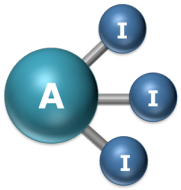

Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery

AI and Machine Learning for Chemical Discovery and Development
18/07/2019 - 19/07/2019
AI^3 Science Discovery Network+, Dial-a-Molecule, Directed Assembly Network &
University of Leeds
Weetwood Hall, Leeds

Dr Bao Nguyen
University of Leeds

30/07/2019

AI3SD-Event-Series:Report-13

AI and Machine Learning for Chemical Discovery and Development
AI3SD-Event-Series:Report-13
30/07/2019
DOI: 10.5258/SOTON/P0017
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 and Machine Learning for Chemical Discovery and Development
Organisers	AI^3 Science Discovery Network+, Dial-a-Molecule, Directed Assembly Network & University of Leeds
Dates	18/07/2019 - 19/07/2019
Programme	Programme
No. Participants	27
Location	Weetwood Hall, Leeds
Committee	Dr Bao Nguyen - University of Leeds, Gillian Smith - Dial-a-Molecule Network, Dr Samantha Kanza - AI^{3}SD & Martin Elliott - Directed Assembly Network

2 Event Summary and Format

This was a joint networks event between AI^{3}SD, Dial-a-Molecule, Directed Assembly Network and the University of Leeds. This was a residential event aiming to bring together stakeholders with different backgrounds, e.g.academic/industry, researchers/data owners, and chemists/engineers/computer scientists, to discuss applications of AI and Machine Learning in Chemical Discovery and Development. The event was made up of some priming talks to stimulate discussion, and a series of structured discussion sessions over the two days to form a general consensus on some key objectives and milestones to deliver the promised impacts of these important tools within the remit of the three networks.

3 Event Background

This joint event was created between the three different networks and the Unviersity of Leeds due to the joint interest in these research areas. This event forms part of the AI3SD Event Series, which aims to bring people together around important areas of using Artificial Intelligence for Scientific Discovery. The attendees of this event were selected through short applications to ensure that there was a good balance of interest among them, and that there was a roughly equal representation from the different networks and research areas.

4 Introduction

We went through the introduction of each attendee and his research interest. BN gave some context to the aims of the workshop, based on the Made Smarter report and outline the wide range of potential application of AI/ML to chemical research.

5 Priming Talks

- AI technologies - Professor David Hogg, University of Leeds
- Learning Chemistry with Machines - Professor Jonathan Goodman, University of Cambridge
- Humanity v the machines - Dr John Mitchell, St Andrews University
- Directed Assembly Grand Challenge - Martin Elliott, Cranfield University

6 Discussion on Areas of Application

The discussion was directed at targeted areas and key challenges for wider application of AI/ML to chemical research. We started with reflection on what chemists are good at, and what chemists are not. This was based on the consensus view that AI/ML should be used to assist chemists, not to replace them. We concluded that chemists are good at:

- Normal retrosynthetic analysis (but habitual use of certain reactions is a problem)
- Recognising large trends
- Predicting standard reactivities
- Intuitive extrapolation

The attendees agreed that chemists/human are poor at:

- Predicting complex effects of small changes or properties of mixtures and blends
- Handling large volume of data
- Quantifying observations and predictions
- Covering chemical and operational space systematically
- Publishing data in useful formats
- Including sensors/detectors for data

Based on this discussion, a number of potential research areas for discussion in Day 2 were suggested by the attendees:

- Data preparation (experimental errors) - dealing with bias - access data - standard
- Smaller datasets
- Confidence evaluation/benchmarking/generating chemical understanding
- Representation of molecules and reactions/interactions
- Route design with constraints/ranking/reaction outcomes
- Molecular discovery with multiple objectives
- Reaction/reagent/catalyst optimisation/impurities
- Predicting crystallisation/isolation
- Engagement between disciplines

These were followed by interest identification of research areas over lunch period and voting later on the day.

7 Discussion on AI/ML methodology

The discussion was based on three areas.

7.1 Method/problem suitability/benchmarking standards

The difficulty in providing benchmarks across the field was highlighted. However, a small number of ***high quality benchmarking datasets***, which are common in the AI/ML community and may be hosted by AI3SD, will be well received in this area of research. These will lower the barrier for researchers from other disciplines, e.g. mathematics and computer science, to joint our field. A common issues in comparing studies with similar objectives is disparity in chemical space coverage of the training sets, which prevents fair comparison. Thus, ***measurement of diversity in the dataset*** must be included. The approach and consideration toward preparing benchmarking datasets will set the standard for researchers and reviewers in this field. Going forward, community-built data standard, e.g. Stepstone project, should be widely adopted as a requirement for grant funding and publication in chemistry journals. This will need leadership and vision from UKRC and the RSC.

Benchmarking and metrics tools to compare models are also required in both academia and industry. An example of these is GuacaMol, an open source python package developed by BenevolentAI for benchmarking of models for de novo molecular design.

While it is appreciated that each research question will require adjustment to the AI/ML methodology, the attendees suggested that some guidelines/standard practice will be helpful for newcomers to the field. These include weighing methods for noisy data, recommended on descriptors: datapoints ratio, balancing between exploration and cost, and standard pairings between problems and AI/ML methods.

7.2 Dealing with non-perfect data, chemical bias, and causal relationship

The attendees generally agree that in chemical context, the current limitation is the data rather than the AI/ML methodology. Due to bias in publication system, the available data in the literature can be incomplete (focused on yield, which consists of several types, instead of reactivity and kinetic data) and biased (focused on successes and not negative results). Data collection by chemists has been non-systematic, sometimes without context (e.g. reactor type, temperature, reaction time, purity of reagents and solvent, etc.) which leads to the need for curation. Certain progresses have been made in these areas, e.g. Inchi code and Stepstone project. However, there is a clear need for leadership in ***setting the standards and formats for data reporting*** (manuscript and ESI), including standard language, metadata and ontology, which will lead to usable data for AI/ML and open knowledge exchange. The adopted format also needs flexibility for expansion as we do not know what data we need to capture in the future. ***Motivation for data deposition***, beyond the goodwill of the research, needs to be provided.

Despite all these drawbacks, there is a large volume of useful data, e.g. physical properties and reactivity, in the literature and tools to make some machine readable. Reaction data can also be useful as starting points if due consideration is given to its inherent bias and incompleteness. These are a potentially valuable sources of data, which do not have complicated legal barrier to assess. Further ***investment in data extraction tools and data curation*** is important for the field to move forward. Some funders do fund ***community-owned-database*** development, but not the EPSRC. Perhaps exception can be made given its importance.

7.3 Confidence and uncertainty

This topic generated a great deal of interest in the room. The consensus is that ***confidence and uncertainty*** are critical and must be put front and centre in the use of AI/ML in chemical research. This will allow informed decision with ***expert interpretation of AI/ML output***. In the context of synthetic route design, they can be used to rank predictions and establish safety margin. Linking uncertainty to variation in input will also provide critical understanding of the process. This is particularly important when evaluating outliers. The required accuracy of the model, however, is application dependent and maximum accuracy is not always the objective.

It is important for researchers to recognise that the maximum accuracy of a model should not exceed the amount of noise in training data. There is a cultural issue with reporting errors in experimental data in synthetic science. Control studies, including DoE, may be employed to evaluate the confidence in the model when required. Background knowledge may also be included in the model to address uncertainty and smaller datasets. Future model may see merging between mechanistic models and AI/ML models toward ***adaptive models*** which can improve as strategic data become available.

8 Discussion on Data

What is 'good' data in your area of interest? There is a diverse range of need depending on the application and given that any data is better than none, no specific barrier was proposed. Generally metadata including context and experimental details is required.

What to do about old data? This was covered above.

Can industrial data be shared? Yes, under certain circumstances. Lhasa has been acting as a trusted intermediary to help company ***sharing information and models based on in-house data***. Companies can share info on patented molecules. These may be further facilitated with encryption and ***standardised data structures and sharing agreements***. It is important for Industry to see solid value proposition and to recognise value in the data they hold through the right ***proof-of-concept projects***. Physical properties data are probably easier to share compared to reaction data.

9 Topics identification through voting

After the discussion on AI/ML methodology and data the topics proposed during lunch break (using 3 post-it's per academics) were put to an informed vote. Those which have more than 1 post-it's were voted on and the top three were taken forward as discussion topics for Day 2. The results are summarised below:

- Model interpretability - 11
- Route design/ranking with constraints - 9
- Data mining/access/standard - 7
- Reaction/process outcome prediction and optimisation - 15
- Predicting reactivity and properties - 13
- Molecular discovery - 9
- Representation of molecules and interaction - 10
- Bias in data and capturing chemical intuition - 5

On Day 2, the workshop was divided into three groups to discuss each of the selected topics in three 1-hour sessions. Attendees were able to freely move between topics as they wish. The guidance was to focus on the challenges and milestones in these areas of research.

10 Discussion of Milestones for Research Area 1: Model interpretability

Three prompts were used to guide the discussion:

- Accessible tools/protocols to gain understanding
- Relevant descriptors/representation of molecules/reaction/ interaction
- Evaluation of confidence

The attendees highlighted the need to understand how the AI/ML model works in order to get the best prediction and balance any inherent bias in the data. AI/ML methods in research and publications should be subjected to the same evaluation as human led studies. The interpretability of the model is essential in gaining confidence in the model, refining it, knowing its limits and understanding the relationship between physical properties of the system. To ensure these, the training data should be relevant and randomised. AI/ML models are built for interpolation and extrapolation must be approached with due consideration. ***Automation and high throughput experimentation*** should be used to assist model interpretation, refinement and validation. Furthermore, descriptive models are sometimes better for interpretability and bias analysis than predictive models.

The use of ***standard datasets*** for comparison between methods is highly valued. In this context of gaining confidence, evaluation of AI/ML against real world problems, instead of idealised ones, is important. Propagation of errors needs to be considered in this context. It is important to understand the research area, its rules and biases, and to ***interpret the research question mathematically*** for AI/ML. This includes multi-objective optimisation, which is common in many chemical research areas.

11 Discussion of Milestones for Research Area 2: Reaction/ Process outcome prediction and optimisation

Five prompts were used to guide the discussion:

- Predicting outcomes (product, impurities, selectivities)
- Optimising with gradient and without gradient
- AI/ML methodologies for different data types/sizes
- Cumulative/adaptive models which grow with data?
- How to approach reactivity and selectivity

Questions were raised on the benefit of AI/ML over ab initio/DFT methods in this context, particularly when one needs to make ***predictions outside the chemical space*** covered by the training data. A suitable method is suggestions by AI/ML, which can be evaluated and decided on by chemists. Again, different level of accuracy in prediction or different questions can be used in different context, e.g. is the reaction going to be clean or messy? A more physical approach, merging mechanistic models and AI/ML models, is to ***predict relative reactivity*** of different functional groups in starting materials, which can be used to predict reaction success as a meter/index. Computational reaction network is another approach using AI/ML to predict organic reaction outcomes.

An added benefit of employing AI/ML in this field is that green chemistry constraints can be built into the system. Development of relevant descriptors for these is still a challenge in this field and ***standard approaches to descriptors development***, including stereochemistry, will be useful for new researchers. Reaction conditions and operational space are often limited by non-chemical factors. Training data do not necessarily reflect reactivity and future ***collection of kinetic/reactivity data***, e.g. HPLC trace at t1/2 with internal standard, in some form, in standardised formats, is critical. Lack of ***negative results*** in the literature is also an issue, which should be addressed going forward and through mining open-access theses.

A holistic and multi-objective approach to include workup, purification and crystallisation may be of more importance in production context. This may include sustainability, yield, cost, robustness, by-product formation and optimisation of a global route instead of individual steps in a synthesis. Transferability of process between reactors/sites is a common problem. Mixtures of solvents.

Given the important of this area of research, community challenges and competitions in combination with benchmarking datasets may significantly speed up its development.

12 Discussion of Milestones for Research Area 3: Predicting reactivity and properties

Five prompts were used to guide the discussion:

- What properties are important and challenging?
- How to represent biological activity?
- Will it work in formulation?
- Accuracy and cost vs first principles techniques
- Reliable benchmarks

Solubility dominates the discussion on chemical properties. This includes ionic compounds, salts, non-aqueous solvents, solvent mixtures. These are linked with formulation, co-crystals an aggregation issues in purification. ***Formulation***, in particular, still heavily rely on experience at the moment. Descriptor development in this area must engage with experts to capture their approach through a theory-guided selection. Another property of interest to chemical development is stability of compounds, e.g. APIs, against light, oxygen, water and themselves. Underpinning many of these predictions, including solid phase/material properties, is ***crystal structure and lattice energy***, which is still currently not accurately predicted. Properties of ionic liquid is another area where AI/ML may help.

Predicting ***biological activity*** is even more complicated, due to its multi-facet and multi-stop nature, e.g. oral availability, transport across membranes, metabolism, protein binding, clearance rate, etc. Minor conformers must be considered. Mathematical description of the binding process, e.g. binding constant, structural and functional changes, is non-trivial. However, data availability and quality is less problematic in this area. The attendees suggested the use of AI/ML to develop standardised fingerprint tests, improve hit rate, and generating potential APIs from fragment data will be important. Part of these are being carried out by Benevolent AI. Some biological targets, however, are more difficult due to lack of structural data. ***Controlling protein-protein interaction*** is still currently very difficult dues to the number of possibilities, but is perhaps an important area for AI/ML in the future.

Most of these targets have been identified for decades. First principles techniques have not been able to effectively address them, but AI/ML has provided success in certain cases such as predicting genotoxicity.

13 Discussion on Wider Engagement

Three prompts were used to start this discussion. Lhasa presented a short summary of their experience working in this field before wider discussion between all attendees.

- Missing expertise
- Training pipeline
- Wider perspectives on application of AI/ML to chemical research and development

The workshop recognised the lack of female attendee. Wider engagement with AI/ML specialists outside chemistry has also been difficult. There is no clear undergraduate training pipeline for future scientists with sufficient training in both chemistry and data science. A number of activities were proposed to address these:

- Summer school (perhaps organised by AI3SD) to provide training to those who wish to be involved in this field and promote standard tools/benchmarks.
- Secondments for data scientists and chemists.
- Community hub for data and code and additional training in the style of 'software carpentry' and 'data carpentry' websites/communities.
- Showcase data/model and competitions to improve trust in the methodology from more traditional chemists.

14 Provisional Roadmap/Milestones

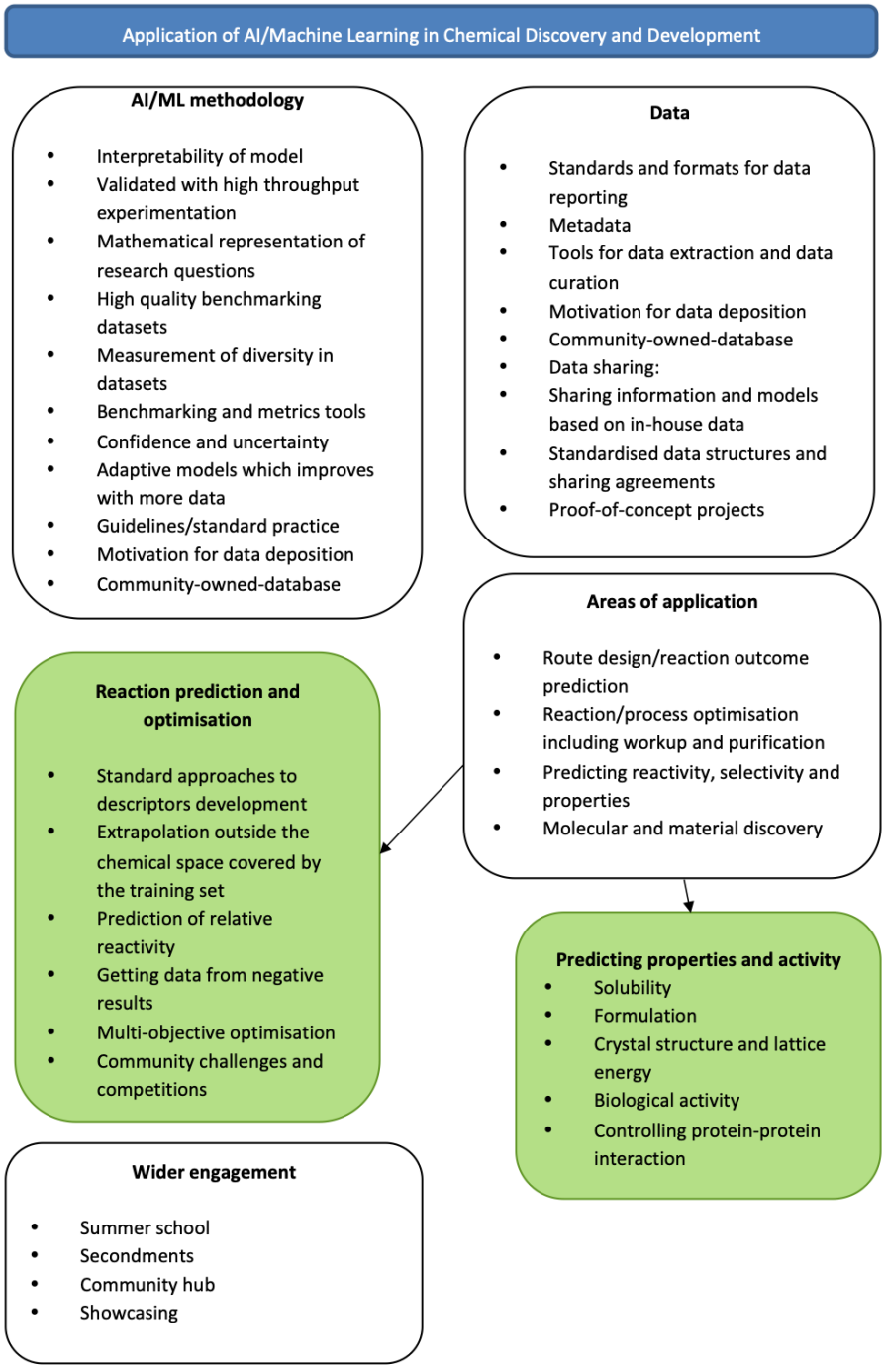

Figure 1: Provisional Roadmap/Milestones

15 Related Events

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