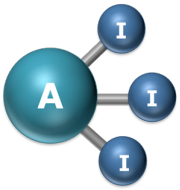

Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery

AI^{3}SD Network$^+$ Conference 2019
18-19/11/2019
AI^3 Science Discovery Network+
Holiday Inn Winchester & Winchester Science Centre

Michelle Pauli
Michelle Pauli Ltd

26/10/2020

AI3SD-Event-Series:Report-15

AI^{3}SD Network$^+$ Conference 2019
AI3SD-Event-Series:Report-15
26/10/2020
DOI: 10.5258/SOTON/P0018
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^{3}SD Network$^+$ Conference 2019
Organisers	AI^3 Science Discovery Network+
Dates	18-19/11/2019
Programme	Programme
No. Participants	80
Location	Holiday Inn Winchester & Winchester Science Centre
Organisation Committee	Dr Samantha Kanza – AI^3 Science Discovery Network+, Dr Nicola Knight – Physical Sciences Data science Service, Professor Jeremy Frey – AI^3 Science Discovery Network+
Session Chairs Leads	Dr Samantha Kanza – AI^3 Science Discovery Network+, Professor Richard Whitby – Dial-a-Molecule Network, Dr Nicola Knight – Physical Sciences Data science Service, Professor Mahesan Niranjan – AI^3 Science Discovery Network+, Dr Colin Bird – AI^3 Science Discovery Network+, Dr Wendy Warr – Wendy Warr & Associates, Professor Jeremy Frey – AI^3 Science Discovery Network+,

Figure 1: The Holiday Inn Winchester

2 Introduction

The AI3SD Network+ (Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery) gathered researchers from a wide range of disciplines for a two-day conference in Winchester, UK.

The Network+ is funded by EPSRC, hosted by the University of Southampton and aims to bring together researchers looking at how cutting edge artificial and augmented intelligence technologies can be used to push the boundaries of scientific discovery.

The conference opened with a half-day workshop on AI Ethics for Scientific Discovery, emphasising the need to place ethics centre stage in AI decision-making. The conference continued with a multidisciplinary blend of presentations and poster sessions, including interim reports from Network-funded projects and reports from EPSRC AI feasibility studies.

In this conference report you will find brief accounts of each of the sessions, along with links to slides and presenter Q+As, where available. The full list of presentations can also be found here.

3 Ethics Workshop

Ethics cannot provide us with answers. That was the stark verdict of Dr Will McNeill, opening the AI3SD Conference 2019 workshop on AI Ethics for Scientific Discovery. However, ethics can offer frameworks that can help us show – and justify – how decisions were made and the trade-offs that are an inevitable part of that decision-making. Dr McNeill's enlightening introduction laid the groundwork for the two-hour session and was followed by Dr Samantha Kanza's introduction to the five themes selected for more in-depth discussion: Data sharing; Decision making; Transparency / explainable AI; Responsible AI / AI 4 Good; and Subverting research with ill intent.

3.1 Dr Will McNeill, Lecturer in philosophy at the University of Southampton: Ethical Frameworks, Ethical Judgements

The EU Commission's 2018 Communication on AI set out guidelines for trustworthy AI: that it needs human agency and oversight, robustness and to be held to higher standards than those we hold each other to; if we replace human activity with machine activity then it needs to be an improvement over human intervention. The Communication highlights privacy but also transparency: we need to show how AI systems work and why they are safe and robust. We must be aware of the potential for bias in how we design these systems and how they operate. There are also wider implications – are these systems good for society, how do they relate to the environment and to wellbeing and so on.

Yet, Dr McNeill points out, there will always be tensions between these different requirements. Take technical robustness and safety. There is clearly no ethical reason to replace a human with a machine unless we believe it is highly robust and safer than a human doing the same job. However, research looking at different forms of machine learning shows a neat inverse correlation between technical robustness and transparency and accountability. A system might be incredibly accurate but impossible to understand or explain to others. At the other end of the spectrum, simpler forms of machine learning may be much more transparent to interpretation but tend to be less reliable and accurate. So there is a clear tension between

desiring transparency and maintaining robustness.

Equally, there is a tension between technical robustness and accountability, and yet another tension between robustness, privacy and data governance. The desire for more data, which can increase effectiveness, rubs up against the ethical requirement that people have a right to keep their data private.

"These are natural tensions that exist and we are asking the wrong questions if we think we can find answers that will show an ethically simple resolution to these tensions," says Dr McNeill.

There will never be a system that can determine for us how our duties towards these different requirements can be balanced. Those tensions will always be there. We can attempt to reconcile them but we cannot resolve them. When we're trying to reconcile difficult ethical questions we need to weigh up the different things we have a duty towards and the result is difficult ethical decisions. To make the end judgements ethical we need to show how those decisions were made – for example, who was in the committee, how long did they sit for, what did they weigh up?

Dr McNeill points to the case of NICE. It uses transparent and non-discriminatory algorithms to decide which drugs can be deployed in the NHS and when. The algorithms show how those tensions are resolved, revealing why NICE has chosen to allow a drug to be released or withheld in the NHS, using the same reasoning as in other cases eg cost-benefit analysis, QALYs etc. But, at the same time, it could be argued that this may be at the cost of moral fairness (when cutoffs to drug availability are arbitrary rather than ethical), human agency and choice and, indeed, to a certain extent, robustness. There is a loss of accuracy on these dimensions by opting for dimensions of transparency and non discrimination. It is an ethical trade off, not a resolution of an ethical problem. It is a decision to favour one set of dimensions over another. Again, it is ultimately a matter of human judgement.

According to Dr McNeill, there is no algorithm that is capable of determining how the tensions between our moral duties are resolved. Instead, ethical frameworks for AI will be required. But ethical frameworks cannot replace ethical judgement; instead they highlight tensions that only human judgement can reconcile.

3.2 Dr Samantha Kanza, Enterprise Fellow at the University of Southampton & AI3SD Network+ Coordinator: Ethics for AI in Scientific Discovery

For Dr Kanza, we're currently in a golden age of AI, especially in relation to how far we can push the boundaries of scientific discovery with AI. But to what extent are the ethics of using AI for scientific discovery being considered?

Dr Kanza has broken the issue down into five main themes, which all come with key discussion questions:

- **Data sharing:** Scientists always need more of the right data in the right format. At what point is it our duty to release that data? Is it an ethical decision to release that data to make progress? But how do we know that it is accurate data? Whose responsibility is it to collect a range of data and diversify it? Homogenous data risks skewing results, for example in medical studies that run the risk of being biased by sex or ethnicity. There is also a need to diversify the people that are collecting the data in order to diversify the

data.

- **Decision-making:** When should humans be making decisions and when should machines be making decisions? In medical treatment, should patient decisions be made by machines, by humans or by both? Should we trust both equally? Augmented intelligence offers the best of human and machine together but is there a tradeoff of reliability and accountability?
- **Transparency / Explainable AI:** How can people make informed decisions about AI for science if we don't understand how AI has made decisions? Are we expecting more from a machine than we do from a human?
- **Responsible AI / AI 4 Good:** Should we be directing AI for scientific discovery towards the greater good? Should we be directing our research to particular areas, such as targeting drug research to specific diseases? Who makes those decisions?
- **Subverting research with ill intent:** How can we protect our research? How does this conflict with transparency? Should there be a point where code and methods are not made available / open source – and where does that point lie?

3.3 Key points from the group discussions

Figure 2: Ethics Workshop Discussion Session

Data sharing

- Trust is the overriding theme.
- Greater awareness is necessary – for example, where drug trials are carried out only on males so they are optimised for men and not for women. Or trials that take place in lower regulation countries. The more open you are at the outset the better the chances of awareness of what is going into the data you collect.

- Missing data can be down to a variety of reasons. Is it a rules-based or data-driven system? Some data will have a skewed distribution but that is not necessarily a bias, it may be intentional. There may be laziness or direct manipulation of data. Customs can also be an issue – 'we've always done it that way' – but a lot of the knowledge is tacit. How do you record that?

Decision-making

- Is it better to have a machine making decisions or influencing a decision a doctor makes – or an augmented system? How much faith should GPs have in AI?
- How would it affect the law? A judge/jury can look into the reasonableness of something but an AI would see it as black or white.
- We are more likely to trust giving personal data to a doctor than a machine.

Transparency

- With issues of accountability and responsibility – transparency v explainability – there can be some distance between the experts who design the systems and the audiences who demand accountability. They often don't speak the same language. There is a need to instill the skills and capabilities to have that conversation. AI3SD is here for that.
- Transparent to whom is a big question. As is explainable – explaining it to whom and at what level? Can we build in that transparency at the design stage – and are we asking those questions as companies like Google go into the NHS? And we need to look at our own research communities and to what extent our actions are guided by practice as well as principles.
- Building in explainability – even if we talk to the people designing the system, we do not necessarily know what a good explanation is and we need to go through some kind of iterative process to understand what people want. Explanation methods that have been created at a technical level may not work at every other level.
- There needs to be an estimation of uncertainty and how this relates to explainability and accountability – explanations create some kind of trust in what the model is doing but models need uncertainty built in, especially in healthcare.

AI 4 Good

- The 'greater good' is subjective – security and police could use AI as surveillance to keep an eye on people they see as suspicious but do we want to live in a surveillance state? What about bias in the data and model training – the model is only as good as the data you put into it.
- Funding is a big factor with researchers less likely to look into greater good AI if they are not going to get funding for it, so who decides on the funding? Are they informed enough? Self-interest also comes into it. The research focus is going to be dependent on the kind of society you're looking at and priorities will be different depending on where you are, such as an emphasis on cancer in the west or malaria in the global south.

Subverting research for ill intent

- What is ill intent? Is it always ill intent or a case that what may be good for one person or group is bad for others? Is it always present to some extent?
- There is general concern about subversion of data and about companies using data in illegitimate ways. At what point should we consider closing systems and reducing transparency in order to stop subversion? But why would data be better if closed?
- Should the use of the word outlier be banned? It's all data! Outliers may, in fact, be the data you need to pay most attention to.

- Amplification bias – just because a dataset has been used by many researchers does it make it more trusted?
- Reproducibility – if you close off your research completely then there is no point doing it – but the counter argument is that's true unless the outcome is worth it eg it produces a completely novel synthesis for a molecule. In which case all the world cares about is that molecule, not how you got there.

4 Session 1: AI3SD

4.1 The AI3SD Network+ – Professor Jeremy Frey, University of Southampton

Jeremy Frey, AI3SD's principal investigator, set the scene for the two-day conference by giving an overview of AI3SD, its work and aims, and laying out one of the key questions for the event: can scientific discovery be automated?

Professor Frey explained that AI3SD brings together those pushing the forefront of AI and scientific discovery to work together and learn from each other. There has been a huge growth in this area but there remain significant gaps in our understanding of what it means for science to use AI and for AI to attempt discovery. The AI3SD network's belief is that augmented intelligence – getting the best out of human and machine intelligence – is the best way forward.

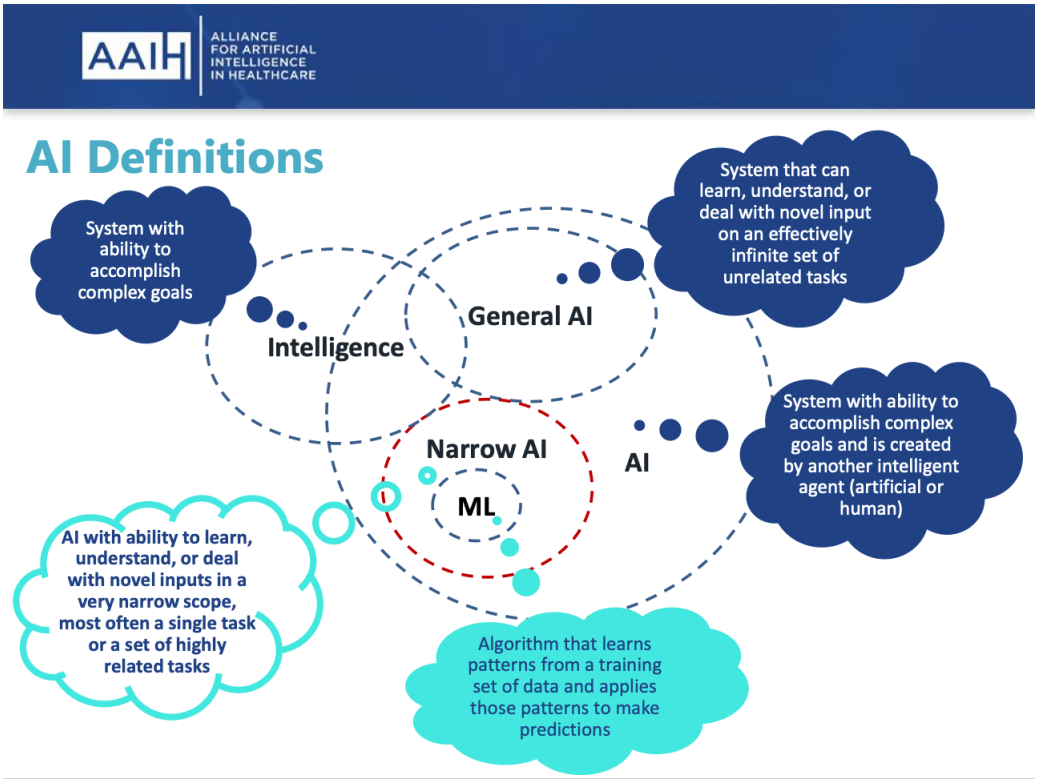

Figure 3: Slide taken from Professor Frey's presentation - original taken from: https://www.theaaih.org/

Professor Frey outlined some of the broader questions that should be in everyone's minds when thinking about AI and scientific discovery:

What does AI need to do to enable discovery and what does it say about how we train people to do science? Industry is facing severe skills shortages so how do we promote the right sort of training for the next generation of researchers and retraining of the current set of researchers?

Does the 'intellectual debt' issue – that we get the answer from AI but we have no idea how it gets that answer – matter?

If AI means that we "do stupid things faster with more energy" does it give us the same answer as before or a different answer? How good is that answer, and how applicable?

What is the environmental impact, given that this work can consume significant resource? Are we wasting time and effort, especially at discovery stage? Could we do it more efficiently?

Bio | Slides

5 Session 2: Data, AI, Molecules and Materials

5.1 Learning with Complex Priors and Interactions – Professor John Shawe-Taylor, UCL

Professor Shawe-Taylor's starting point is that while we have seen remarkable success in learning from large quantities of raw data, for many scenarios this seems to be a wasteful approach, in that we should be able to leverage prior knowledge and models, or a sequence of interactions, in order to make learning more efficient. He reviewed work that has looked at methods and analysis of such approaches.

He started by looking at the multi-armed bandit problem (MAB), which is modelled on a casino with slot machines – the player does not know which has a better reward rate so the only way to find out is to start playing. However, we do not know from a single play which is good or not – the assumption is that each machine has fixed reward distribution so we will eventually get to understand them – and the goal is to maximise the reward. It encapsulates the problems of exploration v exploitation. Performance is usually measured by regret – the extent to which you failed to achieve the optimal. The idea is to devise algorithms that keep the regret under control.

It is possible to do that using a Bayesian approach and then Thomson sampling: sample the probabilities from those posterior distributions then choose the arm that has the maximum of the sample, which is a neat way of trading exploration v exploitation. Professor Shawe-Taylor then discussed upper confidence bounds strategies (UCB), which can be shown to trade two constraints. He explained that it is a framework that can be used to analyse how we can interact and learn from experience but with a good, rigorous analysis at the back of it. It has been shown to work in practical experiments. By comparison, he showed a relatively naïve approach – the 'greedy approach'.

Professor Shawe-Taylor considered the LINREL algorithm – linear associative bandits and the upper confidence tree algorithm, which is used in some Go applications, before moving on to the PAC-Bayes theorem, which was first proved by McAllester in 1999 and which does not rely on the assumptions of Bayesian analysis but follows the inspiration of the Bayesian approach. It has a prior distribution over a set of classifiers but you do not have to believe in that as being the correct prior, and the posterior does not need to be the posterior you would compute using Bayes theorem. It shows the tightness of the bounds these approaches can get

on various datasets. Finally, Professor Shawe-Taylor looked at stochastic differential equation (SDE) models, which model uncertainty.

He was asked when this approach would be used and replied that it is appropriate when you have some structural knowledge about the type of solutions you would expect – eg modelling weather – and have some prior expectation of how things evolve over, for example, time but you also know that the modelling is not exact, so need the data to do the fine tuning. Learning with no prior knowledge and only the data would be difficult. If you put the two together then you might be able to do a better job.

The bandit problem is where your actions do not change the state. If the side information is unaffected by your previous actions, they do not affect the situation you are in then the bandit is the way.

He argued that the hypothesis would be in choosing the particular coherence function: *"I think I will be able to find the correct solution to this problem with this class of functions and by interacting will find the solution".*

Bio | Slides

5.2 Data driven models that predict protein function from sequence – Dr Lucy Colwell, University of Cambridge

Dr Colwell set out the challenge of predicting the functional properties of a protein from its sequence to (i) discover new proteins with specific required functionality and (ii) better understand the functional effect of genomic mutations. Experimental breakthroughs in our ability to read and write DNA enable the rapid acquisition of the data required to train and validate machine learning models that predict protein function from sequence. Because in many cases phenotypic changes are controlled by more than one amino acid, the mutations that separate one phenotype from another may not be independent, requiring us to build models that take into account the correlation structure of the data.

Dr Colwell argued that proteins are not sufficiently studied in this area but there are interesting protein design questions to be asked. She set out how we currently find new proteins: *"Essentially a random local search and selection. So it's a search problem. It's very powerful but it's still challenging to find new proteins. There is a lot of motivation to work on this problem as methods are yielding exciting results."* She offered the example of finding a potential new cure for ebola by mining the immune response of people with the disease: it took a long time to find these antibodies. The space is too large to search exhaustively, to do that screening. There needs to be a shortcut.

One possible solution is to use machine learning to capitalise on the information collected in databases and transfer information between different learning spaces – an active learning loop with wet lab experiments. The proposal is to add machine learning to directed evolution – again back to explore v exploit – to optimise resources and write DNA sequences as well as read them, to make exactly the protein sequence that's needed.

She showed that such models rival the accuracy of existing hidden Markov models at sequence annotation, even when given relatively little training data. The representation of sequence space learned by the model can be used to build families that the model was not trained on.

She reported experimental confirmation that machine learning models can accurately identify

variants of the AAV capsid protein that assemble integral capsids and package their genome, for potential application in gene therapy.

Dr Colwell was asked about the number of layers (fairly lightweight models. Kernel size), how to test fitness of particular candidate (printed all sequences and tested all of them, about 300,000 sequences); prior information, improve ability, smaller search space? (good idea. Tried using more informed representation, but didn't make much difference); and environment of sequence, eg solution (didn't take into account, "surprised it works as well as it does").

Bio | Slides not available

5.3 Materials Development in the Energy and Electronics Sectors through Combinatorial Synthesis, High-Throughput Screening and Machine Learning – Professor Brian Hayden, University of Southampton

Professor Hayden's starting point was that the combinatorial synthesis of solid-state materials combined with high-throughput characterisation and screening provides an opportunity to develop increasingly large materials databases. Machine learning approaches are crucial in several aspects of the building, interpretation and exploitation of such databases. These can also be constructed to include physical and chemical descriptors from, for example, ab-initio calculations. The challenge is to ensure reliable and auditable content and consistent format of the data.

He presented examples of how machine learning is being developed in the interpretation of raw data sets using data from high-throughput investigations of electrocatalysts mediating the oxygen reduction reaction (ORR) and oxygen evolution reaction (OER) for the development of reversible fuel cells and rechargeable metal-air batteries.

A key issue is that the solid state structure can change suddenly with composition, for example there are lots of perovskites whose structures depend critically on composition and these materials are relevant to many different solid state technologies. Suppose a customer would like a new material with specific desired properties. The starting point is often in literature and then we go around the loop, with combinatorial synthesis and screening, usually with a fast approximate initial screen followed by more time consuming detailed screening. Synthetic methods can be relied on to generate a range of stoichiometries using a compositional gradient. The atom effusion methods we use avoid the problems of the standard mixing methods which have high activation barriers and so often do not generate in between phases and these methods are further hindered by the existence of metastable states. Ab initio calculations also often miss structure changes, so the need experiment is paramount.

The method is to take raw data from a database, undertake high throughput characterisation data and apply supervised learning: regressions (different types) and principal component analysis. Generate a thin film of combinatorially synthesised materials and measure properties relevant to application, eg for solid state lithium batteries, ion impedance. Datasets in solid state materials are complex and as the film structure can differ from the bulk there can be systematic differences in properties.

Fuel cell example. Fix structure as perovskite. Looking for oxygen evolution and reduction, but do not usually get both. ML has not yet extracted the data they want. Multi-Task Metric Learning output compared to brute force analysis identified a number of clusters lying along a line in the phase diagram, plus other regions away from the line. Can do electrochemistry (cyclic voltammetry) to identify redox couple, such as $Ni^{2}+/Ni^{3+}$ and Mn^{3+}/Mn^{4+}.

Along the line, one can see anti-correlation: one area good for oxygen reduction, one for evolution, then pick out good catalysts for both. Ni^{2+} and Ni^{3+} different ionic radius, so looking for ions to substitute, such as Ca^{2+}. The results provide an insight into the potential opportunities of machine learning in the future in the predictive development of functional materials.

Professor Hayden was asked about the differences between surface and bulk of material, and change over time (can get surface aggregation, especially if anneal. Don't work in time domain, not necessarily looking at equilibrium states; some are metastable); what was the key human contribution (would compare ML results with existing knowledge to ensure getting realistic results. Also want to understand the science); and how can ML help with accessing experimentally something predicted (one of the challenges at the moment, making materials according to prediction).

Bio | Slides not available

6 Session 3: Network-Funded Projects Interim Reports

6.1 Predicting the activity of drug candidates where there is no target – Professor Matthew Todd, UCL

Professor Todd noted that he was a "beneficiary of AI machine learning, not a practitioner" before recounting the fascinating story of how his project is tackling a common issue in modern drug discovery: the knowledge that a molecule could be useful but does not have a defined target. Medicinal chemists may have the data but not know what to do about it. Could AI be used to predict the next move?

Professor Todd briefly covered his team's work in open source drug discovery, specifically malaria. The process mimics OS development: work cannot be patented; work happens on github using its issue tracker; the bedrock is freely available online lab notebooks; all the molecules made in the project are put on a google sheet (currently a few hundred); social media is used to communicate and bring in new people. Publication is achieved through zipped lab notebooks on repositories, snapshots of blog posts and so on.

The project aims to "look for things that kill malaria and not people". A current series of work is looking promising because two of the identified molecules work in the mice model –

"that's a late stage in the drug discovery process to still be unpatented. If we can get it into pre-clinical trials we could answer the question of how to fund an open series of molecules through trials without usual patent protection – that would be the first time that had happened and we're getting close to that point."

However, the process is frustrating and expensive. While even negative data is useful, *"it's £2000 a shot and eats through grant money. So, with 300-400 molecules..."*. So, the project turned to machine learning.

The problem was membrane protein in the parasite, which controls sodium ion transport. Correlate ion concentrations means one lock, so many keys. The process is to induce artificial resistance by culturing in low concentration of drug; look at changes in parasite DNA that imbue resistance.

Round 0 was the first attempt at a model, overlaying molecules that worked and those that didn't and showed no predictive power.

Round 1, in 2016/17, involved other people and produced six suggestions, all different. Contributors were given the SMILES strings of all the molecules: could they spot the actives amidst the rest? Typically, they could spot three out of the 20, which was not bad but didn't inspire confidence, though it was a good community event.

Round 2 ran the competition again using AI with extra data. Nine solutions came in, some by companies. Everyone can see the models on github. Winners get a cash prize and predict two new molecules, one similar, one different, that will be synthesised in the lab, evaluated in Dundee and there will be a post-mortem meeting in 2020.

The results will be written up and published as a challenge paper with all data in github.

Professor Todd was asked whether resistance mutants are predictors of mode of action (and replied that molecules with similar structures can have different modes of action) and whether the drug changes the membrane (it is compensatory, effect depends on something you have done).

Bio | Slides | Read a Q+A with Professor Todd

6.2 'Next-next' Generation Quantum DNA Sequencing with Chemical Surface Design and Capsule Nets – Professor Tim Albrecht, University of Birmingham

Professor Albrecht discussed his project, which combines quantum tunnelling-based biosensing with advanced machine learning methods. DNA sequencing based on quantum mechanical tunnelling in principle allows for the label-free identification of nucleotides, based on their intrinsic electronic properties, and thus in some ways constitutes the ultimate limit in single-molecule sensing and sequencing. While the sensor performance is affected by many factors, including the design of the tunnelling junction and the surface chemistry of the (metal) electrodes, in this project the main focus is on maximising the level of information that is extracted from the data.

For example, the project has been able to demonstrate a significantly improved error rate when employing convolutional neural networks (CNN) for "base calling", compared to support vector machines (SVM), and is now exploring capsule nets for further improvements.

Professor Albrecht began by dividing AI-based approaches into two: data classification (unsupervised, unlabelled); and prediction / sensing (known classes, labelled) using deep learning. His project is concerned with the latter.

DNA sequencing using quantum mechanic tunnelling is not a new idea; the first example can be seen in 1982. Established sequencing technologies are accurate above 90% today. But how do you make a tunnelling junction in a platform that is high throughput compatible?

In collaboration with City University, Professor Albrecht's team realised that if an SVM can do much better than conventional method then what about a convolutional neural network (CNN) and how does it compare to an SVM? They achieved an accuracy of 98%.

With AI3SD's involvement, the project started looking at other techniques, such as capsule nets. While CNNs do ok with the 'Picasso problem' (recognising components of a face that may not necessarily be in the usual places), capsule nets do better.

On simulated nucleotide data the capsule net outperformed CNNs. It's more robust towards events on the edge. But if centre the events then CNN does significantly better. Increasing the number of components from 16 to 32 resulted in an improvement.

The project is now looking at further optimising data analysis protocols through experimental work in progress and, in the longer term, reintegrating into high throughput.

Bio | Slides | Read a Q+A with Professor Albrecht

6.3 Deep Learning Enhanced Quantum Chemistry: Pushing the limits of Materials Discovery – Dr Reinhard J Maurer, University of Warwick

Atomistic simulation based on quantum mechanics (QM) is currently being revolutionised by artificial intelligence and machine-learning (ML) methods. This involves approaches to efficiently predict materials and molecules with specific properties within the vast space of possible chemical compounds. It also involves efficient regression in high-dimensional parameter spaces to accelerate computationally demanding quantum chemical calculations of molecular properties such as the thermodynamic stability or spectroscopic signatures while retaining the predictive power of QM. Most previous approaches have used ML to predict measurable observables that arise from the QM wave function of molecules. However, all properties derive from the wave function, therefore an AI model able to predict the wave function has the potential to predict all molecular properties.

Dr Maurer explored ML approaches to directly represent wave functions for the purpose of developing methods that use AI and quantum chemistry in synergy. He presented approaches to encode physical symmetries into deep learning infrastructures, along with recent efforts to use data-driven deep learning to develop a highly efficient Density-Functional Tight-binding simulation method to describe hybrid metal-organic materials.

ML is used to parameterise quantum mechanical data: results from solving Schrödinger equation, build models, molecular geometry encoded. The idea is for ML to learn quantum mechanics. Feed in geometry, local environment of each atom; construct pair-wise interactions between atoms; matrix holds them, but essential to capture local environment. Diagonalise the matrix. It is inverse chemical design. Chemistry is usually based on structure-property relationships, so what is the structure that gives these properties? ML-based representations of chemical model, integrate into existing quantum chemistry methods. Evaluate integral once then use in mode, accepting the differences.

As Dr Maurer explained, the goals of his AI3SD project are to establish deep-learning based Hamiltonian; integrate ML quantum model into existing QC methods; and make it tight-binding – calculate once and not suffer so many errors while having the same benefits. The result has been an important proof of concept that does not – yet – answer everything

Dr Maurer was asked if he encoded all symmetries. Yes, initially but should not have to.

Bio | Slides | Read a Q+A with Dr Maurer

7 Session 4: Machine Learning and Deep Learning for Scientific Discovery

7.1 Non-equilibrium Physics and Machine Learning – Professor Juan P Garrahan, University of Nottingham

Professor Garrahan discussed his work at the interface of current questions in non-equilibrium physics and machine learning methods, focusing on the general statistical mechanics issue of accessing and characterising rare dynamical events in stochastic systems.

He described the connection between trajectory ensemble methods – often based on the mathematics of large deviations – and reinforcement learning (RL) in Markov decision processes (MDPs). He also explained how the problem of "making rare events typical" in a stochastic system corresponds to finding the optimal dynamics in an MDP. His results illustrate the many possible synergies between statistical physics and machine and deep learning.

Bio | Slides not available

7.2 Deep Machine Learning of Quantum Chemical Hamiltonians – Professor David Yaron, Carnegie Mellon University

The high computational cost of ab initio electronic structure calculations remains a challenge for computational design of molecules and materials. Semi-empirical models, such as Density Functional Tight Binding Theory (DFTB), can compute electronic structure at a greatly reduced cost. However, the accuracy of such models is insufficient for many applications.

Professor Yaron presented a means to systematically improve the accuracy of such models while maintaining their low computational cost. The key enabling technology is implementation of the DFTB as a layer that can be integrated into deep learning models of machine learning. This layer takes, as input, DFTB parameters generated from a standard deep learning network and generates, as output, electronic properties from self-consistent solutions of the DFTB model. The quantum chemical layer allows back propagation, such that the system can be trained efficiently on data on electronic properties.

Professor Yaron began by examining why we use ML in computational chemistry: to develop predictive models and explore chemical space. The more predictive our models become the better we will be at exploring chemical space. The greatest success of ML in chemistry is taking better advantage of molecular chemistry – to get detailed data from ab initio computations faster and at a lower cost. One frequently asked question is how much physics should this model have? Where, indeed, is the physics?

The top level goals for Professor Yaron's two projects were:

- Data science-produced insights
- Machine learning produces predictions
- AI produces actions

Greatest success was seen in the Behler Parrinnello (BP) neural net. It has very simple physics – a neural network predicts atomic energies – sum to get molecular energy. To the extent that there is physics in the model, it is of two types that can sum together. This simple model works well. 2017 saw the first paper to achieve this – PB net for Organic Molecules:

AN1 – it used big data, reproducing quantum chemistry but with almost no physics in it.

Model Hamiltonians go back to 1980 and worked pretty well. DFTB is more complex as it looks at lots of data on molecules and figures out a correction and changes the philosophy of where the electronic part comes from. It's physics based.

Professor Yaron's work is taking the DFTB model and fitting it empirically. It is interpretable (orbitals, electron densities, population analyses etc), builds in more physics so may require less reference data to learn and is empirical at short range, physics-based at long range (more reliable transfer to large systems). It is compatible with an initio quantum chemistry.

The goal is quantum layer for deep learning and the key is to back propagate through the quantum chemistry.

Bio | Slides not available

7.3 AlphaFold: Improved protein structure prediction using potentials from Deep Learning – Dr Andrew Senior, DeepMind

DeepMind's AlphaFold protein structure prediction system was recently ranked first in free-modelling at the CASP13 (Critical Assessment of Protein Structure Prediction) Biennial blind assessment of prediction methods. The system relies upon prediction of inter-residue distances by a very deep neural network. Using these distance distributions and a reference distribution from a similar neural network, we construct a potential and show that we can optimise this potential by a simple application of gradient descent, as well as with a more conventional fragment assembly / simulated annealing algorithm.

Dr Senior set out a system his team at DeepMind built to predict protein structure. Proteins are large molecules built up of a chain of amino acids, backbone of nitrogen, carbon and carbon atoms, then have a side chain coming off the amino acid. There are 20 possible amino acids which determine the protein. The function of the protein is determined by the structure: if one works out the structure then one can work out its function.

However, only a limited number of structures have been determined. So the goal is to discover the structure of proteins and, ideally, do so purely via computer. Most of the structure is captured in the backbone structure so it is key to predict the structure of the backbone.

Why is DeepMind interested? Its core mission is to solve artificial general intelligence and, Dr Senior says, it is making progress on that goal and using the results to solve important problems, including a range of science problems.

Protein folding is a fundamental problem in biology and the expectation is that if one can make progress here then one can make progress on downstream problems. It is a good problem for deep learning using neural networks and optimising by gradient descent to change parameters.

Why use machine learning for this problem? It's a complex problem; hard to model all the complex interactions in a long molecule; there is data thanks to experimental structure techniques, 150,000 entries in the protein data bank of solved structures. But many have been solved multiple times. So after having reduced the redundancy there are only a few thousand examples of training compared to 15m examples in image banks. CASP assessment provides a benchmark with well-defined goals.

Dr Senior's team is doing this by using a very large number of distributional predictions from a neural network. Individual predictions are detailed, calibrated and smooth. Averaging the agreement scores over large numbers of distributional predictions gives an accurate and smooth scoring function. Although only 150,000 structures are solved, there are banks with billions of sequences. So, even if you cannot find a solved structure, you can still look at sequences and find correlations.

Inter residue distances are used as one can find distances between pairs and draw a distance matrix. Where there are interactions there are short distances. Helix is characterised by a broad diagonal. It takes about four days to train one of the networks and the project used cropping to speed things up – operate network on 64x64 crop and can fit a batch of them into memory consistently. And change the offset so that from one protein can get thousands of training examples – called data augmenting. At test time tile crops to cover the whole distance map and average across alternative offsets.

At CASP13 in 2018 AlphaFold came out as a clear winner in one domain but also quite good at template-based model.

Bio | Slides

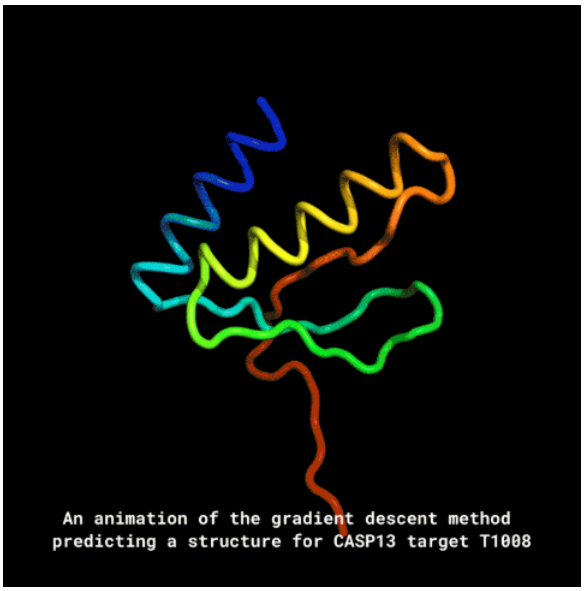

Figure 4: Image taken from Dr Andrew Senior's presentation: An animation of the gradient descent method predicting a structure for CASP13 target T1008.

8 Session 5: AI and Scientific Discovery

8.1 The UKRI Review of Support for AI – Dr Renée Van de Locht, EPSRC

UKRI, and its research and innovation communities, have a key role to play in the development and application of world leading AI technologies. Dr Van de Locht, who manages EPSRC's Natural Language Processing and Artificial Intelligence strategy, explained how UKRI's AI landscape review aims to understand current UK support for AI research and innovation, and engage key stakeholders in identifying the opportunities and future directions for AI research and innovation in the UK.

The UKRI AI review, which began in March 2019, recognised the need for all the research councils to have a common approach to this topic as it is multidisciplinary in its very nature in application, implications and design. AI will be the first area to have a council-wide approach and strategy emerging from the outcomes of the review.

AI has a high political profile. As Dr Van de Locht explained, "The fact that the government is so aware of it is an opportunity but there is also the risk that AI is seen as 'magic fairy dust' and the understanding is not quite there of what it can and can't do."

UKRI has set up a new theme to review the UKRI landscape and assess what it has, what it wants to achieve and how it fits with the current landscape. It will identify opportunities and any gaps in provision and recommend future support strategies that will enable AI to meet its full potential. UKRI is engaging with key stakeholders to try to help formulate desired outcomes, as well as working with the Robotics Growth Partnership and Office for AI, understanding UK's position internationally etc. It will set out a strategy for appropriately supporting AI.

UKRI's AI portfolio is currently £1bn invested in UK AI research and innovation: £460m in research and innovation projects, £180m in skills and training and £410m in key strategic investments.

Deep dive topics for the report are collected under three headings: the environment for research and innovation; crosscutting research; and applications and implications in AI. Particularly interesting, under 'crosscutting research', are 'next generation AI', 'AI for science and research' and 'robust, explainable, secure, ethical AI'. Several areas of importance within 'next generation AI' were identified such as human-AI partnerships and interfaces, human-like AI, cross-domain AI, and future AI architectures.

Anticipated outputs are a review report covering direction, external leadership, coherence of understanding and view, clear definition of what we mean by AI, international comparison, key risks and vulnerabilities, data risks, skills and training, collaborative opportunities, infrastructure provision. Bio | Slides

8.2 Explainable AI and Scientific Discovery – Dr Richard Tomsett, IBM

A broad variety of industries are interested in the potential of AI (particularly machine learning) technologies for supporting business decisions. However, many companies are hesitant to invest in machine learning systems as they are not easily interpretable. At least part of this hesitance is driven by regulations that require firms to be able to explain certain kinds of decision (for example, the so-called "right to an explanation" under GDPR). Such concerns have stimulated investment in "explainable AI" research, leading to an explosion in methods for explaining the behaviour of black-box machine learning models.

To whom does AI need to be explainable and why? It's a fascinating topic and one that Dr Tomsett, who works in the Emerging Technology team in IBM Research, explores in research that comes under the goal of developing fundamental science that will help military coalitions operate more effectively in 10-20 years time. Explainable AI is critical here – when a military commander has to explain every situation, no part of the system can be a black box.

Explainable AI – here referring to explainable machine learning – is not a new area, according to Dr Tomsett, but industry is increasingly concerned with it as machines are used more and

more to support and augment human activities and decision making. It is essential to understand and trust the model to do what you intended it to do. Trusted AI can be divided up into five areas: accountability; value alignment; fairness; robustness; and explainability.

But explainable to whom, exactly? People have different roles in relation to an AI machine (creators, the system itself, operators, executors (who might be the same as the operators), decision subjects, and then might have data subjects and then examiner eg an auditor who might need to come and check that your system is not working in a biased way). Most scientists will be in the creator role and not just building for their own understanding but also for fellow scientists' understanding, so they might be operators, and executors might be funders.

"Explainability is not just for you," emphasised Dr Tomsett.

He laid out a variety of explanation models, which include PDP, ICE, ALE plots, local surrogate, global surrogate, game-theoretic, permutation-based methods and backpropagation-based methods.

LIME is a generic method that can be used for any kind of data, not just images. There is also layer-wise relevance propagation and spectral relevance analysis. However, there are challenges and pitfalls to using these techniques – such as with LIME producing a number of different explanations with the same model and same input. Layerwise relevance propagation variants also has issues. Dr Tomsett offered one way to evaluate explanations: the Area over Perturbation Curve. He also referenced a paper in Nature Machine Intelligence (1, 206-215 (2019)) by Cynthis Rudin: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.

Bio | Slides

9 Session 6: Flash Talks for Online Posters

The following posters were presented and can be viewed on the links below.

- Ontologies for Chemistry - *Dr Colin Batchelor*
- The Physical Sciences Data-Science Service - *Dr Nicola Knight*
- Translation level regulation of cellular protein concentrations - *Mrs Pratheeba Jeyananthan*
- Machine Learning with Coarse-Grained Molecules to Determine Thermodynamic Properties - *Mr Kezheng Zhu*
- Saliency Map on Cnns for Protein Secondary Structure Prediction - *Mr Guillermo Romero Moreno*
- Transfer Learning Across Species on Single Cell Datasets Using a Neural Network - *Miss Xin Du*
- Predictive Modelling of Post-Translational Regulation During Human Cell Cycle Progression - *Mr Gregory Parkes*
- Robust Subspace Methods for Anomaly Detection in High Dimensional Datasets - *Mr Omar Shetta*
- Machine-Learning-Based Density Functional Tight Binding Parametrisation for Hybrid Organic-Metallic Systems - *Dr Adam McSloy*

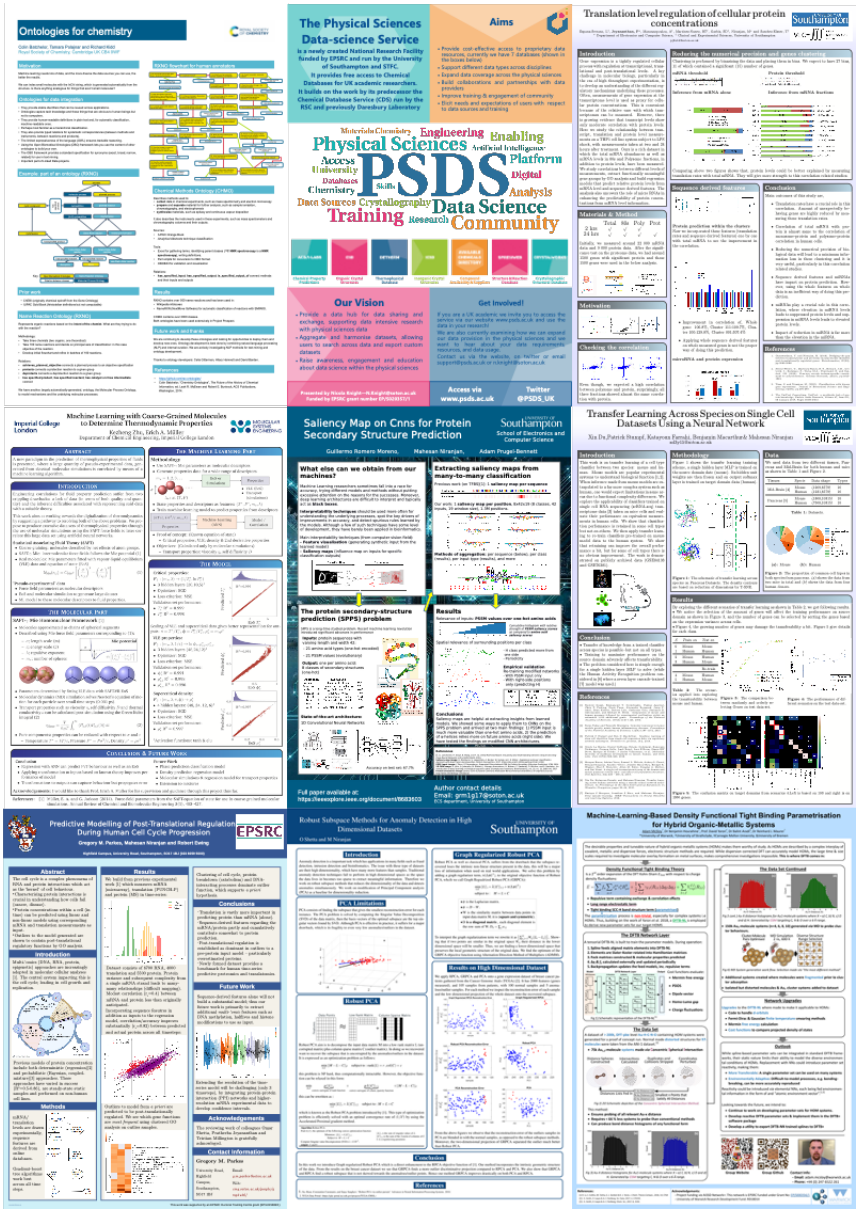

Figure 5: The AI3SD Network Conference Posters

10 Session 7: Talks from EPSRC AI Feasibility Studies

10.1 Machine Learning for Modelling Microstructure Evolution – Professor Nigel Clarke, University of Sheffield

"Microstructure evolution is an exemplar of a rich, complex spatio-temporal problem. We want to understand what we can learn from machine learning about the underlying physics," began Professor Clarke. He proceeded to explain that the parameter space is too large for even powerful supercomputers but machine learning can help us to explore that space.

Professor Clarke's work focuses on Gaussian processes (GPs), a popular non-parametric class of models used extensively in machine learning and uncertainty quantification, which have well documented predictive abilities.

Measuring microstructure is tricky. Real space 3D imaging is hard as organic components

have poor contrast and length scales are often too small for real space dynamic experiments – numerical data, power spectrum of the Fourier transform. Other characteristics – length scale / domain size, volume of phases, interface area, interface curvature, connectivity – are hard to get at but we want to know them. Is this information hidden in the scattering data?

In Professor Clarke's preliminary studies, he applied existing GP methodology to microstructure evolution to determine the feasibility of generating an emulator to supplement more traditional, computationally intense, approaches. As an exemplar, he focused on the non-linear Cahn-Hilliard equation for describing phase separation in blends.

As spatio-temporal problems are particularly challenging for ML due to their high dimensionality, Professor Clarke used machine learning to predict video images, based on the idea of light cones, in which the present is only dependent on the past in the immediate spatial neighbourhood, analogous to real-space time-stepping numerical schemes for PDEs.

He found that, *"Gaussian processes have the lowest error almost by a factor of two and are fantastic."*

He concluded that microstructure evolution is an exemplar for using machine learning to improve our knowledge of underlying physics. Benchmarking Cahn Hilliard parameter learning problem comparing grid search, gradient descent and Bayesian optimisation can be utilised to learn from experimental data. More informed assumptions on the GP prior representing the loss. More flexible physical models that are still constrained by known physics, eg, conservation of material.

Bio | Slides

10.2 The Automation of Science: Robot Scientists for Chemistry and Biology – Professor Ross King, Chalmers Technical University

According to Physics Nobel Frank Wilczek, "in 100 years' time the best physicist will be a machine". And if that comes to pass, agrees Professor King, *"it will transform our understanding of science and the universe. But collaboration between humans and robots will produce better science than either alone."*

In a whirlwind tour of the history of the robot scientist, Professor King set out the advantages of AI systems – such as their superhuman powers around flawlessly remembering facts and learning from vast amounts of data – and explained why science is a wonderful application area for AI. Scientific problems are abstract but involve the real world, problems are restricted in scope, there is no need to know about 'cabbages and kings'. Nature is honest ("unlike economics or games or social networks!") and a worthy object of study.

The concept of a robot scientist is a machine capable of originating its own experiments, physically executing them, interpreting the results and then repeating the cycle. The motivations for creating such a machine range from the philosophical (what is science? Can we fully understand a phenomenon if we can't build a machine that reproduces it?) to the technological. By making science more efficient, robot scientists could increase the productivity of science, be more accurate, work longer, and be easily multiplied. They have the potential to improve the quality of science "by enabling the description of experiments in greater detail and semantic clarity".

The Robot Scientist 'Adam' (2004-2011) was the first machine to autonomously discover novel

scientific knowledge, hypothesise and experimentally confirm it. Adam worked in the domain of yeast functional genomics.

The Robot Scientist 'Eve' (2008-2015, 2015-2019) is the second generation and was originally developed to automate early-stage drug development: active machine learning for Quantitative Structure Activity Relationship (QSAR) learning. The target application was neglected tropical diseases. More recently, Professor King and his colleagues have adapted Eve to work on yeast systems biology, and cancer. Eve's QSAR approach currently uses Gaussian process models, has advantages of being generative and outputting probabilities. There is active learning. Eve has successfully found lead compounds against neglected tropical diseases.

Bio | Slides

10.3 Multi-fidelity Statistical Learning Approach for Organic Molecular Crystal Structure Prediction – Dr Roohoolah Hafizi and Dr Olga Egorova, University of Southampton

The discovery of new crystalline materials is important in many application areas, including healthcare, energy generation and storage. Crystal structure can be crucial for drugs. Conductivity, melting point, dissolution rate, physical and chemical stability, etc are sensitive to molecular packing. The discovery of new crystal forms can be guided by computational methods for crystal structure prediction (CSP), which typically involves a global search of the lattice energy surface, followed by energy ranking of the local energy minima, which correspond to possible crystal structures.

According to Dr Hafizi and Dr Egorova, this method has advantages over experiments: reagents can be expensive, characterisations can be very time intensive and, with Ostwald's rule of stages, the least stable polymorph should be the first isolated in any crystallisation. Completeness is hard to guarantee.

Because of the weak nature of packing forces in molecular crystals, the energy differences between structures are small and energy ranking should be performed using a high level of theory, such as hybrid-functional solid state density functional theory. However, the number of energy evaluations in a typical calculation is prohibitive, and a more efficient way of energy ranking is required.

Dr Hafizi and Dr Egorova explained how statistical learning can be applied to upgrade the energy ranking provided by efficient force field methods. They have explored the feasibility and potential of cost-effective methodology of approximating expensive codes for the purposes of crystal structure prediction.

They start from an affordable atomistic energy model to provide the distinct crystal structures and their energies, then collect more accurate energy data points at various levels using solid state quantum mechanical methods, and use it as a training set for learning the difference between different levels of theory.

Gaussian process regression is applied to learn corrections to the energy differences between different levels of theory, using descriptors of local atomic environments to define similarities between crystal structures. The results demonstrate that high level energy ranking of structures can be achieved at low cost.

They propose further work, including extending the approach on other molecules and

exploring the landscape of the crystal structure – incorporating Bayesian optimisation in the minima search procedure. Bio | Slides

11 Session 8: Contributed Talks

11.1 Isometric classifications of periodic crystals – Dr Vitaliy Kurlin, University of Liverpool

Solid crystalline materials (crystals) have numerous applications from high-temperature superconductors to gas capture. A periodic crystal is an infinite arrangement of atoms or molecules obtained by translating a unit cell (a non-rectangular box) in three independent directions. The Crystal Structure Prediction (CSP) aims to discover solid crystal that is based on a given chemical composition and has several target properties, most importantly the energy of its thermodynamic stability.

The state-of-the-art in CSP has been described as an "embarrassment of over-prediction", because modern software outputs thousands of simulated crystals without identifying only the few most promising candidates for synthesising in a lab. The underlying problem is the enormous ambiguity of crystal representation via conventional unit cells, because many different unit cells can define identical (up to a rigid motion) or nearly identical crystals. Reduced cells compare crystals only exactly (giving an answer yes/no) without quantifying a similarity between crystals in a continuous way.

Dr Kurlin proposes a new classification of crystals by geometric invariants that are continuously changing under perturbations (atomic vibrations of atoms). These invariants will provide a well-defined distance between crystals that can be used for visualising large datasets of simulated crystals by continuously varying a threshold for similarity.

Bio | Slides

11.2 Practical applications of deep learning to imputation of drug discovery data – Dr Benedict Irwin, Optibrium

Dr Irwin, senior scientist at Optibrium, described Alchemite, a novel deep learning method for completing sparse data matrices that accepts both molecular descriptors and sparse experimental data as inputs to exploit the correlations between experimentally measured endpoints, as well as structure-activity relationships (SAR).

He explained that, for a machine learning method to be practically useful in QSAR, it should handle missing values, noisy data, multiple endpoints and data changing with time. The problem is, most algorithms cannot handle missing inputs, pharma data is inherently noisy, input data may not be 'true', models output a number with no context and, with many columns in project data, a model can't be trained for each one. Data evolves as projects continue, chemical space changes, there are activity changes, data sparsity changes (more ADME, less HTS) and uncertainty changes (new assay concentration, finer resolution).

Alchemite, a method for deep multiple imputation, was originally used to design new materials at the University of Cambridge, UK. Optibrium has been optimising the algorithm and applying it to drug discovery data.

Dr Irwin's project sought to compare the accuracy of the Alchemite model to conventional QSAR models; compare models built on all data simultaneously with those built on individual

projects and subsets of data; and evaluate Alchemite's ability to estimate confidence in individual predictions and target the most accurate results

Three sets of models were generated: two Alchemite models of the individual project data sets; a single Alchemite model covering the combined activity and ADME data from both projects; and conventional QSAR models of the individual endpoints. The result was that the single Alchemite model of data for both projects, including biochemical and cell-based activities, and ADME properties, significantly outperforms QSAR models. The single Alchemite model performed equivalently to models of individual projects and subsets of the data and can combine data from multiple chemistries and types of endpoints in a single model. Alchemite can target focus on the most confident and accurate results to use as the basis for decisions.

The conclusions so far are that Alchemite is a practical application of deep learning that can handle missing data and makes the most of extreme levels of sparsity. It provides robust uncertainty estimates on predictions; one model trained for all project data simultaneously exploits assay-assay correlations and is retrainable to handle all stages of a project which changes in time. Alchemite can focus on the most confident and accurate results, and models improve as data is added in a realistic chronological project series.

The practical applications to drug discovery, including pharma-scale collections comprised of up to one million compounds as well as smaller, project-specific data sets, is that the filling in of missing data, combined with the ability to focus on the most confident predictions, guides the selection of compounds and prioritisation of experimental resources in hit-to-lead and lead optimisation.

Next steps are the application to new compounds and data as the project progresses. A follow-on project is considering an additional 874 compounds with many additional ADMET data points.

Bio | Slides

11.3 Dense periodic packings in the light of crystal structure prediction – Miloslav Torda, Leverhulme Research Centre for Functional Materials Design

One of the methods in the design of new materials is to predict the crystal structure of a new compound from its molecular composition. This process involves generating many hypothetical structures based on lattice energy optimisation. A different approach based only on the geometry of a molecule offers the potential to speed up classical crystal structure prediction computations.

Miloslav Torda set out his project's preliminary results with regard to the periodic packing of a geometric representation of the pentacene molecule using Monte-Carlo molecular dynamic simulations.

Around every pentacene atom Torda put 14 points uniformly placed on a sphere with the radius 0.5573/2 then computed the convex hull of the resulting point cloud. The resulting polyhedron was defined by a triangulation with 58 vertices, 112 edges and 168 faces.

The radius was computed from a CSP dataset containing 586 pentacene structures, where the minimum euclidean distance between pentacene molecules within every crystal structure was

computed.

With the input of the polyhedron specifications – vertices, edges and faces – and initial configuration of the system, using an algorithm, the output was density, unit cell parameters, coordinates and rotations of all particles in the unit cell.

Future work will look at the problem that most of the results of optimal packings were obtained by ad hoc methods and Torda would like the packing procedure to be more general: variety of bodies and space groups. Although the objective function is smooth, the constraints are not differentiable. A possible solution: transfer function f (x) to be optimized to a function F(θ) defined on the space of probability measures and then perform a natural gradient descent of F(θ) over a statistical manifold.

Bio | Slides

11.4 Data Science and the Physical Sciences Data-Science Service – Dr Nicola Knight, University of Southampton

The Physical Sciences Data-science Service (PSDS) is a newly funded national research facility to provide access to high-quality curated data resources for UK researchers in the Physical Sciences. Dr Knight introduced the new service and its data resources.

Run by a team from the University of Southampton and STFC, PSDS provides access to data resources within the physical sciences area. The vision is to provide a data hub for data sharing and exchange, to support data-intensive research across the physical sciences (and beyond); provide cost-effective access to high-quality, curated datasets; support a variety of data resources across disciplines; raise awareness, engagement and education of data science; harmonise datasets, allowing users to search across datasets and export customised results; and build collaborations and partnerships with data providers.

PSDS.ac.uk. provides access to the different databases, information about the resources and some training guides. Access is for education purposes in the UK, so users need .ac.uk address.

There are seven separate resources:

- Chemical Property Predictions
- Organic Crystal Structures
- Thermophysical Database
- Inorganic Crystal Structures
- Compound Availability and Suppliers
- Structure and Reaction Database
- Crystallographic Structural Database

Dr Knight was asked about Chembl and she noted that it has a lot of data, is easily searchable, has API access and is well structured – qualities that are lacking in a lot of chemistry datasets. There are some areas of crossover so she hopes to work in partnership with Chembl and EBI and build on their work rather than redo it.

Bio | Slides

11.5 Ellipsoids as a new descriptor for materials – Dr James Cumby, University of Edinburgh

A key challenge in applying machine learning to crystalline materials is to generate materials 'descriptors' that accurately and concisely capture the atomic arrangements in a manner allowing comparisons between different structures. This is particularly true of extended solids, where the lack of a molecular boundary can pose a problem. Many of the existing approaches are based on atom-centred functions such as radial distributions or SOAP (smooth overlap of atomic positions) which, although useful, are not necessarily interpretable. Such methods can also provide huge feature spaces, which have negative consequences for machine learning applications where the amount of data is limited. As such, there is a need for new atom-centred descriptors (applicable to both periodic and non-periodic problems) which result in few parameters.

Dr Cumby's talk focused on short-range coordination environments as the building blocks of materials, and discuss a new method based on minimum-bounding ellipsoids for comparing different bonding environments on an equal scale. The method can also quantify the distortions which are commonly present in such coordination polyhedra, which can lead to many important physical properties such as bulk polarisation or magnetic ordering.

This ellipsoidal approach has currently been applied to metal oxides, revealing an understanding of ferroelectric phase transitions and a potential new area in multiferroic materials, but is more generally applicable to any atomic environment where the geometry of nearest neighbours can be defined.

Bio | Slides

All the conference material is available at https://www.ai3sd.org/conference2019
The presentations can be found at: https://www.ai3sd.org/conference2019/presentations