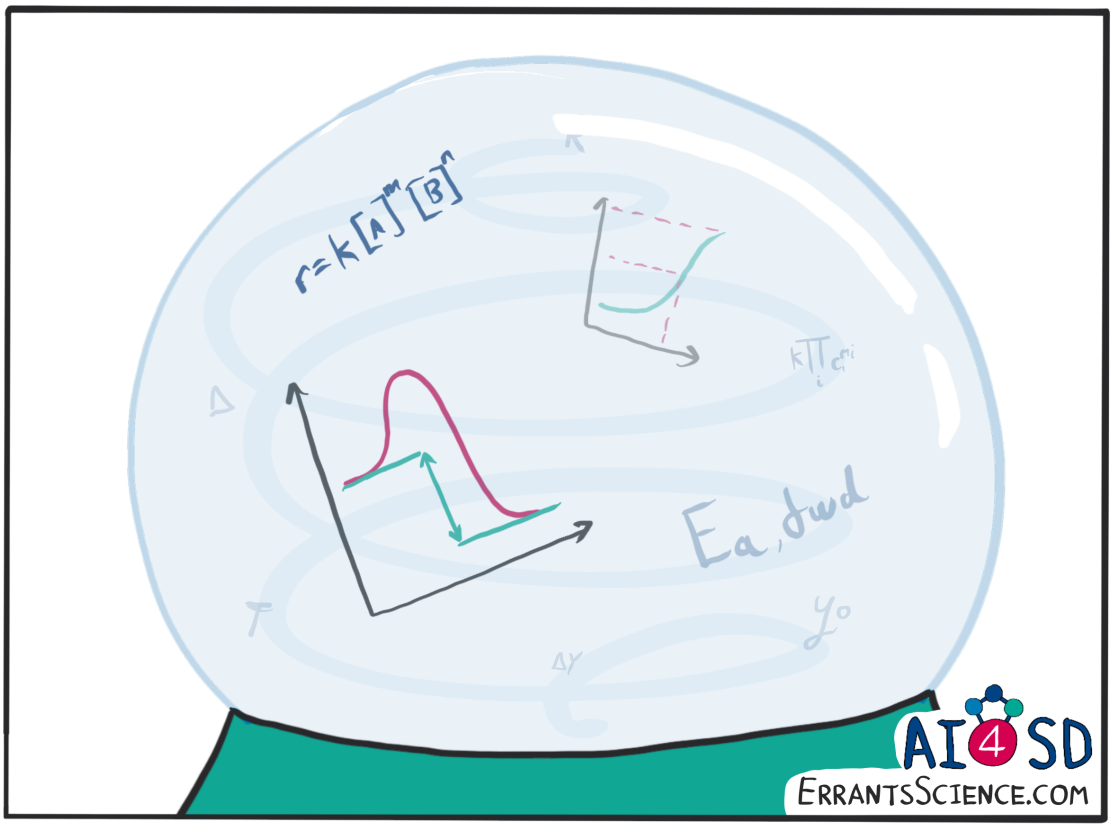
AI 4 SCIENTIFIC DISCOVERY

AI 4 Science Discovery Network+

Active Learning for Cost-Efficient Reaction Prediction using Kinetic Data Final 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: 06/07/2021

AI4SD-Project-Series:Report6_Dingwall_Final

Active Learning for Cost-Efficient Reaction Prediction using Kinetic Data
AI4SD-Project-Series:Report6_Dingwall_Final
Report Date: 06/07/2021
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	AI4SD-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	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	N/A

2.3 Researchers & Collaborators

Aaron McNeill (PDRA, Queen's University Belfast) is the postdoctoral research associate original working on the project collecting kinetic data in the Dingwall lab. Aaron graduated with First Class Honors in Chemistry from Queen's University Belfast in 2016 and completed his PhD thesis under Prof. Andrew Mills on the development and testing of ultrathin flexible photocatalytic films.

Gavin Irvine (PDRA, Queen's University Belfast) has recently joined the project as part of the QUB Covid Recovery Fund. He will continue Aaron McNeill's work on the collection of experimental kinetic data. Gavin graduated with First Class Honours in Chemistry with Medicinal Chemistry and External Placement from the University of St Andrews in 2017. He completed his Year in Industry in Hull working on Nurofen products at RB. He then started a PhD with Dr Efrosyni Themistou at QUB looking at biocompatible amphiphilic block copolymers for drug/DNA delivery.

Gavin Lennon (PhD student, Queen's University Belfast) has provided experimental support for the collection of kinetic data during the project.

Dr. Anh Le (Independent, University of Transport, Vietnam) obtained his PhD in Computer Science at the University of Heidelberg, Germany in 2014. He is working in the Machine Learning part of the project including prediction algorithms and active learning strategy.

Dr. Ha Mai (Independent, University of Transport, Vietnam) obtained her PhD in Chemistry at the University of Grenoble Apes, France in 2014. She serves as a bridge to connect the Machine Learning and Chemistry sections due to her extra background in Machine Learning.

3 Publicity Summary

Predicting an optimal set of conditions for a given reaction is a challenging task. This becomes even more challenging when also trying to predict the performance of a reaction, such as the final yield of product or how long it will take. 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. Our particular focus is on the area of homogeneous catalysis, for which there is a relative lack of application of machine learning techniques 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 time, throughput, cost, 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, and along with them, a prediction of the turnover determining states of a reaction. Doing so will allow us to demonstrate proof of principle of machine learning as a physical organic mechanistic probe. The result of this is particularly interesting as it would allow us to correlate change in molecular features to change in mechanism. 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, turnover determining states, and 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 QUB 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 concentration vs time.

7 Results

- A standard set of experimental conditions (time, temperature, concentration ranges, stirring speed, stirrer bar geometry, glassware, reaction volumes) has been explored and identified. 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. Analytically, in situ IR 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 (e.g., k-Means, k-Medoids, hierarchical clustering, density-based clustering) and similarity measures (e.g., Euclidean, Manhattan, Cosine, weighted Lp metrics) have been implemented and compared on the provided molecular descriptors to look for intrinsic structures of input reactions. A data visualization tool has been built for presenting and assessing the results. A good data clustering

method will organize reaction into meaningful structures to serve as an underlying mechanism to minimize the total number of reactions needed for building the kinetic prediction model. Our experiments have shown that k-Means, k-Medoids and weighted Lp metrics are the best combinations for identifying cluster structures on the input reactions.

- 201 kinetic profiles have been collected so far to be used for the active learning approach.
- The general machine learning framework has been built successfully. This framework serves as a baseline for using different machine learning methods for predicting kinetic profiles. The framework has been enriched with many different prediction methods (e.g., Random Forest, Support Vector Regression, Nearest Neighbour Regression). Experiments on existing kinetic profiles have been shown that Random Forest Regression provides best prediction performance. Advanced deep learning models should be a potential target in the future for improving prediction performance. However, these models require a huge amount of training data which is not possible at the moment due to both time and budget constraints. Visualization tools were built for deeply analysing the results in different aspects.
- Our active learning is divided into two separate parts: pre-training and post-training. In the first phase, we use the clustering results to select cluster centres as initial reactions for obtaining kinetic profiles as the initial set of training data. Our experiments have shown that this scheme helps to significantly lower the total amount of profiles needed to acquire the same level of prediction accuracy compared to the randomized selection scheme when we just choose random reactions to work on. We further enhance the method by incorporating a min-max strategy to additionally cover the special reactions in the search space which cannot be covered well by clustering methods. This help to enhances the performance further. In the post-training part, we employ several different prediction methods to predict the reactions and choose the one with lowest agreements among prediction methods as a target for building additional kinetic profiles to train the prediction models iteratively. By this scheme, we expect to acquire the same prediction accuracy with a lower number of kinetic curves, thus reducing time and cost for overall model building. Further strategies are being developed, in particular boosting and bagging mechanisms.

8 Outputs

Data visualization tools have been constructed for visualizing kinetic structures and prediction performances.

Work is still ongoing on the collection of experimental kinetic data. Once this is complete, our unique dataset will be published online in a freely available data repository.

9 Conclusions

- The first available experimentally generated dataset of kinetic profiles of a homogeneously catalysed reaction has been created.
- Using unsupervised learning provides a very good mechanism to build a small initial set of kinetic profiles to train the prediction model. Combined with Min-Max strategy, it provides a best-known method to minimize the starting training set for our problem.
- The project has successfully enabled the first collaboration, hopefully of many, between the School of Chemistry and Chemical Engineering and the School of Electronics, Electrical Engineering and Computer Science at QUB.

- Our study has been conducted on a small set of only 200 data points, highlighting the power of this approach (typical single timepoint yield approaches require 2000-4000 data-points).

10 Future Plans

- We plan to collect experimental data for a further 6 months. Random data collecting is complete and we will collect active learning data for the second-phase to demonstrate the ability to provide the same prediction accuracy with much smaller training data. This is a significantly advantages that lead to reduce the overall times and costs to generate training kinetic profiles considerably
- With a predictive model in hand, we wish to demonstrate the use of this approach as a physical organic tool for exploring reaction mechanism landscapes
- Once these steps are complete, we will publish our findings
- We plan to pursue further research opportunities in this area beyond the support of this grant
- This grant, and the publication, will be used to support a grant application to the EPSRC to continue the work in this exciting area

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

Various data visualization tools have been constructed and are currently used for analysing the results and developing improved prediction methods.

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.

Work is still ongoing on the collection of experimental kinetic data. Once this is complete, our unique dataset will be published online in a freely available data repository.