Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery

AI for Materials Discovery Workshop
19/03/2019
University of Southampton

Dr Nicola Knight & Dr Colin Bird
University of Southampton

28/06/2019

AI3SD-Event-Series:Report-8

AI for Materials Discovery Workshop
AI3SD-Event-Series:Report-8
28/06/2019
DOI: 10.5258/SOTON/P0009
Published by University of Southampton

Network: Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery
This Network+ is EPSRC Funded under Grant No: EP/S000356/1

Principal Investigator: *Professor Jeremy Frey*
Co-Investigator: *Professor Mahesan Niranjan*
Network+ Coordinator: *Dr Samantha Kanza*

Contents

1 Event Details

Title	AI for Materials Discovery Workshop
Organisers	AI^3 Science Discovery Network+ (AI^{3}SD)
Dates	19/03/2019
Programme	Eventbrite Programme
No. Participants	65
Location	University of Southampton

2 Event Summary and Format

This workshop was a half-day event hosted at the University of Southampton, with the focus of the event being the use of AI in materials discovery and innovation. The event was structured with an introductory talk followed by 5 keynote sessions from experts in the materials field. These talks were separated by a coffee break giving participants time to network.

Hartley Suite
Please return your badges when you leave the workshop.
Please also return any unwanted programmes & wallets so we can reduce our waste.

Figure 1: The AI3SD organising team at the Materials workshop

3 Event Background

This workshop is one in a series of events hosted by the AI3SD Network designed to provide a platform for discussion, innovation and collaboration for using AI in scientific discovery. The events bring together researchers and experts to disseminate knowledge and allow new connections to flourish.

4 Talks

The session of talks was opened by Professor Jeremy Frey giving a short introduction to the Network+ and the workshop format. In his talk Jeremy outlined the AI3SD Network+, quipping that he would have preferred AI4SD. The aim of the network is to encourage cutting edge AI and science, albeit not the whole of science: the focus is on chemistry and materials. Proposals submitted to the first funding call are being evaluated and projects will be chosen soon. This workshop is the first probe into the discovery of new materials; further contributions to the future work in this area are welcomed.

4.1 Theoretical Studies of CO and CO_2 Hydrogenation to Methanol and Conversion of Methanol to Olefins – Professor Felix Studt

Professor Felix Studt gave our first talk, discussing the use of computational catalysis, outlining some of the processes and examples of the use of these models, with a particular focus on Olefin production. Computational catalysis can model reactions with and without catalysts, using predictive calculations to model from the desired functionality to the required electronic structures.

Felix illustrated the technique with calculations of adsorption energies on a range of catalysts used in syngas conversion for production of methanol. For CuZn or CuZnAl catalysis, the relative intrinsic activity is linked to stacking faults; the catalytic activity correlated well with the probability of such surface defects. With many of the catalysts, activity is very dependent on the CO/CO_2 ratio in the reaction. In CO_2 hydrogenation, CO_2 poisons a Cu/MgO catalyst, but ZnO addition gives a higher yield due to the decreased energy barrier for the hydrogenation reaction but increased the barrier for CO conversion.

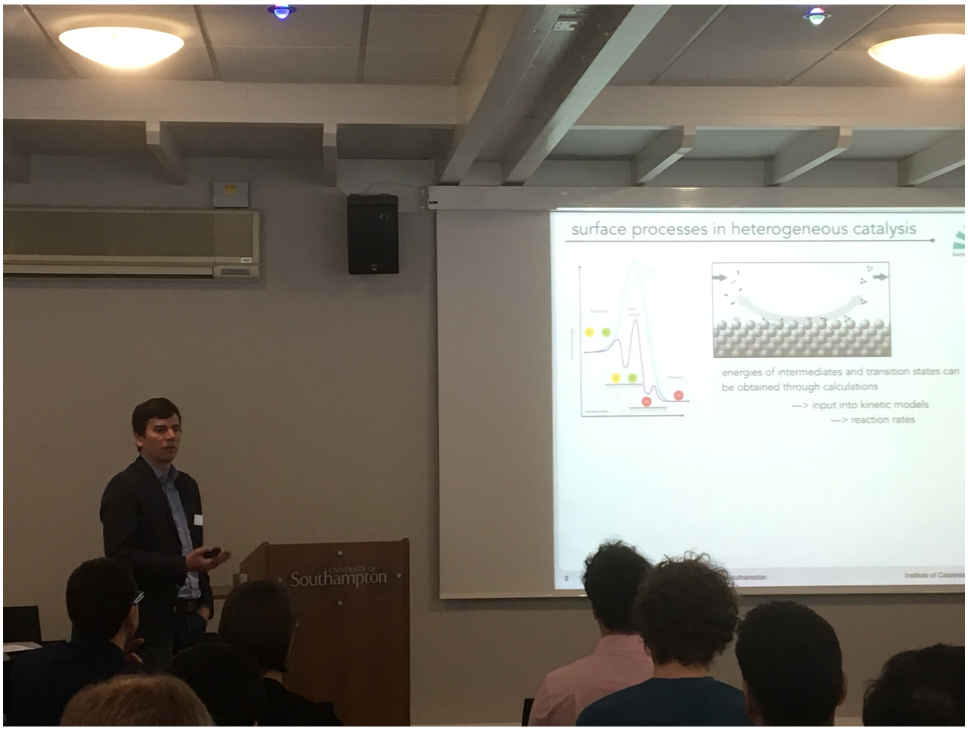

surface processes in heterogeneous catalysis
energies of intermediates and transition states can be obtained through calculations
—> input into kinetic models
—> reaction rates
Southampton

Figure 2: Professor Felix Studt presenting his talk

Using microkinetic modelling for each methanol synthesis reaction stage qualitatively predicts these trends in yield, showing DFT is capable of describing trends correctly. The aim is to find the Sabatier optimum, for which binding is neither too strong (poisoning) nor too weak (insufficient reduction of activation energy). The theoretical studies looked at various surfaces to give an indication of the catalytic effect on CO_2 hydrogenation. Examining the various reaction intermediates it showed good correlation between the binding of OH and the binding of O. This means you can express all energies in terms of the binding of O and use this in screening of catalyst materials.

Felix then discussed the accuracy of the DFT methods, using ammonia synthesis as an example. The uncertainty can be +/- two orders of magnitude which has significant impact on quantities such as reaction rates, which depend on the accuracy. Reproduction of the uncertainty was carried out using BEEF ensemble and can be propagated through to the production of Arrhenius plots and volcano plots. Although the values may have large uncertainties, the trends are still well preserved.

Discussion then moved to the methanol-to-olefins reaction and the olefin cycle. The olefin cycle uses methylation to increase chain length followed by cracking. However, the key stage for the reaction is the formation of the first C-C bond, initiation, which is often facilitated through use of zeolite catalysts. When trying to look at all the different stages for kinetics in batch reactions there are over 100 different elementary reaction steps.

Examining the effect of the zeolite framework shows similar mechanisms across zeotypes but differences due to acidity and dispersion. Studies looked at the Me substitution of the zeolite/aluminophosphate framework with a focus on the methylation of propylene reaction, where you want to make the material as acidic as possible to facilitate catalysis. This compared Si-O-Si (CHA) with Al-O-P (AlPO) with Al-CHA being the most reactive zeolite and Mg-AlPO-34 the most reactive overall. This showed remarkable scaling of transition state energy with ammonia heat of adsorption, which can be extended across modelling different alkenes and different frameworks.

The talk concluded with a summary, the overall conclusion being; the calculations aim to predict as accurately as possibly but can easily be off by two orders of magnitude, the modelling can predict qualitatively, if not quantitatively, allowing identification of trends and improving understanding of the process.

4.2 LAISER: Putting the AI in Laser - Dr Ben Mills

Dr Ben Mills gave the next talk, discussing how lasers can be made ‘smart’ to improve the precision of laser machining, the vision being the development of laser machining with real-time corrective and predictive capabilities. Ben outlines the vision and how use of AI through neural networks (NNs) might help accomplish this in real-time monitoring and control, prediction of outcomes and imaging and sensing.

Neural Networks are a network of non-linear functions with an input layer and an output layer. They are trained on real-world data, as opposed to requiring equations and physical understanding of the transformation. The data driven approach can allow analysis of problems that are too complex to solve through equations. These applications use variants of convolutional neural networks (CNN) as they are great for image analysis.

Real time monitoring: In monitoring the sample is observed by a camera and the image is passed to the CNN, which identifies the material, laser fluence, and the number of pulses. The CNN can process each image much better and faster than a human expert (>80% success in 10ms). The feedback from this analysis can the be used to correct the machining in both shape and position of the beam. The CNN allows simultaneous detection of position and shape and can spot a 200nm beam movement, which is a fraction of a single camera pixel. There is no need for information regarding the underlying processes, as the function is purely pattern recognition; however, a disadvantage is that it is not known how the NN is determining position. This could potentially be applied in quality control and reduction in manufacturing time/costs.

Laser machining prediction: In laser machining a digital micromirror device is used to change the spatial intensity profile of individual laser pulses. Prediction of a depth profile is

Southampton
The Motivation
Lasers are already powerful enough for almost all manufacturing processes
We now need to make lasers "smart", to improve manufacturing precision and efficiency
Southampton

Figure 3: Dr Ben Mills presenting his talk

more complex than simply looking at shape. It involves complex interactions between light and matter, presence of debris, peeling back of material etc. In training, the CNN effectively learns these complex interactions, giving surprisingly accurate results: it is difficult to differentiate between experiment and prediction. The neural net even copes with deformities caused by surface melting on femtosecond timescales. Ben's approach involves training two neural nets simultaneously, adopting the analogy of counterfeiter and policemen, such that each tries to deceive the other. Their first attempts are poor, but eventually the real and fake images cannot be distinguished. The network learns to convert almost any beam shape to an SEM image of the machine surface, getting better at each iteration. Even "discovering" the physical property that we call the diffraction limit. This has potential applications in immediate optimisation of machining parameters.

Ben offered some thoughts about future applications, including simulating the machining process to uncover previously unrealised techniques, suggesting this might allow NN to be creative giving improved task optimisation. Ben also covered an example of NN use in airborne particulate sensing for pollution monitoring and identification. If you shine a light on a sample and record the scattering pattern, can NNs identify the particles? Very good accuracy of real-time, real world pollutant analysis (>90%, 50 milliseconds per identification) through assigning probabilities to each of the possible pollutants. Other areas of application include waterborne pollution & salinity detection, super-resolution imaging through magnification increase and networks of neural nets to make cell growth predictions.

Finally, he asked whether the most complex scientific challenges of the future might be solved without any human understanding, given that neural nets did not need to understand the underlying science. In answer to a question about hyperspectral imaging, Ben replied that the challenge was the collection of labelled data, given that it could amount to terabytes. Other questions were about airborne and waterborne pollution testing.

4.3 Machine learning opportunities in prediction-led discovery of molecular materials – Professor Graeme Day

The third talk came from Graeme Day, discussing the prediction of material properties in material discovery. In contrast with engineering design, of aircraft, for example, with materials it is not straightforward to back track from desired properties to crystal structure. The discovery process tends to rely on trial and error: make and test, which is time consuming and expensive. Could we use a 2D sketch to predict crystal structure and from that predict properties, and could that process be reversed?

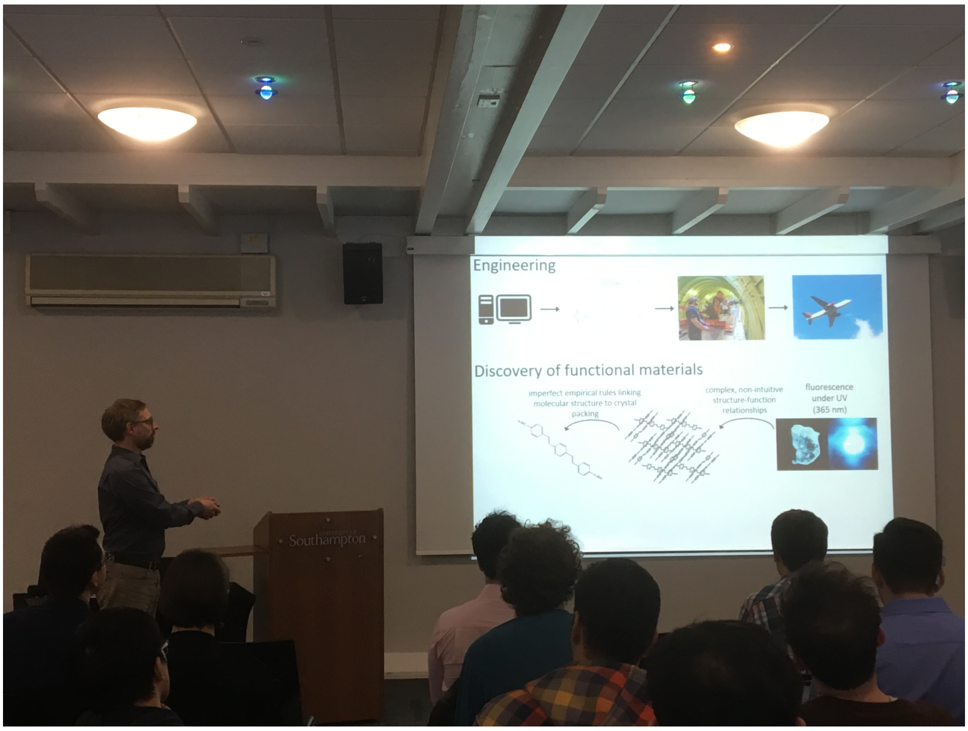

Engineering
Discovery of functional materials
imperfect empirical rules linking molecular structure to crystal packing
complex, non-intuitive structure-function relationships
fluorescence under UV (365 nm)
Southampton

Figure 4: Professor Graeme Day presenting his talk

One area of interest is porous molecular crystals, these have interesting applications but can be challenging to predict. Graeme discusses the process and challenges involved in discovering new materials. Considering methane storage ability in metal-organic frameworks, we are looking for deliverable capacities above 140 cm^3. Could this be achieved without the metal, which adds weight? Which set of candidate molecules give the desired properties?

A set of synthesisable molecules were hypothesised; for each molecule, the crystal structure is predicted, this produces $\sim$450,000 trial structures and 8000 distinct crystal structures. Once looking at energy, you can also layer on other information to create an energy-structure-function map: the promising ones are low-energy structures, in the low-density region. The measured properties of newly discovered materials not only tend to agree with the predictions that led to their discovery but also have unusual structures, not predicted by cheminformatics.

Graeme then discusses several challenges and bottlenecks and how machine learning could help overcome these issues. Computational expense is a challenge, it takes 3 weeks for the largest molecule on 800 processors, most people don't have access to that. It isn't possible to restrict the space groups either as you are looking for porous molecules outside of the normal regions. Hierarchical methods can be used to reduce to computational demand; however, time constraints scale heavily at each level of increased accuracy.

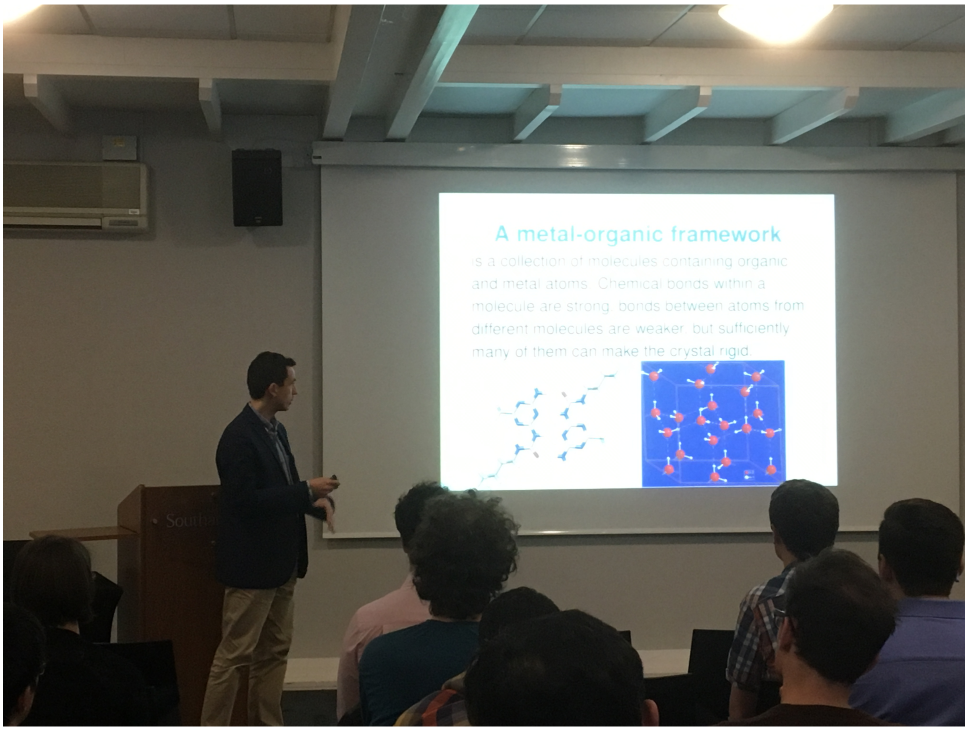
Machine learning methods can help calculations to yield data that leads to new ideas. High level calculations can be performed on a subset of molecules, using the results to predict the rest. This has given promising results, although accuracy lessens with more complex molecules. Another possibility is fragment-based energy models, breaking total energy down into smaller components, and using ML to predict correction terms.

Simulation can be a bottleneck for property prediction. Ranking by neural nets improves with extra training. Considering charge carrier mobility in semiconductors, the hopping model has two parameters, one molecular property and one based on packing. We need better structural descriptors, tailored to the property of interest. All predicted structures are perfect, but sometimes the desired property arises from defects, we need to be able to predict these defects to improve the calculations.

The talk is concluded by summarising some the areas where further investigation could maximise learning. These include packing motif classification and the use of data-driven approaches to obtain clustering. Also considering inverse design: generative models for identifying molecules to target that give desired properties. However, a small change in a molecule can lead to large differences in crystal structure and packing, so properties can be very different. Molecules close in chemical space won't necessarily be close in property space. Graeme highlights the important message that we shouldn't always just jump straight to machine learning, sometimes simplified physics-based models may be better.

4.4 Potential Solutions to Mathematical Challenges for Solid Crystalline Materials – Dr Vitaliy Kurlin

The fourth talk approaches the challenge of crystals and their representation through rigorous application of mathematical concepts. This talk discusses different approaches to the generation of unit cells and some implications of these choices.



A metal-organic framework

is a collection of molecules containing organic and metal atoms. Chemical bonds within a molecule are strong. bonds between atoms from different molecules are weaker. but sufficiently many of them can make the crystal rigid.

Figure 5: Dr Vitaliy Kurlin presenting his talk

The rigidity of metal-oxide frameworks derives from many sufficiently weaker interactions, so we can adopt a topology approach that treats crystals as periodic frameworks, an infinite graph with repeated patterns that form a solid body. Any crystal has a lattice and can be expressed using vectors and linear combination of vectors; however, the unit cell can be ambiguous as many different unit cells can define the same lattice.

Comparing lattices can be a challenge, how do you define if lattices are equivalent? If they are shifted or rotated are they the same? The natural attempt is to find a way to represent a lattice in a unique and non-ambiguous way. This can then be used to compare lattices and also as a non ambiguous input for algorithms. One method that has been used is the Niggli reduced cell (1928), but these can be unstable and real data is noisy, so lattices tend not to compare.

Lattice space is continuous as they can be continually deformed into another, but the space is split into crystal classifications such as cubic or triclinic. This discrete classification means similarity cannot be measured between groups and a continuous space should be used instead. Assuming best rigid motion, matching molecules (15 by default) can be done through root mean squared deviation, but if you increase the number of molecules then this distance continues to increase towards infinity and becomes meaningless.

The Voronoi cell, which looks at the neighbourhood of points that surround a point in a lattice, offers an alternative approach. This approach generates a lattice which is minimised over all rotations; however, they can fail the non-degeneracy axiom if different crystals have the same features. An improvement is to first do a Niggli cell reduction and then calculate the Voronoi cell.

The talk is concluded with a quick summary of the challenges in crystal unit cell representation and a reiteration that the crystals should be classified by continuous invariants. These invariants will map the space of all potential crystals and show which polymorphs are structurally close in the continuous space.

4.5 One million crystal structures: what can we learn? – Dr Angeles Pulido

The final talk in the workshop was presented by Angeles Pulido, presenting the work of the Cambridge Crystallographic Data Centre (CCDC). The CCDC is dedicated to the advancement of chemistry and crystallography through the provision of high-quality data and software across a number of areas including: pharmaceuticals, agrochemicals, pigments and poisons.

Angeles highlights that models will only ever be as good as the data that has gone in. As such they aim to provide high-quality data through the Cambridge Structural Database (CSD) with a high level of data curation and metadata. CSD tries to build a meaningful chemical representation for the molecules; however, chemical interpretation can be a challenge. When structures are submitted, validation is performed with software, but the human touch is also needed at a number of points, so curation is both automated and manual. Although processing speed has increased significantly in the last 10 years, they are still trying to speed this process up.

In the materials arena the CSD has been used to check for risk of unseen polymorphs, in particular with drug materials, as it is difficult to change the drug if a polymorph is discovered once the drug is on the market, for example, Ritonavir. Such studies use interaction data from the CSD to investigate the preferences of hydrogen bond interactions. Using a set of parameters that describe functional group environment, permutations of hydrogen bond networks are built. These give likelihood of hydrogen bond, but also likelihoods under certain conditions. Several examples were given of available drugs on the market. In the ideal case the most favourable polymorph is identified with a large gap between it and the next most favourable. But this is not always the case – which can be dangerous for drugs if there are other polymorphs which are close. Crizotinib is monomorphic, whereas Axitinib has a near neighbour, so is polymorphic. The most stable polymorph was not the one initially found. Machine learning models are also used for solvate prediction, with the finding that driving forces for solvate formation can be quite different for various solvents.

The CCDC also has many other avenues of interest. The database contains more than just structural data with information including; melting points, colours, crystal shapes, bioactivity and oxidation states. The data has been used to predict H, C, N, and O NMR spectra by

CCDC – Database, software & research
CCDC software products aimed to maximized use of crystallographic data in applications in materials and life sciences.
WHAT WE DO
Compile a high quality database
The world repository of small-molecule crystal structures
Develop Scientific Software
Structural knowledge-based and applications software
Perform Collaborative Research
Fundamental research with scientists in academia and industry
Promoting Education and Outreach
Conferences, Workshops,
Training, Teaching Materials
CSD
Organic 43%
Metal-Organic 57%
Not Polymeric 89%
Polymeric: 11%
Single Component 56%
Multi Component 44%
4
Southampton

Figure 6: Dr Angeles Pulido presenting her talk

training ML methods using data from 2000 small molecules. The accuracy of these predictions is fairly well preserved when scaled up to larger molecules. Co-crystal prediction is unfortunately still a challenging area. Molecular descriptors, such as shape and polarity are passed through a machine learning model to predict whether substances will co-crystallize. However, they are still quite far away from being able to accurately predict these.

Angeles also highlighted a number of other current projects by the CCDC. In particular, AI is being used to assess tabletability, the ease with which a crystal can be broken into tablets. Factors are intermolecular interactions, topological rugosity, slip plane, interpenetration, and hydrogen bond bridging. Slip planes are predicted by topological analysis, using CSD data. Tabletability ranks assigned by AI match experimental trends.

Looking forward - There is a lot of information in the CSD that is not used but could be examined and used to predict properties. The CCDC are thinking of moving into the challenging area of putative and experimental crystal structures, which would empower the CSD.

One of the big problems encountered in prediction is the lack of negative results. Angeles comments on the tendency to only report the positive results which inhibits the ability to build models. As a community we need the publication of negative results as well as positive results to help build better models.

5 Participants

This event was well attended with 65 participants from a mixture of backgrounds. The majority of the participants were from an academic background with approximately 85% of the attendees coming from academia. However, within this group the attendees were from a diverse range of departments covering a multitude of disciplines. The remaining attendees were mostly from industry.

6 Conclusions

Professor Frey gave a short summary of the areas covered in the workshop. Jeremy noted that we are modelling the processes, but it is still all about the data. We are exploiting machine learning and making progress, but there is still lots more to do.

7 Related Events

Other events also hosted by AI3SD in this workshop series include:
Molecules, Graphs & AI – 06/02/2019
Semantics & Knowledge learning for chemical design – 1/05/2019

Upcoming events of interest can be found on the AI3SD website events page.
http://www.ai3sd.org/events/ai3sd-events
http://www.ai3sd.org/events/events-of-interest