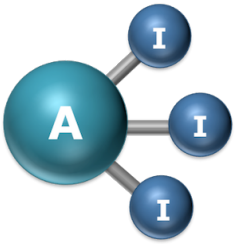

Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery

AI3SD Network+ Launch
05/12/2018
Society for Chemical Industry

Michelle Pauli
Michelle Pauli Ltc

21/01/2019

AI3SD-Event-Series:Report-4

AI3SD Network+ Launch
AI3SD-Event-Series:Report-4
21/01/2019
DOI: 10.5258/SOTON/P0005
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*

Table of Contents

1. Executive Summary

Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery Network+ (AI3SD) launched at the Society of Chemical Industry in London on 5th December 2018 with a large gathering of experts from academia, industry and government.

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

The launch event opened the three-year investigation and featured a series of keynote talks that highlighted some of the opportunities and challenges of artificial intelligence (AI) in these areas.

The **context** in which the Network+ will operate was set out in presentations by Professor Adam Prugel-Bennett, University of Southampton, who explored advances in machine learning; and Professor Michaela Massimi, of the University of Edinburgh, who tackled some of the philosophical questions of what AI means for scientific discovery.

The **opportunities** AI opens up in the field of drug discovery were emphasised by Professor Jackie Hunter, a board director of BenevolentAI, who set out how AI is already transforming R+D to powerful effect. Professor Hunter's colleague at BenevolentAI, Dr Nathan Brown, demonstrated EvoChem, a multi-objective optimisation AI drug design tool with drug-like property prediction. Gábor Csányi, professor of molecular modelling at the University of Cambridge, highlighted how his team's work in force fields is an excellent field for machine learning.

However, there are, of course, **challenges**. Professor John Overington, CIO of the Medicines Discovery Catapult, highlighted the issue of strategies for collating the colossal amounts of data needed for AI – and the level of errors therein. There are also cultural challenges related to the need for interdisciplinary, cross-functional working – a development the Network+ regards as essential if breakthroughs are to be made – and the need to bring together different datasets and the corresponding requirement for interoperability and standardisation.

An exciting three years lies ahead for the Network+ as it seeks to meet these challenges and exploit the opportunities AI offers. The Network's principal investigator Professor Jeremy Frey and coordinator Dr Samantha Kanza outlined the planned **activities of the Network+.** In dialogue with members, a full programme is being devised that encompasses horizon scanning, literature surveys, networking, hackathons and a major funding call in early 2019. Full details will be on the events page of the website.

2. The Network+

Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery Network+ (AI3SD) launched at the Society of Chemical Industry in London on 5th December 2018 with a large gathering of experts from academia, industry and government.

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

The launch event opened the three-year investigation and featured a series of keynote talks that highlighted some of the opportunities and challenges of artificial intelligence (AI) in these areas. There was also discussion and networking, in line with the Network's aim of enabling and facilitating AI and physical science researchers to communicate, collaborate, develop new ideas and discover new processes in support of a sustainable world.

The Network's interdisciplinary approach is critical. It is essential that both scientific and technological human expertise are brought together if we are to make significant advancements in the field of scientific discovery. Equally, both human and artificial intelligence bring different qualities and strengths to the field. These differences must be recognised and both kinds of intelligence harnessed together – each intelligence augmenting the other – if the greatest benefit is to be achieved.

The Network's principal investigator Professor Jeremy Frey and Network+ coordinator Dr Samantha Kanza outlined the planned activities of the Network. Advisory board members were also present to answer questions.

This report sets out the context for the Network's investigations, the opportunity offered in the specific area of drug discovery, the challenges faced and the Network's next steps.

2. The Context

2.1 For AI

The pace of progress in AI over the last few years has been remarkable. This launch took place in the month when the Google's DeepMind announced that its latest AI program, AlphaFold, had taken a significant first step in solving the 'protein folding' problem, coming top of a competition to predict the 3D shapes of proteins.

At the launch, Professor Adam Prugel-Bennett touched on these developments in his fascinating overview of machine learning, particularly the progress DeepMind's AI has made in the Chinese board game Go in the last three years and how it exemplifies the field's sudden, recent change in pace.

AlphaGo hit the headlines in 2016 when it beat the world's best players at a game previously believed to be too complex for a computer to outwit a human. It was followed by AlphaGo Zero, which last year not only out-played AlphaGo but taught itself to do so without ever having seen a human play the game. Now it can also beat human players at chess and the Japanese game Shogi.

However, while these game playing developments are impressive demonstrations of the power of AI, according to DeepMind CEO Demis Hassabis, "it's never been about cracking Go or Atari, it's about developing algorithms for problems exactly like protein folding… a fundamental, very important, real-world scientific problem." It is these very real-world, important scientific problems that AI3SD is also seeking to address.

Whereas in the 1950s AI was envisaged as a machine that could think, feel and sense like a human (a concept known as 'general AI'), Professor Prugel-Bennett was keen to reject the myth that AI is an impeccable logic machine. He emphasised instead that it is "just reducing errors", recognising pattern sets very well and using those to make judgement by reducing errors on some sets. There are certain areas where making fewer errors is clearly of benefit, such as fraud and spam detection, self-driving cars and, of course, medical diagnosis.

Further evidence of the pace of progress being made with this kind of pattern recognition, and image classification in particular, comes with the ImageNet large scale visual recognition challenge. In 2010, the competition's first year, every team got at least 25% wrong; by 2017, 29 out of 38 teams got less than 5% wrong, described by Professor Prugel-Bennett as a "super human performance".

Early work in deep learning used supervised learning where there were labels for the data. In the last three years the excitement has been around unsupervised learning, working with the data alone and learning patterns in data. One of the techniques for doing this is
Generative Adversarial Networks (GANs), the other is Variational Auto-Encoders (VAEs). With object recognition now mainstream, the machine learning community is looking at more demanding tasks, including visual question answering, which involves understanding both natural language and images.

"There is no end in sight – there are regular significant breakthroughs," concluded Professor Prugel-Bennett. "If anything, the pace of progress is increasing. We are just at the beginning."

2.2 For Scientific Discovery

What does AI mean for scientific discovery? This is one the key threads for AI3SD. The use of learning algorithms has the potential to revolutionise the way science develops. Central to the scientific method is the Popperian premise of hypothesis falsifiability, which underpins much of the theoretical background of chemistry, physics and biology. Algorithmic systems offer an alternative phenomenological approach to discovery and prediction, one that is driven by statistical correlations of data. This burgeoning capacity raises fundamental philosophical questions about the

nature of innovation, discovery and prediction, and challenges the very concept of what 'theory' means.

Michela Massimi, professor of philosophy of science at the University of Edinburgh, tackled some of these philosophical questions in a presentation which ranged across centuries and disciplines. While recognising the common desire to see the 'story' in scientific discovery – she referenced Newton's apple and Einstein's beam of light – she urged a different way of thinking about scientific discovery, where progress is measured by ruling out live possibilities, by excluding with high confidence level (95%) certain physically conceivable scenarios and mapping in this way the space of what might be objectively possible in nature.

In 1962 Thomas Kuhn argued that science proceeds through three stages: normal science, crisis and scientific revolutions. He believed that scientific revolutions were triggered by an increasing accumulation of anomalies that the existing scientific paradigm is unable to solve. This Kuhnian view of how science progresses and evolves, via increasing problem solving, fits well with contemporary debates about the use of AI to facilitate scientific discovery, especially in areas such as chemistry and the biomedical sciences.

"I think this is exactly what goes on here in the application of AI to the discovery of new molecules where the exercise consists in sweeping large portions of the chemical space, so to speak, and explore so far unconceived possibilities for chemical synthesis and the engineering of new materials," said Professor Massimi.

"Being able to explore large portions of this huge and very complex space of possibilities is one of the areas where progress can and is being made with AI. I think philosophers of science must pay more attention to this new way of thinking about progress and discovery in science. Even if we are still away from finding the final cure for cancer, the ability to deploy AI to explore the space of possibilities in a more effective way is progress enough in science, and something worth celebrating."

3. The Opportunity: to transform drug discovery

According to BenchSci, in December 2018 there were 117 start-ups using AI in drug discovery across the life cycle, from aggregating information to optimising clinical trials.

One of these is BenevolentAI, which has invested more than $200m in its platform and works across the full range of development activity. Professor Jackie Hunter, one of BenevolentAI's board directors, set out clearly why AI is set to transform this sector: the sheer unsustainable inefficiency of the drug discovery industry. With only 6% of molecules currently reaching market, what other sector would tolerate a failure rate of 94%? It takes 12-15 years to get a drug to market and most people in the industry never work on something that makes it through to clinical trials, let alone patients.

She outlined how AI will reduce costs and improve success rates, describing the "tsunami of evidence" of the last few years and the inability to harness such data at scale until AI is applied to that space. Using machine learning will reduce timelines in all phases of drug discovery: hypothesis generation, drug target validation, lead discovery and lead optimisation. In drug development it will have equally powerful effects, from rapid interrogation of study data, identification of novel biomarkers and enhanced mining of clinical data for patterns of response to better real-world outcomes monitoring.

Professor Hunter made the point that "for an evidence-based industry, drug discovery uses very little evidence" and contrasts that with Benevolent's platform which ingests, 'reads' and contextualises vast quantities of information drawn from written documents, databases and experimental results, infers new knowledge and relationships through deep learning methods and creates a unique knowledge graph and unified data for analysis and machine learning. The result is "infinitely more deductions and inferences across disparate, complex data sources, identifying and creating relationships, trends and patterns in biological networks and chemical space than would be possible for a human alone". She pointed to a case study on the neurogenerative disorder amyotrophic lateral sclerosis where the system flagged around 100 existing compounds as having potential. Five were selected to undergo tests using patient-derived cells, of which four were found to have promise and one was shown to delay neurological symptoms in mice.

This breakthrough is a clear example of the importance of bringing both AI and human intelligence together to find new ways forward. While AI can be very useful in helping us predict and discover potential drug-target paths, including those that a human is unlikely to ever reach because of the sheer quantity of data processing involved, human intelligence is currently unparalleled in being able to see things that computer systems cannot yet detect, and, crucially, be able to pick apart some of the AI's predictions to unravel which ones are viable.

Professor Hunter also pointed to cancer as a pathology where AI could have significant impact by offering more sophistication than a human in assigning categories to tumours, freeing up scientists to concentrate on the more rare and difficult cases and do the 'outside the box' thinking that machine learning cannot.

AI, she said, is "already transforming R+D and the pace will only accelerate. It will require cultural change, different ways of working, different ways of funding and new business models."

Professor Hunter's colleague at Benevolent, Dr Nathan Brown, demonstrated EvoChem, a multi-objective optimisation AI drug design tool with drug-like property prediction. The output ranks a list of AI-designed compounds, along with their metadata. Chemists can then take these molecules to work on live design, alongside modellers, AI scientists and biologists. Once an extracted molecule has been re-designed, it is put back into tool and re-scored for a read-out of physical properties. An AI-powered retrosynthesis tool then describes how to make it.

Meanwhile, Gábor Csányi, professor of molecular modelling at the University of Cambridge, highlighted how his team's work – in generating force-fields that are as accurate as high level quantum mechanics calculations, all the way from elemental compounds to small molecules – is an excellent field for machine learning[1]. His team has developed an open-source GAP-framework docker, which is available for download, and he gave a number of examples of work done on different systems:

- Structure and growth mechanics of amorphous carbon films
- Phosphorus (many different structures)
- TiO_2 (many crystalline forms and amorphous phases)
- GeSbTe storage materials (trying to find the medium range order)
- CH_4 (this is difficult to model without knowing the level of quantum mechanics needed)

Professor Csányi's current work is tackling protein-ligand binding and he gave a brief example of 3-BPA binding to leukotriene hydrolase, using custom force-fields to calculate the binding.

4. The Challenges

One of the challenges for machine learning is the availability of the training data. The question of strategies for collating the colossal amounts of data needed for AI was picked up by Professor John Overington, from the Medicines Discovery Catapult (MDC), a national facility connecting the UK community to accelerate innovative drug discovery. He used his experience with ChEMBL, the world's largest public primary database of medical chemistry data, to delve into the challenges and opportunities of using free, large-scale datasets for AI training and application data, touching on the 'reproducibility problem' along the way.

He was frank about the level of errors in the public datasets he has been involved with – ChEMBL (2.3m compounds, an open-data API is available), SureChEMBL (the public chemical patent resource with 18m structures generated via name-recognition, and available as a client feed) and Unichem (a single chemical integration source). For instance, errors run to 5% of structures; 2-3% of targets; 1% of activity values. There is also variability in the data (the underlying 'reproducibility problem') – about the same 10-fold difference for different orthologues as for different labs, a five to 10-fold variance in cell line data, and the possibility that about 1-2% of compounds from suppliers may not be correct. However, he was clear that, nine years ago when ChEMBL was set up, it was a good design decision to focus on collecting all the data, in case it would have a use in the future and could be cleaned up accordingly.

Working with AI in this field also brings up cultural challenges. Getting rid of silos, whether data or organisational, is one. Cross-functional working is essential. BenevolentAI works in cross-functional teams, with technologists and scientists working side by side, and Professor Hunter admitted that this could be

[1] We thank Stuart Newbold for the report of the meeting he wrote for the CICAG Newsletter for details of this talk http://www.rsccicag.org/index_htm_files/CICAG%20Newsletter%20Winter%202018-19.pdf

uncomfortable to begin with – "scrums and sprints and agile processes!" – but that it was essential when working towards shared goals and objectives.

The need to bring together different datasets and the corresponding need for interoperability and standardisation is a challenge that urgently requires attention and agreement from the community. There is also an increasing imperative to change structures so that people receive reward and recognition for maintaining high quality datasets and making them available.

Further challenges were briefly raised in the Q+A section of the event. These included:

- Using AI simply to do the same things but faster is no longer good enough. Are we making the best use of the theories that we have got, and when and where do we use the appropriate tools? There are professional responsibilities to the scientific method and a need to teach it to new undergraduates.
- The private sector is leading the way with surprising and inspiring applications of AI. However, not enough detail is shared to understand what exactly it is that AI contributes to a project that makes it different or unusual. Detailed project case studies that can be publicly shared would be very helpful.
- The chemistry community should be looking at making an order of magnitude change – real breakthroughs rather than incremental change, an aha moment rather than a solution within a known problem. Has AI produced a new theory as opposed to a hypothesis within an existing theory? What needs to change to make that happen?

5. Next steps for the Network+

An exciting three years lies ahead for the network. In dialogue with members, a full programme is being devised that encompasses horizon scanning, literature surveys, networking, hackathons and a major funding call in early 2019.

Events currently planned include:

- 7 Jan – funding call: the first network pilot project funding call will be formally announced on the AI3SD website. The call will be broad but with areas of priority.
- 22 Jan – Network Town Meeting (at the SCI) to discuss the funding call
- 20th March – 2nd Advisory Board Meeting
- 6th February - Molecules, Graphs and AI Workshop – Ageas Bowl, Southampton
- 6 March – AI in Drug Discovery and Drug Safety, Medicines Discovery Catapult, Cheshire
- 1st May - Semantics and Knowledge Learning for Chemical design – Solent Conference Centre, Southampton
- 17-19th June – Sheffield conference on Chemical Informatics
- June – Leeds symposium on prediction of complex reaction networks with Dial a Molecule Network

- 17-18 July – AI for Reaction pathway prediction (with the Dial a Molecule network)
- 13th September – Hackathon with the Pistoia Alliance
- September – Novel Mathematics for Machine Learning: Topology, Holography for Chemical Discovery (with Joining the Dots EPSRC Project)
- 17th October – AI, Science and Food. Joint workshop with Digital Economy Internet of Food Things Network+ (IoFT+)
- October – second funding call
- November – AI3SD Network Conference, which will include the AI for Scientific Discovery feasibility projects

Full details will be on the events page of the website.

Other workshop areas that are being considered are:

- Statistical Mechanics and Machine Learning
- ML/AI & Molecular Dynamics – from force fields to accelerated prediction, non-equilibrium calculations
- Quantum Computing and Quantum Chemistry
- Intelligent Labs – AI in the Lab
- Re-enforcement Learning: Go for Chemistry
- Provence for and Trust in AI models
- Augmented vs Artificial vs Automation
- AI for Science for Good

Meanwhile, all those interested in the work of the Network+ are warmly invited to become members by joining the AI3SD mailing list. To do this, send an email to listserv@jiscmail.ac.uk

Subject: Subscribe
Message: SUBSCRIBE AI3SD Firstname Lastname

You will then receive a confirmation email. We will send out news and updates relevant to our themes and periodic updates relating to the opportunities offered by the network.

You can also follow the Network+ on Twitter: twitter.com/AISciNet and or join the LinkedIn Interest Group.