-UPAC) [2], my vision of the future comes from the intersection of the idea that what we do is to use scientific approaches to chemistry, and the way we increasingly do this is using a digital technoscape.

Chemists work in Chemical Space! We may start our navigation round this space from different places and take different paths depending on our specialities and aims, from molecules to drugs, compounds to formulations, materials to devices. We are increasingly aiming to be more specific, more precise, whether for personalised medicine or precision agriculture, seeking energy efficiency or enhancing the circular economy. The chemical space we inhabit is vast, and chemistry is also about change so we must not forget the time dimension; we really live in a world of Chemical Space-Time. Chemistry is not just about the properties of molecules, it is also about their transformations and the rates at which these transformations take place—time is important.

How have we coped with this vast, largely unexplored space? Do we need to explore it all? In many cases chemists have been inspired by nature. Nature has had somewhat longer to explore regions of chemical space, by trial and error. But nature is also (probably) restricted by paths traversed long ago. I think most of us also believe that our own imagination and inherent creativity have enabled us to reach out into different regions of chemical space but—do we create or explore or simply navigate paths that already exist? [3]

The Tyranny of Molecules

We need to design and synthesize new molecules and new structures which have new properties. These

[1] Note: The program of WCLM 2021 is available online. A detailed account of WCLM2021 will follow in the next issue of *Chem Int.* at https://iupac.org/event/wclm2021/.

may be intended as drugs to hit specific targets, or assembled to give new materials desired properties to increase the efficiency of devices, *etc.* One of the hopes of computational chemistry has been that with the ability to predict theoretically and computationally the properties of a molecule, (and materials, which is even more difficult), in advance of making it, then we can be more efficient in directing the huge synthetic effort to molecules that are more likely to be successful candidates. However, even with the massive increase in computing power and better algorithms, we cannot yet make these predictions on a sufficiently accurate basis at sufficient scale.

With ideas that trace back to at least the 19th century, of atoms in molecules, functional group additivity, we realise that we may not need to devote all this computational power to all molecules, working with some, and using the experimental data available, and extending the reach of the complex calculation, using types of statistical correlations to unearth and make use of chemical patterns. The key here is that we do need some experimental data, but how much?

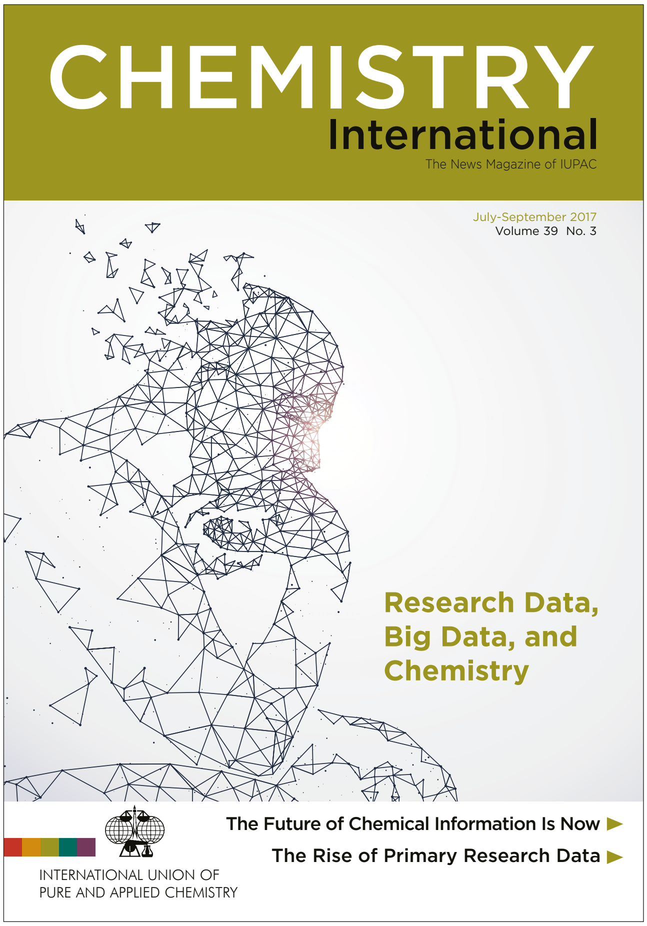
How much data is needed to achieve these aims? Sometimes the chemist already has a very good idea about what aspects of the molecular structure influence the desired properties. These features may be relatively easy to correlate with activity of function. For example, the activity of a drug may be directly related to the solubility in water and lipids. This means the now well-developed and quick calculations of LogP (the partition between octanol and water) can be used, and correlates well with the measured activity. This is where simple Quantitative Structure Activity Relationships (QSAR) models win: simple, easy to understand, and fast.

So, what is the difficulty? Most problems of interest are much more complicated, and no simple obvious descriptors are known in advance. A well-known dictum about unknown unknowns can be used to distinguish between QSAR and Machine Learning models. We have known descriptors, properties that by long experience we know affect the activity of a molecule from which we can build useful and understandable QSAR models. Next in the sequence, we have known unknowns, properties we know could affect activity, but we don't have a clear idea of what types of descriptors we could use. Finally, we have the unknown unknowns where the underlying patterns are not obvious. Perhaps the machine can help, where we use unsupervised learning and reinforcement learning to find models without any specific idea about what the patterns might be.

Machine Learning has shown great promise by producing models that give excellent predictions and this means we can hope to use Machine Learning to make use of underlying chemical patterns. Indeed, generative models can make use of these underlying patterns to 'invent' suitable molecules so that the models become more than just filters of the likely useful one from a pre-existing list. Models for synthetic pathways can then be incorporated so that suggested molecules are likely possible to make.

However, to train the (current) Machine Learning models require lots of data. To train image recognition models requires thousands of images, to train language models, millions of words, (but a human child learns these with much less input) and AlphaGo [4] has played more games than any human has ever played (I am slightly guessing here but with the speed AlphaGo can play itself I am pretty sure it is a good guess).

To get good predictions about which molecules to make from our ML model we are likely to need data on lots of molecules. So here is the tyranny—to be selective about the molecules we make, and make more of the "right ones," we need a good model. To have a good model we need experimental data and to obtain the necessary experimental data, we need to make more molecules. We need to make more molecules, different molecules, to be able in the end, to make fewer molecules.



CHEMISTRY International

The News Magazine of IUPAC

July-September 2017
Volume 39 No. 3

Research Data, Big Data, and Chemistry

The Future of Chemical Information Is Now ▶

The Rise of Primary Research Data ▶

INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY

The emergence and development of digital information technologies have inspired a new look at how research outputs are managed and disseminated. In 2017, CI released a special issue on Big Data. This year, the special issue on Cheminformatics is about to be published in Pure and Applied Chemistry.

The Way Forward

Nevertheless, the situation does not look to be impossible. Chemical knowledge, physical principles, prior knowledge can help here. But we do need to get accurate data and realistically, this means we need to get even better at making small amounts of lots of different molecules and have highly effective ways to characterise them and measure their properties on perhaps nanoscopic amounts in an automated and in a highly reproducible manner. It is therefore important that both research in synthetic and characterisation methods continue in step with the developments of Machine Learning.

From the perspective of a chemist who wants to understand the principles, we can ask what can we learn from ML models? How do we extract understanding from the complexities of a neural network? The rise of explainable AI is driven by the need to be able to understand the reason for the prediction, driven in part by the needs to ensure the model can be trusted and biases (as there will almost always be bias) understood.

Many of the successful graph-driven ML models lead to an internal representation of molecules in a chemical space, but will we be able to codify these representations and give them names as we do for the observed molecular structure? Perhaps we do not need to, but if we are to reliably use these systems then we must be able to accurately describe the way the models operate to convey them to others; defining the standards to ensure we can do this is another critically important role for IUPAC.

Should students give up with traditional chemistry? Will all chemical problems be solved by computer as a data-driven ML exercise? To quote Derek Lowe in his article for *Chemistry World*:

> *"And do you know who will find those things out? Not our AI and ML systems, although I'm sure they'll help whenever possible. No, it's going to be us. Just like it always has been".* [5]

But the culture of chemistry is changing and as far back as 1950 Robert Heinlein suggested that:

> *"When chemistry becomes a discipline, mathematical chemists will design new materials, predict their properties, and tell engineers how to make them—without ever entering a laboratory".* [6]

We've still got some way to go on that one! Chemistry may be changing but understanding chemistry is still a worthwhile objective and a challenge. We simply have new tools, and the necessary skills to use these tools will require changes in the way we train the next generation of chemists.

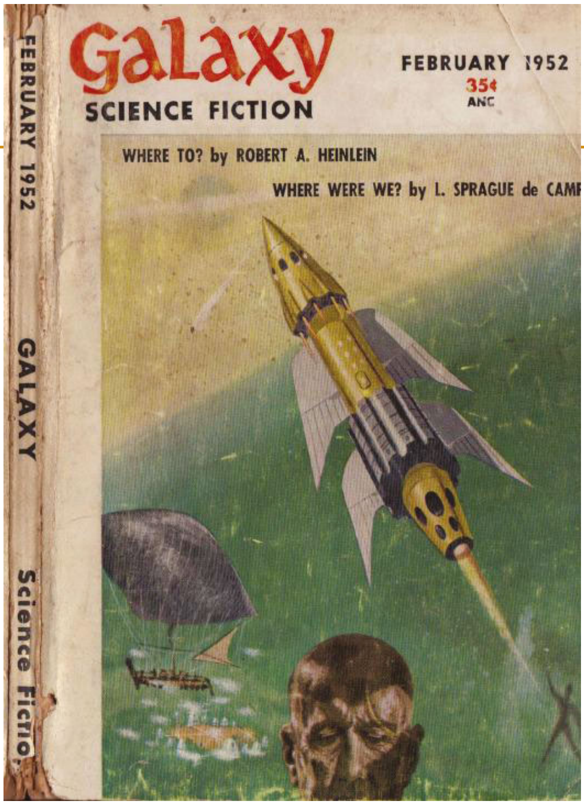
This is an exciting future but what does IUPAC need to do?

As a global community we must have the standards to describe the data, the related provenance and uncertainty of the data and the systems to enable computers to 'understand' the relationships between the data. That is, we need a comprehensive digital ontology to describe the data as only then can we be more certain that automatically generated models will contain some elements of chemical sense.

IUPAC needs to take and is taking a central role in this effort, and it is and will be an enormous effort. It is not just an academic exercise, as I am sure industrial colleagues who need to integrate data will recognise, but without sustained effort and collaboration across the whole of the discipline we will not succeed, and others will try and do a poor job. However, the pace of change is such that the old ways of working find it hard to rise to these challenges. For more resources (human and funding) are needed and raising them, raising the profile, must be one of the most important challenges for the whole organisation. If we can convince the wider society of the need and the rewards then we have a chance to make a step change in the way we respond to the digital, machine learning age.

One of the most significant global challenges for IUPAC is sustainability. This often requires consideration of multiple disciplines and their complex interactions. IUPAC's role in ensuring that chemistry can be clearly and reliably conveyed is extremely important in facilitating discussions that involve chemistry. With the involvement of AI/ML, the concepts used by chemists need to be even more precisely and unambiguously articulated and explained, and must be suitable for computational consumption.

We are still in a liminal period of the transition to digital chemistry and the rise of AI/ML. As these new techniques become imbedded to a greater and greater extent in our discipline, the key worry is that over-reliance on AI may put us in intellectual debt and less able to address the challenges ahead. But then, similar things have often been said about new technologies, however, AI may be rather different. Nevertheless, as I hope I have made clear, the successful, trustworthy and useful adoption of AI/ML by the chemistry community needs the same careful considerations of nomenclature, terminology, units, symbols, and international standards that have been the core concerns of IUPAC in its first 100 years, the only difference (and it's a big one) is that now we have a computer audience as well as a human one.



Robert Heinlein, featured in this issue of February, 1952 Galaxy Science Fiction *magazine, was pondering the meaning of artificial intelligence. ref.6, retrieved from <https://archive.org/details/galaxymagazine-1952-02/page/n13/mode/2up>*

References

1. The 17 Goals, United Nations: https://sdgs.un.org/goals
2. Frey, J.G. "Digital IUPAC: A Vision and a Necessity for the 21st Century", *Chem. Int*. 2014, vol. 36, no. 1, pp. 14-16; https://doi.org/10.1515/ci.2014.36.1.14
3. Jansen, M. and J. Christian Schön ""Design" in Chemical Synthesis – An Illusion?", *Angewandte Chemie*, International Edition 2006, 45(21), 3406–3412; https://doi.org/10.1002/anie.200504510
4. AlphaGo, DeepMind making history: https://deepmind.com/research/case-studies/alphago-the-story-so-far
5. Lowe, D. "The law of conservation of data, Derek Lowe", *Chemistry World*, 11 Jan 2022 https://www.chemistryworld.com/opinion/the-law-of-conservation-of-data/4014927.article
6. Heinlein, R.A. "Where to?" *Galaxy Science Fiction*, February 1952, pp. 13-23 <https://archive.org/details/galaxymagazine-1952-02/page/n13/mode/2up>

Jeremy G. Frey <j.g.frey@soton.ac.uk> is Professor of Physical (and Digital) Chemistry at the University of Southampton, UK. In IUPAC, he is a member of the Physical and Biophysical Chemistry Division, of the Commission on Physicochemical Symbols, Terminology, and Units (responsible for the IUPAC Green Book), the Interdivisional Committee on Terminology, Nomenclature and Symbols, the Committee on Publications and Cheminformatics Data Standards; and the Joint Subcommittee on the IUPAC Gold Book; ORCID.org/0000-0003-0842-4302; @Profechem