Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery

Active Learning for Cost-Efficient Reaction Prediction using Kinetic Data

Interim Report
Project Dates: 26/10/2020 - 30/04/2021
Queen's University Belfast

Dr Paul Dingwall and Dr Son Mai
Queen's University Belfast

Report Date: 29/01/2021

AI3SD-Project-Series:Report6_Dingwall_Interim

Active Learning for Cost-Efficient Reaction Prediction using Kinetic Data
AI3SD-Project-Series:Report6_Dingwall_Interim
Report Date: 29/01/2021
DOI: 10.5258/SOTON/P0039

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 Project Details

Title	Active Learning for Cost-Efficient Reaction Prediction using Kinetic Data
Funding reference	AI3SD-FundingCall2_004
Lead Institution	Queen's University Belfast
Project Dates	26/10/2020 - 30/04/2021
Website	https://www.dingwall-lab.com/
Keywords	Kinetics, homogeneous catalysis, active learning, reaction prediction

2 Project Team

2.1 Principal Investigator

Name and Title	Dr Paul Dingwall
Employer name / University Department Name	School of Chemistry and Chemical Engineering, Queen's University Belfast
Work Email	p.dingwall@qub.ac.uk
Website Link (if available)	https://www.dingwall-lab.com/

2.2 Co-Investigators

Name and Title	Dr Thai Son Mai
Employer name / University Department Name	School of Chemistry and Chemical Engineering, Queen's University Belfast
Work Email	ThaiSon.Mai@qub.ac.uk
Website Link (if available)	N/A

2.3 Researchers & Collaborators

Aaron McNeill (PDRA, Queen's University Belfast), Dr. Anh Le and Dr. Ha Mai (independent, University of Transport, Vietnam), Gavin Lennon (PhD student, Queen's University Belfast)

3 Publicity Summary

Predicting the performance of a given reaction, the final yield of a product or even how long it will take, is a challenging task, even more so when also attempting to predict an optimal set of conditions for the reaction. Existing approaches to this problem use machine learning algorithms that are fed a large volume of single timepoint yield data (the amount of product generated after a set time). Using this data is problematic; if a reaction only reached 50% would it reach 99% if run for longer? How much longer? Was there a slow catalysts activation period followed by a rapid reaction? Or was a catalyst poisoned, stopping the reaction midway?

These are factors described by the reaction kinetics. We propose that using reaction kinetic data will lead to better predictive models. However, collecting kinetic data is significantly more costly and time consuming than collecting single timepoint yields. To minimise these issues, we will be using a cutting-edge machine learning technique called active learning. Rather than performing every possible reaction combination to build our model, active learning first picks the most important set of conditions. The user collects this data, and the active learning process is updated and begun again, resulting in a highly cost-efficient procedure minimising experimental requirements.

4 Executive Summary

We propose that kinetic data might outperform single timepoint yield data in predicting the performance of homogeneously catalysed reactions. The particular focus is on the area of homogeneous catalysis, for which there is a relative lack of application of machine learning techniques in the literature compared to other areas of chemistry. Rather than building a model which simply predicts a numerical value for yield, as must be the case when using single timepoint yield data, we will be predicting a kinetic profile, meaning reaction yield or conversion (and by extension cost, throughput, etc) can be calculated for a given reaction time. Reactions performed at different concentrations will allow the model to be trained on species concentration behaviour (i.e. component orders) which will be correlated to molecular descriptors. Component orders can then be predicted, allowing prediction of the turnover determining states of a reaction. A successful forward model, as just described, will predict how a reaction would proceed; a successful inverse model will predict optimal ‘above-the-arrow’ conditions (ligand, solvent, additive, etc.) recommended on rate behaviour and underpinned by mechanistic inference. Insights gained will allow design of novel catalysts tailored for both rate and mechanistic behaviour; the molecular descriptors of which can be calculated, and performance predicted prior to synthesis. By investigating many conditions (different substrates, catalysts, additives, solvents, etc.) in the course of model training, underlying trends for entire families of reaction will become clear, experimentally elucidating the so-called genome of a reaction. To ensure this study has realistic outcomes, we have focussed on validation of the approach through the study of a simple model system. The prospect of explanative and predictive reaction performance models based on kinetic data has the potential to accelerate the application of novel reactions in an industrial setting, aid in the discovery of new reactivity, and develop understanding more generally in homogeneous catalysis.

5 Aims and Objectives

We aim to develop a method to predict the performance of, as well as optimal ‘above-the-arrow’ conditions for, a homogeneously catalysed reaction through prediction of kinetic behaviour.

This will be achieved via the following objectives:

1. Generation of molecular descriptors for selected model reaction
2. Construction of machine learning model beginning with a simple iterative ensemble model that combines multiple supervised learning algorithms and moving on to an active learning method using both supervised and unsupervised techniques.
3. Collection of kinetic data guided by active learning algorithm.
4. Comparison of predictive models based on kinetic vs single timepoint yield data.
5. Publication of a summary article in a high impact, peer reviewed journal.

6 Methodology

6.1 Scientific Methodology

Experimental Chemistry Methods
The reaction chosen for study is the Copper/TEMPO catalysed oxidation of alcohols (*J. Am. Chem. Soc.* 2011, *133*, 16901; *J. Am. Chem. Soc.* 2013, *135*, 2357; *J. Am. Chem. Soc.* 2013, *135*, 15742). The reaction is an ideal model system for several reasons: 1) it is operationally extremely straightforward, fast, and highly selective, it is simple enough to be used as part of a third year undergraduate lab at Queen's and produces no by-products; 2) it is highly modular, allowing for many different sets of conditions to be chosen from; 3) the components are cheap and commercially available, requiring no time consuming synthesis; 4) most of the components are small and/or constrained molecules, allowing for straightforward computation of the molecular descriptors; 5) the mechanism is well studied, allowing for a degree of confidence that it is suitable for study and allowing us to compare out initial results with what exists in the literature. All experiments were conducted under identical standardised conditions of concentration and temperature. Reactions were sampled at specific time intervals and analysed via NMR using an internal standard to produce a kinetic profile over time.

Computational Chemistry Methods
All compounds were optimised using Gaussian 16 and confirmed as stationary points through the lack of an imaginary frequency. Sterimol parameters were calculated using the command line Python program Sterimol.py. For molecules containing one or more rotatable bonds, weighted sterimol parameters were calculated using the wSterimol programme (*ACS Catal.* 2019, 2313). Buried volume calculations were performed using the online SambVca 2.1 application (*Nature Chemistry* 2019, *11*, 872).

6.2 AI Methodology

Machine learning techniques applied in existing approaches that utilise single timepoint yield data will not be appropriate for handling kinetic profiles; these data represent a trajectory of time-correlated points rather than a single stationary point. Further, our aim is to predict entirely new kinetic profiles while existing techniques focus on predicting single points on an existing trajectory. An additional experimental hurdle also exists in that here, kinetic data will be collected manually, meaning both low data volumes and throughput. Active learning represents an optimal and highly cost-effective machine learning approach in this context. We will use this and other techniques to aid our implementation:

- *Model Training* We propose to develop an iterative ensemble approach that combines multiple supervised learning methods (e.g., Support Vector Machines or Neural Networks) in a single model instead of using only one learning algorithm, An additional novelty will be in combining this with unsupervised and active learning techniques. This is expected to bring a significant boost to performance and accuracy, especially when dealing with small training sets. Importantly, sub-optimal reaction conditions will be included in the data gathered.
- *Unsupervised Learning* Data clustering algorithms, such as K-Medoids will group molecules based on the similarity of their molecular descriptors, a direct measure of their chemical similarity. These groupings help to describe data and reveal hidden patterns, allowing learning algorithms in model training to create better boundaries between groups and to avoid training overfitting, leading to an improved predictive model.
- *Active Learning* The most significant element in our approach. Starting with a small training set, our algorithm will suggest experiments to gather new kinetic profiles as it

progresses. The user can collect this data and the active learning process is iterated. Selection criteria are based on differentiation between the groups built during unsupervised learning and compounds resulting in the most uncertain profiles during the model training phase. This active learning approach results in a cost-effective procedure that will minimize the total number of kinetic profiles used for training models, thus reducing experimental costs. It can also enhance prediction accuracy by strengthening decision boundaries between close groups.

- *Predictive Model* After being trained, our model can be used to predict a full kinetic profile of rate across a substrate concentration range.

To the best of our knowledge, such a mixed learning approach has not been fully explored in Machine Learning communities before.

7 Interim Results

- A standard set of experimental conditions (time, temperature, concentration ranges, stirring speed, stirrer bar geometry, glassware, reaction volumes) has been explored and decided upon. Specifically, this has allowed us to rule out experimental issues such as possible mass transfer limitation of the kinetics (the reaction uses air as a co-oxidant), confirm there is no sensitivity to water content, and that the range of conditions are viable (ie no catalyst deactivation or non-starter conditions). Analytically, NMR has been chosen as the most general and applicable method for monitoring the reaction and the data returned has been shown to be reliable and reproducible.
- A set of 93 compounds (including substrates, ligands, radical, and base) have been selected and molecular descriptors calculated.
- Various unsupervised data clustering methods and similarity measures have been implemented and compared on the provided molecular descriptors to look for intrinsic structures of input reactions. A data visualization tool and website has been built for presenting and assessing the results.
- The general machine learning framework has been built successfully. This framework serves as baseline for using different machine learning methods for predicting kinetic profiles.
- An active learning scheme has been developed based on entropy condition of the results.

8 Outputs

An online data visualization tool has been constructed and is available at http://203.205.25.9:8050/.

9 Progress Summary

- The project start was significantly delayed due to COVID-19, with the PDRA starting work on the project on 26th October 2020. AI3SD have been extremely kind and helpful in working with us to extend deadlines and overcome these obstacles. The project has encountered a series of lockdowns and changing government guidelines on working during the pandemic. However, we have managed to continue working as normal in the laboratory five days a week, albeit with facemasks and 1m+ social distancing guidelines in place.
- Two of the five objectives outlined above have been met: molecular descriptors have been calculated and a prototype active learning algorithm and data clustering and visualization have been written.

- We are beginning with the third objective this month, the collection of experimental kinetic data, having made good progress in the outline of standard experimental procedures. With the time left available to us, we hope to collect in the region of 400 full kinetic profiles to use in training our active learning algorithm. It is worth noting that the single timepoint yield data required for comparison is a part of these results, as we will choose a single timepoint from each profile. The profiles will be collected as a dataset and will be uploaded to a repository on publication of this work.

10 Next Steps

- Collect experimental kinetic data to train the active learning algorithm.
- Continuously refining the prediction models based on newly collected kinetic data.
- Compare the predictive powers of models built on kinetic vs single timepoint yield data.
- Publish results in a high impact, peer reviewed journal.

Ideally, this work will show proof of principle of our ideas. We aim then to use these outputs to strengthen future grant applications and continue our collaboration at the boundaries of chemistry and computer science.

11 References

A by no means exhaustive list is below:

1. Steves, J. E.; Stahl, S. S., Copper(I)/ABNO-Catalyzed Aerobic Alcohol Oxidation: Alleviating Steric and Electronic Constraints of Cu/TEMPO Catalyst Systems. *J. Am. Chem. Soc.* **2013**, *135* (42), 15742.
2. Hoover, J. M.; Ryland, B. L.; Stahl, S. S., Mechanism of Copper(I)/TEMPO-Catalyzed Aerobic Alcohol Oxidation. *J. Am. Chem. Soc.* **2013**, *135* (6), 2357.
3. Hoover, J. M.; Stahl, S. S., Highly Practical Copper(I)/TEMPO Catalyst System for Chemoselective Aerobic Oxidation of Primary Alcohols. *J. Am. Chem. Soc.* **2011**, *133* (42), 16901.
4. AlsBrethomé, A. V.; Fletcher, S. P.; Paton, R. S., Conformational Effects on Physical-Organic Descriptors: The Case of Sterimol Steric Parameters. *ACS Catal.* **2019**, 2313.
5. Falivene, L.; Cao, Z.; Petta, A.; Serra, L.; Poater, A.; Oliva, R.; Scarano, V.; Cavallo, L., Towards the online computer-aided design of catalytic pockets. *Nature Chemistry* **2019**, *11* (10), 872.
6. Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G., Predicting reaction performance in C-N cross-coupling using machine learning. *Science* **2018**, *360* (6385), 186.
7. Estrada, J. G.; Ahneman, D. T.; Sheridan, R. P.; Dreher, S. D.; Doyle, A. G., Response to Comment on "Predicting reaction performance in C-N cross-coupling using machine learning". Science **2018**, *362* (6416).
8. Nielsen, M. K.; Ahneman, D. T.; Riera, O.; Doyle, A. G., Deoxyfluorination with Sulfonyl Fluorides: Navigating Reaction Space with Machine Learning. *J. Am. Chem. Soc.* **2018**, *140* (15), 5004.
9. Granda, J. M.; Donina, L.; Dragone, V.; Long, D.-L.; Cronin, L., Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature* **2018**, *559* (7714), 377.
10. Zhou, Z.; Li, X.; Zare, R. N., Optimizing Chemical Reactions with Deep Reinforcement Learning. *ACS Cent. Sci.* **2017**, *3* (12), 1337.

11. Aggarwal, C. C.; Kong, X.; Gu, Q.; Han, J.; Philip, S. Y., Active learning: A survey. In *Data Classification*, Chapman and Hall/CRC: 2014.
12. Han, J.; Pei, J.; Kamber, M., *Data mining: concepts and techniques.* 2011.
13. Settles, B. *Active learning literature survey*; 2009.

12 Data & Software Links

An online data visualization tool has been constructed and is available at http://203.205.25.9:8050/

GitHub link for the project: https://github.com/anhlvq/chemreacpred. The GitHub link will be made open at the end of the project. We plan to publish the full project source codes for other researchers.