Using Active Learning to Develop Machine Learning Models for Reaction Yield Prediction

Simon Viet Johansson[1,2,*], *Hampus Gummesson Svensson*[1,2,†,*], *Esben Bjerrum*[1], *Alexander Schliep*[3], *Morteza Haghir Chehreghani*[2], *Christian Tyrchan*[4], *Ola Engkvist*[1]

1 Molecular AI, Discovery Sciences, R&D, AstraZeneca, Gothenburg, Sweden

2 Department of Computer Science and Engineering, Chalmers University of Technology, Gothenburg, Sweden

3 Department of Computer Science and Engineering, University of Gothenburg, Gothenburg, Sweden

4 Medicinal Chemistry, Research and Early Development, Respiratory and Immunology (R&I), BioPharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden

** Contributed equally*

† Corresponding author: hamsven@chalmers.se

Abstract

Computer aided synthesis planning is a rapidly growing field for suggesting synthetic routes for molecules of interest. The methods used are usually dependent on access to large datasets for training, but with a finite experimental budget there are limitations on how much data can be obtained from experiments. Active learning, which has been used in recent studies with success, is a strategy to identify which data points impact model accuracy the most. However, little has been done to explore the robustness of the methods predicting reaction yield. This study aims to investigate the influence of machine learning algorithms and the number of initial data points on reaction yield prediction for two public high-throughput experimentation datasets. Our results show that active learning based on output margin reached a pre-defined accuracy (AUROC) faster than using passive learning. Feature importance analysis of the trained machine learning models suggested active learning had larger influence on the model accuracy when only a few features were important for the model prediction.

Introduction

The recent advances in computer aided synthesis planning (CASP) [1–3] have made it a promising tool for finding and assessing plausible chemical routes. Accurately finding and assessing chemical routes can help to reduce the time required to find novel drugs and materials [4]. One subset of CASP, called forward prediction, is focused on the problem of predicting reaction outcomes in a forward synthesis manner. Forward prediction is mainly used to assess reaction plausibility and can help rank routes depending on whether they are predicted to be successful or not. This ranking can increase the likelihood of finding efficient routes to produce the target compound.

High-throughput experimentation (HTE) has emerged as a time- and material efficient technique for producing large amounts of chemical reaction data [5, 6]. HTE is thus a reasonable approach to yield datasets for CASP, which requires large training sets. Recent developments in HTE, and automation in general, have enabled platforms that can conduct and analyse thousands of experiments per day as demonstrated for batch chemistry [7, 8] as well as for flow chemistry [9]. However, it is not feasible to use HTE to investigate all necessary permutations of reaction variables in a typical reaction [10]. Therefore, it is important to identify the most informative data points, e.g., finding the smallest subset of data points that provides the most information about a reaction to a given machine learning model.

Active learning (AL) is a subfield of machine learning exploring different strategies of finding the most informative data points [11]. The aim is to determine which data points maximise the learning and accuracy of a machine learning model. During training, the learning model queries points to be labelled on its own to achieve this goal [11]. One approach to active learning, called pool-based sampling, assumes a small pool of labelled data *L* and a large pool of unlabelled data *U* [12]. Labels of unlabelled data are obtained by querying the labels of data points of *U* from an oracle, e.g., previously known results or conducted experiments. These data points are then added to the pool of labelled data *L* and used to improve the model, as illustrated in Figure 1. The goal in this case is to find the optimal model with the smallest possible pool of labelled data since each query of labels is associated with a cost, such as the cost of conducting an experiment. The problem is then to determine the data points that are most informative and whose labels therefore should be queried from the oracle. The data points that should be included in each query are determined by an acquisition function that estimates the informativeness of each data point. One popular type of active learning strategy is uncertainty sampling where the most uncertain data points, i.e., the data points the learning model knows the least about, are seen as most informative. A pre-determined acquisition function then determines the most uncertain batch of unlabelled data points whose labels should be queried from the oracle. Thus, the data points that the model is already confident about are avoided.

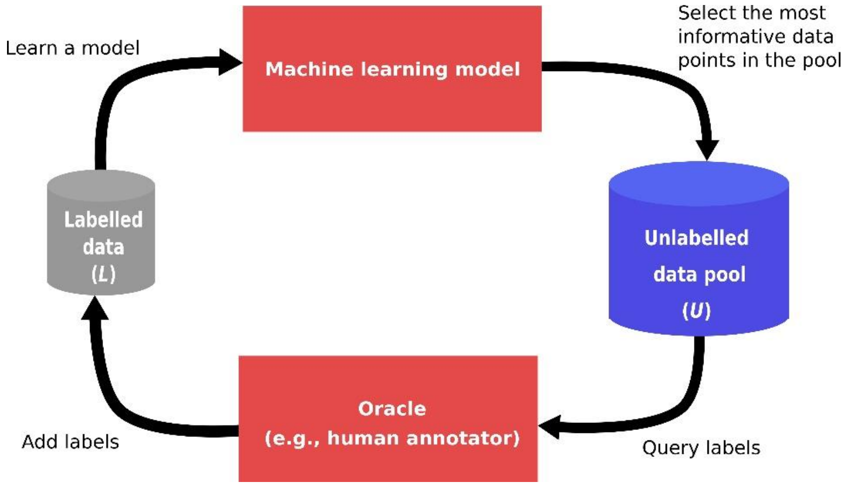
Learn a model
Machine learning model
Select the most informative data points in the pool
Labelled data (*L*)
Unlabelled data pool (*U*)
Oracle (e.g., human annotator)
Add labels
Query labels

Figure 1. The pool-based active learning cycle

For computational chemistry, active learning has been applied for various applications including drug design [12, 13]. For HTE, recent studies have applied active learning to select data points for neural network models. These had positive impact on reducing the number of experiments needed to generate a training set for predicting the reaction yield [14]. However, active learning still struggles to show a significant performance gain compared to randomly selecting data points to query, so called random sampling, when only a few data points have been labelled [14]. It is also possible that active learning performs differently in different settings, such as the initial pool of labelled data points and machine learning algorithm.

Here we have explored different settings of active learning in HTE to better understand when and if active learning can help a model reach a pre-defined level of accuracy faster compared to randomly selecting data points to label. We have done this in a forward prediction setting where reactions are predicted to be either successful or unsuccessful depending on the reaction yield. The rationale for using a binary classification model is that in discovery chemistry a reaction only needs to provide sufficient yield. This is different from process chemistry where the objective is to maximize the yield. In particular, we have investigated the uncertainty-based active learning strategy margin on different machine learning models and different numbers of initially labelled data points in the reaction data set.

Methods

We explored different settings for active learning for predicting the reaction yield of two combinatorially generated datasets. The model used for prediction and the number of initially labelled points were varied. The goal is to develop a prediction model of a predefined accuracy for binary classification of the reaction yield with the labels “successful” reaction and “unsuccessful” reaction.

Datasets

Previous studies from HTE provide fully combinatorial datasets that can be used to benchmark active learning strategies. Thus, this is a retrospective study where all true labels are known beforehand. However, we assume that only the true labels of the initial (labelled) training data points and queried data points are known at each iteration. We explored a Buchwald-Hartwig reaction data set [8] and a Suzuki reaction data set [9]. The Suzuki data set consists of 5769 Suzuki-Miyaura couplings with five varied reaction variables, namely, reactant 1, reactant 2, ligand, base and solvent. We discarded the fourth choice of reactant 2 to obtain a fully combinatorial dataset which consisted of 4608 data points. The Buchwald-Hartwig data set consists of 4608 cross-couplings of aryl halides with four reaction variables, aryl halide, additive, ligand and base.

One-hot encodings were used to represent the different combinations of reaction variables described above in the datasets. Both datasets consisted of the reaction yield of every combination of reaction variables. The distribution of reaction yields for all 4608 reactions in each dataset are displayed in Figure 2. A hard threshold of 20% yield was used to determine the label of each data point, i.e., one combination of reaction variables. A reaction with a yield above this threshold was labelled as a “successful” reaction, encoded as class 1. In the same way, a reaction with a yield below 20% was labelled as an “unsuccessful” reaction, encoded as class 0. The ratios of “successful” reactions are 54% and 65% for the Buchwald-Hartwig and Suzuki data set respectively. In order to evaluate the performance on these datasets, each dataset was randomly split into a training and test set consisting of 80% and 20% respectively, of all data. We evaluated the performance in terms of the area under the receiver operating characteristic curve (AUROC) with the goal of reaching the largest AUROC with the least number of labelled training data points from each dataset. To be robust against fluctuations in the predictions, we used a moving average to determine if an experiment reached the pre-defined desired AUROC. A run was determined to have reached the (desired) AUROC after k queries if:

$$\frac{\sum_{i=k-n}^{k+n} \mathrm{AUROC}_i}{2n+1} \geq \mathrm{AUROC\_target},$$

where AUROC_i is the AUROC after the *i*-th query, AUROC_target is desired AUROC and *n* is the number of values of the moving average considered before and after the k-th query. In this study, $k = 3$ was used. Zero-padding was used to obtain a moving average for all number of queries.

Initial pool of labels

In order to investigate how the initial pool of labelled data points affects the performance of active learning, we evaluated three different sizes of the initial pool of labelled training data points for active learning, namely, 10, 100 and 1000 labelled training data points. Five sets were randomly selected for each number of initial data points, which gives 15 different initial pools per data set.

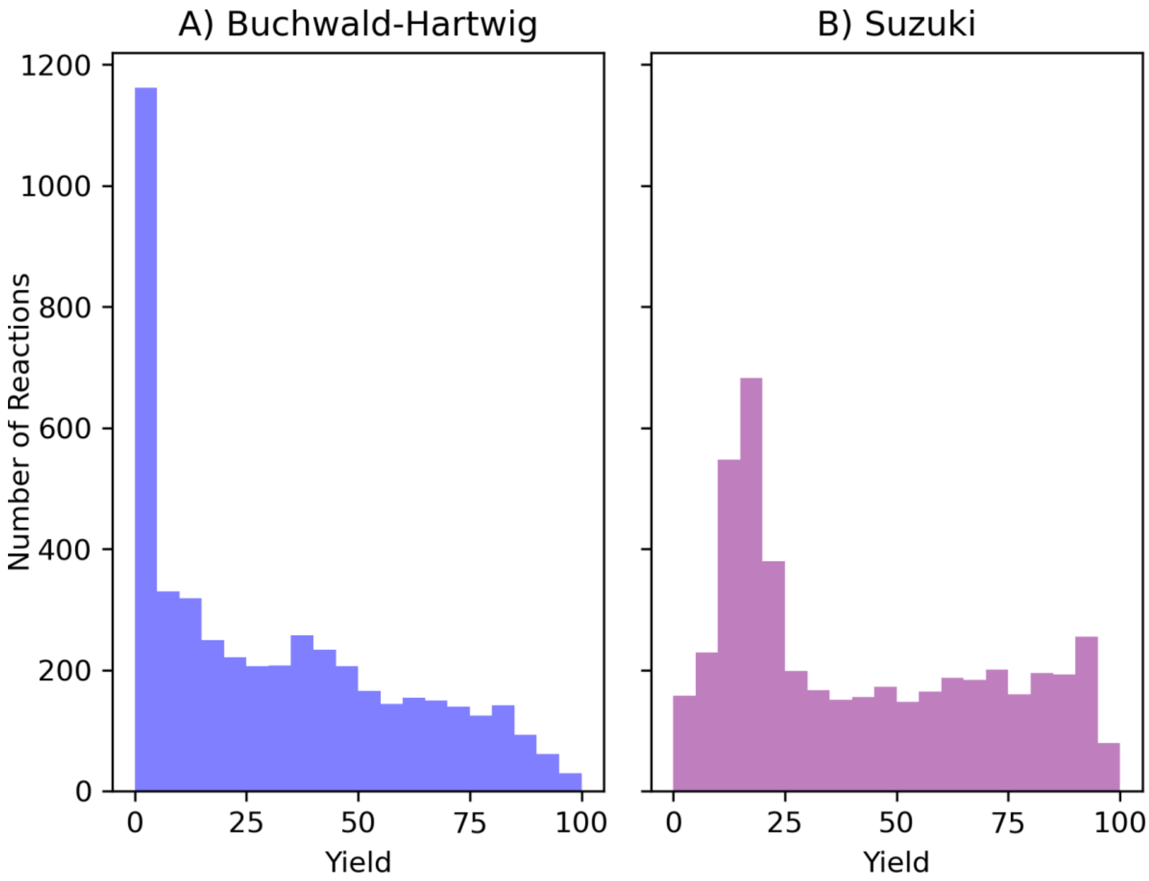

Figure 2. The distributions of reaction yields of the 4068 reactions in the Buchwald-Hartwig and Suzuki datasets.

Models

For both datasets, we investigated four different models: (1) simple neural network; (2) complex neural network; (3) Bayesian matrix factorization model [15]; (4) random forest classifier [16, 17] from scikit-learn [18]. The models are visualised in Figure 3. We trained the models using cumulative learning, i.e., each model was retrained after every query. Cumulative learning usually obtains better results in active learning compared to incremental learning [19].

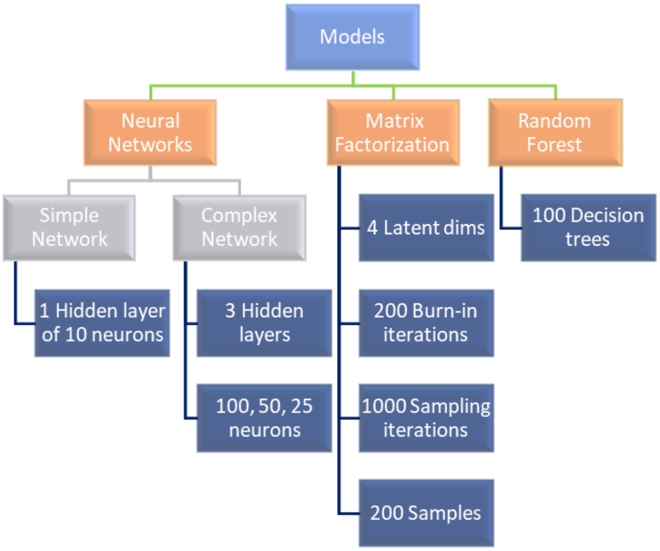

Figure 3. Schematic illustration of the different machine learning models that were investigated.

Neural Networks

The neural networks were developed using PyTorch 1.3.1 [20] together with the PyTorch wrapper PyTorch Lightning 1.0.8 [21]. The experiments were conducted using Nvidia K80 GPUs with driver version 450.80.02, CUDA 11.0. The simple neural network had one hidden layer with 10 neurons; while the complex neural network had three hidden layers with 100, 50 and 25 neurons, respectively. The output layers consisted of two neurons (one for each class) and the input layer corresponded to the one-hot encoding. The complex neural network used dropout with probability 0.5 while the simple neural network used no dropout. Moreover, the networks used Leaky ReLU [22] as activation functions for the hidden layers and softmax as the activation function of the output. Optimization of the parameters was performed using AdamW [23] with a learning rate of 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.99$, ε e $= 10^{-8}$, a weight decay of 0.01 and a mini-batch size of 8. The networks were trained for 50 epochs after each query of active learning. For each set of a specific number of initially labelled data points, the same initial weights of the initial epoch were utilized.

Bayesian Matrix Factorization

The matrix factorization method Macau [24] in the Smurff 0.16.0 [15] framework was used for Bayesian Matrix factorization. The model used four latent dimensions and 1200 training iterations using a probit noise model, of which 200 where burn-in samples and 1000 used for Bayesian sampling. We sampled every 5th iteration, which yielded 200 predictions. The probability used for a label was the frequency that the label had been estimated to be the likely label in the 200 predictions.

Random Forest

The random forest classifier in Scikit-learn 0.24.2 was used [18]. The random forest model consisted of 100 decision trees. The model uses the Gini impurity criterion for assessing the quality of the node splits. On initialisation of each iteration, the individual trees were given bootstrapped datasets.

Active learning

We investigated an uncertainty-based active learning strategy called *Margin* [11]. Margin queries data points based on the output margin

$$x^* = \underset{x}{\operatorname{argmin}}[P_\theta(\hat{y}_1|\, x) - P_\theta(\hat{y}_2|\, x)],$$

where $P_\theta(y|x)$ is the probability, of an arbitrary classification algorithm, that a data point x belongs to class y, and $\hat{y}_1$ and $\hat{y}_2$ are the most and second most probable predictions, respectively, of the algorithm. Previous works have shown better performance of Margin compared to other strategies [19, 25].

Margin was compared to random sampling where random unlabelled data points are labelled for each query. Furthermore, we evaluated the active learning setting where a batch of one data point is labelled for each query. We investigated how the different combinations of model, query strategy (either random sampling or margin), and initial pool of labelled data points affect the number of labelled data points required to achieve specific pre-defined levels of AUROC. The pre-defined levels of AUROC were 0.800, 0.850, 0.900 and 0.950.

For each combination of model, query strategy and set of initially labelled training data points, we investigated the performance of up to 1500 training data points that had been labelled in total. Moreover, each combination was run five times to investigate the stochastic behaviours of the models and query strategies. Also, for each of the five different runs for a specific initial pool of labelled training data points, the same initial weights of the neural networks were used. Since training was cumulative, this means that the same initial weights were used to retrain the neural networks after each query.

Results

Figures 4(a)-(c) show boxplots that illustrate the required number of labelled data points to reach an AUROC of 0.800, 0.850 and 0.900 for the Buchwald-Hartwig reaction when either 10 or 100 data points are initially labelled. The settings when 1000 data points were initially labelled reached the target AUROC scores using only the initially labelled data points and are, therefore, not shown in the figures. All 25 runs obtained an AUROC of 0.800, 0.850 and 0.900. When utilizing complex and simple neural network with the Buchwald-Hartwig data, starting with 10 randomly labelled data points seems to require a lower number of labelled data points to achieve an AUROC of 0.800 or 0.850. This is compared to starting with 100 randomly labelled data points. Matrix factorization with 10 initial data points shows a lower number of required labelled data points, compared to starting with 100 initial data points to achieve the target AUROC. No substantial difference is displayed when utilizing random sampling compared to margin to obtain an AUROC of 0.800 and 0.850. For a pre-defined target AUROC of 0.900, utilizing margin require a lower number of labelled data points when using random forest. Figures 4(d)-(e) display boxplots of the required number of labelled data points to reach an AUROC of 0.950 and 0.975, respectively, when either 10, 100 or 1000 data points are initially labelled. The numbers above the boxplots show the number of runs that reach the AUROC if not all runs did. Both figures show a lower number of required labelled data points when using Margin with 10 or 100 initial data points, compared to using random sampling. When utilizing margin with 1000 initial data points, more runs reach the AUROC of 0.975 compared to the runs when random sampling was used.

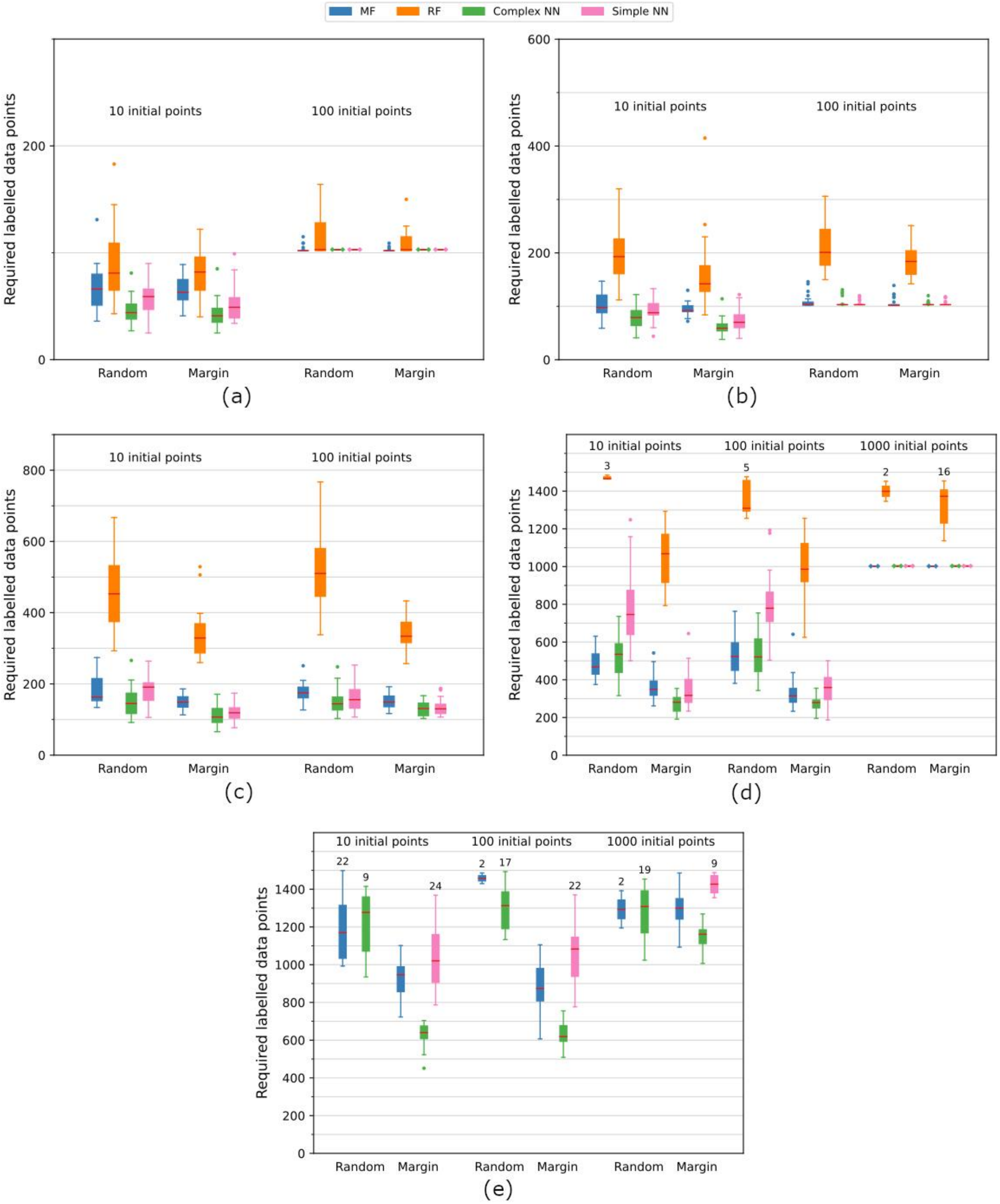

Figure 4. Boxplots showing the number of required labelled data points for the Buchwald-Hartwig reaction to achieve an AUROC of (a) 0.800, (b) 0.850, (c) 0.900, (d) 0.950 and (e) 0.975. When using 1000 initially labelled data points, all models reached the AUROC of 0.800, 0.850 and 0.900 by using only the initial labels and, therefore, these models are not shown.

Figures 5(a)-(d) display boxplots showing the required number of labelled data points to obtain an AUROC of 0.800, 0.850, 0.900 and 0.950, respectively, for different settings for the Suzuki reaction data set. None of the setting obtained an AUROC of 0.975. Therefore, no figure is displayed for this AUROC for the Suzuki data set. For pre-defined levels of AUROC of 0.800, 0.850 and 0.900, settings with 1000 initial data points reached these levels by only using the initial data points. Thus, these are not shown in the figures. For settings where not all of the 25 runs reached the pre-defined level of AUROC, a number denoting the number of models that succeeded is shown above the respective boxplots. Comparing random sampling and margin, no substantial difference in the required number of labelled data points is seen for an AUROC of 0.800, 0.850 or 0.900. As seen in Figure 5(d), the complex neural network is the only model that reaches an AUROC of 0.950 at least once for each setting. Moreover, the complex neural network has more runs that reaches the desired AUROC when margin is utilized compared to random sampling. For 10 initial data points, matrix factorization and the complex neural network utilizing random sampling is able to reach an AUROC of 0.950; while random forest and the complex neural network are able to reach this AUROC when utilizing margin. Furthermore, when using 100 initial data points and random sampling, the complex neural network reaches an AUROC 0.950. When utilizing margin, the target is obtained by random forest on five runs, the complex neural network on 23 runs and the simple neural network on one run. For 1000 initial data points and utilizing random sampling, only one run of the complex neural network obtains an AUROC of 0.950; while 20 runs utilizing margin reaches this AUROC.

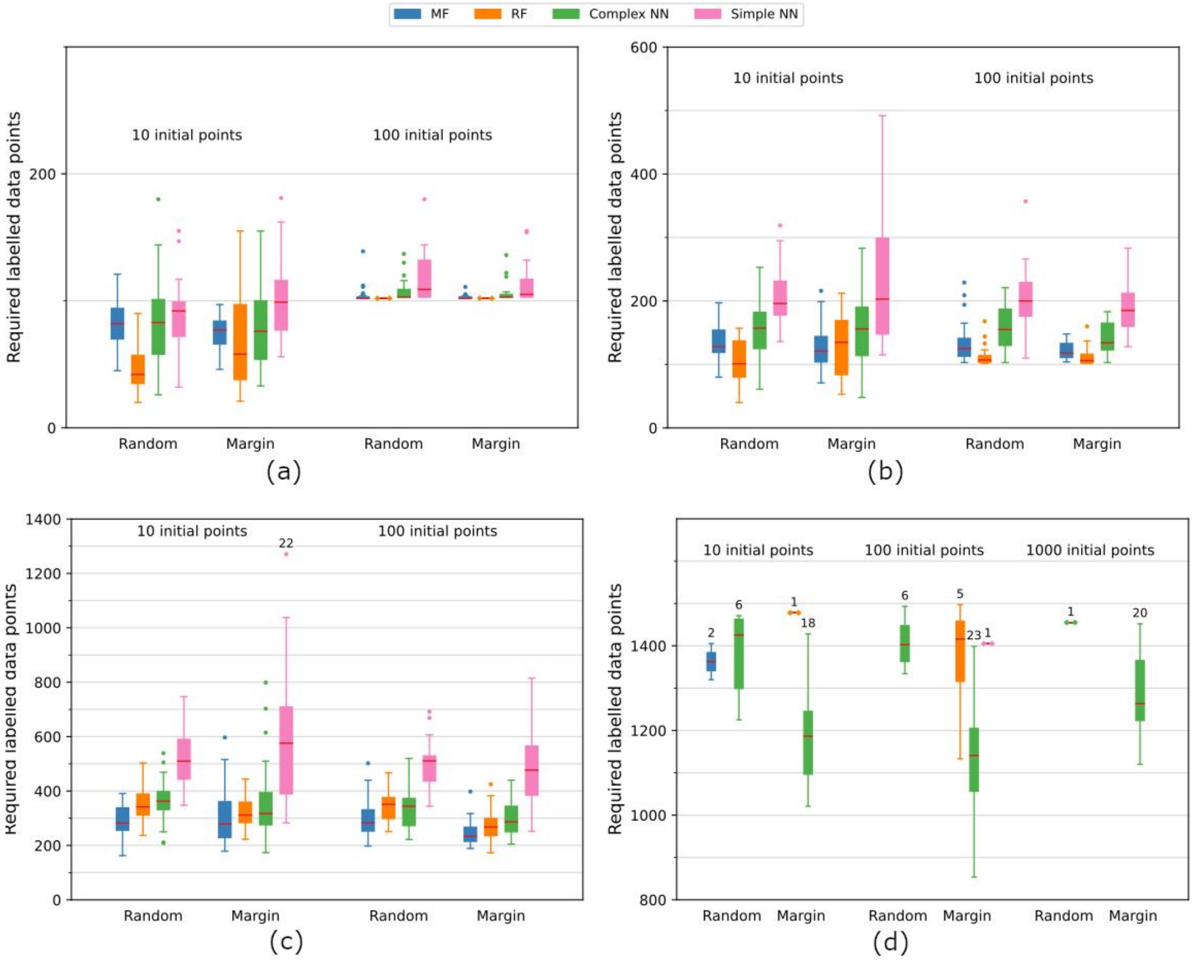

Figure 5. Boxplots showing the number of required labelled data points of the Suzuki reaction data set to achieve an AUROC of (a) 0.800, (b) 0.850, (c) 0.900 and (d) 0.950. When using 1000 initially labelled data points, all models reached an AUROC of 0.800, 0.850 and 0.900 by using only the initial labels and, therefore, these setting are not displayed. No setting reached an AUROC score of 0.975.

Discussions

The results show that there might be a relation between the machine learning models learning the observed reaction data and how well the margin acquisition function performed. As indicated by the Buchwald-Hartwig dataset, the more accurate the models became, the better query by margin sampling performs. This is consistent with earlier studies [14] showing good results after a certain number of labelled data points. However, we can already observe the effect at ca. 400 labelled data points. This number of queries is relatively large compared to 10 and 100 labelled data points. Consequently, there is no substantial difference in performance when starting with 10 or 100 labelled points. For the Suzuki reaction case it was more difficult for the matrix factorization, random forest and simple neural network models to show a substantial difference between the learning strategies. It was only for the complex neural network that an impact from the Margin acquisition function was observed, by more runs successfully reaching the 0.950 threshold. The AUROC and the performance of the active learning are indications that the Suzuki dataset is more difficult to build an accurate reaction yield prediction model using the studied machine learning algorithms.

We tried to investigate the underlying reasons of why there is a difference in both model performance and impact of learning strategy between the data sets. Firstly, we can observe that the Suzuki dataset has five discrete parameters that varied in the combinatorial space as opposed to the four for Buchwald-Hartwig. The extra parameter leads to additional interactive effects from a model building perspective. Secondly, the dataset was more skewed towards the "successful" label, but there was not a severe imbalance between the two labels. The dependency on each feature of the dataset with respect to the labels was investigated in the following way. For each reaction in the datasets, we fixed all features except one and computed the frequency that the label would remain the same while varying the non-fixed feature. We repeated this frequency computation across all features in both datasets. This yields a number between 0.5 and 1, where a frequency of 1 represents all reactions having the same label when a given feature was changed, implying that the feature would be completely interchangeable and uninformative for the classification task. The frequency 0.5 represents the case when all reactions have an even 50% split of labels. There is not a linear relationship between lower frequency and more information. These mean frequencies across each feature are presented in Table 1.

Table 1. The mean frequencies that the label (either a successful or unsuccessful reaction) would remain the same when fixating one feature and varying the other features.

Dataset	Feature	Frequency
Buchwald-Hartwig	Aryl halide	0.61
	Ligand	0.83
	Base	0.81
	Additive	0.74
Suzuki	Reactant 1	0.66
	Reactant 2	0.65
	Ligand	0.75
	Reagent	0.88
	Solvent	0.73

It can be observed that the bases and ligands for Buchwald-Hartwig seem to have a relatively high frequency of retaining a label across variations, as does the reagent for the Suzuki reaction in the studied space. This cannot explain the whole nature of interactive effects for these parameters, as they are not independent. For the random forest model, the feature importance was analysed for both datasets with models trained with 1500 labelled data points, which were obtained by using 10 initially labelled data points and utilizing Margin to query labels, see Table 2. This analysis shows that while the more heterogenous features were directly related to a higher importance to the model predictions in the Buchwald-Hartwig reactions, the same was not true for the Suzuki counterparts. Here the ligand was overwhelmingly the most important factor for the model outcome, and the other four features played a smaller role, although the reactants do stand out as being of the least importance to the model. To do the same analysis for the neural network, we computed the Integrated Gradients (IGs) [26], with the zero vector as baseline, for each reaction of the Buchwald-Hartwig and Suzuki data sets. Similarly for the random forest, this was done with 1500 labelled data points, which were obtained by using 10 initially labelled data points and utilizing Margin to query labels. The IGs were averaged over all data to obtain a single value for each feature and the value were normalised which is displayed in Table 2. This provides a way to compare the relative importance of the features. For the Buchwald-Hartwig data, we see that the aryl halide and additive are the most important features which is consistent with the feature importance obtained by the random forest model. For the Suzuki data, reactant 1 seems to be the most important feature when inspecting the average IGs. However, parallels can be drawn between the model feature importance and the explainable variance in a principal component analysis. This allows us to conclude that for the random forest model and complex neural network, two features in the Buchwald-Hartwig dataset accounted for ca. 80% of the prediction outcome with reasonably a good score, which is a good indication that the dataset was easy to learn for both models. On the other hand, the random forest model for the Suzuki dataset does not reach the same level of feature importance even when using the three most important features. As each feature individually plays a larger role – the difference in feature importance outside of the ligand feature is small, and is indicated by the average IGs for the Suzuki datasets. Since a higher dimension space of knowledge is needed to make the decision, it can be concluded that the Suzuki reaction data set is more complex to learn using the observed models.

Table 2. Feature importance (FI) of random forest and average Integrated Gradients (IG) of the complex neural network over all reactions in the Buchwald-Hartwig and Suzuki data sets. Note that the average values of the integrated gradients are normalized. The FI and IG were derived from models trained with 1500 labelled data points, which was obtained by using 10 initially labelled data points and utilizing margin to query labels.

Dataset	Feature	Feature importance Random Forest	Average Integrated Gradients (normalized)
Buchwald-Hartwig	Aryl halide	0.47	0.58
	Ligand	0.11	0.04
	Base	0.09	0.01
	Additive	0.33	0.37
Suzuki	Reactant 1	0.14	0.40
	Reactant 2	0.10	0.18
	Ligand	0.38	0.25
	Reagent	0.21	0.07
	Solvent	0.17	0.10

Besides showing that there might be a relation between the models learning the observed space and how well that margin performed, the results of this study imply that there is no drawback to implementing an active learning strategy starting at a small number of labelled data points. There was no observed threshold where the margin query scheme performed *worse* than a random scheme. As both learning schemes operate within the same space of data, fully trained models should converge towards the same AUROC, at which point it is no longer interesting to use active learning. The benchmarking used in this study focused on minimizing the experimental cost by looking at which point a desired AUROC was met and found an observable advantage for active learning. It is important to note however, that these experiments were computed with a query consisting of a batch of one data point. The experimental designs of full batches of HTE experiments as used in the Buchwald-Hartwig dataset are typically of sizes 96, 384 or 1536. Thus, the results of this study might not be directly translatable to this setup. If analysis time can be reduced to same time-scale of the conducted experiments, the results could be of interest in flow-process design, such as that used in the generation of the Suzuki dataset.

Conclusions

We have explored active learning on two different high-throughput experimentation datasets. The aim was to investigate if active learning can be used to efficiently learn to predict reaction yields with a certain AUROC. We focused on comparing the uncertainty-based active learning strategy Margin to random sampling, and investigated matrix factorization, random forest, simple and complex neural networks. In our study we explored initial pools of 10, 100 and 1000 labelled data points.

We have observed that active learning, in particular the strategy Margin, can provide a model of pre-defined AUROC with a smaller number of labelled data points. In fact, the better model AUROC we want to achieve, the more evident is the gain in AUROC that can be obtained by using active learning. Moreover, we have observed that the reduction in number of required labelled data points with active learning differs between different data sets. This could be due to different complexity of the data sets. Our feature importance analysis suggests that the models reach a higher AUROC when only few features were important.

Using active learning to create optimal training sets for building machine learning models for reaction yield prediction is an efficient way to reduce the experimental efforts needed.

Code Availability

The source code used to run the experiments is available through a GitHub repository (https://github.com/hampusgs/AL-for-reaction-yield-prediction).

Conflict of interest

The authors declare no conflict of interest.

Funding

SJ, HGS, AS and MHC are supported by the Wallenberg Artificial Intelligence, Autonomous Systems, and Software Program (WASP), funded by the Knut and Alice Wallenberg Foundation, Sweden.

Acknowledgements

The authors want to thank Mr. Carl Blomgren for scientific discussion and input. The authors further thank Mr. Samuel Espley for his efforts in widening our perspective for HTE and the challenges for integrating an active learning feedback loop into a HTE batch workflow. Also, the authors want to thank Mr. Amol Thakkar for constructive criticism of the manuscript and many scientific discussions. Finally, the authors thank the Molecular AI team at AstraZeneca for numerous discussions and feedback.

References

1. Johansson S, Thakkar A, Kogej T, Bjerrum E, Genheden S, Bastys T, Kannas C, Schliep A, Chen H, Engkvist O (2019) AI-assisted synthesis prediction. Drug Discov Today Technol 32–33:65–72

2. Engkvist O, Norrby P-O, Selmi N, Lam Y, Peng Z, Sherer EC, Amberg W, Erhard T, Smyth LA (2018) Computational prediction of chemical reactions: current status and outlook. Drug Discov Today 23:1203–1218

3. Coley CW, Green WH, Jensen KF (2018) Machine Learning in Computer-Aided Synthesis Planning. Acc Chem Res 51:1281–1289

4. Kirman J, Johnston A, Kuntz DA, Askerka M, Gao Y, Todorović P, Ma D, Privé GG, Sargent EH (2020) Machine-Learning-Accelerated Perovskite Crystallization. Matter 2:938–947

5. Isbrandt ES, Sullivan RJ, Newman SG (2019) High Throughput Strategies for the Discovery and Optimization of Catalytic Reactions. Angew Chemie Int Ed 58:7180–7191

6. Mahjour B, Shen Y, Cernak T (2021) Ultrahigh-Throughput Experimentation for Information-Rich Chemical Synthesis. Acc Chem Res. https://doi.org/10.1021/acs.accounts.1c00119

7. Lin S, Dikler S, Blincoe WD, et al (2018) Mapping the dark space of chemical reactions with extended nanomole synthesis and MALDI-TOF MS. Science (80-) 361:eaar6236

8. Ahneman DT, Estrada JG, Lin S, Dreher SD, Doyle AG (2018) Predicting reaction performance in C–N cross-coupling using machine learning. Science (80-) 360:186–190

9. Perera D, Tucker JW, Brahmbhatt S, Helal CJ, Chong A, Farrell W, Richardson P, Sach NW (2018) A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. Science (80-) 359:429 LP-434

10. Murray PM, Tyler SNG, Moseley JD (2013) Beyond the Numbers: Charting Chemical Reaction Space. Org Process Res Dev 17:40–46

11. Valiant LG (1984) A Theory of the Learnable. Commun ACM 27:1134–1142

12. Settles B (2012) Active Learning. Synth Lect Artif Intell Mach Learn 6:1–114

13. Lewis DD, Gale WA (1994) A Sequential Algorithm for Training Text Classifiers. In: SIGIR ’94. Springer London, London, pp 3–12

14. Eyke NS, Green WH, Jensen KF (2020) Iterative experimental design based on active machine learning reduces the experimental burden associated with reaction screening. React Chem Eng 5:1963–1972

15. Aa T Vander, Chakroun I, Ashby TJ, et al (2019) SMURFF: a High-Performance Framework for

Matrix Factorization.

16. Ho TK (1995) Random Decision Forests. In: Proc. Third Int. Conf. Doc. Anal. Recognit. (Volume 1) - Vol. 1. IEEE Computer Society, USA, p 278

17. Breiman L (2001) Random Forests. Mach Learn 45:5–32

18. Pedregosa F, Varoquaux G, Gramfort A, et al (2011) Scikit-learn: Machine Learning in Python. J Mach Learn Res 12:2825–2830

19. Bossér JD, Sörstadius E, Chehreghani MH (2020) Model-Centric and Data-Centric Aspects of Active Learning for Neural Network Models.

20. Paszke A, Gross S, Massa F, et al (2019) PyTorch: An Imperative Style, High-Performance Deep Learning Library. Adv. Neural Inf. Process. Syst. 32:

21. Falcon W, Borovec J, Wälchli A, et al (2020) PyTorchLightning/pytorch-lightning: 0.7.6 release. https://doi.org/10.5281/ZENODO.3828935

22. Maas AL, Hannun AY, Ng AY (2013) Rectifier Nonlinearities Improve Neural Network Acoustic Models.

23. Loshchilov I, Hutter F (2017) Decoupled Weight Decay Regularization.

24. Simm J, Arany A, Zakeri P, Haber T, Wegner JK, Chupakhin V, Ceulemans H, Moreau Y (2015) Macau: Scalable Bayesian Multi-relational Factorization with Side Information using MCMC.

25. Körner C, Wrobel S (2006) Multi-class Ensemble-Based Active Learning. pp 687–694

26. Sundararajan M, Taly A, Yan Q (2017) Axiomatic Attribution for Deep Networks.