Machine learning for yield prediction for chemical reactions using *in situ* sensors

Joseph C. Davies[a], David Pattison[b], Jonathan D. Hirst[a*]

[a]*School of Chemistry, University of Nottingham, University Park, Nottingham, NG7 2RD, UK*

[b]*DeepMatter Group PLC, 38 Queen St, Glasgow, G1 3DX, UK*

[*]*jonathan.hirst@nottingham.ac.uk*

Abstract

Machine learning models were developed to predict product formation from time-series reaction data for ten Buchwald-Hartwig coupling reactions. The data was provided by DeepMatter and was collected in their DigitalGlassware cloud platform. The reaction probe has 12 sensors to measure properties of interest, including temperature, pressure, and colour. Colour was a good predictor of product formation for this reaction and machine learning models were able to learn which of the properties were important. Predictions for the current product formation (in terms of % yield) had a mean absolute error of 1.2%. For predicting 30, 60 and 120 minutes ahead the error rose to 3.4, 4.1 and 4.6%, respectively. The work here presents an example into the insight that can be obtained from applying machine learning methods to sensor data in synthetic chemistry.

Keywords: Buchwald-Hartwig cross-coupling; Long short-term memory neural network; Reaction monitoring; Time series data

Introduction

Machine learning can be applied to identify patterns and trends in large volumes of data with a trained model that generalises and is able to make predictions for unseen examples. Machine learning applied to chemistry is a growing field of research.[1] Many factors, such as advancements in graphics processing units, larger dataset collections and new algorithms have contributed to this renaissance.[2] Baum *et al.* observe that this growth is not uniform and postulate the reason fields such as analytical chemistry have seen faster developments compared to others like organic synthesis is due to the availability of large training datasets in areas which are traditionally more data intensive.[1, 3] For example, in analytical chemistry, machine learning algorithms have been used to find chemical species at concentrations below the usual limit of detection by finding hidden patterns of signals within the noise.[4, 5] However, considerable progress has been made in applying machine learning to organic chemistry. For example, machine learning techniques have been used effectively in computer aided synthesis planning by training on reaction data from Reaxys or the United States Patent and Trademark Office dataset.[6-8]

Sensor data such as *in-situ* temperature and pH can be a good source for machine learning algorithms in time series modelling. Interest in sensor usage has grown with the development of the *internet of things* (IoT) – which is the exchange of data between internet-connected devices, and can be extended to include chemistry equipment and chemically relevant data.[9, 10] The concepts of *cloud chemistry* or *telechemistry* have been introduced and involve remotely monitoring reactions by uploading the results from analytical equipment to the cloud.[11, 12] IoT is part of the wider concept of industry 4.0, which has received attention in recent years and refers to the current trends of interconnectivity, data, and automation. [13]

The application of machine learning techniques to real-time data, including sensor data, for predictive maintenance is an example of industry 4.0 practice.[14] Within process chemistry, these principles have been applied to enable predictive maintenance for pilot plants and self-optimisation of reactions.[15, 16,17] The use of sensors and other in-line methods for process monitoring was part of the vision for process analytical techniques set out by the United States Food and Drug Administration in 2004.[18] These methods are typically non-destructive and real-time, offering advantages over traditional sampling methods.[18-20]

The use of sensors in organic chemistry is an emerging area, fuelled by advances in flow and automated synthesis.[21-23] Using sensors inside organic chemistry reactions generates data which could be valuable in the development of machine learning tools to augment the synthesis process, ultimately helping the chemist. In the field of reaction kinetics, hybrid models, which combine machine learning with traditional modelling methods, have found success at predicting the chemo- and regioselectivity of substitution reactions.[24, 25] Sensors in chemistry can be used to monitor the reaction or to control the processes involved in performing the reaction. Mettler Toledo's ReactIR™ is used for monitoring reactions in real time using infrared (IR) spectroscopy, while FlowIR™ is an adaptation of this designed for use in flow chemistry where additional sensors measure pressure and temperature to monitor the flow within the system. [26-28] In automation, conductivity sensors have been used to detect the phase boundary between two immiscible solvents during extraction.[29] There has been work to standardise the hardware and code used to run experiments to improve reproducibility of results and data sharing.[30] Automation has been used effectively to explore unknown chemical space and predict the reactivity using the NMR spectra from before and after the reaction which has led to the discovery of novel transformations.[22] Despite significant advances in automation, most synthetic chemistry in the lab is done manually in

glassware as batch chemistry. In this study, we explore the utility of using sensor data gathered from hand-performed reactions. We use data collected by DeepMatter's DigitalGlassware platform and a mix of proprietary and original equipment manufacturer sensor devices. Specifically, these consist of a DeviceX™ reaction probe (temperature, ultraviolet (UV), pressure, stir rate and camera) and a Vernier thermometer, both suitable for a multi-necked flask, and an environmental sensor which would go adjacent to the reaction setup. The sensors used in this study recorded and more importantly saved the data in an open extensible markup language format.

The proposed utility focuses on predicting current and future product formation to track the reaction progress. NMR and chromatography are two methods commonly used for monitoring reaction progress. To be quantitative, UV chromatography requires calibration of the UV detector with the analyte. This can be a challenge because authentic samples of the analyte may not be available. Alternative detection methods such as evaporative light scattering detection and chemiluminescent nitrogen detection are less accurate, but do not need calibration with the sample.[31, 32] NMR provides quantitative results with use of an internal standard but can be more challenging to interpret and may require more extensive sample preparation and interpretation. In hand-performed reactions, these methods take time for the chemist to perform and valuable machine time. Sample preparation means these methods are limited to being applied during chemists' worktime, whereas reactions often run overnight or over non-workdays.

Once enough data has been collected for a reaction, sensors could be employed to send data to a trained model which can monitor the product formation and predict the future product formation. This would inform the chemist when the reaction is nearing completion or if the

reaction has stagnated. Sensor choice can be tailored to the specific chemistry. For example, pH sensors could be used in a pH-dependent reaction or colour data could be used if a colour change occurs in the reaction. This can be seen as a step towards automating, quantifying, and recording the data from some of the simpler tasks a human chemist does to understand their reaction. A litmus paper pH test is replaced by a quantitative pH sensor with a time-series and the same for a colour change, qualitatively observable by eye but can be quantitatively defined by red, green, and blue (RGB) colour values.[33, 34] Sensors also offer the possibility for monitoring aspects which cannot be monitored easily in traditional ways. Data which can typically be harder to discern includes activity status of catalyst or reagents which may have poor ionisation, high reactivity, or not show under UV and common thin-layer chromatography stains.

Chemists can use a variety of methods and intuition to analyse the reaction mixture to predict progression at the current instant in time or the future progression. For example, it may be determined a reaction has plateaued if two measurements in succession show no change in the reaction profile. Alternatively, a chemist may have experience with the reaction and develop an empirically based intuition of when the reaction has ended. For example, a reaction may typically reach completion at a variable conversion but within the same timeframe. The research here aims to test the suitability of machine learning for this objective.

Machine learning for yield prediction using the reaction scheme and conditions is an approach which has been implemented by others.[35, 36] This research has been applied to palladium catalysed reactions, including Suzuki and Buchwald-Hartwig cross-coupling reactions.[37] However, the methodology struggled when applied to patent data which had too much

inconsistency for accurate yield prediction.[36] Two issues with chemical reaction data that make predictive tasks challenging are the sparsity of the data within chemical space, and a lack of reported failed experiments.[38] This approach is complementary to the one we developed in this research as both have different aims. Our strategy aims to tackle the variability in yield that can be seen for a specific reaction. This can be a non-trivial problem for reactions that suffer from poor reproducibility. The method we have developed treats each repeat of a reaction as a single instance that can be distinguished from other instances by differences in the time-series data. Our hypothesis is that patterns within these differences can be exploited by a suitably trained machine learning model, thereby allowing accurate predictions of product formation.

The reaction studied in this work is a Buchwald-Hartwig cross-coupling reaction between benzophenone hydrazone and 4-chlorotoluene. Two experienced chemists at a contract research organisation carried out 15 repeats of this reaction and used the hardware described earlier, provided by DeepMatter to generate time-series data for each repeat. The dataset will be processed to a suitable format for training and testing machine learning models to predict the current and future product formation. A series of machine learning algorithms will be used to evaluate the suitability of different models for the predictive task. Product formation is a continuous variable; hence, regression models are of interest including linear and polynomial regression models, decision tree models, and neural networks. The main limitation of the models being developed is they will only relate to this dataset, and therefore, only this specific reaction. However, extending this work to include larger datasets of multiple related reactions would be worth exploring in future work.

Methods

We begin by describing the experiments that were performed to generate the raw data that forms the basis of our study. A DeviceX™ reaction probe and a Vernier thermometer were inserted into the reaction vessel for each reaction (computer-rendered image shown in Figure 1). The DeviceX™ contains a series of immersed sensors, exposed sensors, and a camera at the end for recording colour. Additionally, an environmental sensor, placed in the fumehood, measured the ambient temperature, humidity, and light levels. The data collected from this equipment was saved to a cloud storage and comprises several time-series.

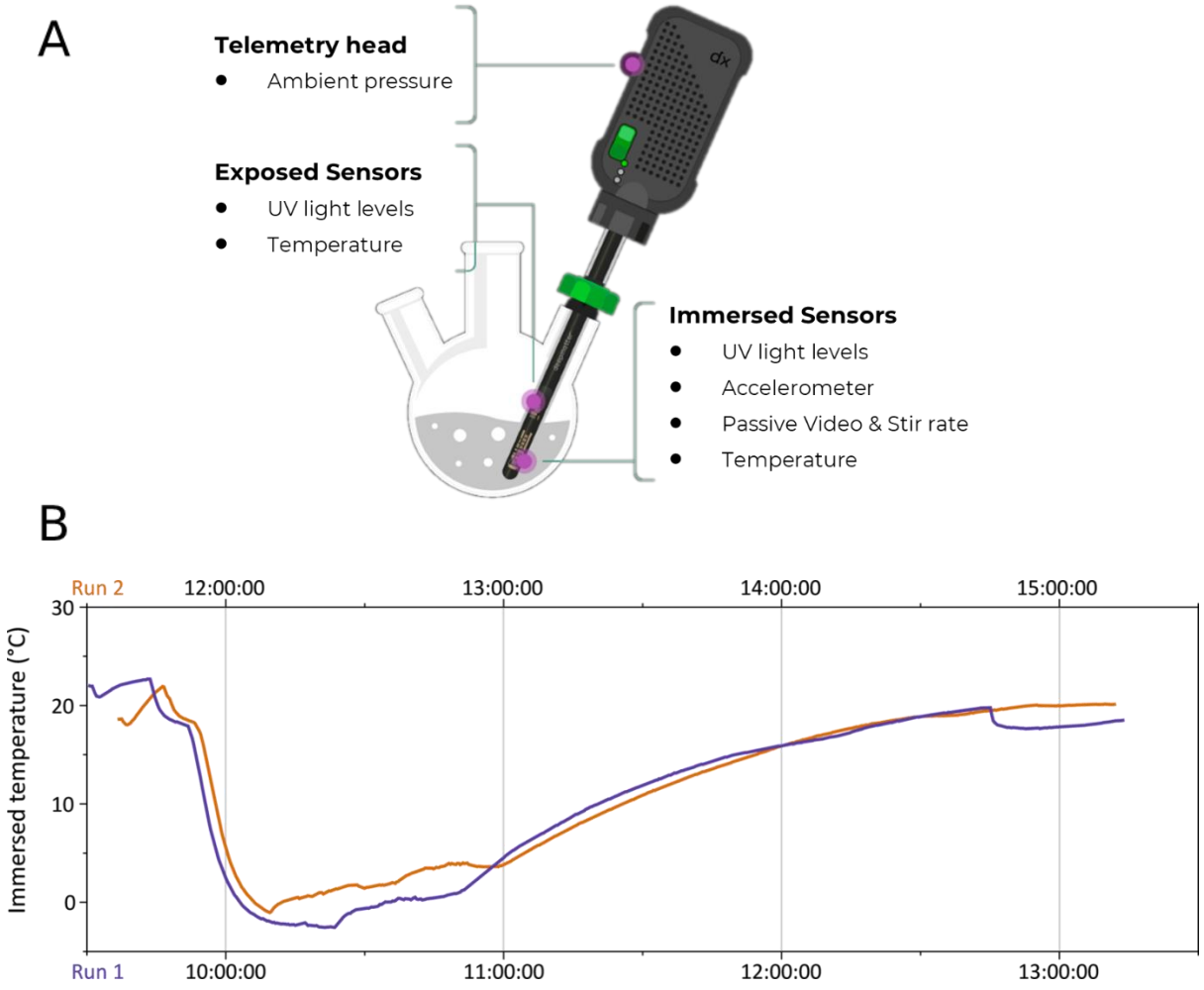

Figure 1: The DeviceX™ is a probe which is immersed in the reaction mixture. A) A schematic of this probe is shown. B) An example of temperature measurements for two discrete runs shown over a time-course.

The 15 reactions in the dataset are Buchwald-Hartwig coupling aminations. The reaction here is the cross-coupling coupling of benzophenone hydrazine and 4-chlorotoluene to form a new C-N bond (Figure 2).

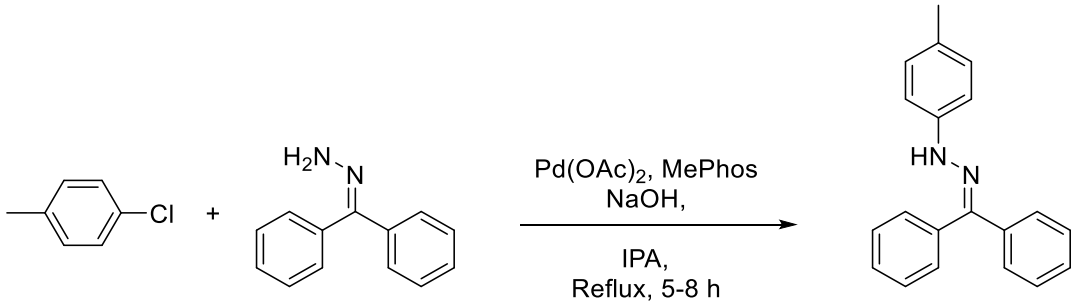

Figure 2: Buchwald reaction between 4-chlorotoluene and benzophenone hydrazone with a palladium acetate catalyst, MePhos ligand, sodium hydroxide base and either an isopropyl alcohol or tert-amyl alcohol solvent.

The experimental procedure was based on an optimised reaction obtained from the literature.[39] This exact literature route was repeated four times. Then the modified version shown in Figure 2 was used in the next 11 runs. The original route used *tert*-amyl alcohol (TAA) in place of isopropyl alcohol (IPA) and only ran for two hours. One difference between the solvents is that IPA's boiling point is 82.5 °C compared to 102 °C for TAA. Other variations between reactions included: the version of DeviceX™ probe which was used and what sensor data was recorded, the type of thermometer used and the rate of hydrazine addition. All of these are variations are displayed in Table S1 in the Supplementary Information. The first four experiments (001 to 011) and 024 were excluded, because the reaction probe used in these did not collect colour data (Figure S2 in the SI). This left ten reactions to use for modelling and evaluation of the models. The first eight runs ran for a duration of five hours and the last two were left for longer (eight hours) as there was evidence that the reaction was still progressing.

The sensor data were collected by four different instruments. The reaction probe measured UV A and UV B wavelengths, stir rate, temperature, and pressure. The system extracted the average RGB components of the images captured by the submersed camera. The environmental sensor measured light, humidity, temperature, and pressure. The Vernier thermometer provided a more accurate measure of temperature than the reaction probe (±1 °C versus ±3 °C). This temperature data was therefore used in preference to the temperature data from the probe.

The dataset also included liquid chromatography-mass spectrometry (LC-MS) data collected at 30-minute intervals to determine the conversion of the reaction. The peaks were assigned by their molecular weights. Specifically, these were benzophenone hydrazine (starting material), the product, and an impurity from hydrazine reacting with IPA which was identified by ^{1}H NMR (Figure S1 in the Supplementary Information). All the monitored reaction components contain the highly conjugated phenylhydrazone group. Therefore, it was expected their UV absorbance coefficients would be similar enough to allow the percentage area of the peak integrals to be directly compared and used for monitoring reaction progression.

The data was processed by changing the timestamps in coordinated universal time to time in seconds relative to the start of the reaction. Datapoints were kept if they fell in the window between the first and final LC-MS measurement. The sensor data was down-sampled to every ten seconds and the LC-MS outcome data was up-sampled by linear interpolation also to every ten seconds.

The problem was approached as a regression problem with a continuous spread of outputs representing product formation. Many widely used time-series specific approaches, such as

autoregressive integrated moving average, were unsuitable for this problem as they assume a static output with only seasonal variations. This contrasts with the task described here that involves predicting an ever-increasing product amount. However, the approaches we use, such as recurrent neural networks, are suitable for application to time-series and do not assume a stationary output.

Due to the time-series nature of the data, cumulative and change values between measures were calculated. Cumulative values were calculated by successive addition of each time instance of a feature. Change values were calculated by measuring the difference between the previous and current time instance of a feature. Power transformations were applied to the features and to the cumulative features to increase the linearity of their relationship with product formation.

Models were made using linear regression, cubic polynomial regression, random forest, gradient boosting regression, and long short-term memory (LSTM) neural networks. The linear regression, random forest, and gradient boosting regression models were all implemented in Python using conventional sci-kit learn libraries. The cubic polynomial regression model was constructed using a custom function in Python. The decision tree models each used 1000 trees and the same random state was used. The LSTM neural networks were made using the Python package Keras. Two LSTMs were constructed. The first LSTM predicts the current product. A second LSTM was constructed to predict future product formation. Both LSTMs used a sliding window of data to make predictions.

Features of interest were identified by manually calculating Pearson correlation coefficients with the product and by visual inspection. Four datasets were generated based on manual

analysis. These were denoted as *all features, green cumulative, Yeo-Johnson transformed green cumulative* and *non-colour features.*

To evaluate the predictive performance of the models, the data was divided into a training and test set, according to run rather than aggregated samples. Due to the small number of runs within the dataset, the test set consisted of a single run and all datapoints from a run were kept together within the same partition to prevent leakage from the training set into the test set which would give an overly optimistic prediction. Uniform Manifold Approximation and Projection (UMAP) was used in conjunction with K-means clustering algorithm to visualise and group similar runs together.[40]

Results & Discussion

The sensor data from the reaction data was averaged across the runs. From these averages the Pearson correlation coefficients, *r,* between the features and product formation were calculated (Figure 3). The coefficients between a feature and product formation were used as a starting point to determine the utility of a feature in modelling.

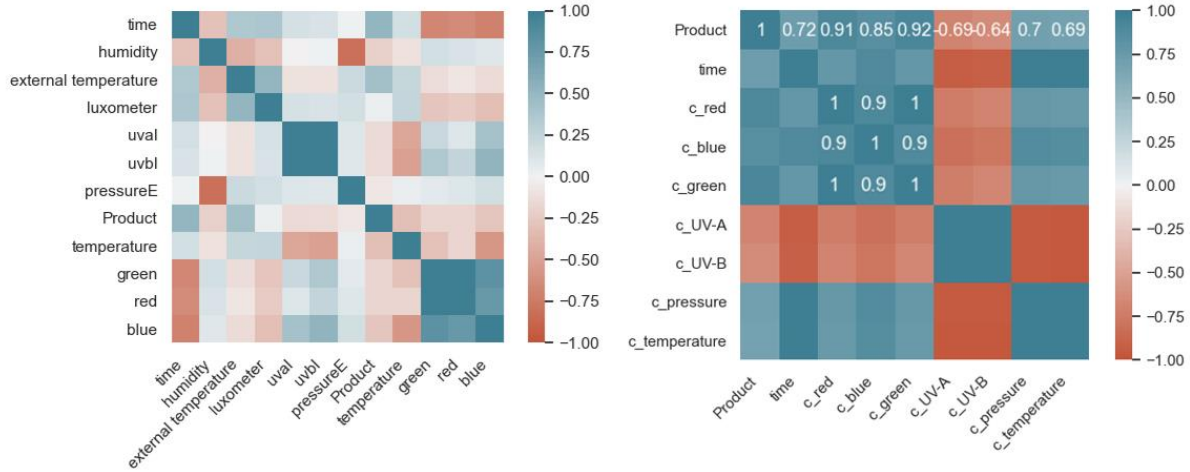

Figure 3: The left panel shows the Pearson correlation values for the unmodified sensor data. The right panel shows the Pearson correlation values for cumulative sensor values, where c_features refers to cumulative features.

To contextualise the coefficients, most cumulative features were expected to show some correlation to product, as if the feature is always positive or negative, both product and the cumulative feature will continually increase in value until product formation has stopped. The features most correlated with product formation are the cumulative colour features, particularly green and red, and a more negative correlation for blue. From the average values across the runs of these three features for the first five hours of the reaction, it can be observed that green and red steeply decrease over the course of a reaction and blue shows a more shallow decrease (Figure 4). A hypothesis for this is the reaction tends to black (where RGB would be 0,0,0) as the palladium catalyst is lost from the reaction and is oxidised to palladium (0) black from the activate palladium (II) in the palladium acetate and the palladium ligand complex. This fits with the chemists' observations of the initial mixture being described as cream coloured but turning black or darker brown later.

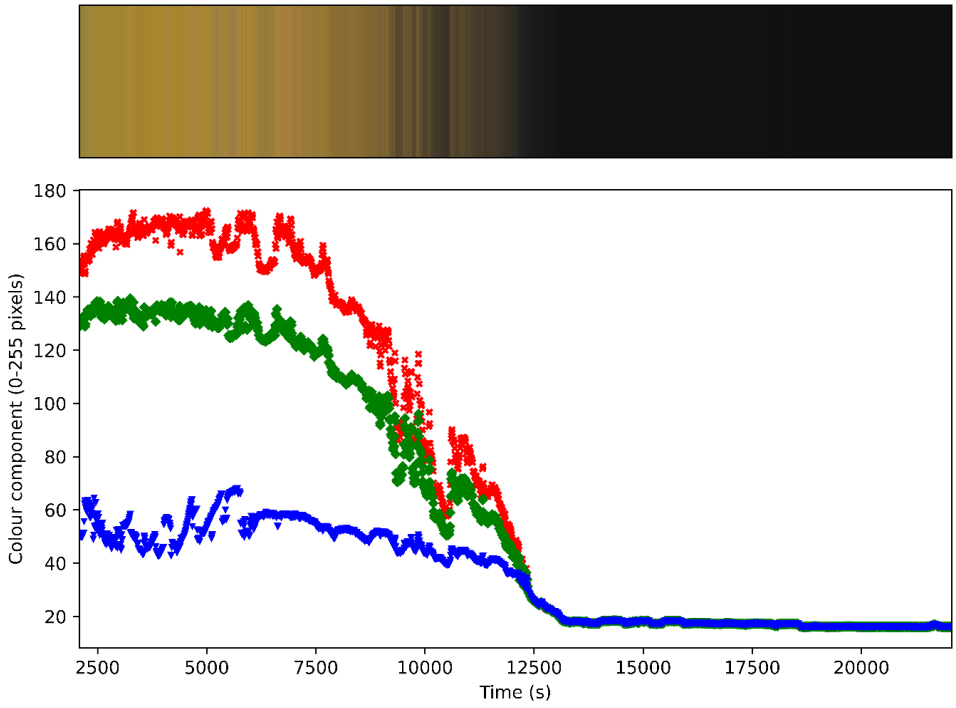

Figure 4: Variation over time of the recorded colour pixel component in the reaction mixture for example run Crocus-020. The top panel shows the calculated colour from the combined RGB values, and the bottom panel shows the individual RGB values. Each colour represents the colour shown, the crosses are red, diamonds green and downward arrows blue.

The runs were kept whole during the 9:1 train/test split to prevent data leakage and more accurately replicate a real-life scenario where the model would not be exposed to any data from the unseen test run. Initially, a linear regression model was developed on the cumulative green values from nine runs to predict the product for the tenth run. Assessing the model using cross-fold validation provides a robust estimate, in data-limited situations, of the accuracy of the model and its ability to generalise by testing its predictions on all the datapoints. Models were evaluated by measuring the mean absolute error (MAE) of the predictions. All product yields are as percentage points Thus, despite MAE being reported with units of % it is an absolute and not a relative error.

Linear regression is a natural starting point. However, it was observed the relationship between the cumulative green and product formation was not linear. To remedy this different power transformations were applied, and they were evaluated by the improvement in accuracy of the transformed features in a linear regression model compared to the non-transformed feature. After evaluating the different power transformations, a Yeo-Johnson transformation was selected as the most promising. The purpose of a Yeo-Johnson transformation is to reduce the skewness and make the distribution of the data more Gaussian.[41] Applying the Yeo-Johnson transformation to cumulative green yielded a more Gaussian-like distribution, as can be observed in Figure 5.

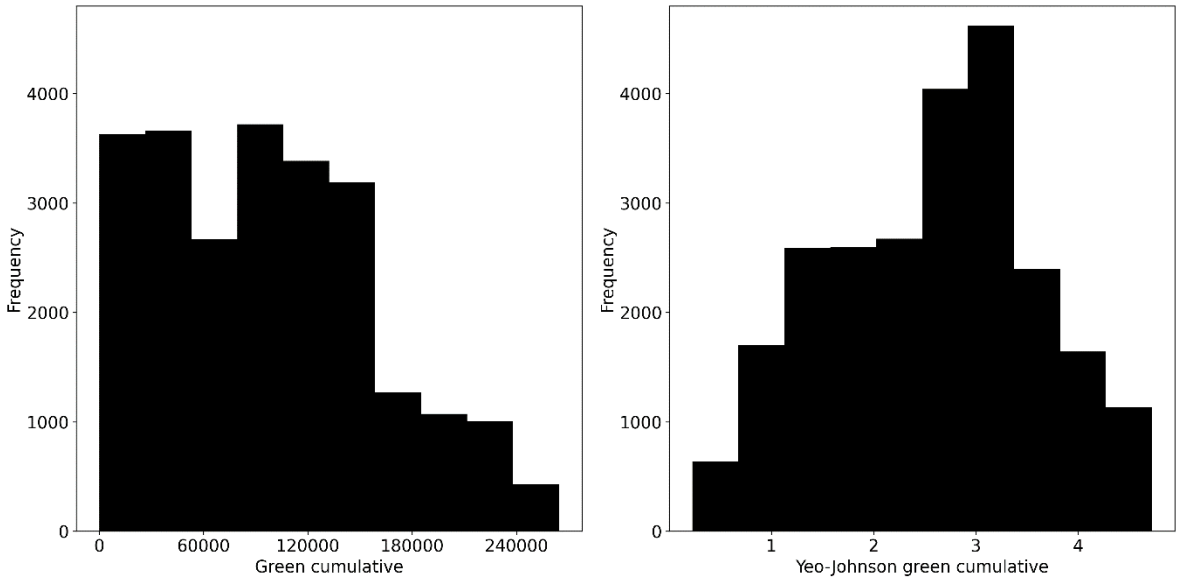

Figure 5: The left panel is the distribution of green cumulative values, and the right panel is the distribution after applying a Yeo-Johnson transformation.

Other models explored included polynomial regression. The polynomial used was a cubic function. This model was considered suitable due to the non-linear relationship between cumulative green and product. Random forest and gradient boosting regression, both based on decision tree models, were also explored.

The four datasets, each containing different features, (*all features, green cumulative, Yeo-Johnson transformed green cumulative* and *non-colour features)* were developed to investigate the effects and significance of the power transform, colour, and other sensor data. Use of *all features* would allow an evaluation of predictions without prior manual identification of features. *Green cumulative* and *Yeo-Johnson transformed green cumulative* compare the effect of the power transform and to what extent a single feature can provide accurate predictions. The *non-colour* dataset investigates how predictive the non-colour data is.

As a baseline, we calculated the MAE of predictions obtained from just the mean product values of the training runs. This gave an MAE of 7.0%. The four datasets were combined with the four models to produce all possible combinations. The results of these are shown in Figure 6. The Yeo-Johnson transformation improved results when using a linear regression model (Untransformed: 4.2% and transformed 3.8% MAE). The results for each algorithm with the *no colour* dataset (MAE 7.0, 7.1, and 8.2%) were all comparable to or worse than the baseline results (7.0%). This suggests the Pearson correlation coefficients identified the useful features in this context, and there is little value in the non-colour data in isolation. The best predictions were obtained using polynomial regression with cumulative green (MAE 3.6%). However, this required the prior step of feature identification. A gradient boosting regression model made predictions which were nearly as accurate (MAE 3.9%) but with all the features. These predictions were more accurate than those obtained using a gradient boosting regression model with cumulative green as the sole feature in contrast to all other models, suggesting there is predictive value in the remaining data when using a suitable model. Using all features has the advantage of obviating the need for prior feature selection. This does not interfere with the interpretability of the model either, as feature importance values can be extracted from the decision tree models. Feature importance using all features in a gradient boosting regression model (Figure S3 in the Supplementary Information) further confirms colour to be the most informative feature and suggests that it is likely all four models identify similar patterns in the relationship between cumulative green and product formation.

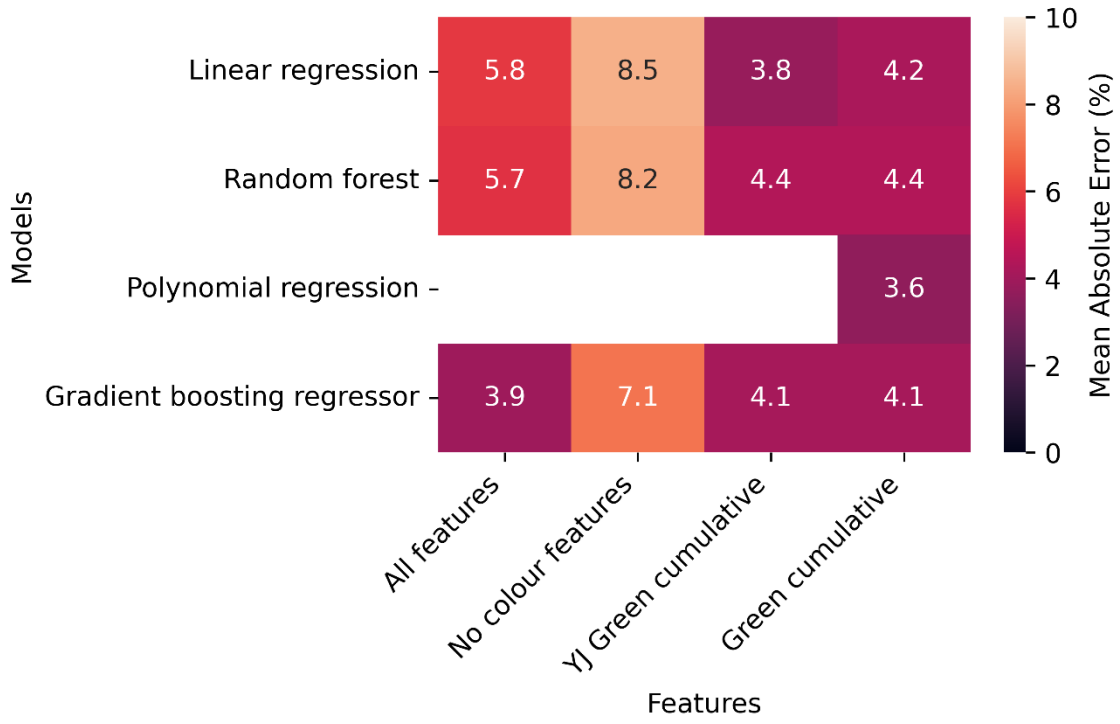

Figure 6: Comparison of the mean absolute errors (indicated by the colour bar) for the different combinations of features (x axis) and models (y axis). YJ stands for Yeo-Johnson. The polynomial regression model developed was not suitable for all feature sets, and hence was only used in conjunction with green cumulative.

None of the above models is time-based, and no explicit concept of time has been encoded in the associated datasets. Given that the input data is a time-series signal, this raises the question of whether a time-series model would be more accurate.

Recurrent neural networks, such as LSTMs, are well suited to time-series problems and other sequential tasks such as natural language processing and have been successfully applied in other areas of chemistry.[42, 43] The LSTM model constructed uses a sliding window approach, whereby a window of fixed size is formed over the data, and this window slides over the data to capture different portions of it. This method ensures the volume of data used in the model remains consistent. The method was employed to use the previous 20 minutes of sensor data. Changing to an LSTM model using *all features*, improved prediction accuracy with a MAE of

1.2%; this is compared to a best of 3.6% for the non-neural-network models. These results can be contextualised by comparison to the yield range in addition to the earlier described baseline. The observed yield was between 7.5 and 38.1%, giving a range of 30.6%. Therefore, a predictive accuracy with a MAE of 1.2% was considered useful in this context.

Following on from the promising results obtained for the instantaneous predictions from the LSTM, a second LSTM was designed for the more ambitious aim for predicting the future product. The second model uses sensor data and predictions from the first model between times $t_1 - y$ and t_1, where t_1 is the current time and y is a fixed time interval, to predict product at time $t_1 + z$ where z is a variable time interval.

To balance ambition (i.e., the extent of the forward time interval), accuracy, and computational cost, y = 2 hours, and z = 0, 30, 60, or 120 minutes. These time interval values can be put into context by comparison to reaction duration which was between five and eight hours (300-480 minutes). The expected behaviour is the larger the value of z, the harder it is to predict product formation. A range of values were selected for z to allow an evaluation of the relationship between the MAE and z. Because of the greater challenge associated facing the second model, y was assigned a higher value than x. This meant sensor data over a longer duration of time was used in the second model. In this task, it is more important to be able to see the larger context of the reaction progress. Having two models was advantageous, as it allowed the assessment of what would have been the intermediate output when $t_3 = 0$.

Since the LSTM gave a significant improvement, this model was chosen for the more ambitious target of predicting future product formation. Because current product can be predicted relatively accurately (MAE of 1.2%), a time-series of predicted product values was

obtained. These predicted product values and the sensor data are then fed into a second model to predict the product conversion in the future. The results from the cross-validation of the two LSTM models are shown in Figure 7. There is a large (but unsurprising) increase in error when predicting just 30 minutes ahead.

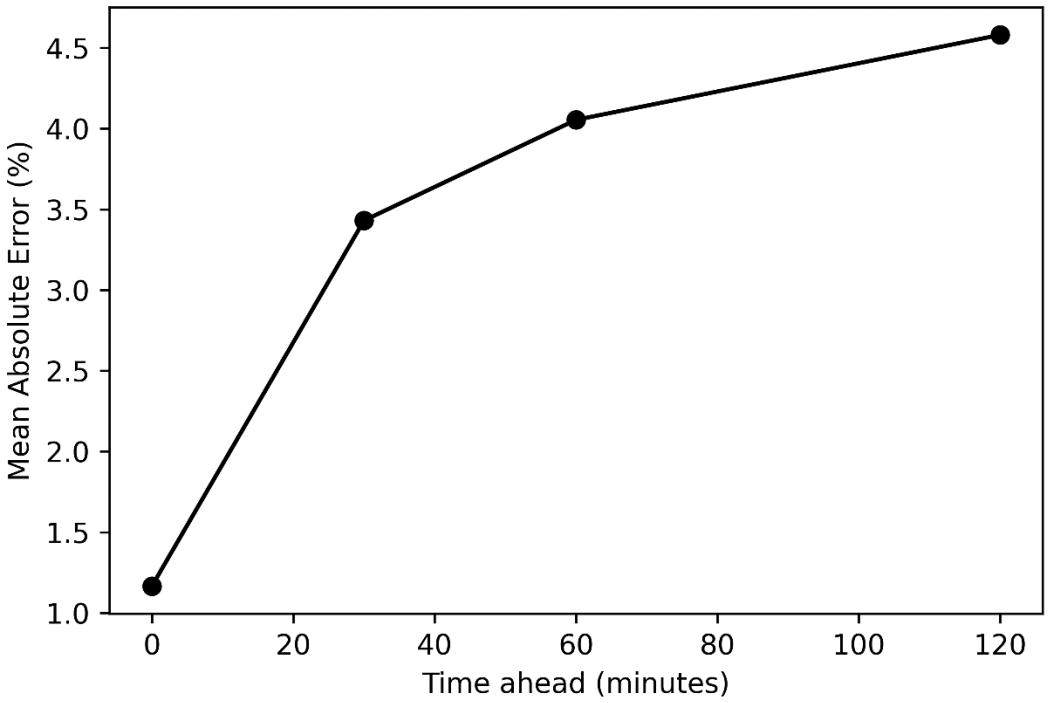

Figure 7: The mean absolute error for product formation predictions from an LSTM model predicting 0, 30, 60 or 120 minutes ahead.

Use Case 1

To assess the model, a series of real-world scenarios were identified. One scenario was to split the data, so the last chronological run was the test data, and the previous runs were the training data. This corresponds to the training data that would exist whilst the final reaction was being performed. The trend of the results from this use case closely align with those from cross-validation (Figure S4 in the Supplementary Information). Compared to the cross-validation, the model performed slightly better for all values of *z* greater than zero.

Use Case 2

Another example use-case would be for predicting when product formation has stagnated in a reaction at a lower conversion than expected. The goal in this scenario would be to detect a failed reaction at an earlier stage, and the live data in combination with predictions would give an indication of this to a chemist. This can be rationalised as, if the colour change occurs to indicate the catalyst has been consumed, no more product will be formed. Reactions 16 and 17 were both low yielding and the chemists performing these reactions observed a potential exposure of the hydrazine to atmospheric moisture.

An 8:2 train: test split was used and reactions 17 and 25 were used as the test runs. The rationale for using this split is twofold. The first and primary aim is to demonstrate that if a failed reaction, 16, is included in the training set, the model can identify future failed runs, such as 17, early. The second aim is keep run 25 in the test set to ensure the model is still capable of generalising and accurately predicting successful runs.

A weakness identified in the model was the imbalanced dataset due to the training data containing only one failed reaction. To address this, oversampling was implemented. A condensed version of the reaction data was interpreted by a SMOGN (Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise) algorithm.[44] The SMOGN algorithm can be particularly useful when the values in the interest of predicting are rare or uncommon, such as the failed reactions in this scenario. The algorithm identifies under-represented reaction runs from the condensed dataframe and returns a new dataframe with oversampling of under-represented data and undersampling of over-represented data. The algorithm performed as expected, with the newly returned dataframe oversampling the only low yielding reaction in the training set, reaction 16. The results for this can be seen in Figure

8. The predictions for run 25 are worse compared to those shown in Figure S4 in the Supplementary Information; however, they are within an acceptable range and the predictions for run 17 demonstrate an ability to predict a low-yielding reaction.

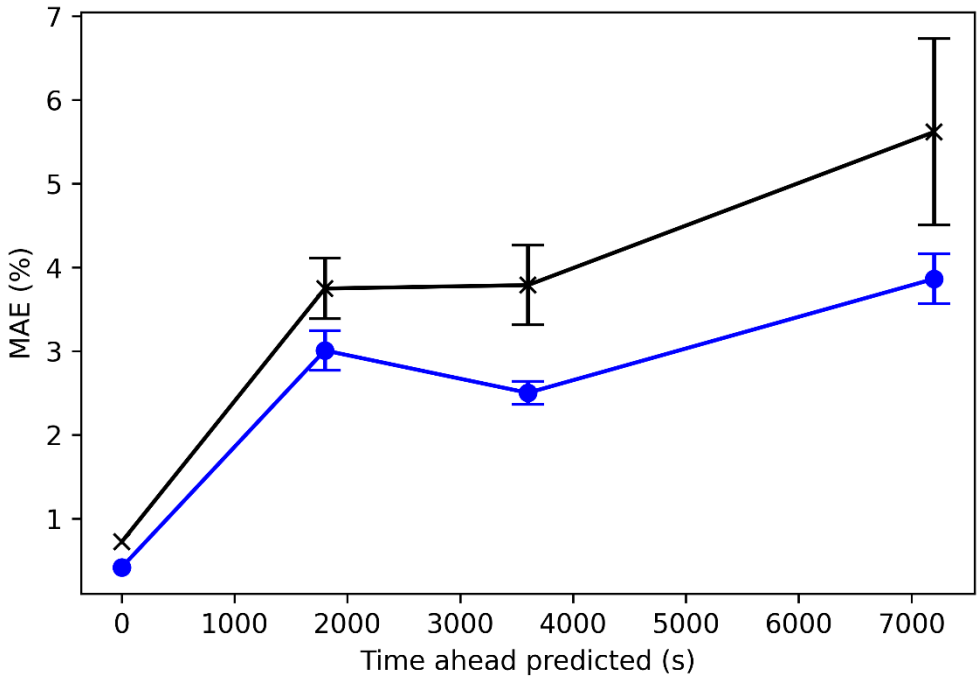

Figure 8: Blue circles show mean error in predictions for reaction 17 and black crosses for reaction 25. The standard error bars are calculated from ten repeats.

Use Case 3

A third use-case is to test the model chronologically and explore the relationship between the size of the training data and prediction accuracy. To test the model chronologically, this would mean training only with runs that come chronologically before the test run. This is a useful test to do, because it closely mimics the real-life scenario of a developing dataset wherein the chemist may be learning subtleties of the problem as they proceed. The model produced reasonable results when $t_3 = 0$. However, in early runs with little training data, the model was unable to make accurate predictions when $t_3 > 0$. Interestingly, the relationship between the size of the training data and prediction accuracy was inconsistent with the MAE increasing as

the training size increases before decreasing again. The reason for this could be due to the differences between runs, with some being more challenging to predict than others. This is also seen in the results from cross-validation. Figure 9 compares the chronological and cross-validation results. The cross-validation results show the error at time zero for each run with a training set size of nine other runs, whereas the chronological results show the error at time zero for each run with a training set of an ascending size only including runs which were performed earlier chronologically. In both scenarios runs 22 and 23 have the greatest MAE. This suggests that these runs are more challenging to predict. Some runs perform better in the chronological results, despite having a smaller training dataset. This could be due to the runs in the dataset being more similar. This can be observed with run 17 as it is similar to run 16, since both had poor product formation.

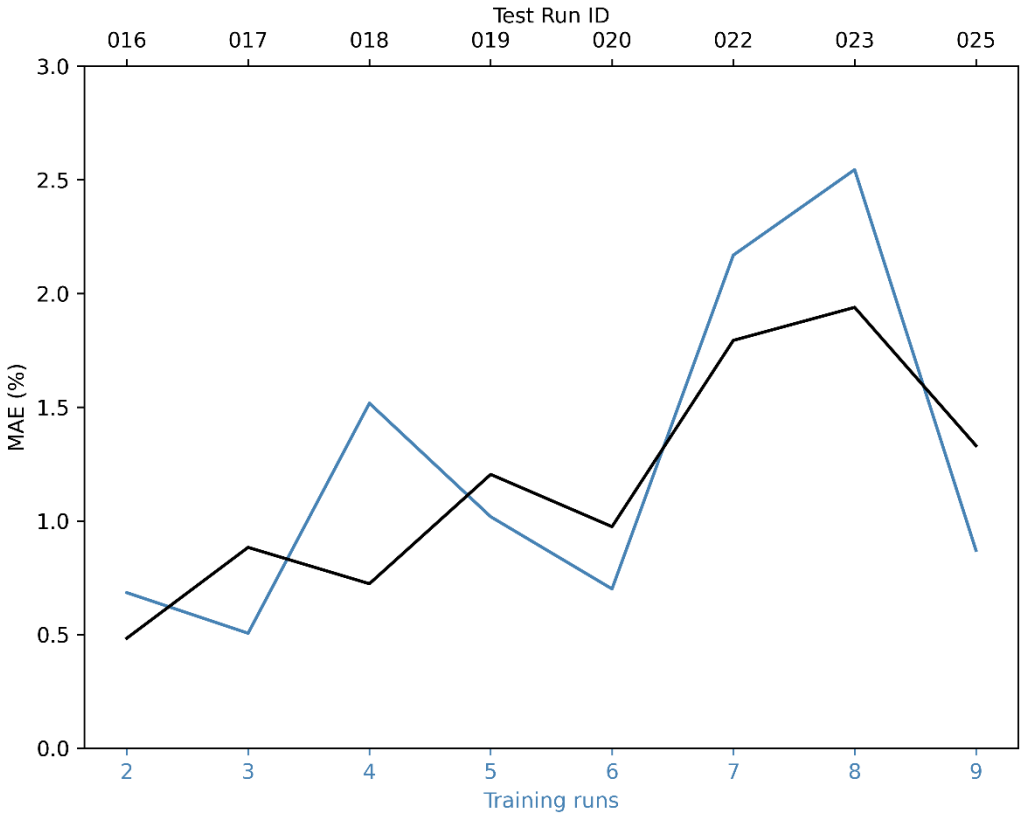

Figure 9: The blue line is the mean absolute error for predictions of a model trained only on runs which were chronologically before the test run, the number of which is shown on the bottom x axis. The black line is the mean absolute error for predictions from the cross-validation where a model was developed to predict the run by training on *all* other runs, regardless of chronological ordering.

An alternative approach to investigate the relationship between training dataset size without the additional variable of different test runs, was to keep run 25 as the test run and increase the training size in chronological order (Figure 10). This gives a more expected relationship with smaller increases in MAE as the training size increases. The other variable that may affect the relationship here is the different runs added to the training set and how useful they are for learning how to predict the product formation in run 25.

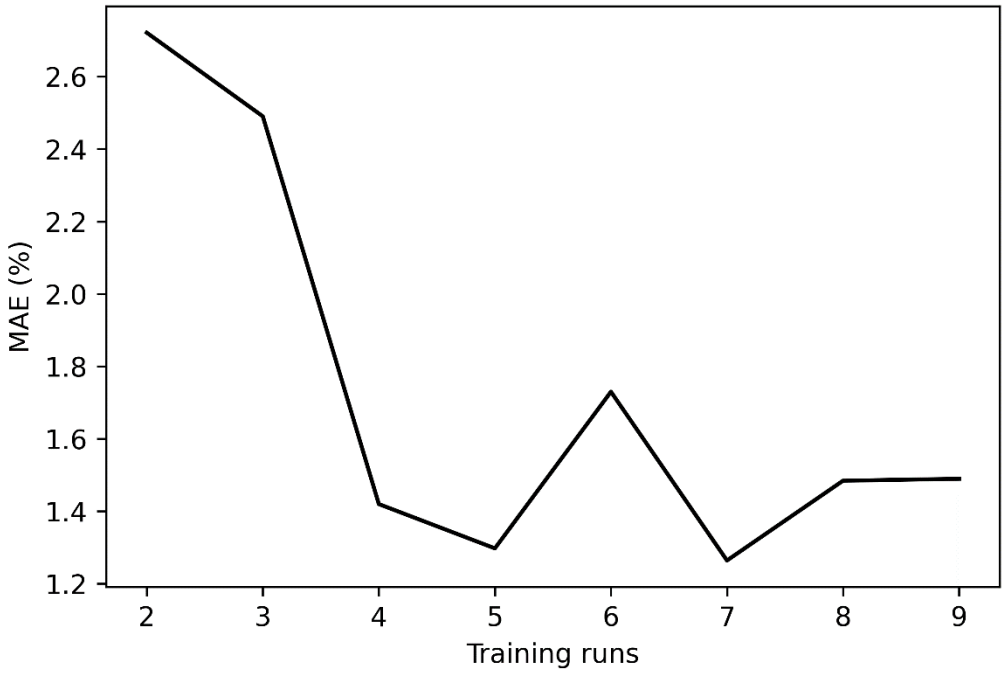

Figure 10: The relationship between the number of runs in the training data and the accuracy of predictions for product formation in run 25.

It was hypothesised that individual runs could be tracked within the 2D projection and similar runs could be near one another To investigate this the data was projected onto a two-dimensional manifold using UMAP and a K-means clustering algorithm applied to obtain distinct groups. After trying different values of *k*, *k=3* was used and the three clusters obtained from this can be observed in Figure 11. Runs 16 and 17 demonstrate similar runs occupy similar space, as both have poor product formation (below 20%) and are mostly situated in

the top right corner of the UMAP projection. In comparison, run 25 initially occupies space in the top right but as product formation increases, it has more datapoints in the middle and bottom left cluster. (Figure S5 in the Supplementary Information). This spread of data suggests the model may struggle to generalise, as runs all occupy different spaces, and hence may struggle to predict runs further from the space of the training data.

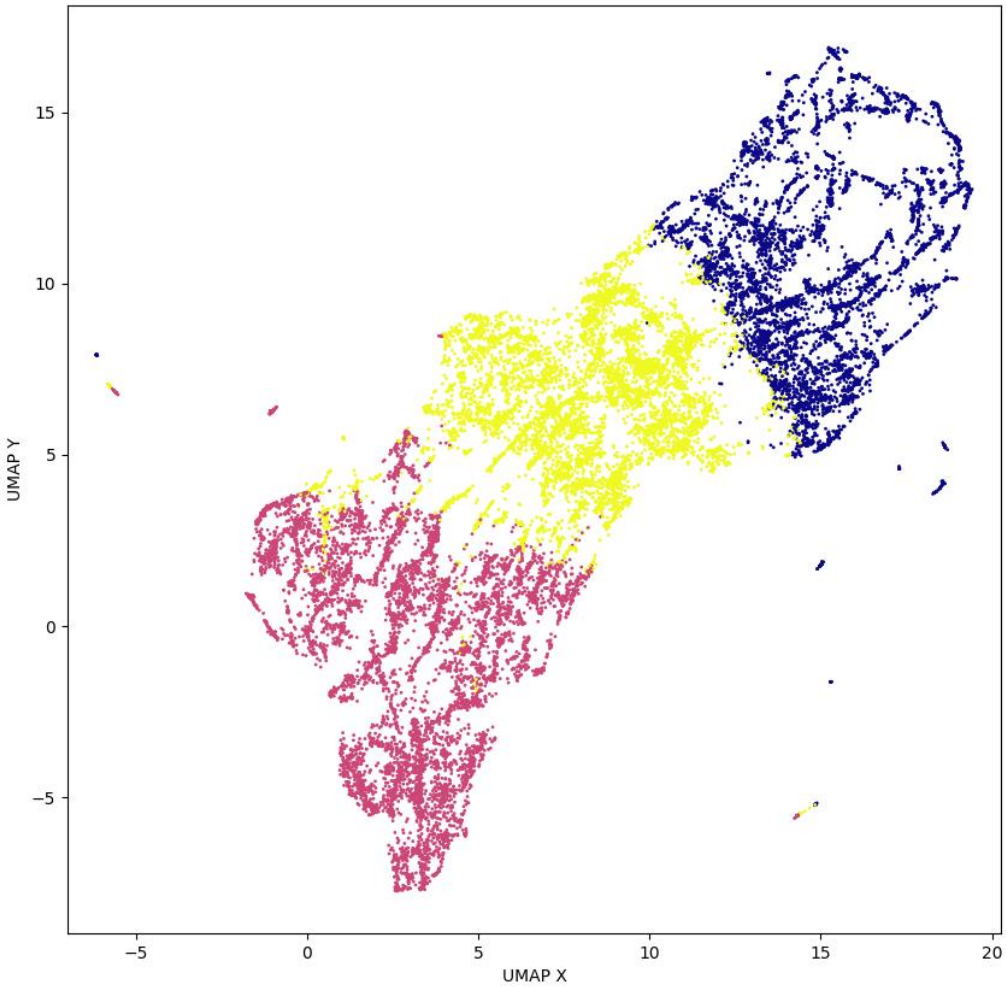

Figure 11: *k*=3 clustering of a UMAP projection of the reaction data for all runs used in machine learning models.

In a larger dataset, it could be envisaged that this could allow similar runs or reactions to be identified. This could allow the use of tailored models which would seek to place emphasis within the training data on runs identified as similar to the run of interest and this could create models which could give more accurate results. A small-scale example of this can be observed

earlier in the accurate results of run 17 in the chronological use-case, which included also failed run 16 in the small volume of training data.

Conclusion

Models to predict the current and future product formation from reaction sensor data collected by an *in-situ* reaction probe were constructed and evaluated. In this work we have demonstrated different use cases for these models. By using cross-validation and three realistic use cases, the predictive accuracies were assessed.

The reaction used here was suitable due to the reliable curve shape and profile of the reaction. Mild changes between runs, such as the rate of hydrazine addition, did not have a noticeable impact on prediction. However, large changes may; concentration or temperature could affect the rate of reaction. Future work will investigate how more significant changes affect the accuracy of the model. To assess further the potential and utility of using machine learning to predict product, more examples would need to be examined. This methodology could enable AI augmentation of reaction monitoring to assist synthetic chemists and facilitate a greater understanding of the reaction by identification of correlations between sensor features and reaction outcomes. Insights into the chemistry being performed could also be developed, for example, the correlation between cumulative green and product formation in this work, providing a quantitative description of the colour change in the reaction.

As use of sensors in synthetic organic chemistry grows, more data will become available and allow greater insights into the chemistry. This will also permit a more thorough investigation into the relationship between dataset size and accuracy. The models demonstrate that

information recorded by specialised reaction probes can be exploited by a neural network for product prediction.

Data and software availability statement:

Code and data for replicating the work reported here can be found on GitHub.

https://github.com/JoeDavies-6/ML-for-product-prediction

Acknowledgements

J. H. is supported by the Royal Academy of Engineering under the Chairs in Emerging Technologies scheme. We thank our colleagues Simon Preston and Dave Parry for helpful discussions.

References

1. Z. J. Baum, X. Yu, P. Y. Ayala, Y. Zhao, S. P. Watkins and Q. Zhou, *J. Chem. Inf. Model.*, 2021, **61**, 3197-3212.
2. A. Karthikeyan and U. D. Priyakumar, *J. Chem. Sci.*, 2022, **134**, 2.
3. L. B. Ayres, F. J. V. Gomez, J. R. Linton, M. F. Silva and C. D. Garcia, *Anal. Chim. Acta*, 2021, **1161**, 338403.
4. S.-Y. Cho, Y. Lee, S. Lee, H. Kang, J. Kim, J. Choi, J. Ryu, H. Joo, H.-T. Jung and J. Kim, *Anal. Chem.*, 2020, **92**, 6529-6537.
5. B. Debus, H. Parastar, P. Harrington and D. Kirsanov, *TrAC, Trends Anal. Chem.*, 2021, **145**, 116459.
6. M. H. S. Segler, M. Preuss and M. P. Waller, *Nature*, 2018, **555**, 604-610.
7. S. Genheden, A. Thakkar, V. Chadimová, J.-L. Reymond, O. Engkvist and E. Bjerrum, *J. Cheminformatics*, 2020, **12**, 70.
8. Y. Mo, Y. Guan, P. Verma, J. Guo, M. E. Fortunato, Z. Lu, C. W. Coley and K. F. Jensen, *Chem. Sci.*, 2021, **12**, 1469-1478.
9. A. Cadeado, C. Machado, G. Oliveira, D. E Silva, R. Muñoz and S. Silva, *J. Braz. Chem. Soc*, 2022, **33**, 681-692.

10. M. Mayer and A. J. Baeumner, *Chem. Rev.*, 2019, **119**, 7996-8027.
11. G. R. D. Prabhu, H. A. Witek and P. L. Urban, *React. Chem. Eng.*, 2019, **4**, 1616-1622.
12. R. A. Skilton, R. A. Bourne, Z. Amara, R. Horvath, J. Jin, M. J. Scully, E. Streng, S. L. Y. Tang, P. A. Summers, J. Wang, E. Pérez, N. Asfaw, G. L. P. Aydos, J. Dupont, G. Comak, M. W. George and M. Poliakoff, *Nat. Chem.*, 2015, **7**, 1-5.
13. I. S. Khan, M. O. Ahmad and J. Majava, *J. Clean. Prod.*, 2021, **297**, 126655.
14. T. Zonta, C. A. da Costa, R. da Rosa Righi, M. J. de Lima, E. S. da Trindade and G. P. Li, *Comput. Ind. Eng.*, 2020, **150**, 106889.
15. E. Alladio, M. Baricco, V. Leogrande, R. Pagliari, F. Pozzi, P. Foglio and M. Vincenti, *Front. Chem.*, 2021, **9**, 734132.
16. A. D. Clayton, J. A. Manson, C. J. Taylor, T. W. Chamberlain, B. A. Taylor, G. Clemens and R. A. Bourne, *React. Chem. Eng.*, 2019, **4**, 1545-1554.
17. J. Ke, C. Gao, A. A. Folgueiras-Amador, K. E. Jolley, O. De Frutos, C. Mateos, J. A. Rincón, R. C. D. Brown, M. Poliakoff and M. W. George, *Appl. Spectrosc.*, 2022, **76**, 38-50.
18. *Food and Drug Administration (FDA) Guidance for Industry: PAT—A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance*, Rockville, MD, Sep 2004.
19. B. R. Kowalski and J. E. Koch, in *Chemical Non-Destructive Evaluation at the Center for Process Analytical Chemistry*, Springer US, 1996, pp. 1-8.
20. A. Chanda, A. M. Daly, D. A. Foley, M. A. LaPack, S. Mukherjee, J. D. Orr, G. L. Reid, D. R. Thompson and H. W. Ward, *Org. Process Res. Dev*, 2015, **19**, 63-83.
21. D. E. Fitzpatrick, C. Battilocchio and S. V. Ley, *Org. Process Res. Dev*, 2016, **20**, 386-394.
22. D. Caramelli, J. M. Granda, S. H. M. Mehr, D. Cambié, A. B. Henson and L. Cronin, *ACS Cent. Sci.*, 2021, **7**, 1821-1830.
23. L. Wilbraham, S. H. M. Mehr and L. Cronin, *Acc. Chem. Res.*, 2021, **54**, 253-262.
24. K. Jorner, T. Brinck, P.-O. Norrby and D. Buttar, *Chem. Sci.*, 2021, **12**, 1163-1175.
25. Y. Guan, C. W. Coley, H. Wu, D. Ranasinghe, E. Heid, T. J. Struble, L. Pattanaik, W. H. Green and K. F. Jensen, *Chem. Sci.*, 2021, **12**, 2198-2208.
26. N. Collins, D. Stout, J.-P. Lim, J. P. Malerich, J. D. White, P. B. Madrid, M. Latendresse, D. Krieger, J. Szeto, V.-A. Vu, K. Rucker, M. Deleo, Y. Gorfu, M. Krummenacker, L. A. Hokama, P. Karp and S. Mallya, *Org. Process Res. Dev*, 2020, **24**, 2064-2077.
27. T. Hardwick and N. Ahmed, *Chem. Sci.*, 2020, **11**, 11973-11988.
28. C. F. Carter, H. Lange, S. V. Ley, I. R. Baxendale, B. Wittkamp, J. G. Goode and N. L. Gaunt, *Org. Process Res. Dev*, 2010, **14**, 393-404.
29. D. Angelone, A. J. S. Hammer, S. Rohrbach, S. Krambeck, J. M. Granda, J. Wolf, S. Zalesskiy, G. Chisholm and L. Cronin, *Nat Chem*, 2021, **13**, 63-69.
30. A. J. S. Hammer, A. I. Leonov, N. L. Bell and L. Cronin, *JACS Au*, 2021, **1**, 1572-1587.
31. S. Heron, M.-G. Maloumbi, M. Dreux, E. Verette and A. Tchapla, *J. Chromatogr. A*, 2007, **1161**, 152-156.
32. W. L. Fitch, A. K. Szardenings and E. M. Fujinari, *Tetrahedron Lett.*, 1997, **38**, 1689-1692.
33. R. Zeng, C. M. Mannaerts and Z. Shang, *Sensors (Basel)*, 2021, **21**, 6699.
34. S. V. Ley, R. J. Ingham, M. O'Brien and D. L. Browne, *Beilstein J. Org. Chem.*, 2013, **9**, 1051-1072.
35. D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher and A. G. Doyle, *Science*, 2018, **360**, 186-190.
36. P. Schwaller, A. C. Vaucher, T. Laino and J.-L. Reymond, *Mach. Learn.: Sci. Technol.*, 2021, **2**, 015016.

37. A. L. Haywood, J. Redshaw, M. W. D. Hanson-Heine, A. Taylor, A. Brown, A. M. Mason, T. Gärtner and J. D. Hirst, *J. Chem. Inf. Model.*, 2022, **62**, 2077-2092.
38. F. Strieth-Kalthoff, F. Sandfort, M. Kühnemund, F. R. Schäfer, H. Kuchen and F. Glorius, *Angew. Chem. Int. Ed.*, 2022, **61**, e202204647.
39. C. Mauger and G. Mignani, *Synth. Commun.*, 2006, **36**, 1123-1129.
40. L. McInnes, J. Healy, N. Saul and L. Großberger, *J. Open Source Softw.*, 2018, **3**, 861.
41. I.-K. Yeo, *Biometrika*, 2000, **87**, 954-959.
42. J. A. Bilbrey, C. O. Marrero, M. Sassi, A. M. Ritzmann, N. J. Henson and M. Schram, *ACS Omega*, 2020, **5**, 4588-4594.
43. W. Bort, I. I. Baskin, T. Gimadiev, A. Mukanov, R. Nugmanov, P. Sidorov, G. Marcou, D. Horvath, O. Klimchuk, T. Madzhidov and A. Varnek, *Sci. Rep.*, 2021, **11**, 3178.
44. N. Kunz, 2020, SMOGN: Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise, https://github.com/nickkunz/smogn, (accessed February 2022).