**Dial-a-Molecule**

**Predicting Reaction Outcomes**

**Cambridge Science Park**

**26th March 2018**

**Summary of Outcomes**

* What is the main problem?
  + DFT limitations (size/precision)?
    - *No problem?! Attendees disagreed. The acceptable level of theory is increasing fast.*
    - *The strategy of increasing the level of theory until the desired outcome is obtained should be avoided.*
    - *However, should be pragmatic – the use of a relatively low level of theory can be justified in specific cases by close correlation with experimental data*
    - *Even B3LYP should not be treated with contempt*
    - *If you are studying a reaction for which charges appear and disappear along the reaction coordinate, good luck! This will be a challenging study.*
    - *Must have good benchmarking for computational methods*
  + Data gathering (spectrometer to publication)?
    - *What data are important enough to keep? Everything?*
    - *ELNs will never be routinely used in academia. Attendees disagreed. What we mean by an ELN is likely to develop.*
  + Data availability (open access / closed data)?
    - *How can we gather reliable negative data? Past knowledge? Continuous observation of experimentalists (for safety reasons)*
    - *Companies need to guard their data and will never be completely open*
    - *We will never have enough data – but we may have enough diversity and quality for a reasonably complete description of reaction outcomes*
    - *New data will always change our ideas of chemistry. New reactions are being discovered faster and faster Peak chemistry is far in the future*
  + Indexing (RInChI)
    - *An effective process for labeling reactions in a consistent and canonical way could revolutionise the practicality of searching large, diverse and distributed datasets*
  + Machine Learning?
    - *Is perfect data enough? No! Need knowledge of environment and context too*
    - *Can we give up on DFT? Not for the moment, probably*
* What to do next?
  + - *Do the right experiment first time – even better PhD students! Need the right human-computer interface / training / culture*
* **Develop a global network for reaction prediction**

**Predicting Reaction Outcomes**

What is the state of the art of predicting reaction outcomes? Seven diverse speakers discussed different aspects of the problem at a meeting in Cambridge on Monday.

Peter Johnson (University of Leeds) spoke about the development of synthesis planning tools and how data-driven reaction prediction is a key part of the process. Kris Ermanis (University of Cambridge) then explained how DFT methods can be used to explain the outcomes of enantioselective catalytic oxy-Michael reactions, and how machine learning can make conformation analysis more tractable. Arthur Smith, from the Office of Scholarly Communication at the University of Cambridge, showed how open data and open research is changing the way chemical discoveries are being used, and how they have the potential to revolutionise research in the near future. Jonathan Goodman (University of Cambridge) discussed how specific insights into structural behaviour can be verified by calculation and used to explain and to predict reaction outcomes. In addition, knowledge from outside the immediate realm of study may be needed in order to best interpret chemical data and so predict future measurements, particularly if it is possible to unite diverse databases using a canonical index such as the RInChI.

Robert Phipps (University of Cambridge) gave an account of how an experimentalist discovers new reactions, with examples from his work on non-covalent interaction controlled catalysis and enantioselective Minisci-type additions. Elizabeth Krenske (University of Queensland) then described how a superficially simple reaction, a base-catalysed silylation, required a worldwide consortium to investigate the mechanism, and then the mechanisms, of this highly complex process, and illustrated the care which is needed to draw conclusions about detailed reaction mechanisms. Ola Engkvist, from AstraZeneca, reported on the possibilities of automating reaction development and the opportunities that this gives for rapid and consistent data acquisition. These data, combined with data from publishers and open sources of information, make it possible to predict reaction outcomes using machine learning. Jeremy Frey (University of Southampton) then highlighted how a new EPSRC network, AI3D, should be able to advance scientific discovery in chemistry by helping people and computers to work effectively together.

The panel discussion covered the reliability of DFT and other simulation methods. Ideally, these should be validated against experimental results. Availability of data is a major issue. We can never have enough, and there are issues both with gathering data and making it available. Effective indexing can make information retrieval more tractable. With plentiful, high-quality data, and appropriate contextual information, machine learning could soon change the way we think about reactions.

The Dial-a-Molecule conference on *Predicting Reaction Outcomes* was held at the Bradfield Centre on the Cambridge Science Park on Monday, March 26th. It was organised by Jonathan Goodman, Richard Whitby, Gill Smith and Jayshree Mistry. Sponsorship from RSC-CICAG and MestreLab is gratefully acknowledged. We intend to present continue discussion of this key area through virtual meetings – more details soon.

* DFT and simulation – what is missing?
* Data Collection, Broadcast, Analysis
* Next steps: how to improve reaction prediction
  + Wait for someone else to do it
  + Work on different aspects
  + Common project: data CBA?

What is a good synthesis?

* Cheap/sustainable/reproducible starting materials
* Low hazard; low waste products
* Familiar reactions – use chemists’ expertise
* No inseparable by-products
* High yields; high stereoselectivity
* Convenient processes
* Get product fast/cheaply/reproducibly
* Suitable for making analogues
* Calculations are accurate, but time-consuming
* Machine learning (*AI*) can help understand data
* Cannot have too much data
* Calculations and data together can:
  + Get more value from current data
  + Help decide which molecules to make next
  + Test the molecules we want to test,   
     not just the molecules we know we can make
* Maximise value from investment
* A company issue or a global issue?
* How can we share data most effectively?