The Effect of Chemical Representation on Active Machine Learning Towards Closed-Loop Optimization

A. Pomberger,[1] A.A. Pedrina McCarthy,[2] A. Khan,[1] S. Sung,[3] C. J. Taylor,[1,4] M.J. Gaunt,[2] L. Colwell,[2] D. Walz[5] and A.A. Lapkin[1,3]

[1]*Department of Chemical Engineering and Biotechnology, Cambridge, CB3 0AS, United Kingdom*

[2]*Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, United Kingdom*

[3]*Cambridge Centre for Advanced Research and Education in Singapore Ltd., CREATE Tower 05-05, 138602 Singapore*

[4]*Astex Pharmaceuticals, 436 Cambridge Science Park, Milton, Cambridge CB4 0QA, United Kingdom*

[5]*BASF SE Data Science for Materials, Carl-Bosch-Strasse 38, 67056 Ludwigshafen am Rhein, Germany*

Abstract

Multivariate chemical reaction optimization involving catalytic systems is a non-trivial task due to the high number of tuneable parameters and discrete choices. Active Machine Learning (ML) represents a powerful strategy for automating reaction optimization. However, the translation of chemical reaction conditions into a machine-readable format requires the identification of highly informative features which accurately capture the factors which determine reaction success. Herein, we compare the efficacy of different calculated chemical descriptors for a high throughput experimentation generated dataset to determine the impact on a supervised ML model when predicting reaction yield. Then, the effect of featurization and size of the initial dataset within a closed-loop reaction optimization was examined. Finally, the balance between descriptor complexity and dataset size was considered. Ultimately, tailored descriptors did not outperform simple generic representations, however, a larger initial dataset accelerated reaction optimization.

***Keywords**: reaction optimization; machine learning; high-throughput experimentation; HTE; molecular parameterization; closed-loop optimization*

Introduction

Identifying the optimal reaction conditions to enact a specific transformation is a major challenge for chemists, particularly in the field of small molecule drug synthesis and natural product synthesis.[1, 2] The field of laboratory automation allows for the rapid and systematic generation of high-quality data, which, when used in combination with ML-directed self-optimization algorithms can become a powerful tool for research.[3-8] In order to apply such tools, a chemical reaction must be represented in a machine-readable format. This representation must be composed of descriptors that are simple and relevant enough to avoid the introduction of undesired noise, yet information-rich, enough to account for properties that impact reaction success such as sterics and electronics.

Unlike the large datasets used for ML in other disciplines (e.g., image recognition), synthetic chemistry datasets are often extremely small and, to compensate, researchers often develop bespoke descriptors, which are based on expert knowledge such as mechanistic understanding or quantum chemical calculations.[9-12] However, it is possible that the descriptors generated contain little relevant information and are simply perceived as distracting noise by the ML model. In this publication, we aim to investigate the relationship between descriptor complexity and ML model performance when predicting the yield of chemical reactions. Furthermore, we aim to explore how the descriptor complexity impacts closed-loop optimization, a strategy that may help to guide synthetic chemists towards optimal reaction conditions.

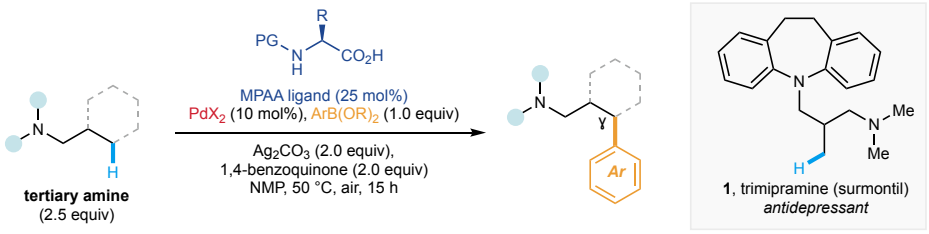
Whilst it can be challenging for humans to identify complex relationships in large datasets, ML relies on building statistical models that adjust to the given input data and has recently proven to be a powerful tool for the successful identification of nuanced patterns in complex data.[13-17] Trained ML models can be used to make predictions when given inputs that lie within the defined parameters of the training dataset, referred to as interpolative tasks. This includes new combinations of already known components, such as catalysts/ligands/additives. In contrast, extrapolative tasks, which are represented by

predictions of inputs which are not represented in the training data are challenging, with the predictive power of an ML model decreasing as structural differences between the training and test data increase. These extrapolative tasks require the ML model to learn about the fundamental chemical properties and, as such, the inputs are generally molecular descriptors that are based on fundamental molecular properties such as atomic distances, orbital energies or charge distributions.[18]

The application of ML as a decision-making tool during reaction optimization represents an effective combination as it accelerates experimental workflows and allows for rapid gains in understanding. Active ML-driven closed-loop optimization uses an initial dataset to make predictions about yet unseen conditions and these predictions can inform decisions about the subsequent experiments. For example, the experiments predicted to deliver the highest yields or the greatest improvement in model performance can be prioritized and conducted. The data gained from running this experiment can be used to re-train the ML model and new predictions are made. This iterative process continues until the desired objective (such as more accurate predictions or increased yield) is fulfilled. ML-based optimizers have the potential to increase the efficiency in which chemical space is navigated, removing operator bias and ultimately reducing the total number of experiments required, thus reducing waste significantly.[7, 19-21]

Previous reports vary in their conclusions on whether the implementation of chemical descriptors, rather than generic one-hot encoding (OHE) representations, truly boosted the predictive performance of their ML models.[19, 22, 23] Pd-catalyzed C(sp^3)–H activation is a powerful reaction manifold that enables the facile introduction of functional complexity in small molecules, in addition to the late-stage functionalization of complex molecules like trimipramine (**1**), a tricyclic antidepressant, as recently demonstrated by one of our groups.[24] Hence we chose to explore the parameterization and featurization of the newly developed tertiary amine directed C(sp^3)–H bond activation with a HTE-generated dataset, comparing tailored descriptors, based on *in silico* studies, to understand the influence of descriptor complexity on supervised ML prediction and closed-loop optimization.

Traditional round-bottom flask chemistry is still the major strategy for reaction optimization; however, it is limited by the capacity of a human experimentalist and experimental set-up can vary between chemists, unintentionally introducing sources of error. Hence, we chose to employ high-throughput experimentation (HTE) to conduct experimental arrays in parallel, increasing the rate of data generation and improving reproducibility. To the best of our knowledge, this is the first application of this chemical transformation in HTE.



Scheme 1 General conditions for catalytic C(sp^3)–H bond activation of tertiary alkylamines. Original conditions: $Pd(OAc)_2$ (10 mol%), (L)-tert-Leucine (25 mol%), $ArB(OH)_2$ (1.0 equiv).

The reaction contains three discrete reaction parameters which were varied in tandem: mono-*N*-protected amino acid (MPAA) ligand, the palladium pre-catalyst, and the aryl boronate (Scheme 1). All combinations of 31 ligands (plus one control), three pre-catalysts, and two boronates results in 186 unique conditions that were each run in quadruplicate on a 125 nmol scale using nanoscale HTE, outliers were eliminated, and the repeats averaged to ensure reproducibility. For the active learning studies, this dataset would serve as the navigable chemical space that experiments would be simulated within.

Materials and Methods

Molecular Parameterization

Developing machine-readable representations that can capture the correlation between structure and reactivity represents a major challenge within computational chemistry.[30] A key consideration when choosing a parameterization method is that, depending on the ML model used, the sparsity of input features (i.e., the location and number of ones and zeros in a bit vector) may influence modelling performance by undermining relevant information within the input vector.

Within the obtained dataset a varied structure-reactivity relationship was demonstrated by the ligands, with >90% of the variation in yield is attributable to ligand choice, and their role as knock-out criteria for the reaction, we focussed primarily on parameterizing them whilst the boronates and pre-catalysts were encoded by means of OHE and Morgan 2 fingerprints only.

MPAA ligands (Figure 1), first popularized by Yu and co-workers in 2008,[25] are uniquely favoured ligands for Pd-catalyzed C–H activation as both the carboxylate and amide motifs have relatively weak coordination strengths when compared to phosphine or NHC-type ligands, and both are able to bind to Pd as L-type (neutral) and X-type (anionic) ligands enabling them to dynamically adjust between coordination modes throughout the catalytic cycle.[26] Diversity within this ligand-set comes primarily from variation of the α-substituent, and with the amide protecting group, enabling a wide range of steric and electronic profiles to be generated. The subtle impact these modifications may have on reactivity, is unintuitive and, as such, it is thought that ML may be able to aid in predicting reaction outcomes for reactions that depend on these ligands.

Although we concede that the conformation of the ligand when alone and when in the catalytic complex differs, we aimed to keep our approach computationally inexpensive and attempted to derive descriptors from the unbound ligands (instead of the ligand-metal complex). We chose three levels of chemical parameterization complexity which contained sequentially more chemical information at the expense of increased computing time. Initially, one-hot encoding (OHE) was used, a computationally simplistic method that merely details the presence or absence of certain reagents whilst encoding no chemical information. For the inclusion of structural information, Morgan fingerprints (radius of 2, see SI for discussion) were chosen to encode the molecular fragments that were present in a given reaction mixture.[27] These bit vectors contain information on atom types, neighbouring connectivity relationships and bond types. Finally, steric and electronic descriptors were calculated for the set of ligands using density functional theory (DFT) from geometry optimized structures (B3LYP functional and 6-31G(d) basis set, see SI for full details). Within this investigation different sub-sets of descriptors were used and combined to develop hybrid molecular representations, containing information of the fingerprints and DFT descriptors.

As discussed previously, there are two main positions of diversity within MPAA ligands: the α-carbon and the acetamide protecting group. Hence, to separate their influence on reaction outcome, both positions were parameterized separately for each ligand rather than using a single descriptor for the entire molecule. The ligand's steric profile was expected to influence reaction yield greatly – preliminary work clearly identified an increase in yield with increasing size of the group on the α-carbon: H → CH_3 → $C(CH_3)_3$, for more detailed insights see Rodrigalvarez *et al.*[24] Two steric descriptors were introduced, Sterimol and percentage buried volume (%VBur).[28, 29] Sterimol descriptors quantify steric demands along different principal axes, making them well-suited to describing the steric effects of unsymmetrical substituents. The percentage buried volume is a descriptor that is traditionally used for catalyst-ligand complexes and describes the percentage of the volume of a sphere that is occupied by a given substituent. We used the α-carbon/[*N*-residue] as centre of our sphere and the calculation was performed considering only the variable residue extending from this position. In addition to this, a number of electronic descriptors were calculated for the ligand molecules as the fine-tuned electronics can impact Lewis basicity of the two binding atoms (N and O) and the aptitude to engage in the key mechanistic step (concerted metalation deprotonation, CMD). To capture the electron density distribution, we calculated the HOMO/LUMO energies and conducted a NBO analysis (Natural Bond Orbital) and a CHELPG analysis (CHarges from ELectrostatic Potentials using a Grid-based method). [30, 31]

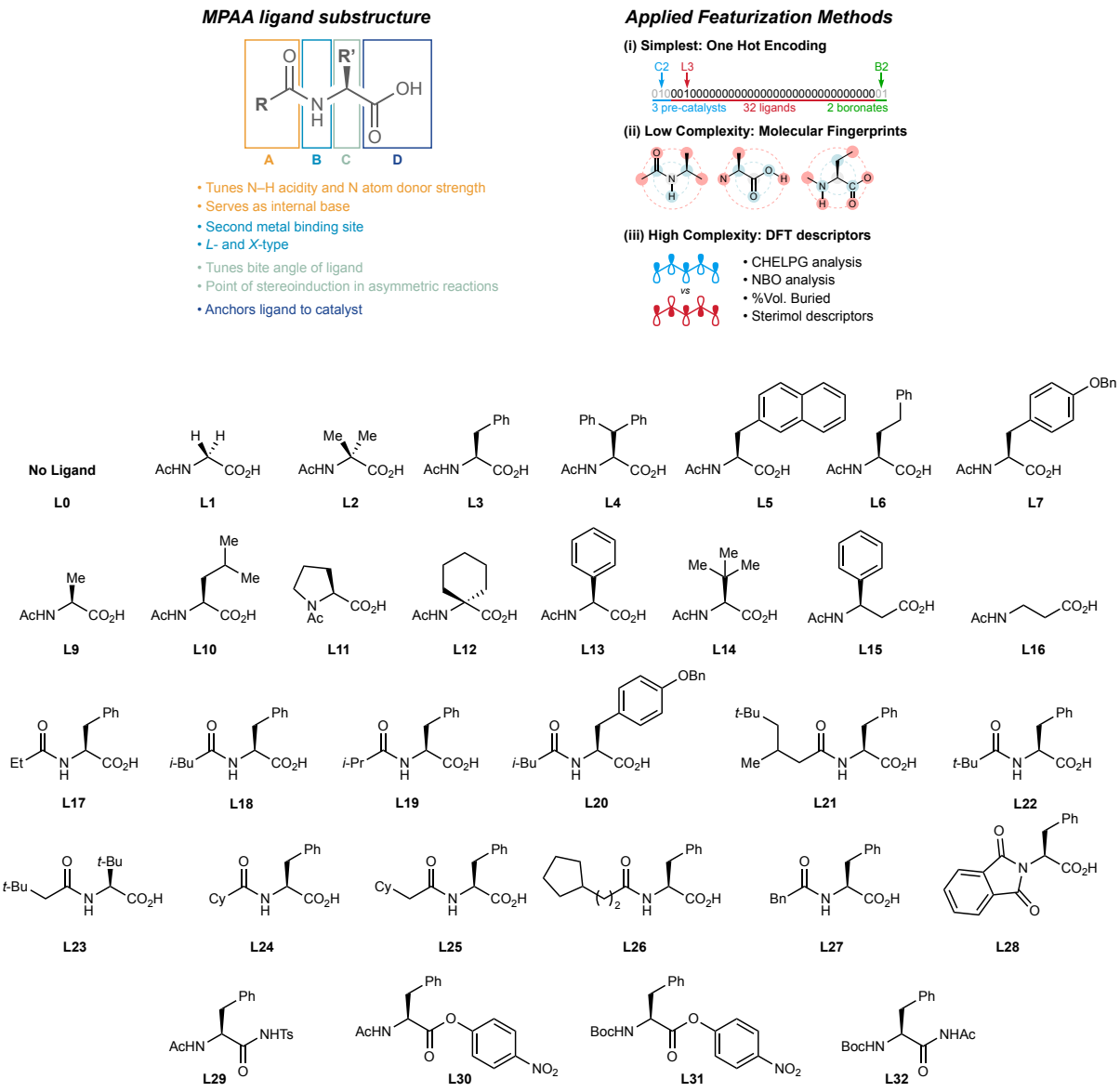


Figure 1. An overview of common data representation methods for organic molecules (top) and the diverse ligand space used for screening within this study (bottom).

Each calculated descriptor was used as a stand-alone input and was also combined with other inputs to develop hybrid features, aiming to deliver a streamlined set of input features which allow for a detailed description of the reaction conditions. A feature importance assessment based on Gini importance method (or mean decrease in impurity, MDI) was conducted and indicated high importance for the ligand fingerprints and the DFT descriptors (see SI for more details).

For more detailed information about different parameterization techniques, we refer the reader to these papers on molecular fingerprints[20, 32, 33,] and on DFT-based descriptors. [7, 9, 11,

34

Machine Learning Surrogate Models

Following feature engineering, we wanted to compare different ML models and assess their performance given a predictive task, mapping reaction conditions to yield and make predictions for unseen conditions. Different data structures and featurization methods can deliver varying performance with different ML models, something which cannot be predicted *a priori*, meaning empirical evaluation is required.

To evaluate the performance of a ML model, the dataset is partitioned into training and testing data. A model is trained (using the training data), before being given the inputs for the, previously unseen, testing data and asked to make predictions on the outputs. The difference between the predictions and the actual values is given with the root mean squared error (RMSE), a common performance metric. To partition the dataset into training and test sets, we applied two different strategies, a random split, and a designed split. When data for the training and testing partitions are chosen at random it is very likely that the training data contains information that is well distributed amongst the dataset and, as such, is in part representative of the test data. Thus, a random split may be considered as an interpolative prediction task. To simulate out-of-sample prediction the training/test partitions can be chosen with the intention of neglecting a specific part of the dataset, for example by excluding one ligand from the training data *via* a leave-one group out (LOGO) cross validation (CV). After the model has been trained, it is given the testing data that was poorly represented in the training data and attempts to make predictions. In this way a train/test partition can be designed to simulate an extrapolative prediction of unseen data. To obtain more general results, many different train-test partitions are used and an average RMSE of the predictions based on the test dataset is calculated and used as a performance metric.

Reaction Optimization

Within this study we applied closed-loop optimization using simulated experiments and assessed the effect of different surrogate models and data representations. In terms of sampling strategy, we conducted both exploitative search (the condition with the highest predicted yield is chosen for experimental evaluation) and Bayesian optimization (BO) based

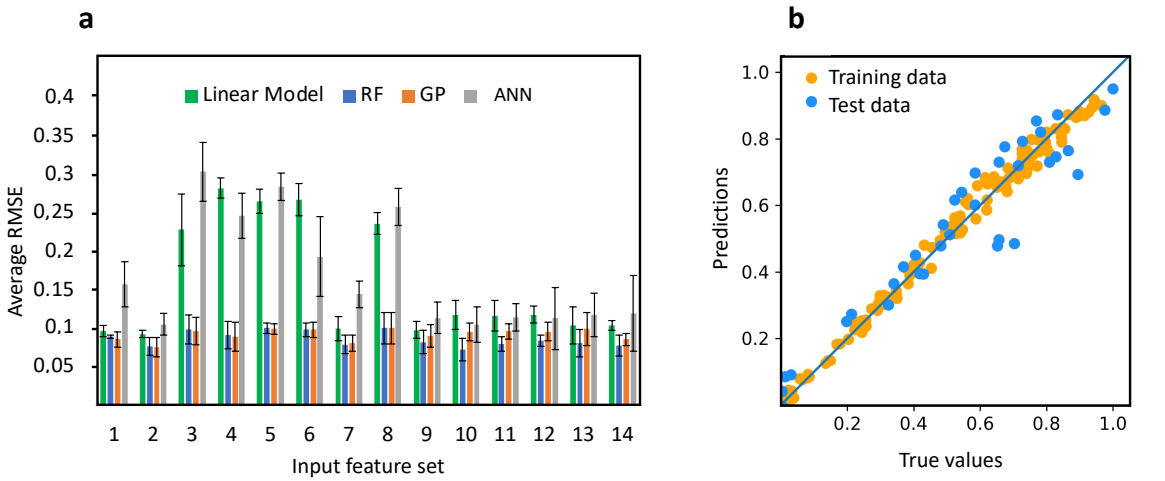
on expected improvement (EI) acquisition function (the condition with the highest expected improvement for yield was chosen for evaluation – incorporating the uncertainty of the prediction). For more detailed information on EI, BO and sampling strategies, refer to the SI and published literature. [19, 35, 36]

Results and Discussion

Preliminary Studies: Supervised ML Modelling Towards Yield Prediction

Random Split – Interpolation

Four different commonly used ML models – linear regression (baseline), random forest (RF),[37] Gaussian processes (GP),[38] and artificial neural networks (ANN),[39] were compared for a regression task and the effect of the input features on model performance was assessed with the goal of determining whether we could boost model performance with hand-crafted DFT based descriptors. The investigation began with simple OHE where the model is not given any chemical information. Subsequently, more features such as fingerprints/DFT descriptors were added to evaluate how a richer source of chemical information influences the model performance. Hybrid features were also generated, consisting of combinations of steric descriptors, electronic descriptors, OHE and principal components of Morgan 2 fingerprints. To compare the performance of different ML models on the given dataset within an interpolative task, the existing data was split randomly into a training (80%) and test (20%) set and the evaluation was repeated six times to generate mean and standard deviation values. This was conducted for each data representation and the mean and standard deviation values were calculated (Figure 2a).



Entry	Feature information	Number of input features
1	One-hot encoding (OHE)	36
2	Morgan fingerprints radius=2 (MFP2)	3072
3	All ligand DFT descriptors (NBO, Sterimol, CHELPG, %V buried)	15
4	Sterimol (ligands only)	6
5	NBO (ligands only)	4
6	First 30 principal components of MFP2 (PCA30_MFP2)	90
7	OHE + %V buried volume	37
8	Sterimol + NBO	10
9	OHE + Sterimol + NBO	46
10	MFP2 + %V buried volume + all ligand DFT descriptors	3088
11	Sterimol + PCA30_MFP2	96
12	Sterimol + PCA30_MFP2 + OHE	132
13	Sterimol + PCA30_MFP2 + OHE + NBO + %V buried	137
14	Sterimol + PCA30_MFP2 + OHE + NBO + CHELPG+ %V buried + N-H	142

Figure 2. Supervised ML of different surrogate models for yield prediction using random split of training and test data (a) A comparison of the used data representations (see table) and models for modelling the initial dataset. Error bars represent the standard deviation (b) Parity plot of the RF regression using feature set 14. RMSE values are reported with respect to yield - between 0 and 1, instead of a 0-100 % scale. Abbreviations: RF, random forest; GP, Gaussian process; ANN, artificial neural network.

More complex features (that were chosen based on prior knowledge of the reaction) such as DFT and fingerprint-derived descriptors delivered only marginal increases in performance compared to OHE which represents the baseline. RF and GP demonstrated almost equal prediction performance, regardless of which features were used. OHE along with a linear model delivered an RMSE of 9.6% ± 0.7% (standard deviation) yield. The best performance was achieved with feature set 2 and 14 using RF, giving an RMSE of 7.6% ± 1.2% and 7.2% ± 1.3% respectively. As visible in Figure 2, feature set 14 is a combined input of steric and

electronic ligand DFT descriptors (Sterimol, %Bur, NBO, CHELPG), principal components of the Morgan 2 fingerprints, OHE and existence/absence of a proton on the amide nitrogen. RMSE is reported with respect to yield - between 0 and 1. Figure 2a shows that different features influence the performance of ANN and the linear model significantly and that their overall performance is worse than RF or GP, which is likely due to the small size of the dataset or choice of the features. Figure 2b illustrates a parity plot of RF regression using feature set 14, illustrating the fitting of the training and test data.

The estimated statistical uncertainty of the datapoints is 2.8%, averaged over all 186 conditions (see SI). This serves as a lower limit to the RMSE in the predictions of any model. It may seem surprising that models with more informative hybrid features did not strongly outperform those with OHE. However, since the latter already allows for a qualitatively good performance, there is little room for improvement when using more descriptive features. As discussed, with a random partition for the training/testing data it is likely that every ligand will be represented in the training data and, as such, it is likely that the good performance of OHE results from this 'data leakage' since the other two parameters (boronate, pre-catalyst) do not influence the outcome significantly. The magnitude of this effect would likely be smaller if the other two parameters had a greater influence on the reaction outcome.

Out-of-sample Prediction – Extrapolation

Aiming to make predictions for reaction conditions that are not as well represented by the training data represents a more significant challenge. To simulate these extrapolative-type tasks, all data points are assigned a group number according to the ligand used and data partitioning was restricted so that all datapoints of the same group can be either in the test or the train set. Thus, the task can be considered as an extrapolation into untrained chemical space. For automating the process of model evaluation, LOGO CV was applied. The ligand was chosen as the variable parameter. Then, the dataset is split up into 31 sections (31 ligands) and the models are trained on all sections except for the single held-out section on which the models are tested.

A graphical representation is shown in Figure 3a, and a more detailed summary of LOGO CV is presented in the SI. After the generation of test RMSE for all data sections, the mean and

standard deviation can be calculated and used as indicators for model performance. Figure 3b illustrates the comparison of different surrogate models and different data representations. Linear regression failed to conduct extrapolative predictions using features containing bit vectors as input and thus was dropped. Expanding the scope of the out-of-sample prediction, we chose to introduce two additional commonly used surrogate models: support vector regression (SVR)[40] and adaptive boosting (AdaBoost)[41] to experimentally test their ability to fit the chemical reaction data.

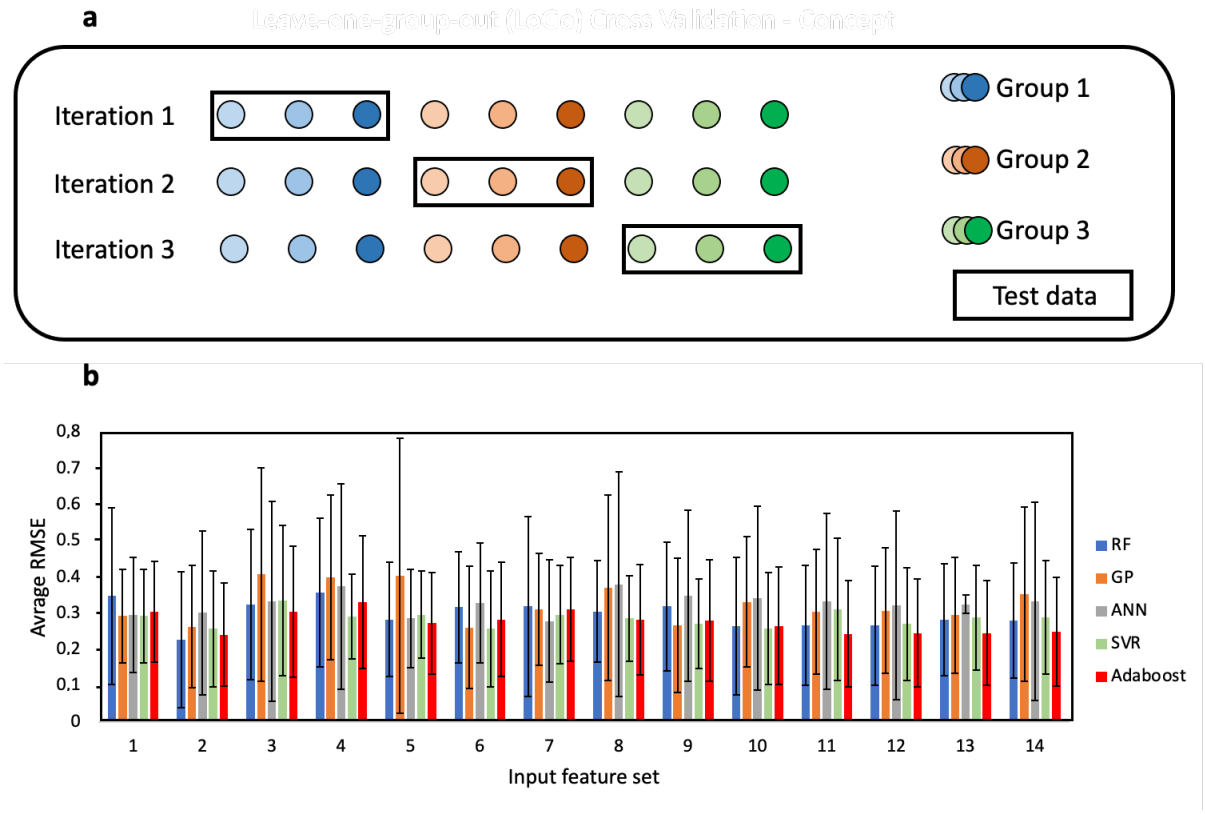


Figure 3. Leave-one-group-out cross validation (CV) (a) A conceptual illustration of LOGO CV. Different shades of blue/brown/green represent datapoints within the same group. For each iteration a different colour is circled which indicates that these datapoints are used as test data for model evaluation and the other remaining datapoints are used as training data (b) LOGO CV results of different ML models with varying input features (for feature description see Figure 2). Error bars represent the standard deviation.

Comparison of the different models suggests that RF delivered the best overall performance (using Morgan 2 fingerprints as input), achieving the lowest average RMSE of 22.7% ± 18.8% (standard deviation). On the other hand, GP seems to deliver the worst performance for out-of-sample predictions across most of the input features. Feature set 2, consisting of

exclusively Morgan 2 fingerprints, delivered the best prediction performance across all models. Additionally, even though the hybrid features (feature sets 11-14) include relevant principal components of the fingerprints and additional steric/electronic information from DFT calculations, they did not outperform the feature set 2. Interestingly, the performance of many of the feature sets was similar to the performance of OHE, which serves as a control (no chemical information), as can be seen in Figure 3a in which all of the error bars of feature sets 1-14 overlap. Thus, it can be concluded that the additional time and expertise required to generate high-fidelity DFT descriptors, based on mechanistic understanding, is unjustified when the improvement in performance of fingerprints alone is only modest.

Overall, these experiments demonstrate promising predictions with interpolative modelling tasks with the lowest RMSE of 7.2% yield, however extrapolative out-of-sample predictions still represents a significant challenge with the lowest RMSE of 25% yield (or 50% of MAE). The latter results confirm and emphasize the lack of predictive power of ML for reaction condition prediction that are not directly represented within the training data, in low data regimes which are of particular relevance to bench chemists. Within the LOGO CV experiment, even if a diverse chemical space is captured in the training data, the average prediction performance for the held-out test data was limited. The preliminary assessment of these investigation allows for benchmarking of combinations of input representations and surrogate models by showing the most appropriate strategy for similar sized datasets generated by the chemical community. We believe that the performance of extrapolative predictions may be better with larger datasets and, also, may vary depending on the reaction mechanism itself since this affects the learning of structure-reactivity relationships by ML models.

Closed-loop Active Machine Learning

Active ML represents a strategy for continuous model optimization through the iterative improvement of the surrogate model by repeatedly retraining on new experimental data as it is collected.[42] An objective, such as yield, can be rapidly maximized through the efficient exploration of chemical space, with experimental prioritization being guided by the surrogate model. Based on our preliminary modelling using different data representations/surrogate models, we aimed to assess how well these surrogate models perform within a closed-loop

optimization framework. To allow for a fair comparison, the initial dataset was shuffled, a random batch was used for model initialization and the remaining data was stored as "Lookup-table" (Figure 4). The initial batch size was varied - if not stated otherwise the models were initialized with 15 datapoints (7.5% of the dataset).

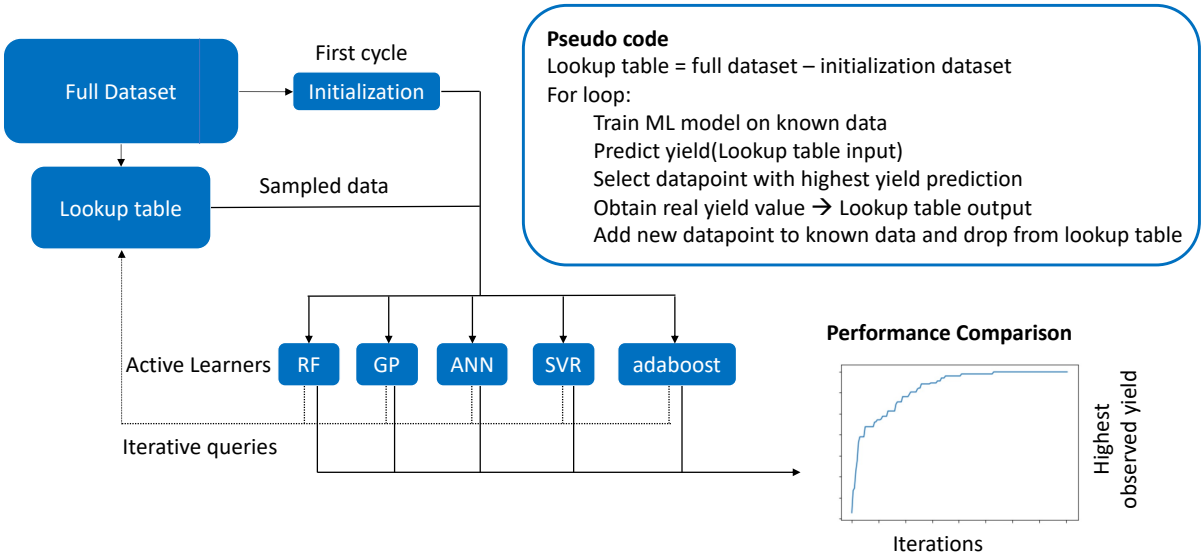


Figure 4. Schematic of the active learning workflow and pseudocode of the optimization loop. To allow for generalizability all closed-loop experiments were conducted 10 times and the average was calculated.

Once the models were trained on the initialization data, yield predictions were generated using the relevant reaction feature-sets. The datapoint (or a batch of datapoints) with the highest yield predictions were selected for "experimental" evaluation and the true yield was transferred from the lookup table (serving as a simulated experiment) to the training dataset. The models are then retrained, and this workflow is repeated until the global optimum reaction yield was identified. We chose to use feature set 14 (unless stated otherwise), for all experiments due to the highest information content. To allow for an easy and fair performance comparison, the initialization was kept the same across all surrogate models. All learning curves shown within this section represent averaged learning trajectories from 10 individual experiments – for insights into standard deviation of those 10 single experiments please refer to SI.

Comparison of Different Surrogate Models for Active Learning

To assess the different surrogate model performances within this iterative optimization strategy and identify their ability to operate under an initial low data regime, we compared 5 different models: RF, ANN, GP, SVR and AdaBoost. Figure 5a shows that the yield distribution within the dataset is evenly distributed between 0% and 100% yield, except for an increased number of samples with 0% yield. The learning curves of the single surrogate models within the closed-loop optimization (Figure 5b), in which the maximum yield observed in each active learning iteration is presented, illustrate performance of the models when searching for the optimal conditions. The experiments were conducted using sequential sampling such that one datapoint was sampled during each iteration using an exploitative acquisition function.

Overall, whilst the rate of improvement in the highest observed yields in the earlier iterations did not significantly vary between the different ML models, the required number of iterations to find the best-performing conditions (=99.9% yield) highlighted the differences between the models. Although the ANN model started initially with the lowest yield, the model achieved the optimal conditions within approximately 60 iterations, the fastest of all models. Tree based models such as AdaBoost and RF required approximately 100 iterations and GP/SVR achieved the ideal conditions within 110 iterations.

We hypothesized that using a combination of active surrogate models within the same optimization strategy may increase performance compared to single models. In detail, we based our investigation on the fact that the current best model (ANN) typically does not perform well at the beginning of the optimization, when very little data is available. Random forest, however, seems to perform better under a low data regime. Within the RF-ANN hybrid model, the rule was set that during the first 10 iterations the decision-making was conducted based on the predictions of the RF and then the ANN continued. Unexpectedly, the performance of the hybrid model did not outperform ANN. Nonetheless, it was the second-best model and achieved the optimal conditions within 70 iterations.

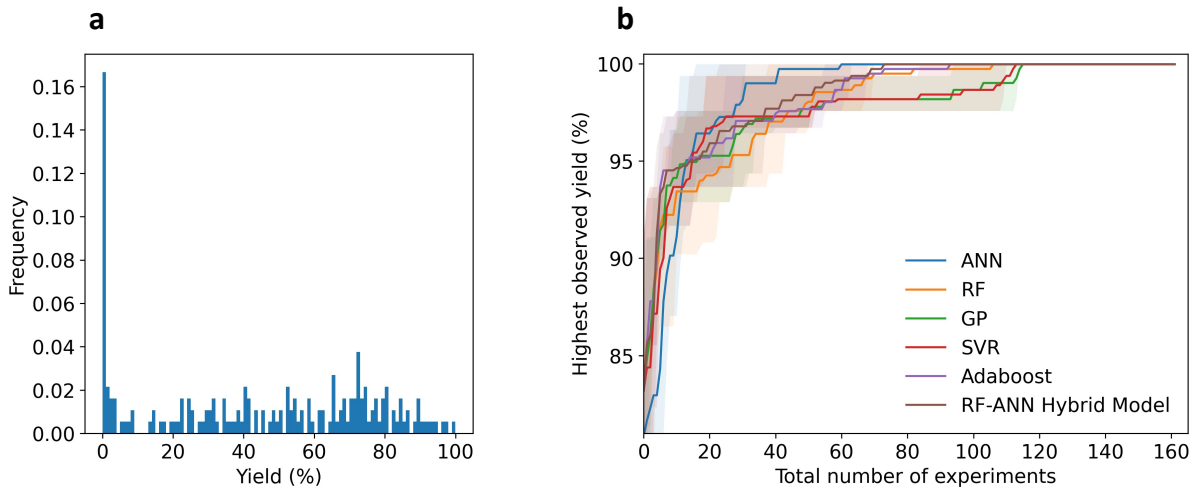


Figure 5. Variation of surrogate models for active learning (a) Distribution of reaction yield over the training data set (b) Comparison of different surrogate models within the active learning loop using feature set 14. The confidence intervals (interquartile range) of the repeated experiments are shown as filled area.

The Effect of Input Features on Active Learning

Intuitively, adding more descriptive input features to a model, such as relevant structural and electronic information, should allow for better modelling and hence better prediction performance, as observed during the preliminary studies for interpolative predicting reaction yields. The effect of adding chemical information within active learning was subsequently studied to observe how ML models perform in initial low data regimes. Four different data representations were compared: OHE, Morgan 2 fingerprints, dimensionality reduced Morgan 2 fingerprints (the first 30 principal components generated after principal component analysis (PCA) of each of the three varying reagents, giving 90 principal components in total) and the full feature set 14 (including OHE, PCA of fingerprints and all DFT features).

It was observed that OHE clearly outperformed the other representations after 10 iterations were achieved and reached the optimal set of conditions in the fewest number of iterations (Figure 6). In a similar manner, recent results by Shields *et al.* observed that OHE delivers approximately equal performance compared to hand-crafted DFT descriptors during Bayesian optimization of organic reaction conditions.[19] Initially, we assumed that the superiority of OHE performance could be due to the full factorial chemical space since all possible parameter combinations could be evaluated. Whilst this effect would benefit all

representations, we hypothesized that the simplicity of OHE along with a full factorial space could be more beneficial when compared to the effect on other input features (*e.g.* fingerprints) that are far more complex and might represent a challenge for the model to detect patterns in the data. To test this assumption, we dropped a random selection of the datapoints of the entire dataset (25%), therefore no longer representing a full factorial chemical space. However, we still observed that OHE outperformed the full feature set (see SI). Another reason for the good performance of OHE might be that during each optimization experiment the model receive datapoints of the same ligand multiple times (in combination with a different pre-catalyst or boronate). As discussed previously, the impact of the ligand is significantly higher to reaction outcome, compared to the other two parameters. As a result, it is likely that OHE captures the variability between ligands and therefore can efficiently identify high yielding reaction conditions.

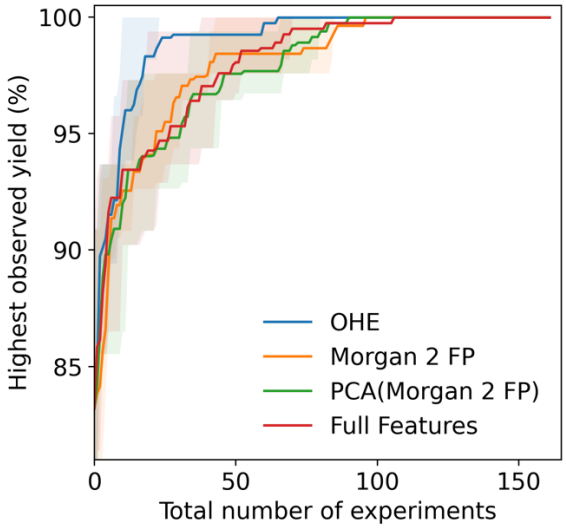


Figure 6. The effect of different input features on active learning using RF surrogate model. The confidence intervals (interquartile range) of the repeated experiments are shown as filled area. FP: Fingerprints, PC: Principal Component.

All previous experiments were conducted in a sequential design - during each iteration of the active ML algorithm, one decision was made and one single datapoint was sampled from the lookup table. In a real world setting this means that after each single experiment the yield is evaluated and then added to the ML training data. However, typically organic chemists conduct multiple experiments in parallel to accelerate the process of finding optimal conditions. Therefore, batch-sequential sampling within active learning represents a more realistic approach and was also investigated. The ideal batch size is a trade-off: a large batch

size typically brings benefits to experimental workflows as HTE equipment can be applied, for example screening 96 conditions at a time. If the experimenter gains more useful data per HTE run, this will often lead to minimizing the total number of HTE runs and be less time intensive. However, large batch sizes at the beginning of the active learning strategy could lead to the acquisition of chemically redundant data as the active learning model may make low informative predictions due to the initially limited number of datapoints used for training. Conversely, a smaller batch size allows the training data of the ML model to be updated more frequently and thus enables better quality predictions. As shown in detail in the SI (Figure S18), the batch size did not significantly vary model performance (most learning trajectories of different batch sizes are overlapping) and thus we propose they should be chosen in accordance with experimental workflows.

To conclude, we believe that prospective research in this area should consider the required complexity of molecular parameterization, due to increased computational time and expense. It could be possible that the low data regime in combination with complex features does not allow the models to efficiently learn from the data and likely over-complicates the task. Moreover, we conclude that instead of over-allocating resources on feature generation, it may be more strategic and resourceful to increase experimental data generation capabilities.

The Impact of Initialization of the Closed-Loop Optimization

The success of closed-loop optimization algorithms strongly depends on the information included in the initialization data on which the initial model is trained. To assess the effect of the data used for initialization, we conducted a case study where the optimization is initialized using: (*i*) a broader set of reaction conditions from multiple ligands, and (*ii*) a restricted dataset that contains only reaction information from three ligands. Generally, ML models deliver better prediction for areas in the chemical space that are close to or within the training data. In Figure 7a, two different extreme situations were compared – initializing the active learning either with local data (the dataset contains datapoints of only three ligands) or with random data (on average the dataset contains information of 7 ligands). By restricting the dataset we intentionally introduce biases (by showing the model only a very restricted part

of the chemical space) in order to assess the impact on the closed-loop optimization. It is apparent that even though the local initialization possesses restricted knowledge, within ten iterations the model performance is approximately equal to an initialization dataset which is more diverse. These findings may be very beneficial for experimental chemists that want to start their optimization workflow with a restricted set of chemicals (e.g., ligands) before purchasing much more diverse, potentially inadequate, chemicals. The results show that restricted initialization data can rapidly catch up with diverse initialization data when predicting experimental yields.

Another important factor for initializing active learning is the size of the initial training data. The choice of the size of the initial dataset represents a trade-off between showing the model enough information so that initial predictions are useful and keeping the dataset sufficiently small to limit the amount of experimental time and resources used. We chose four different sizes of initialization with five, 10, 15 and 20 random datapoints. As shown in Figure 7b, the results indicate that larger sized initial datasets allow the model to predict conditions that give >99% yield in fewer iterations. Using 20 random datapoints for initializations allowed finding the global optimum on average within 40 iterations while using only five initial datapoints required up to 100 iterations. In total 186 datapoints are available. Whilst the choice of the adequate size of the initialization dataset might vary on the parameter space of the dataset as well as the size and complexity of the prediction space, we assume that a minimum of 15-20 datapoints should be chosen – here this choice allowed for identification of optimal conditions in approximately 45 iterations. Overall, we suspect that the smaller sized initialization datasets are detrimental as biases may get introduced from the beginning, particularly since a greedy search (no exploration) was used. This leads to negative impacts when the active ML is conducting mostly extrapolative predictions (see the low performance of extrapolation described previously). In the case of the small (*e.g.*, five datapoints) initialization datasets, this effect was severe for the learning curves. However, being restricted to a local initialization dataset (data from three ligands) led to very steep initial learning curves and a performance similar to the initialization with a more diverse dataset, thus demonstrating the power of successful navigation through the chemical space by active ML driven closed-loop optimization.

Based on the previous findings on the size of the initial dataset and the different chemical representations, a direct comparison was conducted to assess the performance between complexity of parametrization and size of the initial dataset. Initialization datasets of 10, 15 and 20 datapoints were chosen along with OHE, Morgan 2 fingerprints and hybrid full feature representation (feature set 14, see Figure 2 for more details). When using different sizes of initialization dataset the remaining data limits the number of possible active learning iterations. To allow for easy comparison of different sized initialization datasets, the data was normalized and, as such, at every location in the x-axis the different models have access to the same number of datapoints and so the effect of initialization dataset is represented.

By comparing extremes such as having no chemical information, but a larger initialization set (Figure 7c, OHE 20) to using a smaller fully parametrized initialization dataset (Full features 10), the effects of parametrization complexity and initialization data size could be more clearly identified. Within the case study, the results clearly indicated that having a dataset parameterized to higher complexity only delivers acceptable learning curves when the size of the initial dataset is sufficiently large. When using 10 datapoints for initialization, the OHE dataset clearly outperformed the fully parameterized dataset, however, when using 20 datapoints for initialization we found less of a difference in performance. When comparing size of the initial dataset against complexity of parameterization, we found that OHE 20 reached the maximum yield within 40 experiments whereas initialization with only 10 datapoints with full features required more than 110 experiments - almost more than three times more experiments where required. Based on these insights, we believe that it is relevant to consider the trade-off between feature complexity and size of the dataset (i.e., number of experiments) when conducting reaction optimization with HTE and active ML. A more detailed case study can be found in the SI.

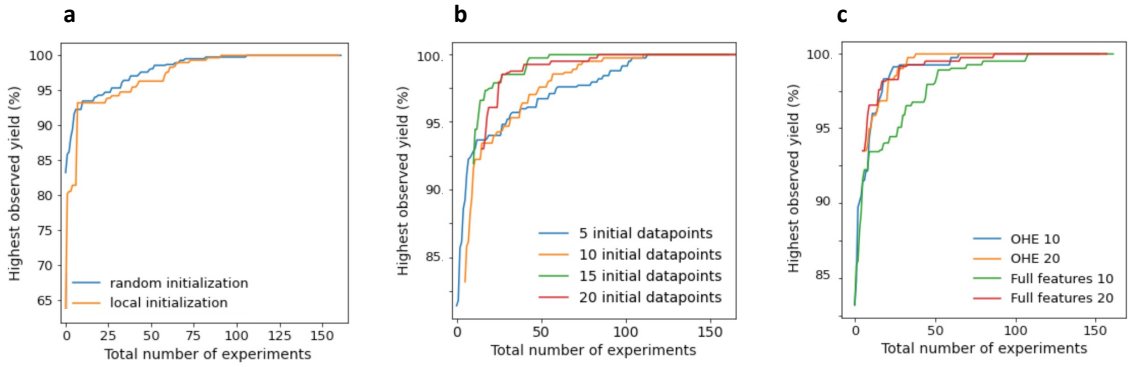


Figure 7. Comparison of different learning trajectories by variation of initialization and chemical representation (a) Random initialization vs local initialization using RF and feature set 14. (b) Different sizes of the initialization dataset (c) Variation of complexity of parameterization and size of the initial dataset.

The Effect of Incorporating an Uncertainty Metric for Active Learning: Exploitative Search vs. Expected Improvement

So far, all presented active learning strategies operated under a pure exploitation regime. While it is not feasible to directly identify the prediction uncertainty for all surrogate models, which is required for exploration, GP models were chosen due to their intrinsic ability to deliver variance for each prediction. To allow for a controlled trade-off between exploitation and exploration, different acquisition functions can be applied for sampling of subsequent datapoints. Within this comparison, the expected improvement (EI) acquisition function was chosen.[38] Figure 8a illustrates a comparison between exploitation, EI and a random search (baseline), starting with the same initialization. While a random search clearly delivered the lowest optimization performance, the differences between EI and exploitation become more obvious after the initial rise of the learning trajectory, with pure exploitation discovering the global optimum after more iterations. Figure 8b-c provides insights into how the active learning algorithms explore the chemical space, where the graphs illustrate the true yield of every sampled condition, i.e., the experimental yield of a selected input parameter selection. In an ideal case, the graph should indicate the highest values in the beginning and the lowest values at the end, thus indicating that the algorithm picks the condition which will deliver a high yield during the first iterations. Of course, this is unrealistic as the model requires a certain number of iterations to screen the chemical space and understand in which region the maximum is located. The plot for exploitation (Figure 8b) demonstrates that the initial search started in a region of the chemical space which delivered high yields and the model seem to exploit this area. However, since no exploration was used for sampling, the global maximum (slightly higher than the datapoints which were sampled in the beginning) could only be found after more than 100 iterations. The two peaks indicate that the model only found these two high yielding regions after the area around the initial data was exploited. By contrast, Figure 8c illustrates that EI samples *a priori* over a broader space (many high and low values are

sampled and the curve is noisier) due to the explorative character and then more steadily reaches low yielding areas of the chemical space. In a direct comparison, this method often allows for finding the optimal conditions in fewer experiments than just exploitative search.

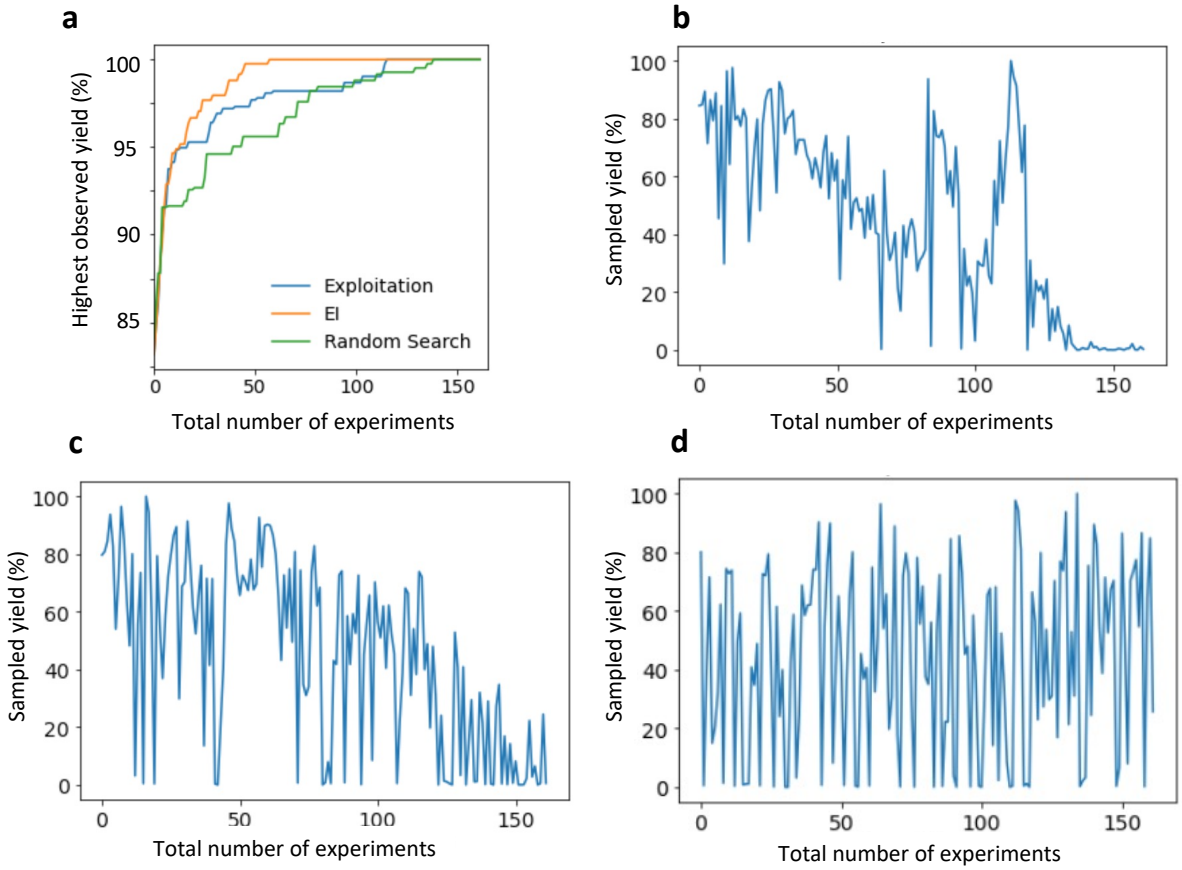


Figure 8. Comparison of different search methods for active learning(a) Exploitative search vs expected improvement using GP and feature set 14 vs random search (b) Query trajectory of the exploitative search over the yield (GP) (c) Query trajectory of the expected improvement over the yield (GP) (d) Query trajectory of the random search.

Conclusions

Using an HTE-generated dataset of conditions for the Pd-catalysed C(sp^3)–H bond activation of tertiary alkylamines, we investigated the role of parameterization for simulated active ML closed-loop optimization. By using different complexity levels of data representation, we identified the optimum modelling regimes for fitting moderately sized chemical datasets. When using a random split of the data for training/testing partitions, we found that simple OHE delivered already high-quality predictions for reaction yield, however, by adding more

complex chemical descriptors we achieved slightly lower prediction error. For out-of-sample predictions we learned that neither fingerprints nor complex DFT-derived descriptors delivered significantly better performance compared to OHE, even though the descriptors were chosen based on mechanistic insights.

Then, based on these preliminary findings, we conducted simulated closed-loop optimization experiments wherein the impact of feature complexity on active learning performance was assessed. Unexpectedly, OHE outperformed complex parameterizations that incorporated chemical information even in low data regimes which are used to initialize active learning models. To understand the impact of initialization of the closed-loop optimization, different sized initialization datasets and differently sampled data (random, out-of-sample) were used, showing that initialization with minimal data led to ineffective optimization whilst initialization with out-of-sample data still allows the active ML model to rapidly find ideal conditions. Most importantly, when comparing initialization of the closed-loop optimization with data that included the full feature set (fingerprints, DFT descriptors) to a double-sized dataset that was encoded with OHE (no chemical information), the latter identified the highest yield conditions in fewer experiments. Moreover, we found that increasing complexity of the parameterization requires a larger initialization dataset to deliver comparable performance.

The results of this study clearly indicate that current methods for parameterization are not descriptive enough to capture the factors that govern reaction success even when based on specific and relevant mechanistic insights. It must be noted that the success of different feature sets and models depends on the complexity of chemistry, the dimensionality of the design space and the number of variables. Given a different chemical design space with a larger number of ligands it might be possible that DFT-based descriptors start to outperform OHE because the number of OHE features increase whereas the number of descriptors stays constant. We believe that this work should serve as a challenge for the chemical community, and stimulate discussions about the trade-off between the development of more tailored parameterization methods or more exhaustive screening as two key factors for efficient reaction optimization.

Author contributions

A. Pomberger developed the research question, conducted the molecular parameterization and ML modelling under supervision of A.A. Lapkin. S. Sung, A. Khan, L. Colwell and D. Walz supported the project with their expertise in ML. A.A. Pedrina McCarthy and M.J. Gaunt designed, conducted and analyzed the HTE experiments. C.J. Taylor helped with the structure of the manuscript. Figures were generated by A. Pomberger and A.A. Pedrina McCarthy. All authors discussed the results and prepared the final manuscript.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgments

We are grateful to BASF SE and EPSRC Centre for Doctoral Training, SynTech (EP/S024220) for PhD studentship to A.P., Saudi Aramco for PhD studentship to A.K. and the EPSRC and GSK for a PhD studentship to A.A.M.P. Position of S. Sung was funded by Pharma Innovation Partnership Singapore (PIPS) *via* "C4" project. We thank Dr. Jesus Rodrigalvarez for useful discussions of the CH activation chemistry. We thank Robert van Putten, Kobi Felton, Daniel Wigh and Ferdinand Kossmann for helpful discussions and providing feedback on the manuscript.

References

1. A. Y. S. Lam and V. O. K. Li, *Memetic Comp.*, 2012, **4**, 3-17.
2. A. Cernijenko, R. Risgaard and P. S. Baran, *J. Am. Chem. Soc.*, 2016, **138**, 9425-9428.
3. C. W. Coley, N. S. Eyke and K. F. Jensen, *Angew. Chem. Int. Ed.*, 2020, **59**, 22858-22893.
4. C. W. Coley, D. A. Thomas, J. A. M. Lummiss, J. N. Jaworski, C. P. Breen, V. Schultz, T. Hart, J. S. Fishman, L. Rogers, H. Gao, R. W. Hicklin, P. P. Plehiers, J. Byington, J. S. Piotti, W. H. Green, A. J. Hart, T. F. Jamison and K. F. Jensen, *Science*, 2019, **365**.
5. J. M. Granda, L. Donina, V. Dragone, D.-L. Long and L. Cronin, *Nature*, 2018, **559**, 377-381.
6. S. Steiner, J. Wolf, S. Glatzel, A. Andreou, J. M. Granda, G. Keenan, T. Hinkley, G. Aragon-Camarasa, P. J. Kitson, D. Angelone and L. Cronin, *Science*, 2019, **363**.
7. Y. Amar, Artur M. Schweidtmann, P. Deutsch, L. Cao and A. Lapkin, *Chem. Sci.*, 2019, **10**, 6697-6706.
8. A. Echtermeyer, Y. Amar, J. Zakrzewski and A. Lapkin, *Beilstein J. Org. Chem.*, 2017, **13**, 150-163.
9. C. B. Santiago, J.-Y. Guo and M. S. Sigman, *Chem. Sci.*, 2018, **9**, 2398-2412.
10. W. Beker, E. P. Gajewska, T. Badowski and B. A. Grzybowski, *Angew. Chem. Int. Ed.*, 2019, **58**, 4515-4519.
11. D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher and A. G. Doyle, *Science*, 2018, **360**, 186.
12. G. Skoraczyński, P. Dittwald, B. Miasojedow, S. Szymkuć, E. P. Gajewska, B. A. Grzybowski and A. Gambin, *Sci. Rep.*, 2017, **7**, 3582.
13. C. W. Coley, W. H. Green and K. F. Jensen, *Acc. Chem. Res.*, 2018, **51**, 1281-1289.
14. E. A. Gerlein, M. McGinnity, A. Belatreche and S. Coleman, *Exp. Sys. App.*, 2016, **54**, 193-207.
15. H. Rafiei Mohammad and H. Adeli, *J. Constr. Eng. Manag.*, 2016, **142**.
16. A. L. Tarca, V. J. Carey, X. Chen, R. Romero and S. Drăghici, *PLOS Comp. Biol.*, 2007, **3**.
17. J. VanderPlas, A. J. Connolly, I. Ž and A. Gray, *arXiv preprint*, 2014, DOI: arXiv:1411.5039v1.
18. M. McCartney, M. Haeringer and W. Polifke, *J. Eng. Gas Turb. Power*, 2020, **142**, 061009-061001 - 061009-061010.
19. B. J. Shields, J. Stevens, J. Li, M. Parasram, F. Damani, J. I. M. Alvarado, J. M. Janey, R. P. Adams and A. G. Doyle, *Nature*, 2021, **590**, 89-96.
20. N. S. Eyke, W. H. Green and K. F. Jensen, *React. Chem. Eng.*, 2020, **5**, 1963-1972.
21. P. Jorayev, D. Russo, J. D. Tibbetts, A. M. Schweidtmann, P. Deutsch, S. D. Bull and A. A. Lapkin, *Chem. Eng. Sci.*, 2021, **247**, 116938.
22. B. Zagidullin, Z. Wang, Y. Guan, E. Pitkänen and J. Tang, *Brief. Bioinformatics*, 2021, **22**, 1-15.
23. F. Sandfort, F. Strieth-Kalthoff, M. Kühnemund, C. Beecks and F. Glorius, *Chem*, 2020, **6**, 1379-1390.
24. J. Rodrigalvarez, M. Nappi, H. Azuma, N. J. Flodén, M. E. Burns and M. J. Gaunt, *Nat. Chem.*, 2020, **12**, 76-81.
25. B.-F. Shi, N. Maugel, Y.-H. Zhang and J.-Q. Yu, *Angew. Chem. Int. Ed.*, 2008, **47**, 4882-4886.
26. K. M. Engle, *Pure Appl. Chem.*, 2016, **88**, 119-138.

27. H. L. Morgan, *J. Chem. Doc.*, 1965, **5**, 107-113.
28. A. V. Brethomé, S. P. Fletcher and R. S. Paton, *ACS Catal.*, 2019, **9**, 2313-2323.
29. L. Falivene, Z. Cao, A. Petta, L. Serra, A. Poater, R. Oliva, V. Scarano and L. Cavallo, *Nat. Chem.*, 2019, **11**, 872-879.
30. F. Weinhold, C. R. Landis and E. D. Glendening, *Int. Rev. Phys. Chem.*, 2016, **35**, 399-440.
31. C. M. Breneman and K. B. Wiberg, *J. Comp. Chem.*, 1990, **11**, 361-373.
32. D. Rogers and M. Hahn, *J. Chem. Inf. Model.*, 2010, **50**, 742-754.
33. K. Bouhedjar, A. Boukelia, A. Khorief Nacereddine, A. Boucheham, A. Belaidi and A. Djerourou, *Chem. Bio. Drug Des.*, 2020, **96**, 961-972.
34. J. De Jesus Silva, M. A. B. Ferreira, A. Fedorov, M. S. Sigman and C. Copéret, *Chem. Sci.*, 2020, **11**, 6717-6723.
35. P. I. Frazier 2018, DOI: arXiv:1807.02811.
36. K. C. Felton, J. G. Rittig and A. A. Lapkin, *Chem. Met.*, 2021, **1**, 116-122.
37. H. Tin Kam, *Proc. 3rd Int. Conf. Doc. Anal. Rec.*, 1995, **1**, 278-282 vol.271.
38. C. E. Rasmussen and C. K. I. Williams, *MIT Press*, 2006.
39. J. Schmidhuber, *Neural Netw.*, 2015, **61**, 85-117.
40. C. Cortes and V. Vapnik, *Machine Learning*, 1995, **20**, 273-297.
41. B. Kégl, *arXiv preprint arXiv:1312.6086*, 2013.
42. B. Settles, *Comp. Sci. Tech. Rep.* 2010, **52**, 3-8.

The Effect of Chemical Representation on Active Machine Learning Towards Closed-Loop Optimization

A. Pomberger[1], A. A. Pedrina McCarthy[2], A. Khan[1], S. Sung[3], C. J. Taylor[1,4], M. J. Gaunt[2], L. Colwell[2], D. Walz[5] and A. Lapkin[1,3]

[1]Department of Chemical Engineering and Biotechnology, Cambridge, CB3 0AS, United Kingdom
[2]Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, United Kingdom
[3]Cambridge Centre for Advanced Research and Education in Singapore Ltd., CREATE Tower 05-05, 138602 Singapore
[4]Astex Pharmaceuticals, 436 Cambridge Science Park, Milton, Cambridge CB4 0QA, United Kingdom
[5]BASF SE Data Science for Materials, Carl-Bosch-Strasse 38, 67056 Ludwigshafen am Rhein, Germany

Supporting information

Table of Contents

General Considerations

Unless mentioned otherwise, all solvents, reagents, and substrates were purchased from commercial suppliers and were used as received, including 2'-deoxyguanosine (Fluorochem), 3-nitropyridine (Sigma-Aldrich), $Ru(bpy)_3(PF_6)_2$ (Sigma-Aldrich), DMSO (Fisher), piperidine (Sigma-Aldrich), and all photocatalysts, additives and amines tested in the high-throughput experiments (various suppliers). Compound names are those generated by ChemDraw 16.0 software (PerkinElmer), following the IUPAC nomenclature.

Analytical Methods

Proton nuclear magnetic resonance (^{1}H NMR) spectra were recorded at ambient temperature on a Bruker Avance III HD spectrometer (400 MHz), a Bruker Avance III HD Smart Probe spectrometer (500 MHz) or a Bruker Avance II+ spectrometer (700 MHz). Chemical shifts (δ) were reported in ppm and quoted to the nearest 0.01 ppm relative to the residual protons in $CDCl_3$ (7.26 ppm), D_2O (4.79 ppm) and DMSO-*d*$_6$ (2.05 ppm) with coupling constants (J) were quoted in Hertz (Hz). Coupling constants were quoted to the nearest 0.1 Hz and multiplicity reported according to the following convention: s = singlet, d = doublet, t = triplet, q = quartet, qnt = quintet, sxt = sextet, spt = septet, oct = octet, m = multiplet, br = broad and associated combinations, e.g. dd = doublet of doublets. Where coincident coupling constants have been observed, the apparent (app) multiplicity of the proton resonance has been reported. Data were reported as follows: chemical shift (multiplicity, coupling constants, number of protons and molecular assignment).

Carbon nuclear magnetic resonance (^{13}C NMR) spectra were recorded at ambient temperature on a 400 MHz Bruker Avance III HD spectrometer (101 MHz) or a 500 MHz Bruker Avance III HD Smart Probe spectrometer (126 MHz). Chemical shifts (δ) were reported in ppm and quoted to the nearest 0.1 ppm relative to the residual solvent peaks in $CDCl_3$ (77.16 ppm) and DMSO-*d*$_6$ (39.52 ppm). DEPT-135, NOE experiments and 2D experiments (COSY, HMBC and HSQC) were used to support assignments when appropriate but were not included herein. Fluorine nuclear magnetic resonance (^{19}F NMR) spectra were recorded at ambient temperature on a 400 MHz Bruker Avance III HD spectrometer (376 MHz). Chemical shifts (δ)

were reported in ppm and quoted to the nearest 0.1 ppm relative to the residual solvent peaks in $CDCl_3$ (77.16 ppm).

Infrared (IR) spectra were collected using a Thermo Fisher Scientific Nicolet Summit Pro equipped with an Everest ATR, with absorption maxima (ν_{max}) quoted in wavenumbers (cm^{-1}).

Analytical thin layer chromatography (TLC) was performed using pre-coated Merck glass-backed silica gel plates (Silica gel 60 F254 0.2 mm). Visualization was achieved using ultraviolet light (254 nm) and chemical staining with basic potassium permanganate solution as appropriate, or otherwise stated. Flash column chromatography was undertaken on Fluka or Material Harvest silica gel (230-400 mesh) under a positive pressure of air unless otherwise stated.

Analytical mobile phases for LC–MS in both projects were A = 2.5 L acetonitrile + 131 mL water + 1.25 mL and formic acid, B = 2.4 L water + 1.50 g ammonium formate + 2.4 mL formic acid. The autosampler was washed between each run with a 1:1 mixture of acetonitrile:water. Gradients were generally 5-95% over 0.8/1.2 min.

Low resolution LC–MS for HT quantification: samples were analysed using a Shimadzu LC–MS; SIL-20AC XR autosampler, 2 × LC-20 AD XR pumps, CBM-20A communicator, SPD-M20A photodiode array (PDA), CTO-20AC column oven and LCMS-2020 mass spectroscopy unit. The 384-well analysis plate was placed into autosampler on a Shimadzu microtiter plate (MTP) rack. All samples were run on a Kinetex® 2.6 µm, 50 × 2.1 mm, 100 Å C18 column (Cat. No. H16-189446). The mass spectrometry unit was set to dual mode (DUIS), in which both atmospheric pressure chemical ionisation (APCI) and electrospray ionisation (ESI) mode are used simultaneously, in the positive mode and set for selective ion monitoring of M+1 for product and internal standard (scan speed 15000u). Data analysis was undertaken using Shimadzu Lab Solutions software (Version 5.97) and exported into Microsoft Excel for further statistical tests and data visualisation.

Nanoscale C–H activation: Nanoscale reactions (50-100 nmol) were run using Corning 1,536-well plates (Corning Echo qualified, Cat. No. 3730, Cyclic Olefin-Copolymer COC, 12.5 µL-wells,

flat bottom, clear) as reaction plates. Reactions at elevated temperatures were ran in Corning 1,536-well White High Base plates (Cyclic Olefin Copolymer Cat. No. 4570) and typically with Axygen 384-well plates (Cat No. P-284-120SQ-C, Polypropylene, 120 µL, V-bottom, translucent) used as solution source plates for stock solutions and for analytical plates on LC–MS equipment. Analysis plates were sealed with gas permeable adhesive sheets (4titude, Cat. No. 4ti-0516/384).

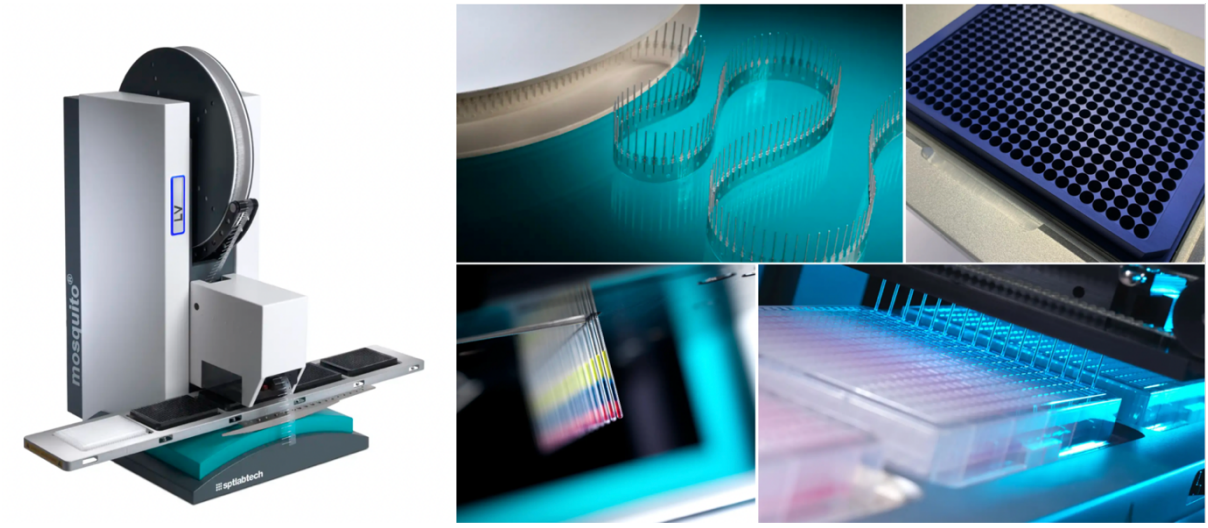


Figure S1 The Mosquito liquid handling robot

For reactions at elevated temperatures: the 1,536-well plates were covered by a perfluoroalkoxy alkane (PFA) mat (0.125 mm thickness, FLONFILM™ 600 PFA film), followed by a neoprene rubber matt (on top and below) and then secured within a custom-built plate-sealing device, which was tightened through gradual even turning of all 14 screws in a crosswise pattern. The entire assembly was heated in an oven for the reaction duration. Once complete the assembly is cooled in a laboratory fridge to ~10 °C, minimising contamination when unsealed. Following this, the plate is removed from the assembly and centrifuged prior to Mosquito dosing, Figure S1 illustrates the Mosquito liquid handling robot. However, commercial plate-sealing alternatives to this are now available from Analytical Sales and Services Inc. (Cat. No. 1626100).

High Throughput Experimentation

Reaction preparation: stock solutions of the reaction components were prepared in *N*-methyl-pyrrolidone (NMP) according to table S1, below. Stock solutions were then charged into a 384-well source plate according to the source plate layouts in Figure S2. The required volume for each source well was calculated to include an additional 20 µL top-up, ensuring that there would be an excess during plate dosing.

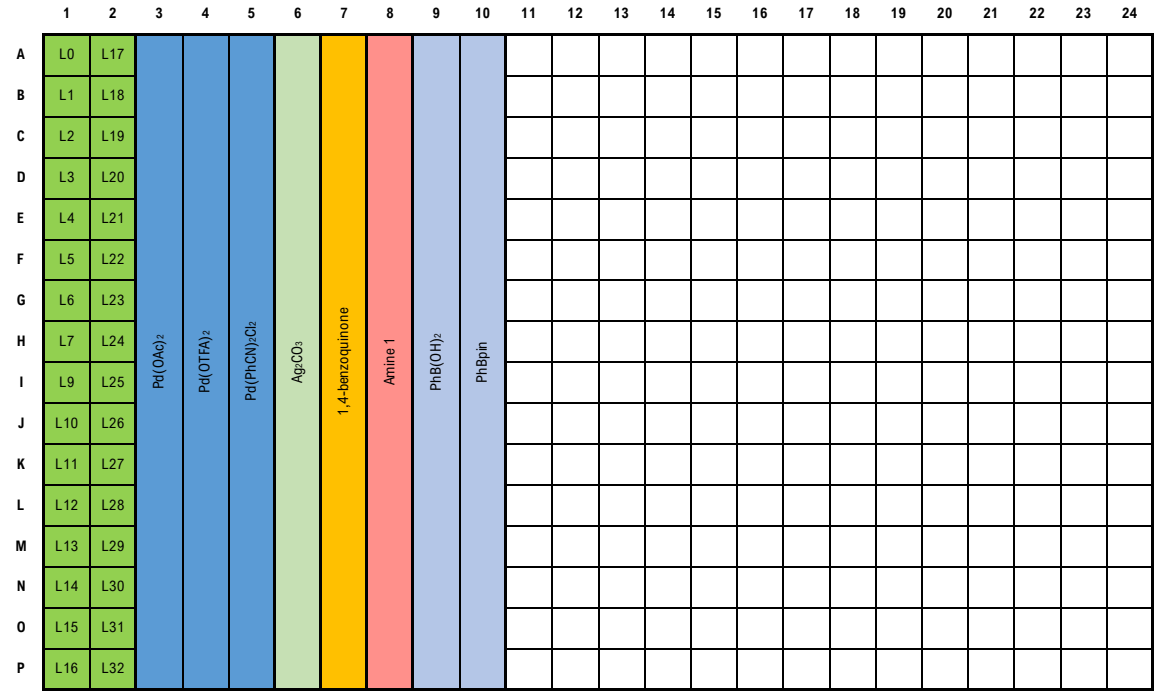


Figure S2 Source plate layout for the high throughput optimization

The Mosquito liquid handling robot was used to sequentially transfer 0.5 µL aliquots of each of the six reaction components from the 384-well source plate to the 1,536-well reaction plate (dosing sequence given in Table S1). Upon dosing the final reagent, three cycles of the Mosquito's dispense mix setting (500 nL, move 0.5 mm), was used to ensure all reagents were evenly distributed. Silver carbonate is completely insoluble in NMP and settles at the bottom of source plate wells in ca. 1 min, blocking pipettes and leading to inconsistent stoichiometries. As such, preparation of the source and reaction plates required an alternate method; i) a slurry of the required concentration in NMP was prepared; ii) under vigorous stirring, using hand-held electronic pipettes, the source plate was charged with the required aliquot (68 µL); iii) the Mosquito was paused after each transfer from source plate to the

reaction plate and wells containing silver carbonate were mixed with the 'aspirate-mix' function (20 µL, 3 rounds) of a multi-channel electronic pipette. This method allowed for consistent dispensation of insoluble reagents.

*Table S1 Dosing table of the HT optimization of C(sp3)–H activation of amine **1***

Dosing sequence	**Reagent**	**Equiv.**	**Concentration / M**	**Min source plate vol. (+ top up) / µL**	**Aliquot / µL**
1	MPAA ligands	0.25	0.060	24 (+20)	0.5
2	Pd pre-catalyst	0.10	0.024	16 (+20)	0.5
3	Ag_2CO_3	2.5	0.60	48 (+20)	0.5
4	1,4-benzoquinone	2.0	0.48	48 (+20)	0.5
5	amine	2.5	0.60	48 (+20)	0.5
6	boronates	1.0	0.24	24 (+20)	0.5
-	**Total**	-	0.04	-	3.0

The reaction plate was then placed into a custom-designed aluminium plate sealer (commercial alternatives now available, i.e., Analytical Sales NanoNest), topped with a chemically inert PFA film (0.125 mm thick) and a silicone gasket. The assembly was secured by gradually tightening 14 screws in a cross-wise pattern, ensuring even compression on all sides of the reaction plate. The entire assembly was then placed in a temperature-controlled laboratory oven set to 55 °C.

Analysis: After the desired reaction time had elapsed, the assembly was removed from the oven, allowed to cool to room temperature and placed in a 10 °C fridge for 10 min prior to opening (generating negative pressure inside the wells and avoiding messy pressure release). The Mosquito was used to aspirate and transfer 100 nL from each reaction well into a 384-well 'analysis plate' which was pre-loaded with 50 µL of a quenching diluent, MeCN:H_2O:formic acid (2:1:1) that contained a known concentration (0.04 mM) of an internal standard (*N*,*N*-dibenzylaniline, DBA). This plate was then diluted further by the addition of 50 µL of the IS-doped diluent, sealed with an adhesive LC–MS autosampler-compatible sealing film and analysed by LC–MS (Shimadzu LC–MS-2020, selective ion monitoring to follow the total ion count of the IS and arylated product). Prior to calibrant and reaction

sample analyses, 200 matrix-matched 'sacrificial samples' were used to pre-condition the LC–MS, ensuring consistent performance between runs.

Calibration samples: Using authentic, independently synthesized, product four known-concentration calibration samples were prepared with product/IS ratios of 0.25, 0.50, 0.75 and 1.0 (Table S2), matching the concentration of the reaction samples. These were analysed immediately preceding and following the reaction samples, repeats were averaged to build a calibration curve for the desired product (Figure S3). Analytical data were pre-processed in the native LabSolutions software (version 5.97), before being transferred to Microsoft Excel for final processing and visualisation.

Table S2 Composition of calibration samples

Calibration Sample	IS conc. / mM	Product Conc. / mM	Calib. (% Yield)	Prod/IS TIC Ratio
1	0.040	0.000	0	0.00
2	0.040	0.010	25	0.26
3	0.040	0.020	50	0.50
4	0.040	0.030	75	0.71
5	0.040	0.040	100	0.97
Reaction Sample	0.040	Unknown	Unknown	Unknown

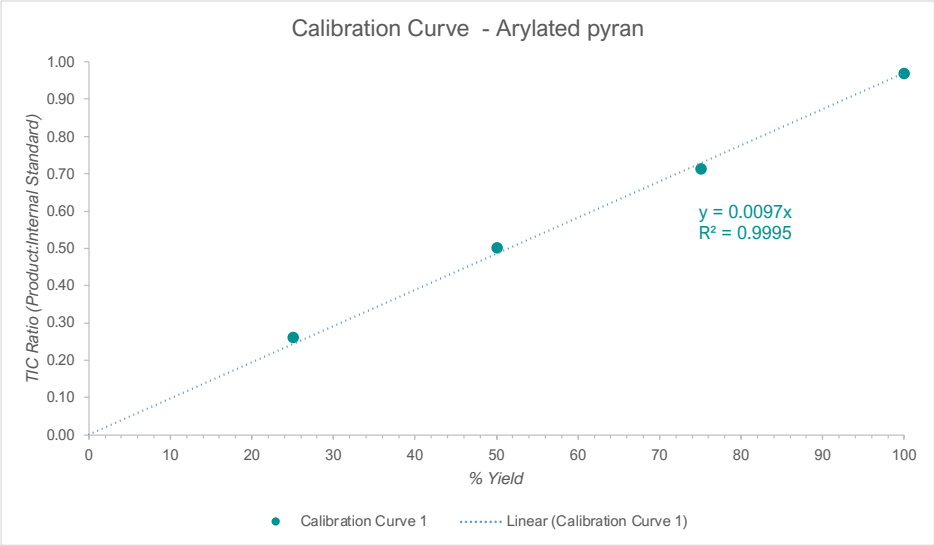


Figure S3 Product calibration curve used to quantify product in the reaction mixture

Reaction Scheme and Ligand Structures

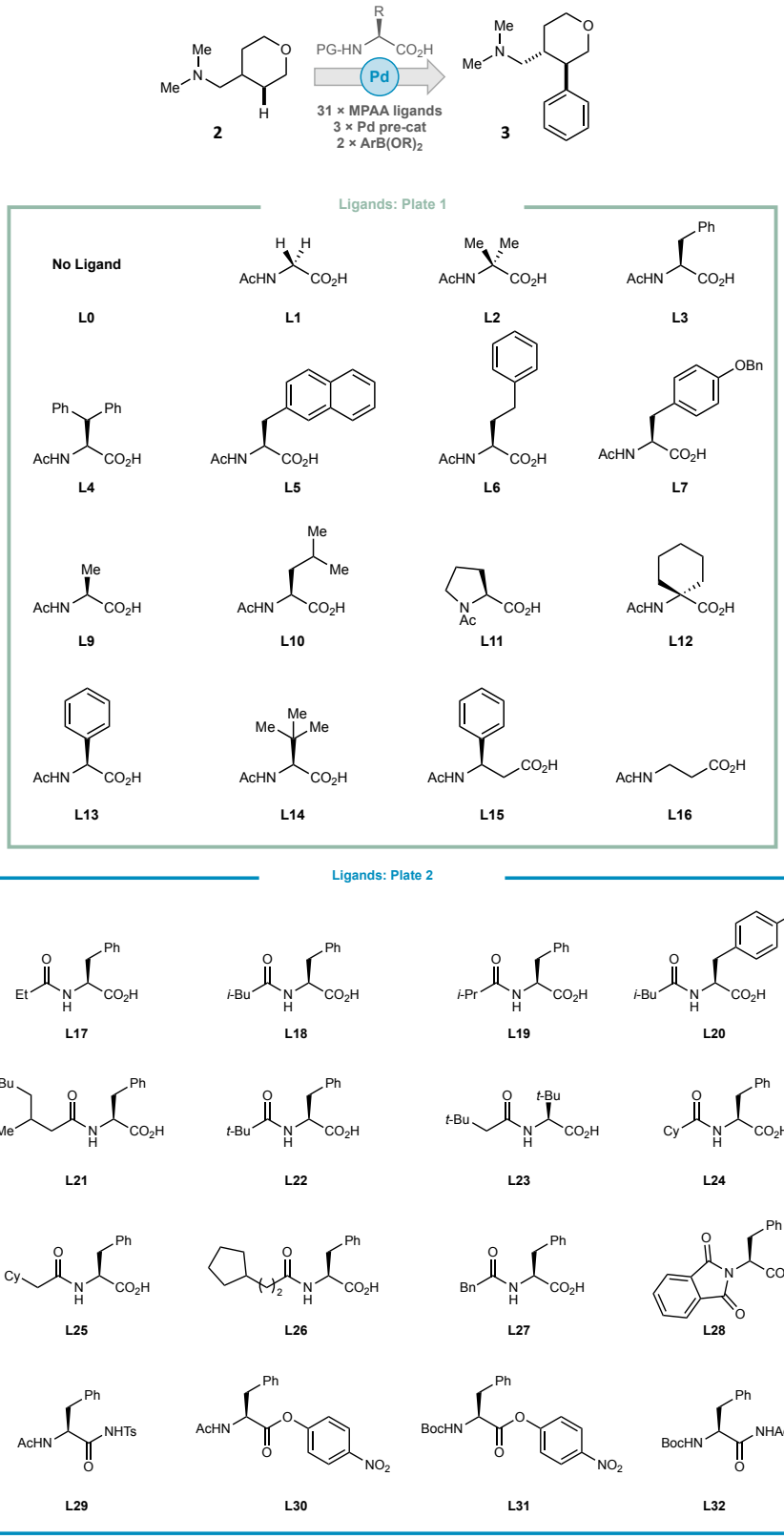
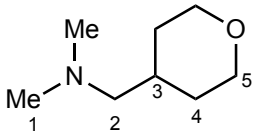


*Figure S4 Ligands explored in the HTE C(sp3)–H activation of tertiary alkylamine **2**. Ligand 8 is not represented in the table since this was initially the blank – for convenience the blank was changed as L0.*

Synthesis of Materials

*N,N-dimethyl-1-(tetrahydro-2H-pyran-4-yl)methanamine (**2**)*



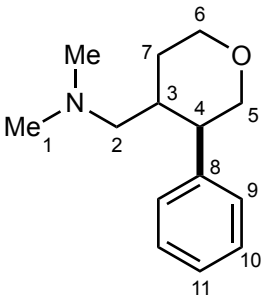
To (tetrahydro-*2H*-pyran-4-yl)methanamine (2.6 mL, 21.9 mmol) was added formaldehyde (5.6 mL, 37% aq., 69.0 mmol) at 0 °C with vigorous stirring. Formic acid (4.7 mL, 91.0 mmol) was added over 5 minutes and the reaction left to rise to room temperature over 1 hour, then heated to 85 °C for 24 hours. The yellow reaction mixture was cooled to room temperature and HCl (20 mL, 3.0 M aq.) was added. The aqueous mixture was washed with diethyl ether (3×40 mL) and the pH adjusted ca. pH 8. The aqueous layer was then extracted with diethyl ether (3×40 mL). The combined organic extracts were dried over $MgSO_4$, filtered and concentrated *in vacuo*. The resulting colourless oil was purified by vacuum distillation (93-95 °C, 75 mbar) to yield **2** as a colourless oil (1.45 g, 46%).

^{1}H NMR (400 MHz, $CDCl_3$) δ 3.96 (dd, *J* = 11.3, 3.7 Hz, 2H, H_{5eq}), 3.38 (td, *J* = 11.8, 2.0 Hz, 2H, H_{5ax}), 2.20 (s, 6H, H_1), 2.11 (d, *J* = 6.9 Hz, 2H, H_2), 1.80 – 1.59 (m, 3H, H_3/H_{4eq}), 1.26 (qd, *J* = 12.0, 4.5 Hz, 1H, H_{4ax}).

^{13}C NMR (101 MHz, $CDCl_3$) δ 67.9 (C_5), 66.4 (C_2), 45.9 (C_1), 33.2 (C_3), 31.6 (C_4).

Analytical data agrees with those reported previously.[1]

*N,N-dimethyl-1-(3-phenyltetrahydro-2H-pyran-4-yl)methanamine (**3**)*



An oven dried 10 mL microwave vial, equipped with a stir bar was charged with $Pd(OAc)_2$ (6.7 mg, 0.03 mmol), *N*-acyl-*L*-*tert*-leucine (10.4 mg, 0.06 mmol), benzoquinone (64.9 mg, 0.60 mmol), silver carbonate (207 mg, 0.75 mmol), *N,N*-dimethyl-1-(tetrahydro-2*H*-pyran-4-yl)methanamine (107 mg, 0.75 mmol) and NMP (6.5 mL). The vial was sealed, heated to 50 °C and stirred at 1000 rpm before benzeneboronic acid (36.6 mg, 0.30 mmol) was added as a solution in NMP (1 mL). The reaction mixture was stirred for 18 h, cooled to room temperature, diluted with diethyl ether (50 mL) and washed with 1% aq. NaOH (5×200 mL). The organic phase was dried over $MgSO_4$, filtered, and concentrated *in vacuo* to yield a brown oil which was purified by column chromatography (silica gel, 99:1 CH_2Cl_2:0.45 M NH_3 in MeOH, R_f = 0.25) to yield **3** as a light brown oil (22.5 mg, 34%).

^{1}H NMR (400 MHz, $CDCl_3$) δ 7.31 (t, *J* = 7.3 Hz, 2H, H_{10}), 7.26 – 7.20 (m, 1H, H_{11}), 7.17 (d, *J* = 6.8 Hz, 2H, H_9), 4.09 (dd, *J* = 11.5, 4.2 Hz, 1H, H_{6eq}), 3.87 (dd, *J* = 11.4, 4.4 Hz, 1H, H_{5eq}), 3.53 (td, *J* = 11.9, 2.1 Hz, 1H, H_{6ax}), 3.35 (t, *J* = 11.2 Hz, 1H, H_{5ax}), 2.51 (td, *J* = 10.8, 4.4 Hz, 1H, H_4), 2.10 (s, 6H, H_1), 2.08 – 1.95 (m, 3H, , H_{7eq} , H_3, $H_{2'}$), 1.89 (dd, *J* = 11.6, 2.5 Hz, 1H, $H_{2''}$), 1.43 (tdd, *J* = 13.4, 10.6, 4.5 Hz, 1H, H_{7ax});

^{13}C NMR (101 MHz, $CDCl_3$) δ 140.6 (C_8), 128.6 (C_{10}), 127.9 (C_9), 126.8 (C_{11}), 73.8 (C_5), 68.4 (C_6), 63.6 (C_2), 48.3 (C_4), 45.8 (C_1), 38.0 (C_3), 31.2 (C_7);

IR ν_{max}/cm^{-1} (thin film) 2945, 2817, 2763, 1602, 1493, 1455, 1455, 1385, 1301, 1266, 1236, 1177, 1129, 1088, 1052, 1041, 1013, 997, 977, 901, 869, 846, 793;

HRMS (m/z): [M]+ calcd for $C_{14}H_{21}NO$, 220.1696; found, 220.1698.

Preparation of the Dataset

The obtained HTE data (see heatmap in Figure S5) was analyzed with respect to the deviation of the single measurements (within the quartet) and eventually used to create the dataset for modelling. Within the heatmap the lines refer to the number of the ligand (L0 = no ligand) and the columns refer to the precatalyst (C1 = palladium(II)acetate, C2 = palladium(II)trifluoroacetate, C3 = bis(benzonitrile)palladium(II)chloride) as well as the boronates. As visible, all experiments were conducted four times and the mean yield of each quartet was used for the modelling. We obtained an average standard error of the mean of 2.8% (Eqn. 3), an average mean absolute deviation of 3.1% (Eqn. 4) and an average maximal deviation of 5.3% (Eqn. 5).

$$\sigma_{x_j} = \sqrt{\frac{1}{n-1}\sum_i^n \left(x_{ij} - \bar{x_j}\right)^2}$$ Eqn. 1 (standard deviation)

$$\sigma_{\bar{x}_j} = \frac{1}{\sqrt{n}}\sigma_{x_j}$$ Eqn. 2 (standard error of the mean)

$$\sigma_{\bar{x}} = \sqrt{\frac{1}{N}\sum_j^N \sigma_{x_j}^2}$$ Eqn. 3 (average standard error of mean)

$$MAD = \frac{1}{N}\sum_j^N \frac{1}{n-1}\sum_i^n \left|x_{ij} - \bar{x_j}\right|$$ Eqn. 4 (average mean absolute deviation)

$$maxAD = \frac{1}{N}\sum_j^N \max_i \left|x_{ij} - \bar{x_j}\right|$$ Eqn. 5 (average max absolute deviation)

Where $n = 4$ is the number of repetitions and $N = 186$ is the number of conditions.

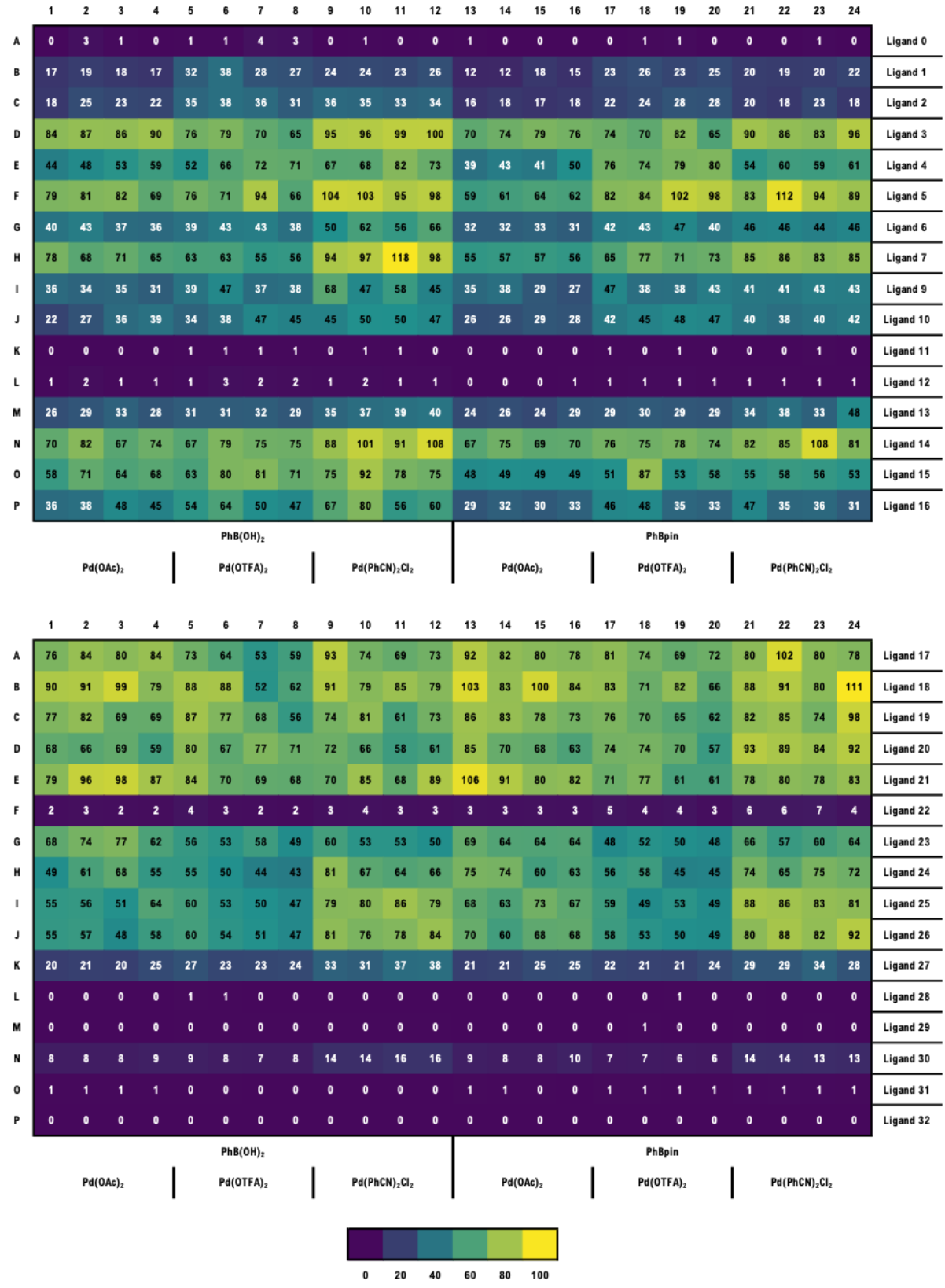


Figure S5 Heatmap of the data obtained from HTE screening, lighter hues indicate higher conversion

It should be noted that individual results that surpassed 100% yield were observed. Although the precise reason for this is extremely challenging to determine, analytical artifacts such as these can arise in the preparation of the analytical samples, the sampling of the analytical plate by the LC–MS autosampler, or during the analytical run. To mitigate the impact of such events on the overall data quality, each condition was run in quadruplicate and averaged. With each datapoint being the average of four repeats the impact that a single erroneous result can have on the overall dataset was thought to be minor and, as such, we decided not to eliminate any artificially high results (e.g., L7-C3) and instead treat all data points uniformly.

To provide a dataset for subsequent modelling, we calculated the average of all measurements which resulted in a dataset of 186 single datapoints. Whilst some single measurements had a yield above 100%, the averaged values did not.

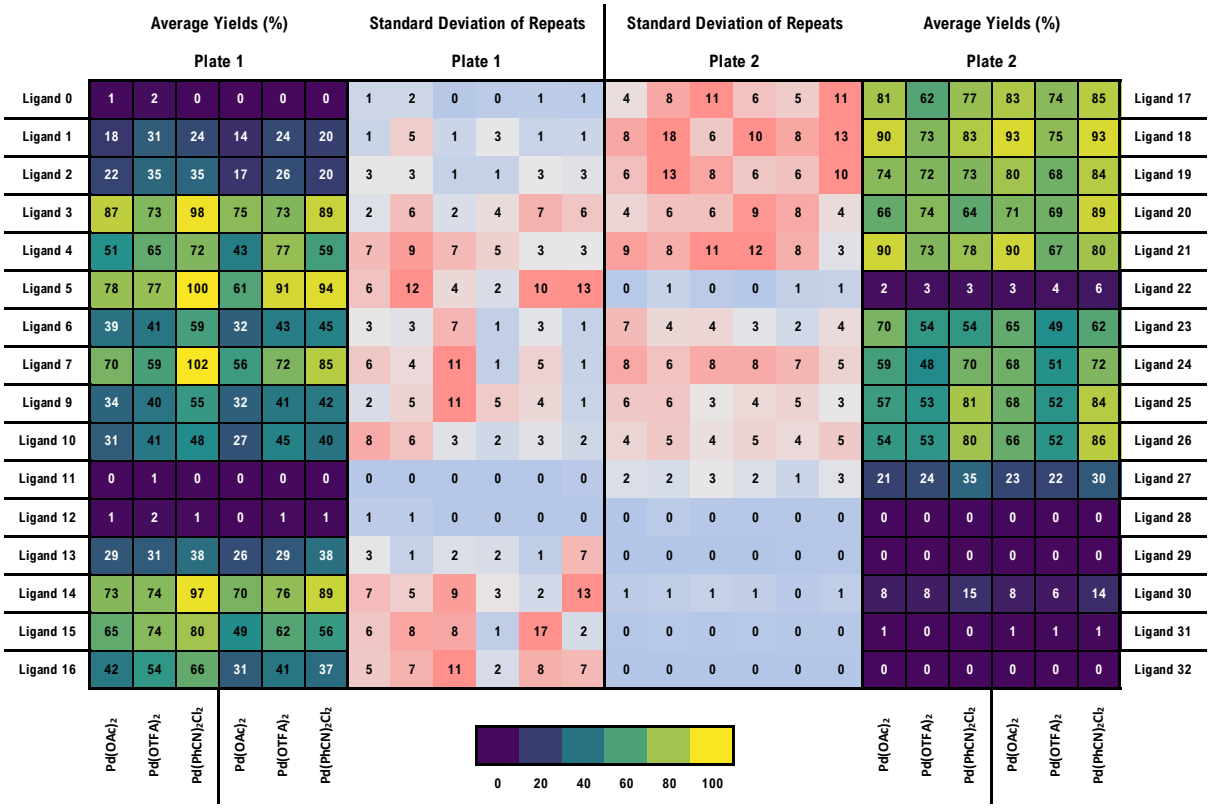


Figure S6 Condensed heatmap alongside the standard deviation of repeated measurements

Generation of Morgan Fingerprints

All reagent molecules (pre-catalyst, ligands, boronates) were drawn using ChemDraw and then the SMILES were generated. The SMILES were canonicalized using the package RDKit in python. Then RDKit (RDKit version 2020.03.2) was used to generate Morgan fingerprints of a radius 2 with a set length of 1024. The three fingerprints were concatenated and saved as Numpy arrays for subsequent modelling. Initially, different radii of circular fingerprints of the ligand molecules were screened, and it was observed that a radius of 2 was ideal due to the lowest model error. Figure S7 illustrates different radii of Morgan fingerprints versus their RMSE of three different ML models (RF, GP, ANN) – the RMSE is averaged from three single evaluation (prediction of yield) of a random split of the data (80% training, 20% test data) and the error bars represent the standard deviation.

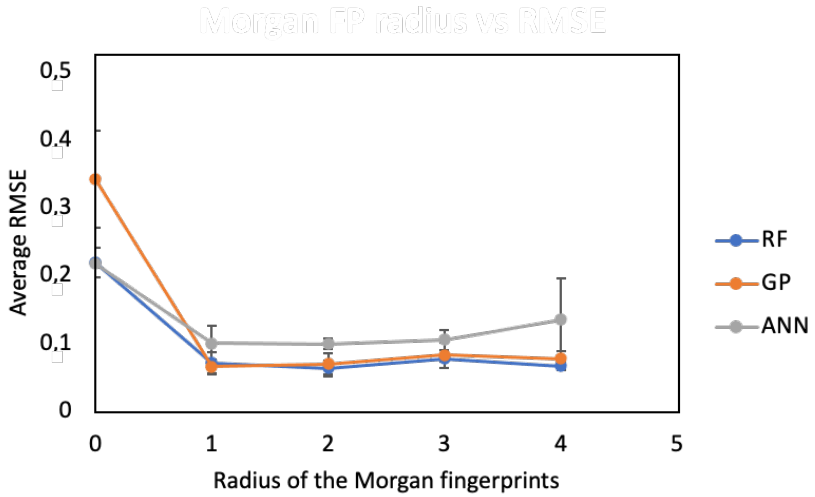


Figure S7 Variation of the used radius for fingerprint generation. The presented values are averaged from three single evaluations using random split (80/20 : train/test) and the error bars represent the standard deviation.

Based on the generated fingerprints we conducted a similarity assessment between all ligands using the Tanimoto similarity index (using RDKit) (see heatmap in Figure S8a). Moreover, hierarchical clustering was conducted (using the SciPy python package) to allow for insights into similarities between the ligands and understand which are structurally close.

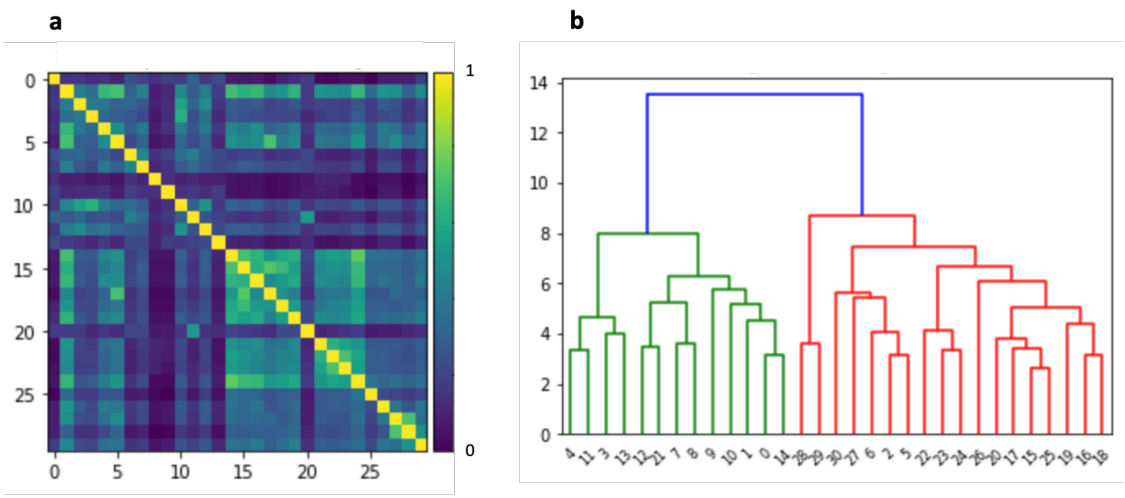


Figure S8 Similarity assessment of the ligand scope (a) Tanimoto similarity index between all ligands (b) Dendrogram showing hierarchical clustering of all ligands

Density Functional Theory (DFT)-based Geometry Optimization

For DFT geometry optimizations we relied on the B3LYP functional and 6-31G(d) basis set (Gaussian 16). As stated in the manuscript, DFT was used for unbound ligand molecules only. This section details the generation of steric and electronic descriptors. In addition to the features explained below we also calculated HOMO/LUMO energies of the ligand molecules. Table S3 displays all DFT derived descriptor values.

Sterimol Parameters

Sterimol descriptors comprise of three single length measurements that capture the steric footprint of a molecule across a specified axis and relative to a fixed point of reference. All calculations were conducted using a python package developed by Brethomé *et al.* and the geometry optimized ligand molecules.[2] Parameterization was separated into the α-carbon residue (Figure S9, R-res) and the acetyl residue (Figure S9, N-res). The arrows in the figure indicate the direction of the reference axis.

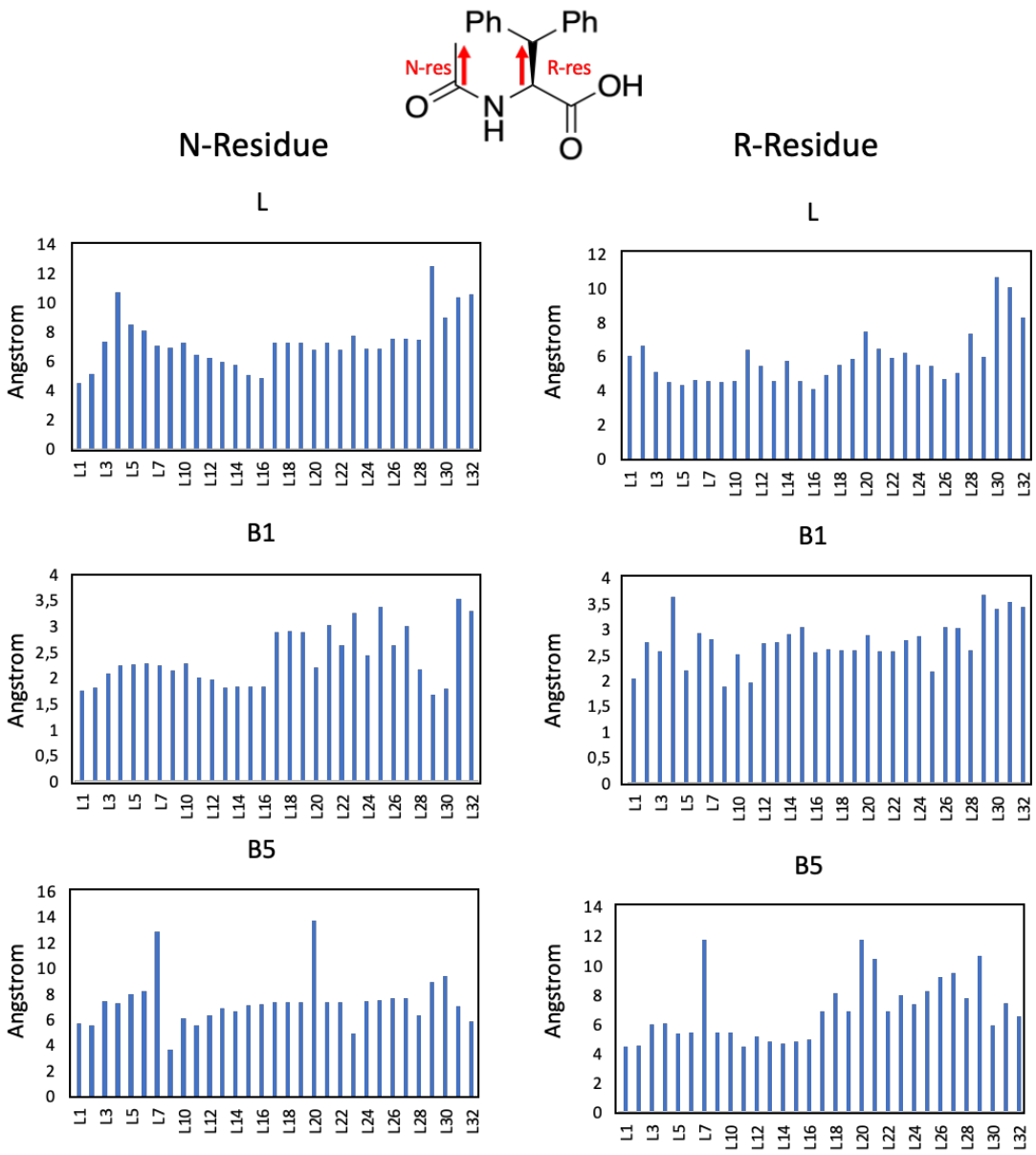


Figure S9 Sterimol parameters of the N-residue and the R-residue. The molecule on top indicates the reference axis for generation of the Sterimol parameters and the six plots show the results of the calculation.

Percentage Buried Volume

The bulkiness of the α-carbon residue was additionally quantified by calculating the percentage of the buried volume, based on geometry optimized ligand molecules. In this case the α-carbon was set as center and the calculation was conducted regarding solely the residue on the α-carbon position, using the SambVca 2.1 web application[3] as shown in Figure S10. The used reference pane of the 4 neighbouring atoms and the direction of the reference axis is shown in Figure S10a. Figure S10b illustrates the part of the molecules (R-res) that was considered for the calculation (here a *tert*-butyl group) and Figure S10c is a two-dimensional steric heatmap of the outcome of the calculation of the % buried volume. The graph (Figure S10d) displays the results of the calculation and allows for steric insights into the bulkiness of the ligand molecule.

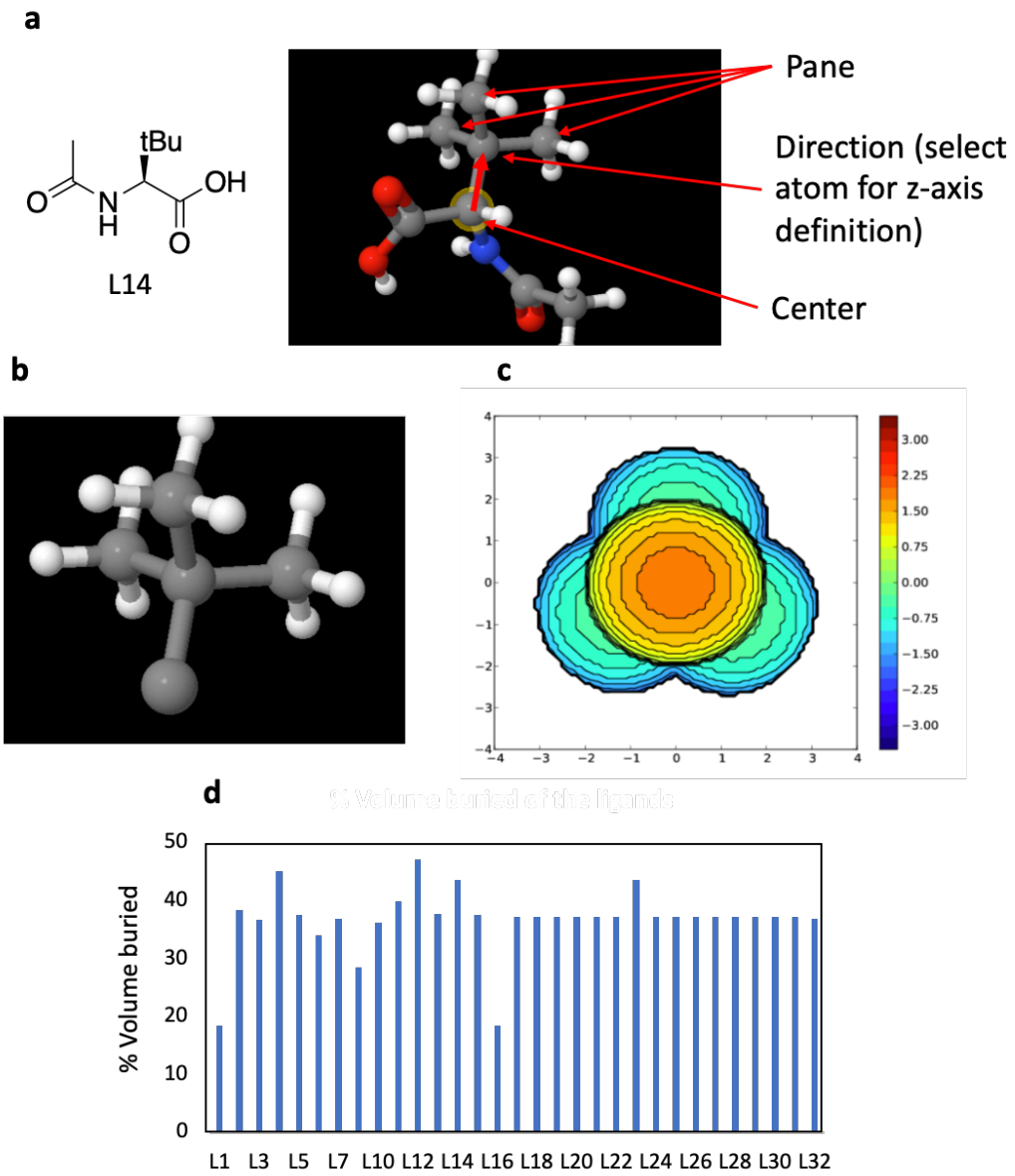


Figure S10 Calculation of the percentage buried volume (a) Structure and DFT optimized geometry of ligand 14. Illustration of the center, reference axis and pane for calculation of the % buried volume (b) Illustration of the actual R-residue (here tert-butyl) which was considered for the calculation (c) Graphic illustration of the two-dimensional steric heatmap of the calculation of the % buried volume from the web-based platform (d) Results of the % buried volume calculation.

Natural Bond Orbital (NBO) Analysis

In order to capture the electron density distribution, we conducted a NBO analysis using Gaussian 16.[4] Figure S11 illustrates the location of the atoms in the ligand molecule which were selected for NBO analysis as well as the results of the calculation.

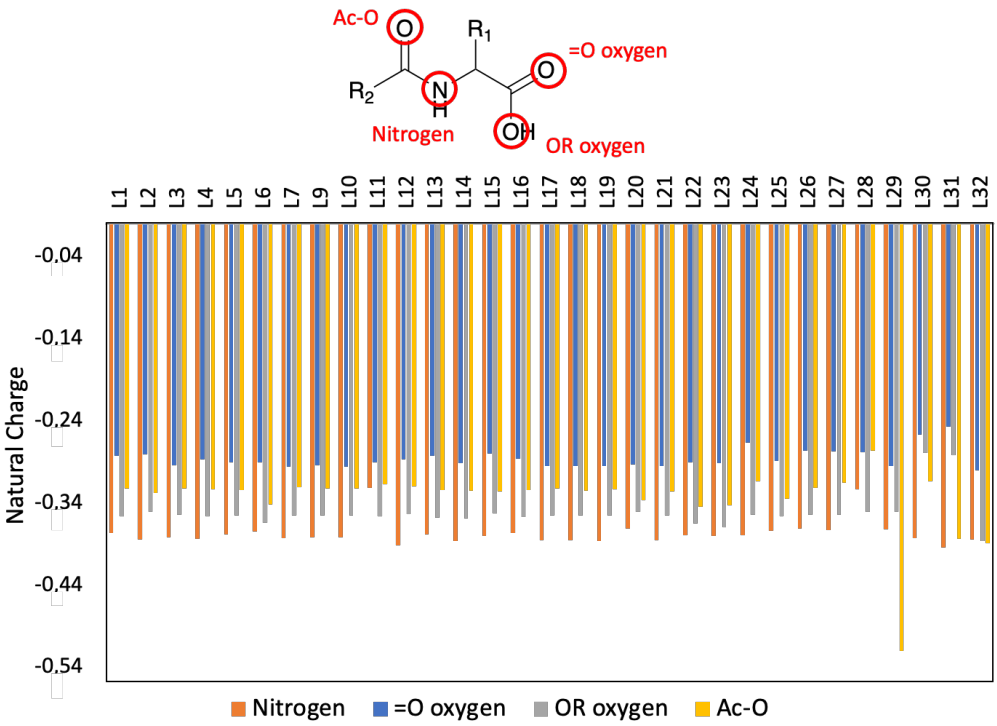


Figure S11 Location of the atoms which were used for the NBO analysis and results of the NBO calculation of all ligands.

CHarges from ELectrostatic Potentials Using a Grid-Based Method (ChELPG) Analysis

CHELPG[5] analysis was conducted for all ligands using Gaussian 16. Figure S12 illustrates the location of the atoms in the ligand molecule which were selected for CHELPG analysis as well as the results of the calculation.

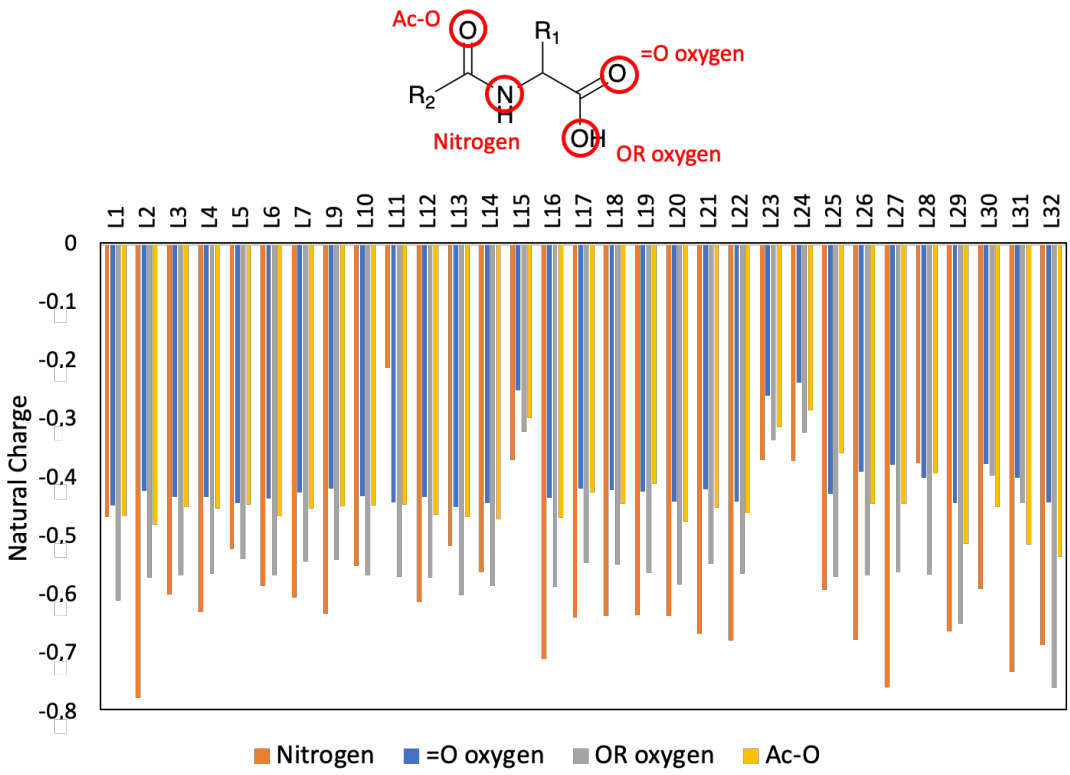


Figure S12 Location of the atoms which were used for the CHELPG analysis and results of the CHELPG calculation of all ligands.

Summary of DFT Descriptor Values

Table S3 includes all values of the DFT based descriptors.

Table S3 Summary of all calculated DFT descriptors

Ligands	Sterimol R-residue			Sterimol N-residue			NBO analysis				CHELPG analysis				% buried volume
	L	B1	B5	L	B1	B5	N	OR	=O	Ac-O	N	OR	=O	Ac-O	
L1	6.06	2.05	4.51	4.54	1.76	5.73	-0.37619	-0.35586	-0.28253	-0.3226	-0.466513	-0.608944	-0.446147	-0.464045	18.4
L2	6.66	2.75	4.56	5.12	1.82	5.56	-0.38442	-0.35069	-0.28116	-0.32745	-0.775537	-0.569357	-0.421585	-0.478563	38.4
L3	5.1	2.58	6.03	7.36	2.09	7.43	-0.38195	-0.35478	-0.29432	-0.32243	-0.599068	-0.565653	-0.432344	-0.448455	36.7
L4	4.49	3.64	6.08	10.74	2.26	7.29	-0.38321	-0.35592	-0.28733	-0.32306	-0.629191	-0.563101	-0.431461	-0.451794	45.1
L5	4.36	2.21	5.37	8.54	2.27	7.96	-0.37789	-0.35516	-0.29067	-0.32381	-0.520643	-0.538302	-0.441716	-0.444796	37.6
L6	4.62	2.93	5.47	8.08	2.29	8.23	-0.37512	-0.36435	-0.29063	-0.34228	-0.583526	-0.565598	-0.433752	-0.464669	34
L7	4.58	2.81	11.79	7.05	2.25	12.85	-0.38234	-0.35532	-0.29576	-0.32062	-0.603403	-0.542356	-0.423286	-0.451632	36.9
L9	4.49	1.9	5.45	6.94	2.15	3.7	-0.38201	-0.35491	-0.29448	-0.32238	-0.630682	-0.538847	-0.417083	-0.448027	28.5
L10	4.6	2.52	5.43	7.3	2.29	6.1	-0.3818	-0.35569	-0.29587	-0.32223	-0.549897	-0.566176	-0.430958	-0.446708	36.3
L11	6.4	1.96	4.51	6.45	2.02	5.55	-0.32148	-0.35629	-0.29017	-0.31724	-0.211833	-0.56818	-0.440953	-0.444532	40
L12	5.49	2.74	5.21	6.28	1.98	6.31	-0.39196	-0.35362	-0.28726	-0.3199	-0.61206	-0.569427	-0.4315	-0.462546	47.2
L13	4.55	2.76	4.87	5.96	1.83	6.92	-0.37844	-0.35757	-0.28296	-0.32388	-0.515195	-0.599674	-0.448818	-0.465924	37.7
L14	5.77	2.92	4.67	5.76	1.85	6.66	-0.38598	-0.35873	-0.29181	-0.32543	-0.560531	-0.584289	-0.442257	-0.470392	43.7
L15	4.56	3.04	4.86	5.07	1.84	7.11	-0.37986	-0.35286	-0.27977	-0.32606	-0.368372	-0.320297	-0.248912	-0.296172	37.6
L16	4.12	2.56	5.01	4.84	1.84	7.23	-0.37648	-0.35736	-0.28612	-0.32406	-0.708605	-0.585613	-0.433567	-0.467369	18.4
L17	4.94	2.61	6.89	7.27	2.9	7.4	-0.38537	-0.35497	-0.29515	-0.32284	-0.637421	-0.544184	-0.417856	-0.424352	37.3
L18	5.53	2.6	8.17	7.27	2.91	7.37	-0.38569	-0.35496	-0.29525	-0.32549	-0.635056	-0.546674	-0.420481	-0.443747	37.3
L19	5.86	2.6	6.93	7.29	2.9	7.38	-0.38591	-0.35521	-0.29524	-0.32314	-0.633977	-0.561779	-0.422134	-0.409802	37.3
L20	7.46	2.89	11.81	6.77	2.22	13.77	-0.37146	-0.35119	-0.29294	-0.33671	-0.635565	-0.581287	-0.440082	-0.474375	37.2
L21	6.47	2.58	10.51	7.27	3.04	7.35	-0.38579	-0.35495	-0.29528	-0.3262	-0.665955	-0.545958	-0.418825	-0.450242	37.3
L22	5.95	2.57	6.93	6.81	2.64	7.34	-0.37901	-0.36471	-0.29035	-0.34504	-0.676505	-0.562611	-0.440287	-0.458779	37.3
L23	6.25	2.79	7.98	7.76	3.26	4.92	-0.3798	-0.36903	-0.29149	-0.34273	-0.369425	-0.334984	-0.259106	-0.312619	43.6
L24	5.51	2.87	7.4	6.9	2.44	7.44	-0.37897	-0.35457	-0.26648	-0.3135	-0.370268	-0.322206	-0.236543	-0.283715	37.3
L25	5.47	2.18	8.31	6.9	3.38	7.49	-0.37417	-0.3565	-0.28921	-0.33472	-0.590294	-0.568786	-0.426612	-0.35735	37.3
L26	4.67	3.04	9.24	7.57	2.65	7.66	-0.37087	-0.3543	-0.27604	-0.32191	-0.67638	-0.565751	-0.388932	-0.443194	37.3
L27	5.06	3.03	9.55	7.56	3.02	7.7	-0.373	-0.35399	-0.27696	-0.31547	-0.756809	-0.560124	-0.377225	-0.44324	37.3
L28	7.32	2.59	7.83	7.48	2.17	6.34	-0.32331	-0.35088	-0.27854	-0.27649	-0.374424	-0.564444	-0.398645	-0.391023	37.3
L29	6	3.68	10.69	12.5	1.68	8.97	-0.372	-0.35088	-0.29467	-0.5198	-0.660768	-0.647813	-0.442392	-0.511759	37.3
L30	10.62	3.41	5.97	8.97	1.8	9.38	-0.38254	-0.27938	-0.25713	-0.31348	-0.588752	-0.395623	-0.375246	-0.448494	37.3
L31	10.04	3.54	7.48	10.4	3.54	7.03	-0.39434	-0.28205	-0.24733	-0.38337	-0.730678	-0.44217	-0.398695	-0.513599	37.3
L32	8.3	3.45	6.56	10.6	3.31	5.84	-0.38476	-0.38625	-0.30004	-0.38919	-0.684506	-0.758856	-0.441578	-0.533879	36.9

Machine Learning

Within this study five different ML surrogate models were used. Information on the chosen hyperparameters, the used software packages and the implementation can be found in this section. ML hyperparameter tuning was conducted for all models using feature set 14 within the initial supervised ML (random split). We observed that tuning hyperparameters separately for each feature set did not deliver significantly better performance than using the hyperparameters obtained when using feature set 14.

Figure S13 illustrates the workflow of parameterizing inputs and subsequentially conducting initial supervised ML and active ML featured closed-loop optimization.

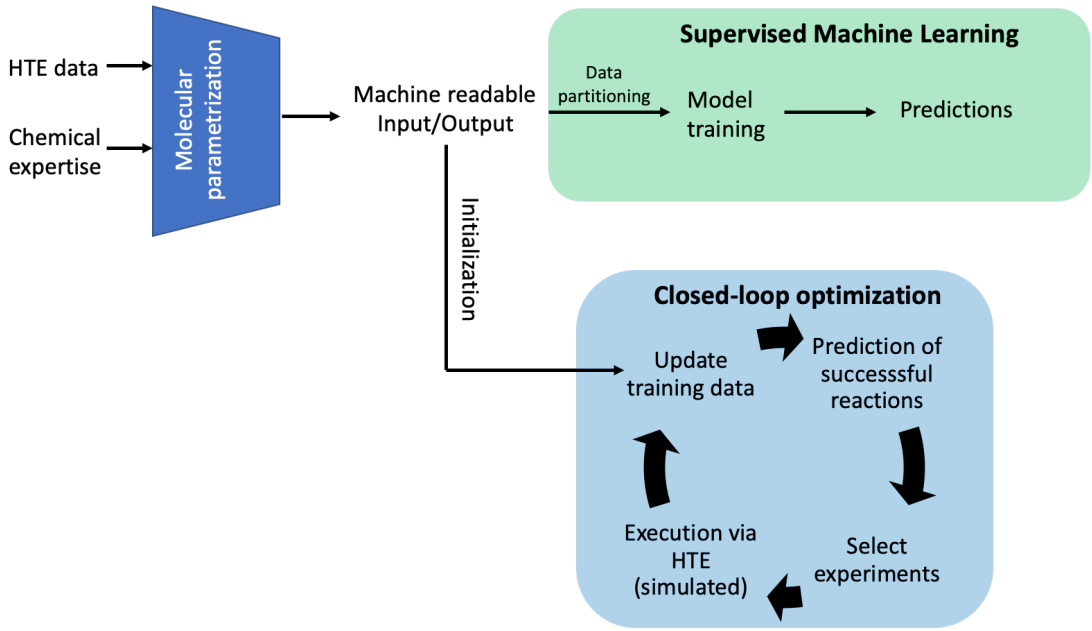


Figure S13 ML workflow in this study: The input data was parameterized and then used for supervised ML studies and for active ML.

Linear Model

We implemented the linear model using the package scikit-learn, version 0.23.0 – this version was used for all subsequent modelling. Unless explicitly stated, all parameters were left at default values.

Random Forest

We implemented the random forest surrogate model using the package scikit-learn. A total number of 200 estimators were used consistently for modelling. Unless stated otherwise all

parameters were kept to default. Unless explicitly stated, all parameters were left at default values.

Gaussian Process

We implemented the Gaussian process model using the package scikit-learn. As covariance function we used a Matèrn 3/2 kernel with a common length scale for all inputs. We observed that adding a white kernel to account for measurement noise did not improve the prediction performance and was hence omitted.

kernel = 1.0 * Matern(length_scale=1.2, nu=1.5)

Artificial Neural Network

The artificial neural network model was implemented using the Tensorflow Keras 2.3.0. The fully connected feed-forward network consisted of six hidden layers of ten nodes each and ReLu activation function. The final layer consisted of one single node. The weights and biases were initialized with the default schemes (Glorot uniform and zeros, respectively). Training was done with RMSProp using default parameters over 1000 epochs with a minibatch size of 32.

Adaptive Boosting Model

We implemented the AdaBoost model using the package scikit-learn. A total number of 200 estimators were used consistently for all modelling.

Support Vector Regression

We implemented the support vector regression model *via* the package scikit-learn using a linear kernel.

Leave-one-group-out (LOGO) Cross Validation (CV)

LOGO CV was used within the study of supervised ML to assess the models' performance to conduct extrapolative predictions. Figure S14 allows for insights into the single folds of the LOGO CV using feature set 14. This plot illustrates the test/train RMSE (y-axis) of the single groups (x-axis, 1-31), highlighting that the modelling performance strongly varies from ligand to ligand. Additionally, it is visible how different models fit train/test data.

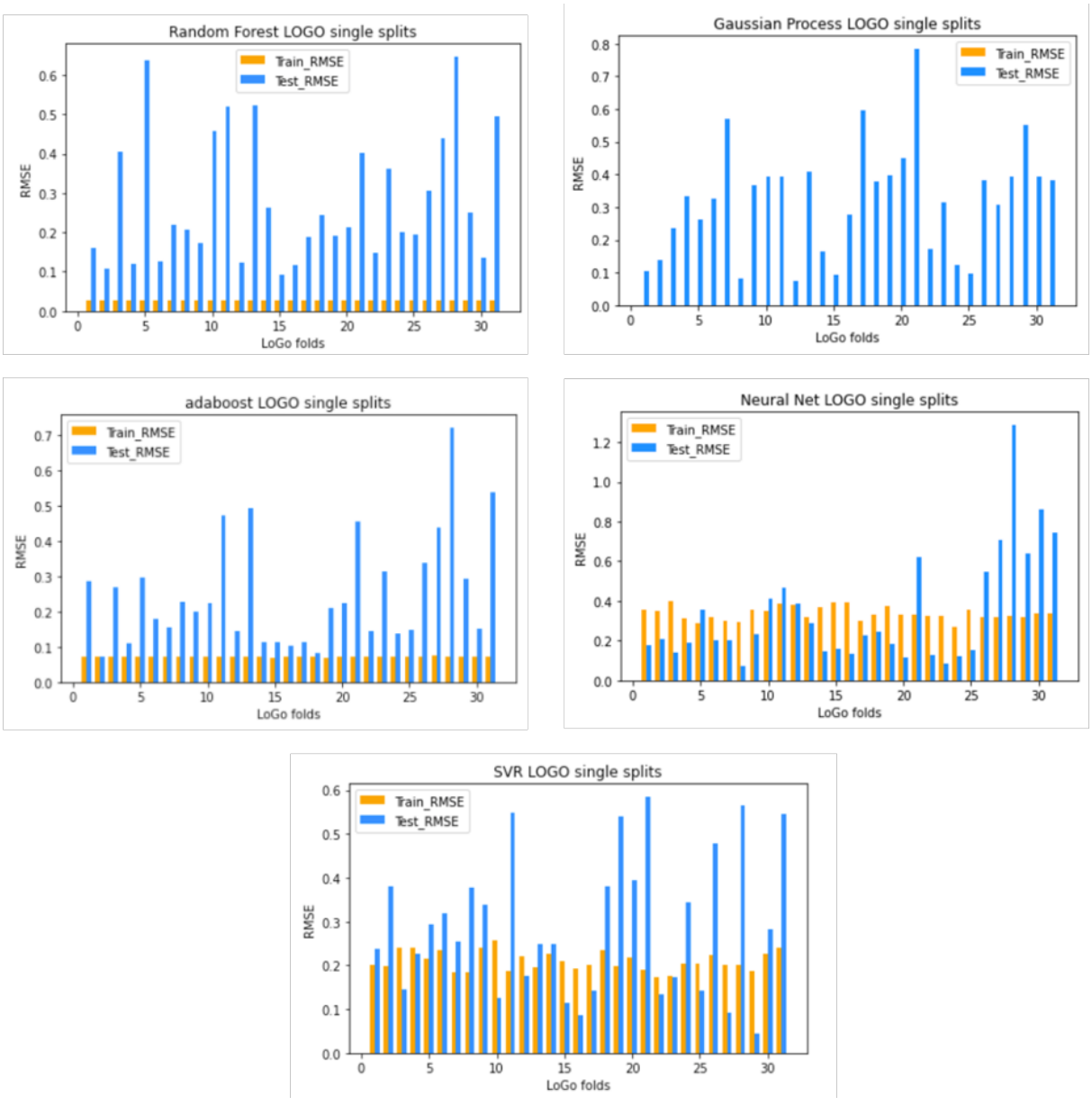


Figure S14 Detailed insights into the LOGO CV of different surrogate models

To investigate a potential correlation of the variation of the different train/test folds on model performance we looked at structural similarity. Using the Tanimoto similarity index,[6] the similarity between the training data (30 of the 31 ligands) and the test data (1 of the 31

ligands) was calculated and then the average of the values was taken for all 31 ligand molecules. Figure S15 shows the relationship between the averaged similarity indices and the RMSE of all 31 folds. When overlaying the results from all 5 models a higher density of datapoints in the lower right quadrant of the plot suggests that higher similarity between test and train data delivers lower model error, as expected. This follows the rational considerations of ML that model performance is typically increased when the training and test data have a higher similarity.

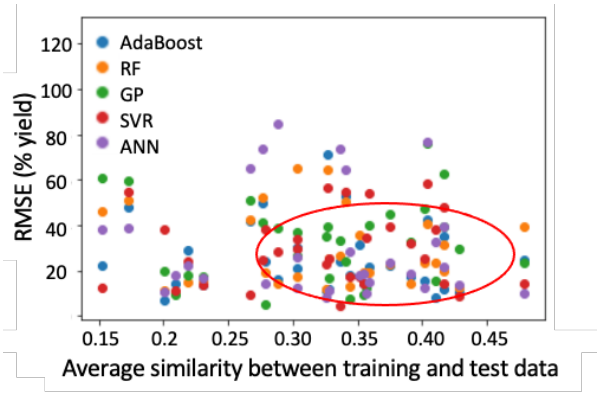
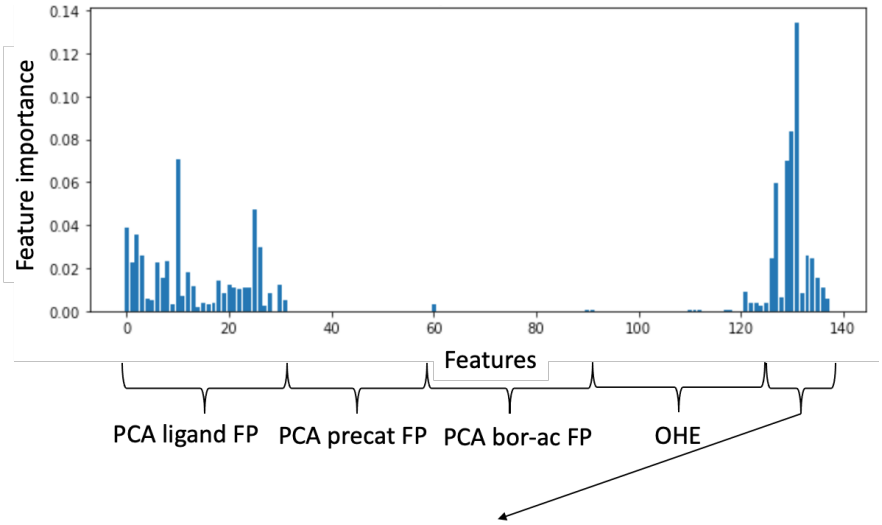


Figure S15 Insights into the LOGO CV evaluation - RMSE of the fold vs train-test data similarity using Tanimoto similarity index (feature set 14). The circle indicates increased density of datapoints, suggesting that a higher similarity leads to a lower RMSE. This follows general considerations of ML that increased similarity of training and test data delivers better performance.

Feature Importance Assessment of the Random Forest

Explainable AI (XAI) is a field which attempts to convert a 'black-box' model into a 'white-box' model throught increased transparency, ultimately allowing for an explanations/justifications as to how certain predictions were made.[7, 8] The application of XAI in synthetic chemistry challenges should therefore allow for increased understanding of the chemical system and enable synthetic chemists to profit from the pattern recognition-based strengths possessed by ML. While black-box models often keep these patterns hidden within the model architecture, highlighting this information can deliver great benefits. Within this project we attempted XAI with a feature importance assessment – here a RF model which was trained on the complete hybrid feature set was subsequently analyzed using Gini importance.[9] Figure S16 illustrates the feature importance, highlighting that the PCA of the fingerprints of the ligands contain relevant information while those of the pre-catalyst/boronic acid seem redundant. Overall, OHE also has a very low importance when combined with the other features in feature set 14.



Feature: 126, Score: 0.02462 → HOMO
Feature: 127, Score: 0.05956 → LUMO
Feature: 128, Score: 0.00629 → NBO_N
Feature: 129, Score: 0.07011 → NBO_OR
Feature: 130, Score: 0.08362 → NBO_O__
Feature: 131, Score: 0.13443 → NBO_AcO
Feature: 132, Score: 0.00859 → chelpg_N
Feature: 133, Score: 0.02560 → chelpg_OR
Feature: 134, Score: 0.02451 → chelpg_O
Feature: 135, Score: 0.01562 → chelpg_AcO
Feature: 136, Score: 0.01073 → vol_bur
Feature: 137, Score: 0.00575 → n_proton

Figure S16 Variation of the feature importance for RF models and detailed insights into relevant features of the hybrid feature set 14

It must be noted that the feature importance slightly changes every time a RF is retrained as the single trees within the forest are populated differently. Nonetheless, the highest importance was predicted to be the NBO analysis of the amide oxygen, thus suggesting that the electron density around the oxygen matters. This observation reflects the known elements of the concerted metalation deprotonation (CMD) mechanism, with the acetamide oxygen functioning as an internal base that abstracts the proton. This demonstrates the ability of XAI to deliver chemical insights on complex systems. Moreover, the insignificance of the features encoding for the pre-catalyst and the boronic acid aligns with the fact that those parameters do have a high impact on reaction yield. As visible in Figure S5, the reaction outcome is mainly influenced by the ligands, rather than the identity of the pre-catalysts or boronates. This is clearly visible as in several rows of the heatmap (row = ligand) the whole row is either dark blue (low yield) or green/yellow (high yield), however, within the columns of the line there is only limited variation in yield.

Closed-loop Optimization

Expected Improvement Acquisition Function

Compared to a purely exploitative search within the closed-loop optimization, using the EI acquisition function allows for a controlled trade-off between exploitation and exploration. Any parameter combination θ delivers a predicted mean $\theta(\mu)$ and a standard deviation $\sigma(\theta)$. Following Eqn. 6, EI can be calculated relatively with respect to the current best condition from previous iterations, referred to as m_{opt}.

$$EI(\theta;\, m_{opt}) = \delta(\theta)\Phi\left(\frac{\delta(\theta)}{\sigma(\theta)}\right) + \sigma(\theta)\phi\left(\frac{\delta(\theta)}{\sigma(\theta)}\right) \qquad \text{Eqn 6.}$$

where $\delta(\theta) = \mu(\theta) - m_{opt}$, Φ is cumulative standard normal,

ϕ is standard normal density

The distance to the best condition is calculated by $\delta(\theta)$ and the search is conducted with the objective to find the θ that maximizes EI. It is noteworthy that not all ML models deliver an uncertainty metric, for example, GPs have built-in variance due to the model design, whereas ANN and RF do not have an uncertainty output. It should be noted that in our study we navigate in a solely discrete optimization space.

De-full Factorization of the Chemical Space Study

We hypothesized that the simplicity of OHE along with a full factorial space could be more beneficial when compared to the effect on other input features (e.g. hybrid inputs) which are far more complex and might represent a challenge for the model to detect patterns in the data. To test this assumption, we dropped a random selection of the datapoints of the entire dataset (25%), therefore no longer representing a full factorial chemical space. However, we still observed that OHE outperformed the full feature set (Figure S17a). Figure S17b illustrates that even though the dataset was reduced, yield was still well distributed.

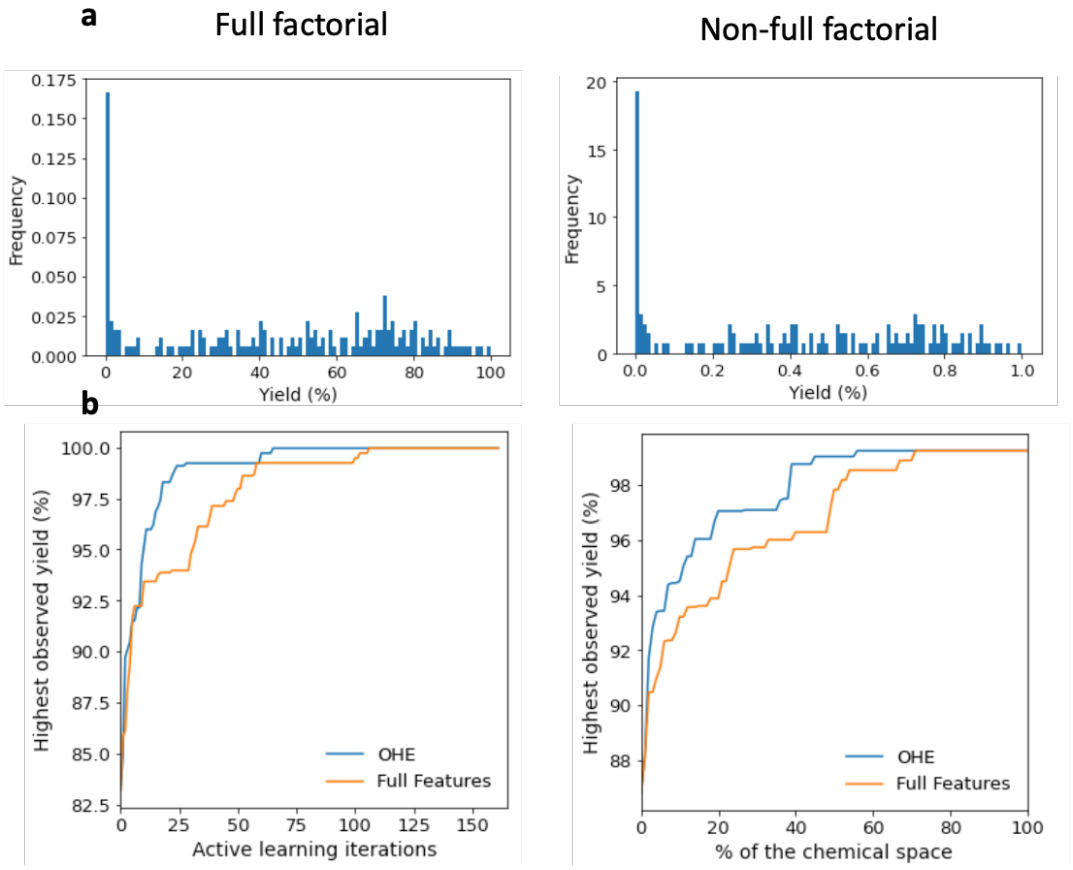


Figure S17 Comparison of active learning model performance of a full factorial and a non-full factorial chemical space (a) Yield distribution of full factorial and a non-full factorial chemical space (b) Active learning curves of full factorial and a non-full factorial chemical space

Batch-Sequential Active Learning

We assessed the impact of using different batch sizes *vs* using sequential sampling. In Figure S 18 (x-axis normalized) the batch size was varied between two and 25 experiments (during each iteration) and sequential sampling (one experiment at a time) is illustrated as a baseline. The minimal differences in the learning curves indicate that smaller batch sizes have favorable learning curves compared to larger batch sizes. We hypothesize that a smaller batch size allows the active ML model to be updated more frequently and thus conduct predictions of slightly higher accuracy. At 40% of the chemical space (Figure 18, x-axis) the active learning strategy using a batch size of 25 iterated two times whereas the batch size of two iterated 30 times. Overall though, it seems that the batch size does not significantly impact the learning trajectory and thus the size should mainly be chosen based on experimental restrictions (e.g. possible number of experiments which can be run in parallel).

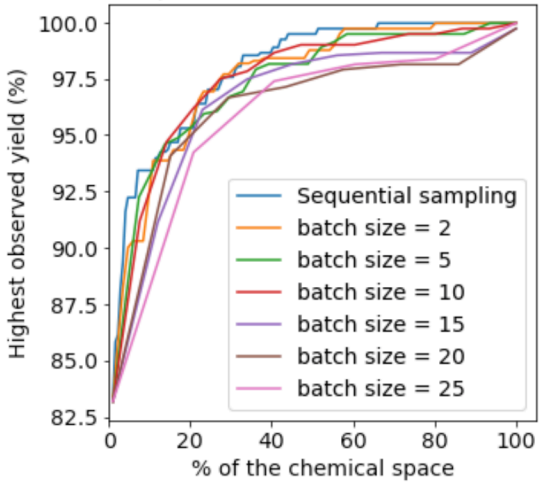


Figure S18 Comparison of different batch sizes for active learning using RF and feature set 14

The Impact of Initialization of the Active Learning

The efficiency of closed-loop optimization algorithms depends on the data on which the very initial model is trained. Herein, we are comparing a broader set of reaction conditions (on average the dataset contains information of 7 ligands) to a restricted dataset (the dataset contains datapoints of only 3 ligands). To allow for general statements, the ligand in the train/test set were varied during 10 single experiments and the average of the learning curves was used for the plot. As visible in Figure S19, the location of the initialization data, and the lookup table (of one experiment) were illustrated in a dimensionality reduced 2D map that was generated using the first two principal components of the Morgan 2 fingerprints of all components. Whilst Figure S19a illustrates the initial data being randomly distributed over the chemical space, Figure S19b has the initial data located close to each other such that it has been intentionally limited to only to three ligands. It must be noted that the plot contains an overlap of datapoints which is a result of the dimensionality reduction. In Figure S19c, it is apparent that even as the local initialization possessed restricted knowledge, within ten iterations the model performance was approximately equal to an initialization dataset which is more diverse.

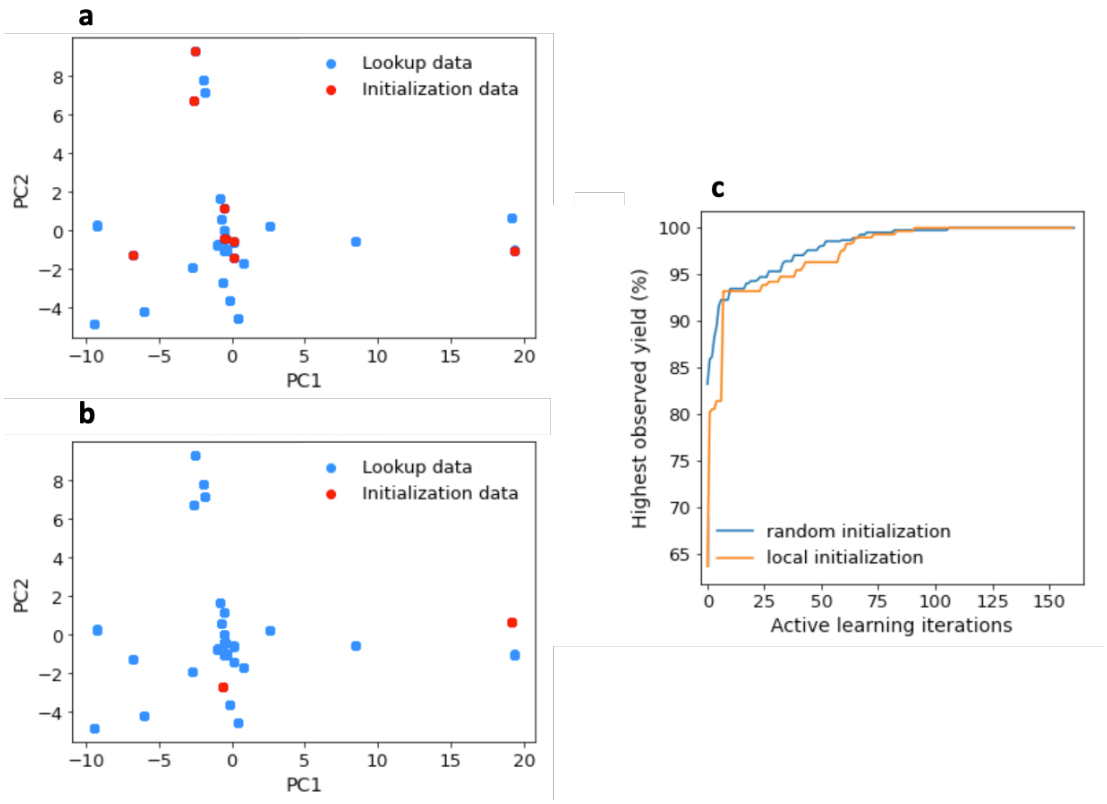


Figure S19 Random versus local initialization (a) Dimensionality reduced plot of the training and test data within random initialization (b) Dimensionality reduced plot of the training and test data within local initialization (c) Learning curves of the random versus the local initialization

The Impact of Initialization: Dataset Size vs. Complexity of Parameterization

We conducted a comparison to understand the performance variation between complexity of the parametrization and the size of the initial dataset. In terms of size of the initialization dataset, 10, 15 and 20 datapoints were chosen along with OHE, Morgan 2 fingerprints and hybrid full feature representation. Figure S20 illustrates all the learning curves of the conducted experiments – in the main manuscript we restricted the plot to 4 trajectories for simplification. The trend of increased performance when using a larger number initialization datapoints using OHE as opposed to a smaller but more complex parameterized initialization dataset could be observed again.

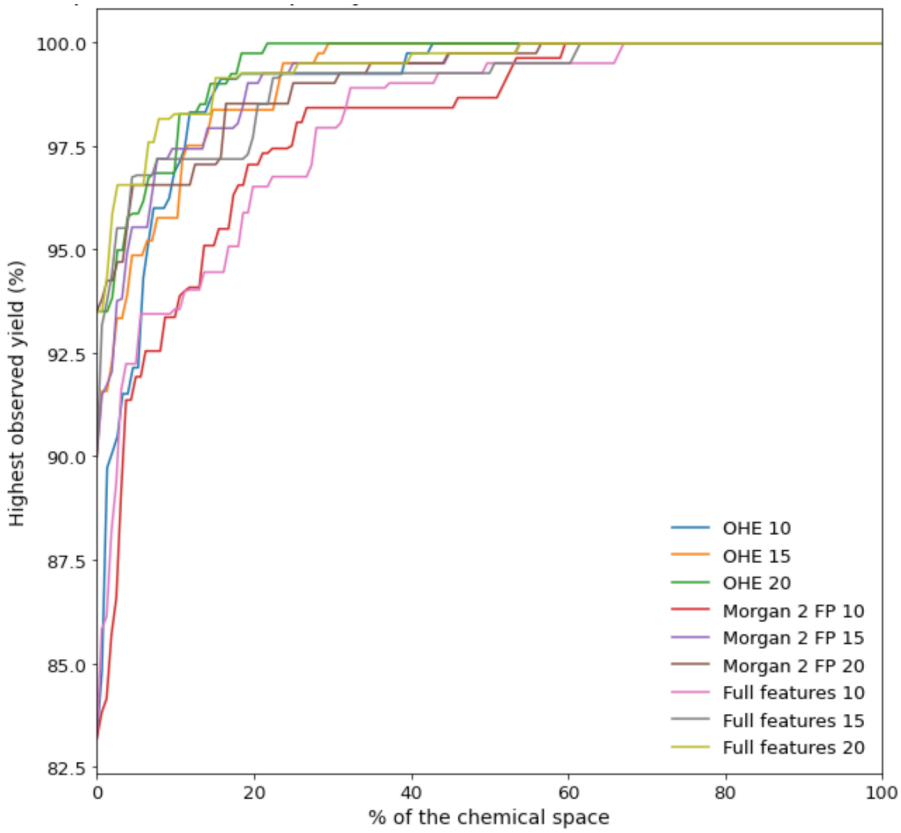


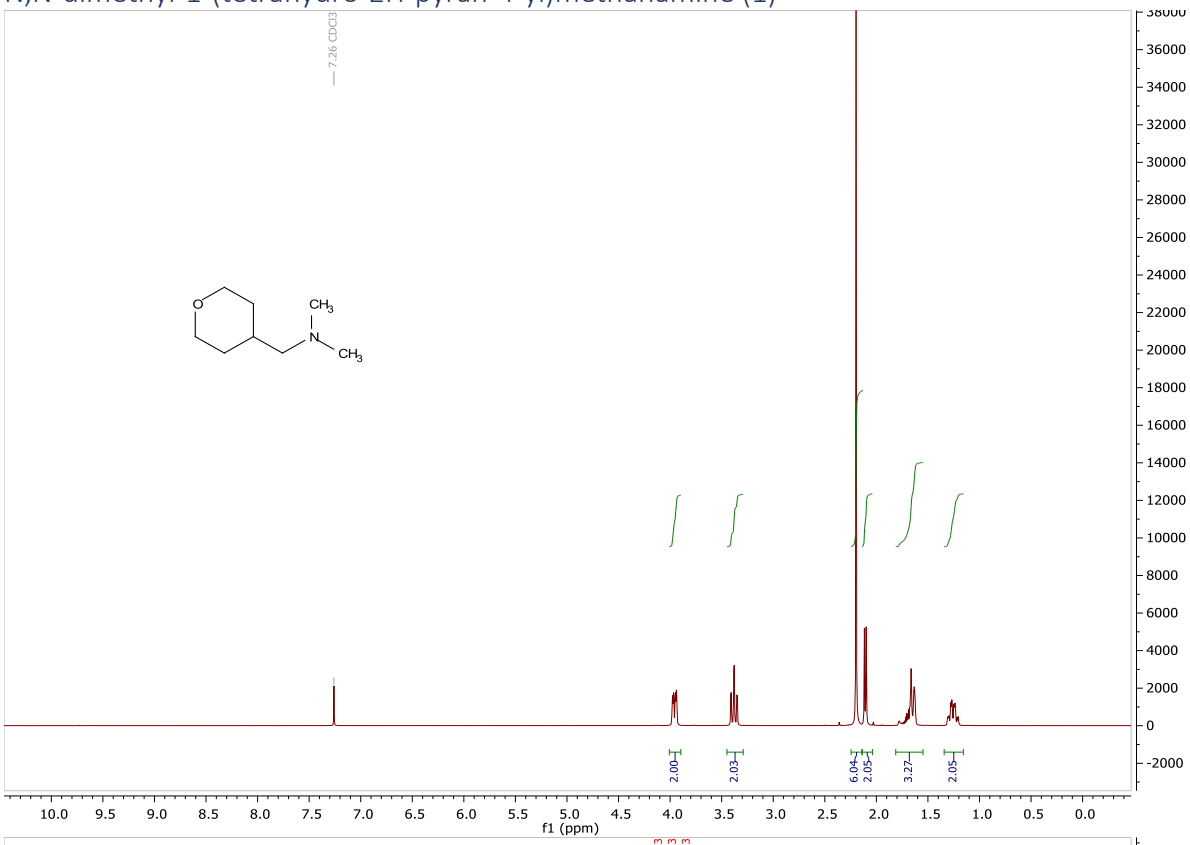
Figure S20 Evaluation of different initialization strategies for the active learning - variation of chemical representation and size of the initialization dataset

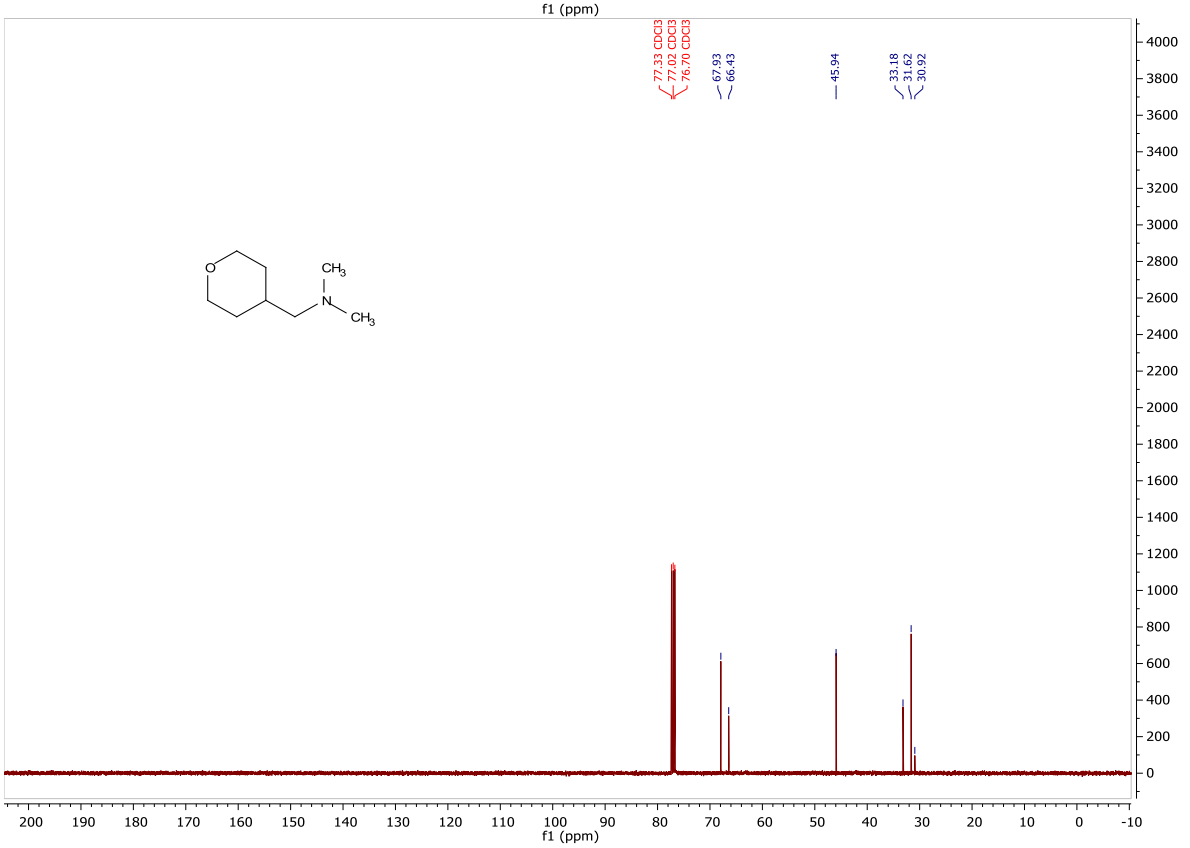
References

1. D. Y. Ong, Z. Yen, A. Yoshii, J. Revillo Imbernon, R. Takita and S. Chiba, *Angew. Chem. Int. Ed.*, 2019, **58**, 4992-4997.
2. A. V. Brethomé, S. P. Fletcher and R. S. Paton, *ACS Catal.*, 2019, **9**, 2313-2323.
3. L. Falivene, Z. Cao, A. Petta, L. Serra, A. Poater, R. Oliva, V. Scarano and L. Cavallo, *Nat. Chem.*, 2019, **11**, 872-879.
4. F. Weinhold, C. R. Landis and E. D. Glendening, *Int. Rev. in Phys. Chem.*, 2016, **35**, 399-440.
5. C. M. Breneman and K. B. Wiberg, *J. of Comp. Chem.*, 1990, **11**, 361-373.
6. G. Maggiora, M. Vogt, D. Stumpfe and J. Bajorath, *J. of Med. Chem.*, 2014, **57**, 3186-3204.
7. J. Feng, J. L. Lansford, M. A. Katsoulakis and D. G. Vlachos, *Sci. Adv.*, 2020, **6**.
8. J. Jiménez-Luna, F. Grisoni and G. Schneider, *Nat. Mach. Intell.*, 2020, **2**, 573-584.
9. B. H. Menze, B. M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich and F. A. Hamprecht, *BMC Bioinform.*, 2009, **10**, 213.

Appendix

N,N-dimethyl-1-(tetrahydro-2H-pyran-4-yl)methanamine (1)





N,N-dimethyl-1-(3-phenyltetrahydro-2H-pyran-4-yl)methanamine (2)