# Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery

Molecules, Graphs & AI Workshop
06/02/2019
The Ageas Bowl, Southampton

Dr Nicola Knight
University of Southampton

09/02/2022

AI3SD-Event-Series:Report-6

Molecules, Graphs & AI Workshop
AI3SD-Event-Series:Report-6
09/02/2022
DOI: 10.5258/SOTON/P0006
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 | Molecules, Graphs & AI Workshop |
| Organisers | AI$^3$ Science Discovery Network+ (AI$^3$SD) |
| Dates | 06/02/2019 |
| Programme | Eventbrite Programme |
| No. Participants | 30 |
| Location | The Ageas Bowl, Southampton |

# 2 Event Summary and Format

This event was one of the first full day workshops hosted by the AI3SD Network. It was hosted by the AI3SD Network at the picturesque Ageas Bowl, Southampton. The workshop was designed to explore the ways in which molecular graphs can be used to drive property and other predictions using Machine Learning and other AI techniques. The programme was made up from a number of keynote talks ranging on the applications of Graph Theory in a variety of areas from biosciences to machine translation. The keynote talks were followed by discussion with the participants and two sessions group talks on a number of themes surrounding the application of graph theory. Refreshments were provided throughout the day which gave plenty of opportunities for attendees to network and talk further about AI topics. The day also concluded with a drinks networking session.

Figure 1: The picturesque Ageas Bowl Southampton

# 3 Event Background

The AI3SD Network aims to drive progress within Machine learning and Artificial intelligence in scientific discovery. Their events are designed to be a platform for facilitating discussion and collaboration within this area. This workshop is the first event in a set of planned workshops covering different topics on the interface of AI in scientific discovery.

# 4 Talks

The event was started off with an introductory talk from Professor Jeremy Frey, Principal Investigator of the AI3SD Network. Frey set the scene for the meeting within this talk with an introduction to the network, the purpose of the meeting and an introduction of each of the participants at the event. This initial talk highlighted that the meeting should prompt discussion around the application of graph theory and formulate ideas that the network can take forward in plans for other events and sessions.

## 4.1 Keynote 1: Unsupervised, multiscale learning through atomistic graphs: From molecules to systems

Professor Sophia Yaliraki (Imperial College London) spoke next, with the first of the keynote sessions. This talk focused on Yaliraki's use of graphs in unsupervised learning and what graphs mean to Yaliraki. The talk took the participants through some methods and applications of graph theory in chemistry. Yaliraki touched on some of the motivations behind using graph theory and some considerations when using graph theory. Yaliraki highlighted that certain approaches to graph theory may not be suitable for all applications and that the different graph constructions methods can produce different answers.

Yaliraki also covered a few examples of applications including healthcare data, twitter analysis and allosteric communication. This last area was covered in more depth, discussing recent developments in Yaliraki's use of graph theory in identification of allosteric sites and allosteric communication. This particularly highlighted the speed with which these methods can be applied, having computations that are on the timescale of seconds.

## 4.2 Keynote 2: Inference from Outliers

Professor Mahesan Niranjan (University of Southampton) gave the second keynote talk with a presentation about his experiences with the applications of machine learning and inference that can be made from outliers. Niranjan began with an outline of his research background and the theory behind machine learning including data inputs and dimensionality reduction.

The talk then moved on to discuss some of the applications of machine learning and the identification of outliers. The applications covered a breadth of areas from machine translation to chemical solubility. However; across all of these topics the goal remained the same, to make accurate predictions and improve knowledge about the problem. Niranjan discusses that it is possible to learn a lot about the problem from data points that are systematic outliers, allowing researchers (with the help of domain experts) to gain knowledge about the underlying chemistry or biology behind the problem.

This talk also covered the low rank approximation method to identify outliers developed by Niranjan which is being incorporated into current research including work on gene expression at multiple levels.

### 4.3 Keynote 3: Source-and-sink models for molecular conduction

Professor Patrick Fowler (University of Sheffield) completed the keynote talks with his discussion on the use of graph theory in molecular conduction prediction. This talk focused on the Fowler's research into the source-and-sink potential model (SSP) which models for the conductivity of a molecule. Fowler discusses Hückel theory and the similarities in concepts between Hückel theory and graph theory, followed by expansion of the insight that can be gained from simple Hückel theory. The matrix applications were discussed for SSP showing how the different connections through the molecule can be modelled and how fragments can be used to calculate the properties of a larger molecules.

Fowler also discussed the selection rules that govern whether a molecule will conduct or not and how the different cases can be classified to show the conductor/insulator properties of a molecule. This work has been extended to all pi molecular devices where there are 81 different cases, currently all but 2 of these cases have been resolved.

## 5 Working Group Discussions

### 5.1 Initial Breakout

In between keynote sessions and initial breakout discussion session was held to get conversation flowing and identify topics that could be discussed further. Three groups were created with broad topics to initiate the discussion. These topics were: problems and questions of graph theory in crystal structures, topology and molecular representation. These breakout discussions generated a lot of different questions and ideas which were captured through the use of posters.

Figure 2: Deep in discussion during breakout groups

### 5.1.1 Graph theory in crystal structures

- Goal: want to identify patterns
- Universal descriptor of crystal structures

  Is this realistic as a goal

  Will depend on what we want it for

  As a starting point for iteratively optimising descriptors for a given problem
- How does property of interest define the graph/descriptor we need?
- How do we explore the space of possible descriptors?

### 5.1.2 Topology

- Networks from Molecular dynamics
- Identification of conformational space
- Labelled graphs
- Modelling water networks
- Energy landscapes (identification of minima + global minima)
- Non-planar graphs
- Symmetry in molecules.

### 5.1.3 Molecular representation

- Representation
- Non-linear interpolation between molecules
- New molecules in chemical space (functional)

  Synthesizability
- RNN on SMILES strings
- Conductions on graphs
- Molecule/target interaction matrix

## 5.2 Second group discussion

In the afternoon a full group discussion was held which covered the points raised in the first breakout session, and following this, topics were decided for a second breakout session for further discussion on four topics:

- Chemical / descriptor space
- Graph representation
- Topology & Topography
- What do you want out of machine learning?

These separate groups covered a huge variety of topics in their discussion with some of the questions and areas of discussion highlighted below.

### 5.2.1 Chemical space / descriptor space

- Distances in chemical/property space can be very different to 'distances' in synthetic space

  How do we measure distances in chemical space, taking into account target property?

  Do we need to be able to define distances?

- How do we navigate?

  What are the best strategies for navigation?

- What is the best approach to optimisation?

- How much control do we have in what we can make?

  E.g. chemical bonding vs. conformer, molecule vs. polymorph.

  Are these (conformers) fully controllable?

  We are getting better (experimentally) at isolating unusual + metastable structures

- Synthesisability

  We want navigation to give us accessible compounds

  Reliably predictable?

### 5.2.2 Graph Representation

- Graph representation is very important

- Can get very different results from different representations

  Not all representations are suitable for all problems

- Need to consider what do you actually learn through the application of graph theory to that specific problem.

- How do you encode multiple features?

  Graph stacking (multi layer)

  Single matrix graph with edge functions

- How do you input into ML methods?

  Graph convolution neural networks – can these be applied to chemical molecules?

  Can bias the ML if the representation encodes features

  Ideally want the simplest form of representation for ML input

- Time series: How can you map graphs over time? Dynamic time warping

- The most appropriate representation is dependent on the question and method being used!

### 5.2.3 What do you want from machine learning? / Applications of ML

- Active learning

- Cutting down search space

- Current challenges

  Scaling

  Imbalance of data

  Interpretability of models

  Evaluations

  Interpolation between discrete observations (molecules)

  Applicability domain

  Best practices

## 6 Participants

There were 30 participants at this workshop, with a broad range of backgrounds including; chemistry, mathematics, cheminformatics and computer science both from an academic and industrial background. The participant list was printed in the event packs handed out at the event.

## 7 Conclusions

The series of talks and the discussions that followed confirmed that graph representations of molecules and crystals are key to the way we can present information to machine learning systems. The integration of knowledge graphs, data graphs and molecular graphs for a combined analysis is still more of an aim than an implementation. These are clearly subjects for further consideration.

## 8 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