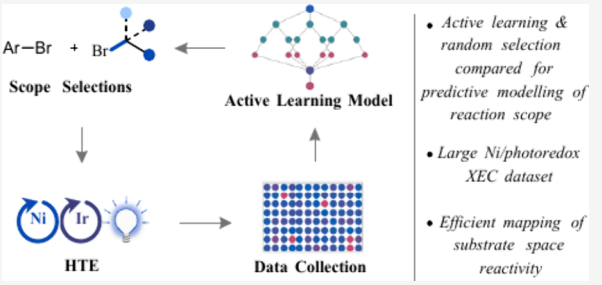
JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
pubs.acs.org/JACS
Article

Applying Active Learning toward Building a Generalizable Model for Ni-Photoredox Cross-Electrophile Coupling of Aryl and Alkyl Bromides

Lucas W. Souza, Nathan D. Ricke,[§] Braden C. Chaffin,[§] Mike E. Fortunato, Shutian Jiang, Cihan Soylu, Thomas C. Caya, Sii Hong Lau, Katherine A. Wieser, Abigail G. Doyle,* and Kian L. Tan*

Cite This: https://doi.org/10.1021/jacs.5c02218
Read Online

ABSTRACT: When developing machine learning models for yield prediction, the two main challenges are effectively exploring condition space and substrate space. In this article, we disclose an approach for mapping the substrate space for Ni/photoredox-catalyzed cross-electrophile coupling of alkyl bromides and aryl bromides in a high-throughput experimentation (HTE) context. This model employs active learning (in particular, uncertainty querying) as a strategy to rapidly construct a yield model. Given the vastness of substrate space, we focused on an approach that builds an initial model and then uses a minimal data set to expand into new chemical spaces. In particular, we built a model for a virtual space of 22,240 compounds using less than 400 data points. We demonstrated that the model can be expanded to 33,312 compounds by adding information around 24 building blocks (<100 additional reactions). Comparing the active learning-based model to one constructed on randomly selected data showed that the active learning model was significantly better at predicting which reactions will be successful. A combination of density function theory (DFT) and difference Morgan fingerprints was employed to construct the random forest model. Feature importance analysis indicates that key DFT features that are related to the reaction mechanism (e.g., alkyl radical LUMO energy) were crucial for model performance and predictions on aryl bromides outside the training set. We anticipate that combining DFT featurization and uncertainty-based querying will help the synthetic organic community build predictive models in a data-efficient manner for other chemical reactions that feature large and diverse scopes.



INTRODUCTION

Supervised machine learning (ML) has emerged as a powerful tool for applications in synthetic organic chemistry. In recent decades, numerous papers have explored retrosynthesis and forward synthesis prediction of complex reaction parameters such as stereoselectivity, reaction yield, and reaction conditions.[1,2] While significant advances have been made in the area of retrosynthesis, generalized forward prediction of yield and reaction success remains a significant challenge for ML.[3]

The early perceptions that big data coupled with machine learning would solve the yield prediction challenge have fallen short of expectations. Attempts at leveraging large publicly available databases such as literature data for the prediction of reaction yields have led to poor model performance (Figure 1).[4,5] While these data sets are large, they introduce a number of confounding variables, often resulting in low-quality models.[6−8] Another large source of chemical data has been sourced from pharmaceutical companies via electronic laboratory notebooks (ELN). These data, however, have been found to suffer from many of the same issues as public databases. Mainly, inaccurately reported yields and large variations in reaction setup introduce confounding variables.[9]

One area of success for machine learning in organic chemistry has been using high-throughput experimentation (HTE) to generate de novo data sets. HTE data sets have been used to build accurate reaction yield models for Suzuki–Miyaura cross-coupling reactions, Buchwald–Hartwig aminations,[9−13] and many other reaction classes.[13−23] However, due to the relatively small substrate scope of these works, models that can generalize to unseen substrates have been challenging to identify. Recently, the Coley lab, in collaboration with AbbVie, published an exhaustive study utilizing an HTE data set for Suzuki reactions spanning over a decade of internal electronic lab notebook (ELN) data from a select few number

Received: February 5, 2025
Revised: April 24, 2025
Accepted: April 28, 2025

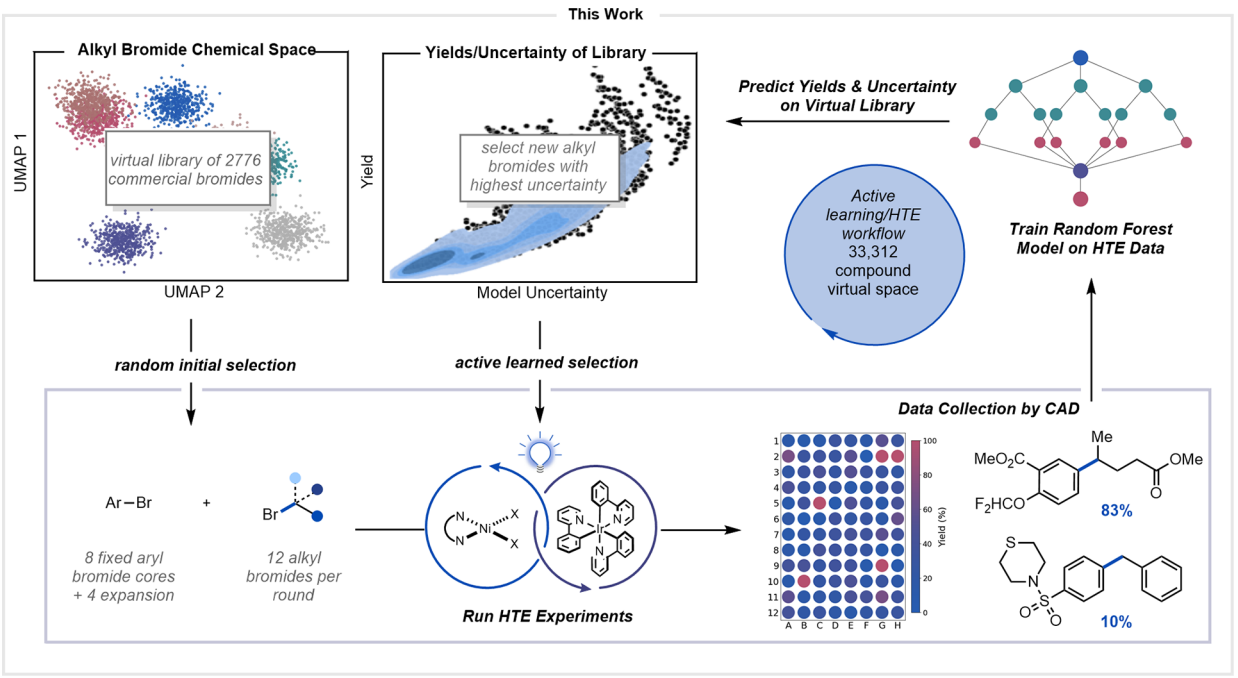


Figure 1. Strategy for the active learning generalization study of Ni/photoredox cross-electrophile coupling.

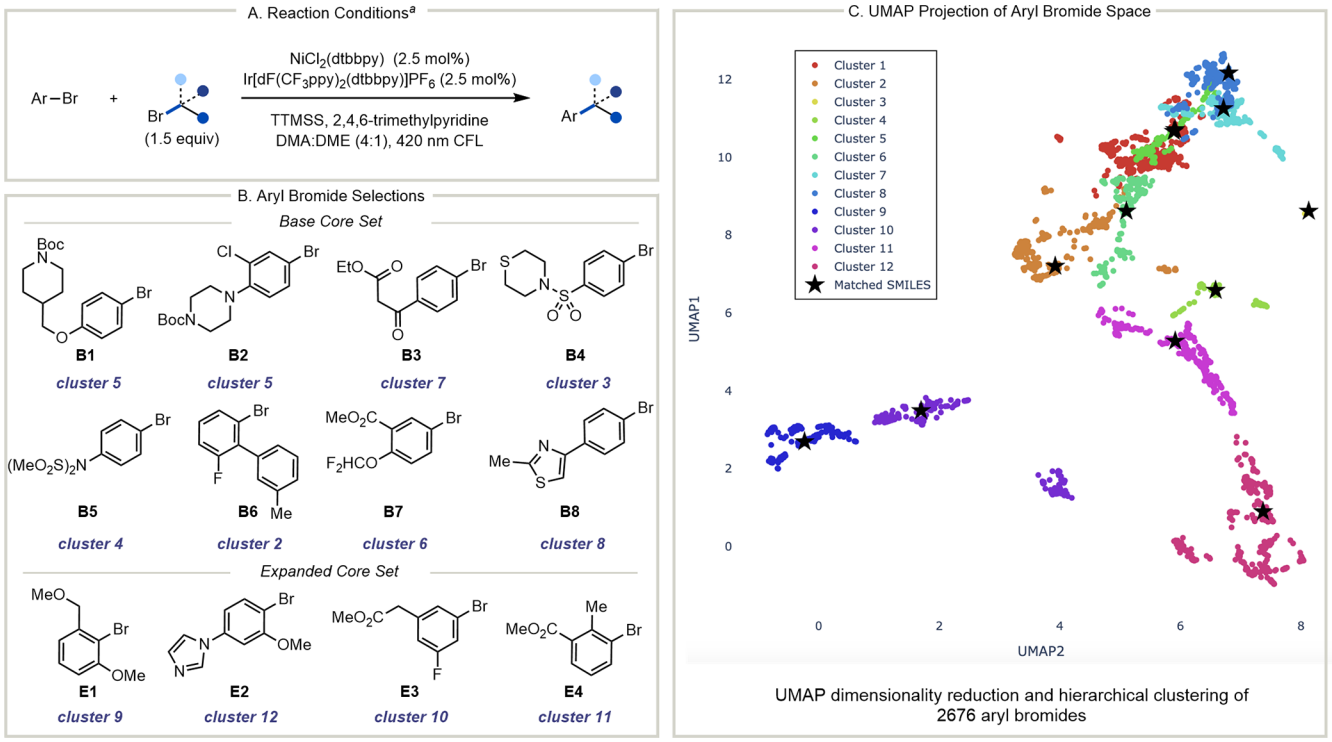


Figure 2. (A) Reaction conditions; (B) aryl bromide selections. Note: two molecules are chosen from cluster 5, (C) UMAP of aryl bromide space. [a]Reaction conditions; aryl bromide (0.033 mmol, 1 equiv), aryl bromide (0.049 mmol 1.5 equiv), $NiCl_2$(dtbpy) (2.5 mol %), Ir[df(CF_3)ppy](dtbbpy)PF_6 (2.5 mol %) 1,1,1,3,3,3-hexamethyl-2-(trimethylsilyl)trisilane (TTMSS, 1.2 equiv 0.039 mmoi), 2,4,6-trimethylpyridine (3 equiv, 0.098 mmol), dimethylacetimide (DMA, 43.6 μL), dimethoxyethane (DME, 174.2 μL), compact fluorescent lamp (CFL).

of chemists.[2,24] While this work showed encouraging levels of generalizability, the rare and highly curated nature of the dataset represents a significant hurdle to access by the broader chemistry community. Additionally, the authors note that changing sets of reaction conditions inherent to data spanning such a long period of time proved challenging and likely impacted the degree of generalization observed.[17] These publications highlight the potential of HTE data to facilitate yield prediction models of coupling reactions but also demonstrate the challenge of modeling the vastness of substrate and/or reaction condition space.

https://doi.org/10.1021/jacs.5c02218
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

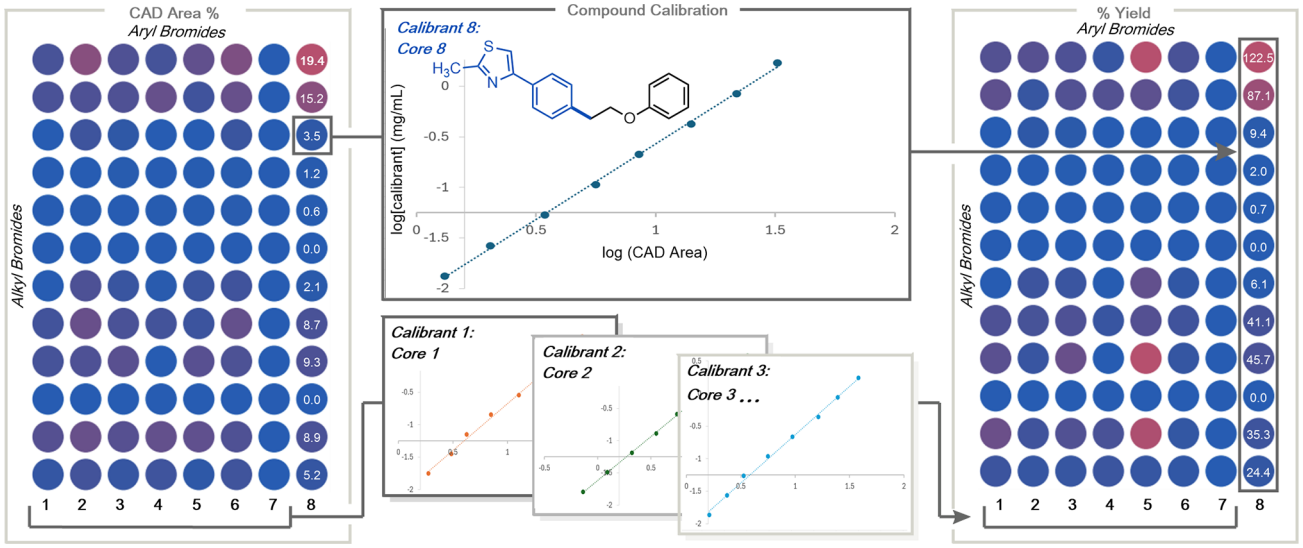


Figure 3. Plate layout and CAD calibration workflow. The % yield of each reaction is calculated using the appropriate curve generated from one of 11 calibrants containing the same aryl bromide core.

The enormity of substrate space likely makes it expensive in terms of time and money to build fully generalized models. Therefore, we have been rethinking how to more efficiently acquire experimental data to train ML models for reaction prediction, as well as reconsidering how to approach the problem of generalization. Traditionally, we would design a set of high-throughput experiments to learn substrate reaction mapping in a static space.[25−28] Instead, in this study, we consider substrate space as dynamic, in which we need to constantly expand the model to learn new spaces that are of interest.[13] From this perspective, the approach to data collection and model generation becomes the focal point rather than the model itself.

In this paper, we investigate the use of active learning (in particular, uncertainty querying) to map out substrate space (Figure 1). Active learning is a branch of machine learning that uses an algorithm in concert with a human user to choose the next most informative experiments to improve model performance. In particular, active learning has been shown to be an effective strategy for building a model with a limited amount of data.[29] This is similar to Bayesian optimization for reaction condition optimization,[30,31] where instead of optimizing for yield, model performance is the objective. Our approach focuses on building an initial model for a reaction class and then using a minimal set of experiments to extend the model to a new desired chemical space.

To explore this approach, we focused our efforts on nickel photoredox cross-electrophile coupling reactions of aryl and alkyl bromides (Figure 2A).[32] This was a particularly appealing reaction class for the outlined approach because, being a relatively new reaction methodology, literature data sets are sparser compared to data on other cross-coupling reactions. Moreover, the rapid adoption of $C(sp^2)-C(sp^3)$ coupling reactions in the pharmaceutical industry (especially within library groups) demonstrates the high value of these new transformations.[33−35] It is well documented that incorporation of $C(sp^3)$ carbon into drugs leads to compounds with better properties, which correlates with a higher success rate in clinical trials.[36] In this work, we used only one set of reaction conditions, enabling the focus on the exploration of substrate space. Overall, we anticipate that both the outlined approach for collecting reaction data sets and the data set itself will be of broad value to the organic chemistry and data science communities.

RESULTS AND DISCUSSION

Active learning has been shown to improve the efficiency of model construction in chemistry.[37−39] To frame the use case of building an initial model with active learning, we defined an initial virtual product space of aryl bromides and alkyl bromides. While the aryl component of Ni $C(sp^2)-C(sp^3)$ coupling reactions tends to be relatively conserved, variation of the alkyl component often drives reactivity, and in many cases even leads to new methodologies.[35] Additionally, in an HTE library context, the aryl bromide tends to be the core scaffold, while alkyl bromides are the diversity element. As such, when considering the design of the data set, we focused on robustly exploring the alkyl bromide space while sampling the aryl bromide space. With this in mind, part of our experimental design was to establish the number of experiments required to expand to new aryl bromides. The initial virtual space was composed of a matrix of 8 aryl bromides × 2776 alkyl bromides (22,208 compounds). A second virtual space was also created to study the expansion of the model, which was defined by 4 new aryl bromides × 2776 alkyl bromides (11,104 compounds).

Compound Selection. The aryl bromide and alkyl bromide chemical spaces were developed in a similar fashion to previous work in the Doyle lab.[40] A search was performed in the "Reaxys" database, and the output was subsequently filtered to include only molecules available from Sigma-Aldrich to ensure that all molecules included in the space were readily available. For the alkyl bromides, this search returned 2776 commercially available primary, secondary, and tertiary alkyl bromides. Density functional theory (DFT) features of the selected alkyl bromide scope were generated using the AutoQchem software developed by the Doyle group.[41] Features selected include both molecular (global) features and atomic features (min/max and C/Br atomic). Redundant features that did not vary across the data set and features with a high correlation to other ones were removed, resulting in a total of 54 features. We employed Uniform Manifold Approximation

https://doi.org/10.1021/jacs.5c02218
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

and Projection (UMAP)[42] for dimensionality reduction of the DFT features. We then used hierarchical clustering to group alike alkyl bromides into 15 clusters within the 10-dimensional UMAP space. The molecules that were closest to the center of these clusters were chosen for the first round of HTE, and subsequent rounds of active learning on this space were used to further guide model construction.

The same initial process used to design the alkyl bromide chemical space was then applied to generate the aryl bromide chemical space. The initial filtered search for aryl bromides generated 2683 molecules. Due to structural irregularities and DFT convergence failures, the number of molecules was reduced to 2676 after featurization and preprocessing. The removal of highly correlated or nonvarying features reduced the total features for the aryl bromides from 168 to 95. We then applied UMAP dimensionality reduction (to 10 dimensions) and hierarchical clustering to subdivide the space into 12 clusters (Figure 2C). From the 12 clusters, eight aryl bromides were selected to serve as our base set of cores (Figure 2B). Four aryl bromides from clusters 9−12 were selected as the expanded core set to be used to understand the extendibility of the model to a new chemical space. Reaction conditions were chosen based on the most popular conditions utilized at Novartis.

Experimental Methodology. All reactions were run in a 96-well plate under a single set of conditions. Product quantities were detected by ultrahigh-pressure liquid chromatography mass spectrometry (UPLC-MS) in combination with charged aerosol detection (CAD) (Figure 3) and are referred to as CAD yield or yield. Although CAD detectors are viewed as universal detectors, we have found that the accuracy of concentration can be improved by generating a calibration curve of a structurally similar compound. Our quantification method is similar to one previously published from our lab.[43] In this study, one product from each aryl bromide was synthesized, and a calibration curve was generated. CAD area under curve measurements for reactions of that aryl bromide with a given alkyl bromide were then used in combination with the requisite curve to calculate the product amounts. In our experience, these CAD measurements have a concentration variance of ±27% (see Supplemental file). Core B3 failed to produce enough product in any reaction in this study to generate a CAD calibration curve. As such, the curve for core B8 was used to calculate the yield for the few reactions that produced products with B3. Area under curve measurements of CAD signals were taken utilizing Virscidian analytical studio. In a limited number of cases, ^{1}H quantitative nuclear magnetic resonance spectroscopy (^{1}H-QNMR) of crude reaction mixtures was used to determine yield with subsequent validation by ^{1}H NMR of purified product. Roughly 5% of crude reaction mixtures had ambiguous CAD traces, and/or crude ^{1}H NMR spectra could not be validated due to difficult isolation of the desired compound. In such circumstances, these data were excluded from our models.

Modeling Methodology. Active learning is a machine learning strategy in which a trained machine learning model is allowed to query unlabeled data points to be labeled and used to update the model. It consists of iterative cycles of data selection, annotation, and model retraining. The machine learning models in this study use a combination of structural features from difference reaction fingerprints and electronic features computed using DFT. These DFT features were included to supplement the difference reaction fingerprints (DFRP) because we anticipated that specific DFT features could serve as effective and inexpensive predictors of reaction yield.[44] We computed the difference fingerprints using RDKit[45] to generate Morgan fingerprints[46] for both the reactants and products with a radius of 4, fixed length folding into 2048 bits, and the chirality flag as true. We then took the difference of these fingerprints as products minus reactants to compute the difference reaction fingerprints.

To maximize the impact of the DFT features, we chose to compute a set of DFT properties that were specifically relevant to the radical intermediates of the reaction (see Supplemental file for the computational workflow). As many of these properties were not available in Auto-Qchem, we used an alternate computational workflow described as follows. First, conformer ensembles were generated for each reactant,[47] then all structures in the ensemble were optimized using xTB,[48,49] and the lowest xTB energy conformer from each ensemble was selected. DFT geometry optimization and subsequent DFT property calculations of the DFT-optimized geometry were then performed. Turbomole[49] was used for all DFT calculations, applying the B3LYP functional,[50−52] Def2-SVP basis set,[53] and cosmo solvation with a dielectric constant of 78.4.[54] The computed properties include both molecular properties such as HOMO/LUMO energies or electron affinity and atomic properties such as NMR (Nuclear Magnetic Resonance) shielding and partial charges. For the alkyl bromides, these properties were computed for both the alkyl bromide and the debrominated radical intermediate that is relevant to the reaction mechanism. To maintain a fixed size, featurization focused on the reaction site, only atomic features for the bromine atom and the single carbon covalently bound to it within the aryl and alkyl bromides were included. The full list of computed properties used as model features is provided in the Supplementary Material.

Throughout the active learning cycles of this study, the DFRP and DFT features were used to train a sci-kit learn[55] implementation of random forest models with 500 trees.[56] This hyperparameter was chosen prior to data generation and was therefore selected based on our experience as a value that could balance model capacity without extraneous complexity for the expected 500−1000 training data points at the completion of the study. To compute uncertainties for active learning, the standard deviation of the output from all decision trees in the random forest for each prediction was determined. This inherent uncertainty estimate available from random forest models in combination with computational efficiency guided the selection of the model architecture used in this study (see Supplementary Material for the evaluation of random forest error calibration in this study).

Batch Selection. In our experimental design, each round of active learning requires a selection of 12 alkyl bromide building blocks (8 cores × 12 alkyl bromides = 96 reactions). As mentioned above, the uncertainty of a single reaction is extracted from the decision tree ensemble (Figure 4A). To compare the information content of the building blocks, an average uncertainty of each building block over the 8 cores was determined (Figure 4B). To favor diversity of feature space within each batch, the Kriging Believer selection process was applied.[57]

In this approach, the alkyl bromide with the highest average uncertainty was selected for testing (Figure 4C), and then the prediction for this building block was used to augment the training data. The model was retrained on this new augmented

https://doi.org/10.1021/jacs.5c02218
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

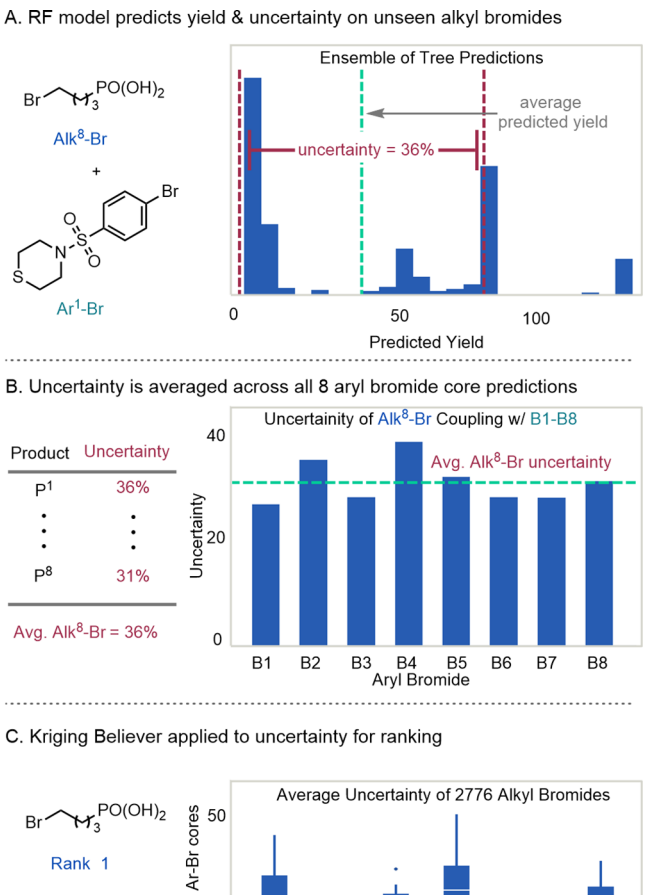


Figure 4. Workflow used to predict the reaction yield uncertainty using a random forest model. For the batch cycles within this work, the model made predictions on each alkyl bromide in the library reacted with a fixed set of aryl bromides. The random forest model, which is composed of an ensemble of decision trees, predicts both the average and standard deviation of reaction yield. Active learning then selected alkyl bromides from the library with the highest average standard deviation of the reaction yield across all products formed by the reaction of an alkyl bromide with the fixed set of aryl bromides.

training dataset, which included both the original experimental training data as well as the model predictions generated during the current batch selection. The process lets the model believe the predictions for the previous selection are correct in order to reduce uncertainty near the feature space around the selection, causing active learning to deprioritize molecules similar to what it has already selected within the batch. This process was repeated, augmenting the training data with model predictions for all previous selections, until all 12 alkyl bromides were selected.

Model Evaluation. To judge model performance after each iteration, a test set was held out, composed of the products of the 8 aryl bromide core set with 26 alkyl bromides (which were excluded from training data selection). For the search space of 2776 alkyl bromides, selecting batches of alkyl bromides 12 at a time, 231 full batches could theoretically be collected, but ideally, an effective data acquisition strategy could produce a useful model with only a few batches. To systematically evaluate the effectiveness of active learning as a data acquisition strategy, we experimentally tested alkyl bromides selected by random sampling for a baseline comparison. As active learning experiments must be conducted in series, the approach can only be justified if it produces a superior model with fewer total experiments. We conducted this comparison by selecting equally sized batches of alkyl bromides with both acquisition methods, using the data from each approach to train separate models that were evaluated against the base core test set (Figure 5). Note that because each batch of 12 alkyl bromides was reacted with 8 aryl bromides, each batch had 96 reactions. This process continued until the model performance stabilized between iterations for both acquisition methods, which occurred at batch 4.

Active learning initially did not outperform random sampling; however, after the third and fourth iterations, the active learning trained model had a lower root mean squared error (RMSE) and higher R^2 values than the randomly trained model when evaluated on the test set. Notably, the leveling of model performance occurred after generating ~250 to 350 unique products within this local chemical space (i.e., after coverage of ~1 to 2% of the virtual space). To investigate the generalizability of the models, we evaluated their performance on the expanded core test set, which is composed of aryl bromides not found in the training data of the first 4 iterative cycles. Although the RMSE for both acquisition methods was higher on the expanded core test set, the active learning model's performance improved significantly with each iteration, while random sampling consistently underperformed. By the fourth cycle, the R^2 value for the expanded core set approached the performance of the base core set, highlighting that the active learning model is predicting new aryl bromides, even without any local core information in the training data.

To investigate the value of adding local core information, a fifth cycle of active learning was performed using the expanded core set as the reactants. Note that active learning for this cycle was allowed to select from both previously tested and untested alkyl bromides with the base core set, as the reactions with novel aryl bromides would produce novel products. In this final batch of training data, active learning selected 24 alkyl bromides with high uncertainty in the expanded core set. These data are indicated by the star in Figure 5. Retraining the model on this batch significantly improved the performance on the expanded core test set while slightly increasing the performance on the base core test set. The larger performance gain on the expanded core test set is consistent with the expectation that adding local chemical information should improve the model performance for that space.

The significant boost in performance persuaded us to investigate further how much data was required to see these improvements. In the original experiment, we added 96 reactions (4 cores × 24 building blocks) in the fifth iteration. Instead of retraining on the entire fifth batch, we investigated adding building blocks to the model one at a time (4 data points/iteration: 1 building block × 4 cores). This process

https://doi.org/10.1021/jacs.5c02218
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

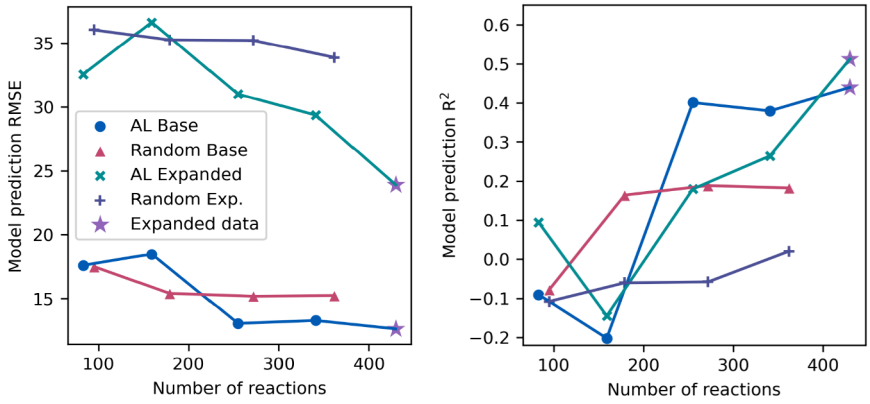


Figure 5. Regression metrics on model performance on base and expanded test sets as each batch was collected. The first four batches of active learning and random sampling used the eight aryl bromides in the base cores, while the last active learning batch used the expanded cores (marked with a purple star).

retained the original selection ordering by Kriging-believer and therefore reflects the predictions of the model if we had chosen to order and test fewer molecules in this batch. Figure 6 shows clearly that collecting data on new reactions improves the model until iteration ~20, at which point the performance appears to level off. This data suggests running ~20 building blocks on a new core would enable expansion of this initial model to a new core.

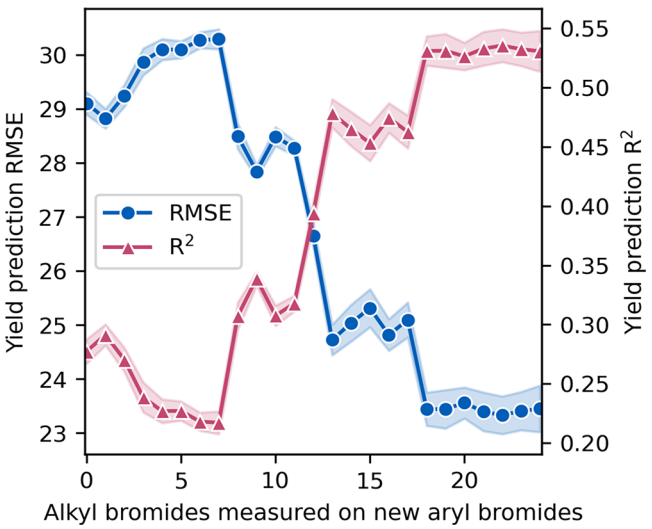


Figure 6. Regression metrics on the model trained on the active learning data set as the Kriging-believer alkyl bromide selections on the expanded core product space are included in the training data one alkyl bromide at a time. The coloration bands depict the standard deviation of the model RMSE when retrained 10 times with different seeds.

One notable difference between active learning and random sampling within this study is that active learning consistently selected alkyl bromides with a higher yield than random sampling. Grouping reaction yield into bins, active learning selected more molecules than random sampling in every >10% yield bin (Figure 7). Despite also having been selected using random sampling, the statistical analysis of the test set indicates that the average yield happened to be higher than random sampling purely by chance, which likely gave a small edge to active learning. Active learning, on the other hand,

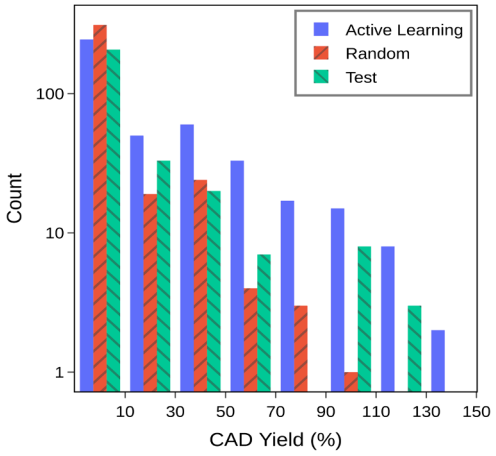


Figure 7. Histogram of experimental reaction CAD yield by each data acquisition method, as well as reactions in the test set. Measured yields in excess of 100% are due to experimental error from CAD limitations and were not modified in order to avoid biasing the model. Note that some bins contain no examples for some categories at higher yields as relatively few reactions had very high CAD yield.

selected alkyl bromides with higher yields by a statistically significant margin (see Supplemental file). We hypothesize this is because active learning selects compounds with the highest uncertainty on an absolute scale, irrespective of how large that uncertainty is relative to the predicted yield, which means the model will select alkyl bromides with predicted yields of 80% ± 20% over 10±10%.

Visualizing the relationship between predicted yield and predicted uncertainty across the entire search space, it is clear that the model has higher predicted uncertainty when the predicted yield is also high (Figure 8). This correlation held even as the Kriging-believer iterations progressed despite the continued selection of alkyl bromides with a high predicted yield. A contributing factor to this trend is that the random forest model uncertainty is computed by the standard deviation of the trees in the ensemble. Low predicted values of yield, such as 1%, are the average of many small values predicted by each tree, which sets a small upper limit on the maximum uncertainty of the predicted yield. As more experimental data was collected in this study, a second contributing factor to this trend arose: the CAD yield measurement has an experimental uncertainty of ~27% relative

https://doi.org/10.1021/jacs.5c02218
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

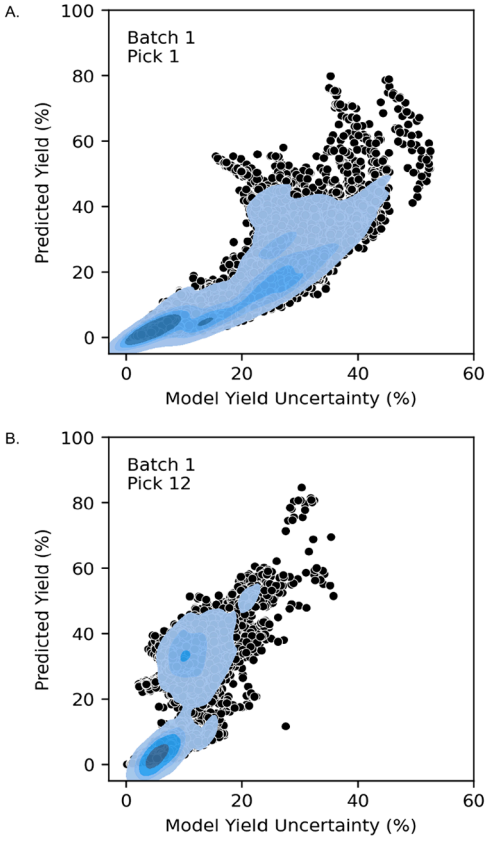


Figure 8. Scatter plot of the model predicted CAD yield and uncertainty across all reactions in the search space during active learning. The overlaying blue surface depicts the density of points to show where predictions are concentrated, with darker blue regions indicate a higher density of points in that area. (A) Model uncertainty and CAD yield during the first selection of the first batch of active learning. (B) Model uncertainty and CAD yield during the 12th selection of the first batch of active learning. Model uncertainty has decreased during Kriging Believer iterations because the predictions for previous selections were used as synthetic training data to increase the diversity of the selections within the batch.

to the value of the measured yield, so an experimental yield of 80% has an uncertainty of about 20%, whereas a measurement of 20% has an uncertainty of about 5%. Although this bias may seem initially undesirable, it is clear that systematically selecting reactions with higher predicted yield led to finding reactions that did have higher experimental CAD yield. Reactions with crude yields greater than 20% by CAD, for example, are of greater practical use to chemists seeking to generate products with downstream applications in an HTE drug discovery screen. This threshold was chosen based on our empirical experience running high-throughput screens for project teams. Typically, in situ yields of roughly 20% translate to successful purification and subsequent entrance of compounds into first-tier biological and ADME assays. However, reactions meeting this threshold appear to be a minority based on the random sampling data. In this context, the propensity of the active learning acquisition method to favor higher CAD yields may have led to generating a training data set with greater practical utility than a data set generated by an acquisition method without such a bias.

Plate Enrichment Study. Although the active learning model performed well in the regression metrics, we wanted to evaluate both models further in our envisioned preliminary application within a drug discovery context: filtering a library of possible reactions for only the ones that would have appreciable yield for new aryl bromides, the model had not been trained on initially.[58] We therefore conducted an experiment where both models were trained only on the base core set data and then used to select alkyl bromides that would have appreciably high yield when reacted with the expanded core set (Figure 9A). To simulate predicting in a novel space, we identified predictions in the expanded product space that diverged significantly in predicted yield between the active learning and random models. For the model trained on active learning data, we first filtered for alkyl bromides with predicted yields >40%, while the model trained on randomly sampled data predicted yields <10% for these same compounds. These threshold values were chosen to maximize the difference in model predictions without narrowing the chemical diversity of candidate alkyl bromides. From a compound selection perspective, the active learning model is designating compounds as synthesizable with appreciable CAD yield while the random model is labeling the compounds as not synthesizable. This constraint yielded 115 alkyl bromides within the search space, which we filtered down to eight using k-means clustering on the model features and selecting the centermost from each cluster (Figure 9B, AL1-AL8). A second set of alkyl bromides was created where the random model predicted a successful reaction and the AL model predicted failure. Because the random model predicts so few high-yielding reactions, the criteria were relaxed for the random model to a yield prediction threshold of >15%, while the model trained on active learning data predicted <10% yield. Even with these relaxed criteria, there were still only 12 alkyl bromides that remained after filtering, of which we selected four to test experimentally because many of the 12 were chemically similar (Figure 9B, R1−R4).

The results of this experiment are outlined in Figure 9B,C. For clarity, the average yield of each alkyl bromide was calculated by averaging the yield of the requisite alkyl bromide reacted with E1−E4 (9B). The yields of each individual reaction are presented as a heatmap in Figure 9C. These results demonstrate that the active learning model succeeded at the task of identifying alkyl bromides with high average yield across the four aryl bromides in the expanded core set: no selections had average yields below 10%, and only two selections had yields between 15 and 18%, while the rest were much higher. At the granularity of individual reactions, 80% of reactions between the active learning model selections had a yield greater than 10%, whereas this metric was only 36% for the random sampling model selections. The active learning model is certainly not perfect: the random sampling model identified two reactions with high yield that the active learning model underestimated, and the active learning model made predictions that were too high for some reactions and too low for others. What is worth keeping in mind during this evaluation, however, is that the majority of reactions in this search space have very low yield. For the expanded core test set, only 23.3% of reactions had yields greater than 10%, whereas for all data randomly sampled for the base core set, only 18.2% had yields greater than 10%. This data strongly suggests that using the active learning model could therefore significantly reduce the number of necessary reactions in high-throughput experimentation within this search space by

https://doi.org/10.1021/jacs.5c02218
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

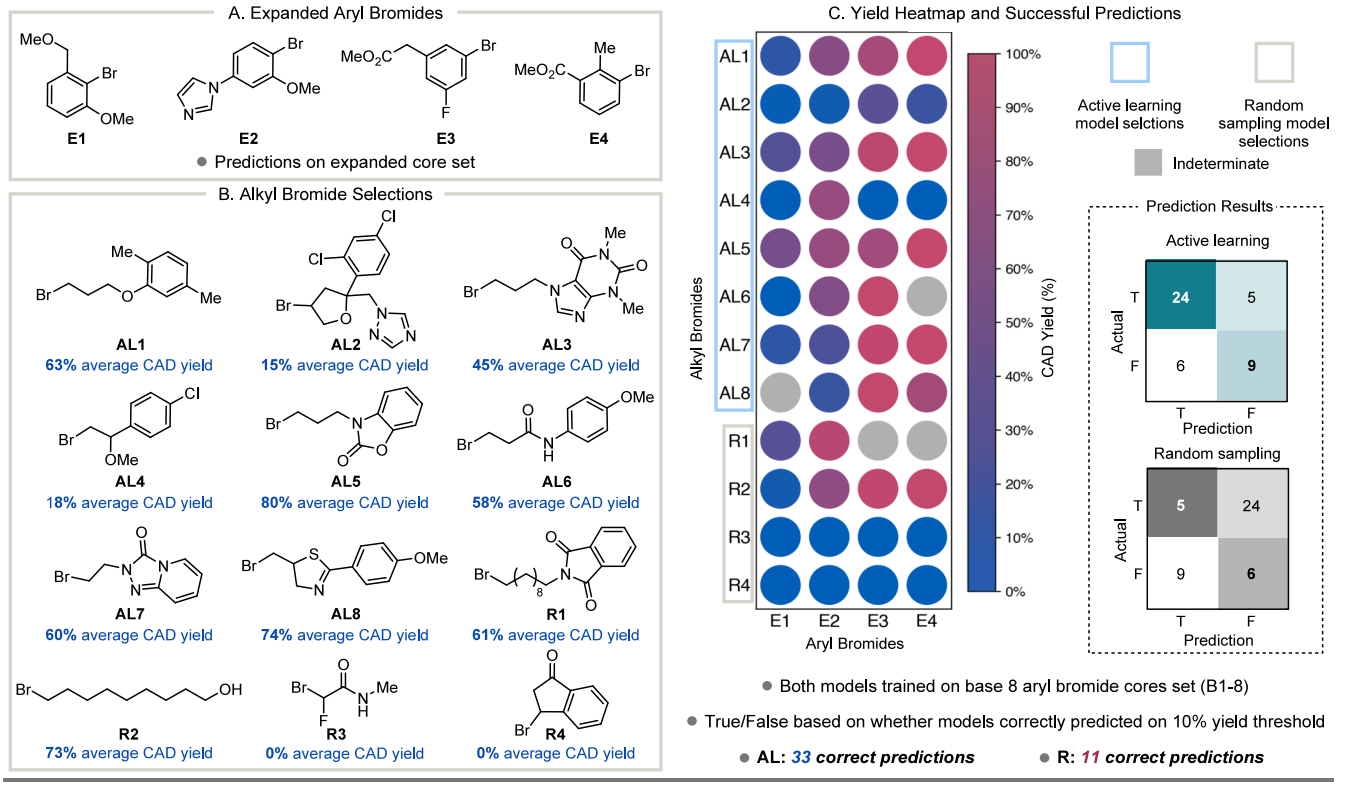


Figure 9. (A) Expanded set of aryl bromides. (B) Alkyl bromides selected for the experiment. AL1-AL8 selected by active learning model for success, R1−R4 selected by random sampling model for success. Average CAD yield represents the average yield of each building block reacted with E1−E4 indicated in blue. These represent the average of the CAD yields presented in 9C. (C) Heat map of individual reaction CAD yields.

prefiltering reactions that are unlikely to produce sufficient yield.

Feature Importance. Prior to data collection, the only way to decide model features and architecture was by intuition. The model architecture remained fixed during active learning iterations to avoid the inclusion of additional noise, but after collecting all the training data, we revisited the model features and hyperparameters to further interrogate feature importance and explore different model architectures. For hyperparameter optimization, we merged the training data from both acquisition methods to form a dataset of 892 training points. We then performed a grid search on a number of trees and the maximum depth of the random forest ensemble, arriving at the best model with 80 trees and a maximum depth of 6 (see Supplemental file).

Using this hyperparameter optimized model architecture, we trained random forest models on several subsets of the training data: the training data from random sampling (random), the active learning training data including only the base core set (AL base set), all active learning data for the base core set and the expanded core set, (AL base set and expanded), and the full set of all training data (AL+random). This time, we evaluated the models at reaction yield thresholds of 10−50%. To reduce noise from model initialization, each model was retrained 50 times and averaged across all replicates (Figure 10A). Although the model trained on random sampling performs relatively well at the 10% threshold, it is again outperformed by the comparable AL base set from active learning. This is notable because the model trained on data collected by active learning selected many fewer reactions below 10% yield in the training data than random sampling, yet the active learning model still performs better at predicting on this threshold. While the set containing all training data performed comparably to the set containing only active learning data on the base core set, it performs notably worse on the expanded core set; the adage that more training data is always better would suggest that nearly doubling the training data would improve the model, but it appears the addition of randomly sampled data from the base core set may hinder the model from learning on the minority of data from the expanded core set. The only metric where including the randomly sampled data appears to improve model performance is on the base core test set at the yield thresholds 10 and 20%, likely because the randomly sampled reactions with the base core set were nearly all low yields. In contrast, adding data from the expanded core set causes a small but consistent improvement in the active learning model across yield thresholds on the base core test set, which could be due to improved model generalization across cores in a similar fashion to the effect of adding the expanded core set data to the active learning model in Figure 5. While evidence of model generalization is encouraging, Figure 10A also shows that including the expanded core set data in the active learning model has a large positive impact on the model performance for the expanded core test set. Model generalization is certainly desirable, but for this reaction model, it has a smaller impact than having training data on at least one of the pair of reactants.

The choice of features is also worth revisiting because the inclusion of DFT features introduces significant computational expense during inference on larger chemical libraries. To this end, we used the AL+random training data set to train models

https://doi.org/10.1021/jacs.5c02218
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

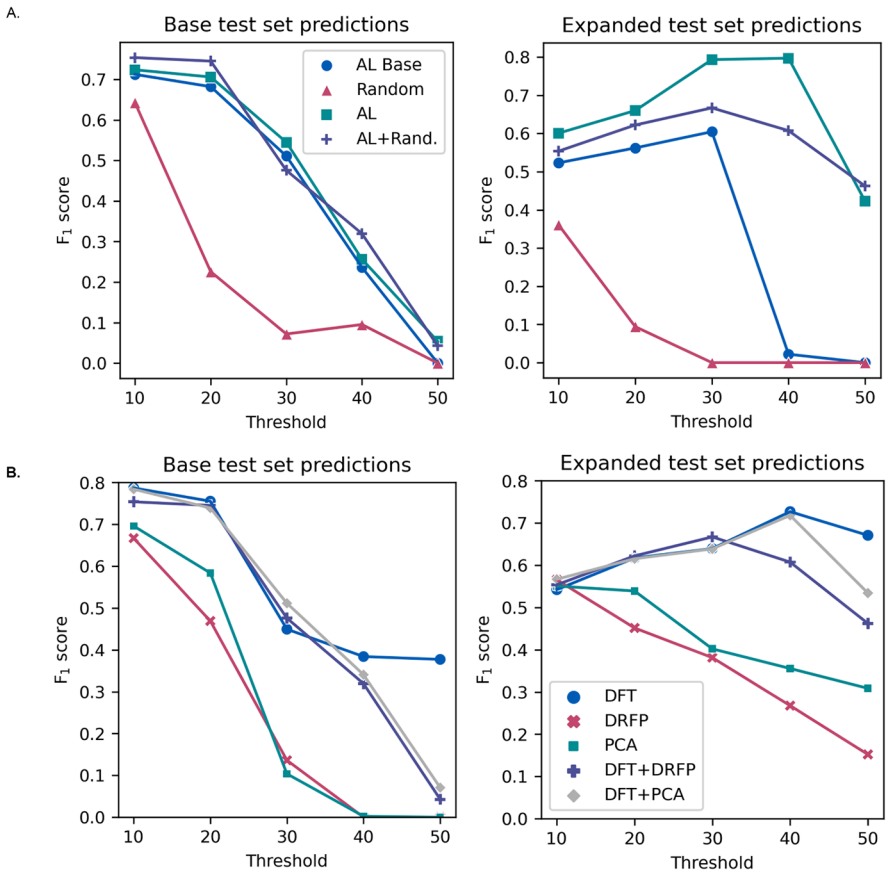


Figure 10. (A) Performance of regression models at predicting reaction yield above or below thresholds between 10 and 50% yield. Each model is trained on a different subset of training data: the label *AL Base* refers to the model trained only on data selected during active learning on the base core set, the label *AL* refers to the model trained on active learning data for both the base core set and the expanded core set. *Random* is trained only on data selected during the random sampling iterations on the base core set, and *AL+Rand.* is trained on all data collected by both active learning and random sampling. (B) Performance of regression models at predicting reaction yield above or below thresholds between 10 and 50% yield. Each model differs by input features: DFT uses only DFT features, DRFP uses only difference reaction fingerprints, DFT+DRFP uses both DFT and difference reaction fingerprints, PCA uses the DRFP features projected onto their largest 15 principal components, and DFT+PCA is a combination of the DFT and PCA features.

using several combinations of input features, two of which were models with only DFT features and only DRFP features (Figure 10B). Among these newly trained models, DFT features performed quite comparably to the full feature space at yield thresholds 10−30%, and better at yield thresholds 40−50%, where training data is scarce. Conversely, DRFP features alone performed worse, especially at yield thresholds of 30 and 40%. This raised the question of whether the 2048-bit fingerprints were too sparse for a training set of 892 reactions, so we evaluated dimensionality reduction of DRFP using principal component analysis (PCA). We conducted a grid search of dimension sizes for the PCA and found the best performance with 15 principal components (see Supplemental file). We then evaluated the performance of PCA features alone and PCA features combined with the DFT features. Overall, the PCA projection of DRFP performs slightly better than the original DRFP, but not nearly as well as the models that include DFT features.

Out of our DFT feature set for this model, the majority of dominant features, including the top 3 in the overall active learning model, correspond to building block features, with two electronic features for the cores appearing with lower importance (Figure 11A). This observation is consistent with the general structure of the data set comprising 12 (8 + 4) aryl bromide cores and ∼60 unique alkyl bromides. The top alkyl bromide features include the energy to remove the Br atom, the LUMO of the alkyl radical, and two charge-based features for the Br and C, respectively. The features correspond both to the alkyl bromide and the resulting radical, consistent with the mechanism of Ni-catalyzed cross-electrophile coupling, where the alkyl bromide is proposed to be activated via halogen abstraction and radical rebound on Ni. The top two features collectively cluster the alkyl bromides into different categories that correspond to radical stability and reactivity, such as allylic/benzylic, alpha to carbonyls, and aliphatic (Figure 11B). As seen in Figure 11C, the active learning selection of alkyl bromides effectively covers this feature space and helped elucidate a hot spot in the top right corner corresponding to high-performing coupling reactions.

Evaluation of Diverse Aryl Cores and Core-Substrate Dependence. To understand the impact of these alkyl bromide features on the reactivity of the individual aryl bromide cores, separate random forest regression models were trained on data from each of the 12 aryl bromide cores. Instead of using the entire data set collected up to this point (which would skew the expanded core set toward the higher-yielding building blocks seen later in the active learning selection process), the subset of building blocks tested on all 12 cores

https://doi.org/10.1021/jacs.5c02218
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

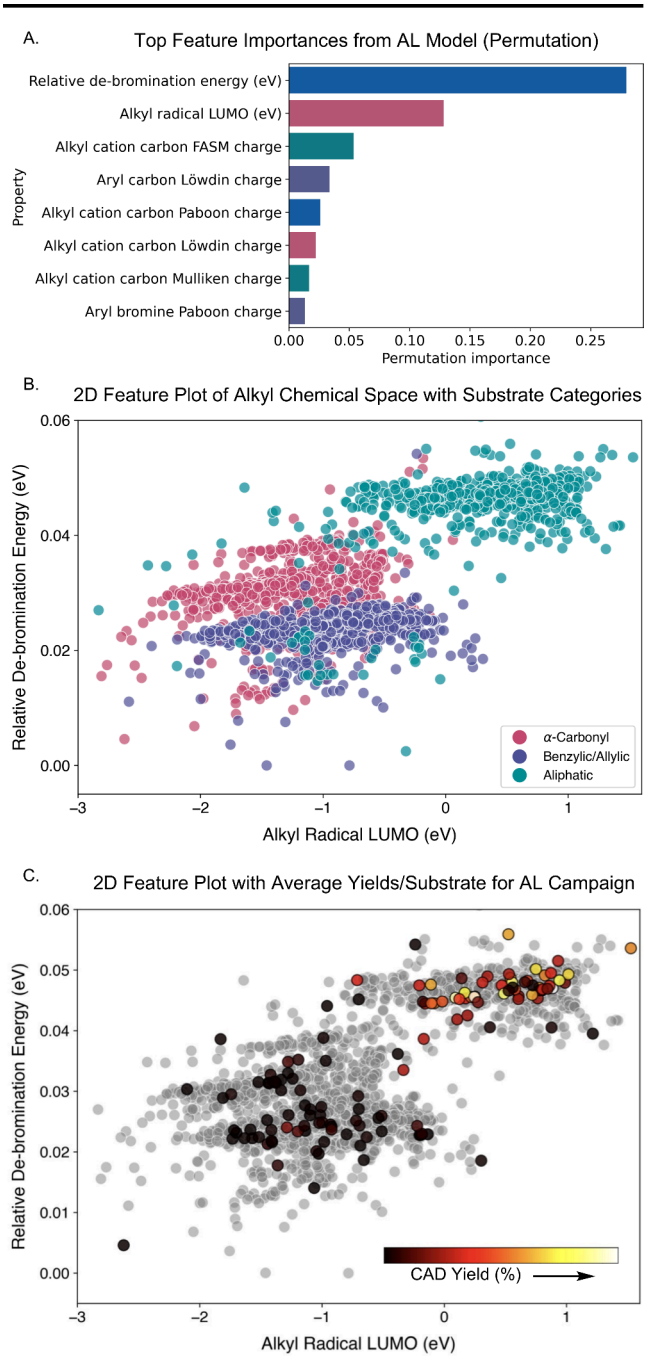


Figure 11. (A) Top feature importance from the active learning model after all rounds of the campaign. Importances measured by permutation importance. Chemical space map for alkyl bromides plotted according to top two features from the active learning model. (C) Chemical space map overlaid with building blocks selected in the AL campaign, colored by average CAD yield per substrate.

were used (~32 per core), primarily from test sets for the models. Additionally, only computed DFT features were used in the training of these models in order to more readily interpret feature importances chemically. For each of these corresponding 12 random forest models, feature importances and model predictions on the other aryl bromide data sets were evaluated in order to determine any major differences between the important alkyl bromide features for a given core.

Figure 12A shows Euclidean distances of feature importance for these models, in comparison to model features. Core E2 stands out in this analysis because it has distinct features from any other model and has the largest difference from the features of the overall model. This divergence makes chemical sense, because Core E2 is an ortho-substituted and electron-rich aryl bromide, which is expected to have distinct reactivity and potentially a different turnover limiting step in the catalytic reaction (Figure 12B). Likewise, core B2 is also very distinct from our expanded core set in terms of features, including a large dissimilarity from E2, consistent with the difference in electronic and steric features surrounding the reactive bromine.

To measure how this difference in feature importances translates to model prediction, we conducted an analysis wherein each core model was tested on another's data (including self-prediction). These data are visualized in the Figure 12C heatmap. Consistent with the discrepancy in DFT features, the model trained on core E2 data had the most difficult time predicting on other cores and had the lowest average R^2 of any model, including our failed core B3. This result demonstrates the significant difference in features correlating to a difference in reactivity. Core E2's *ortho*-substituted and electron-rich nature would correspond to difficult oxidative addition, radical capture, and reductive elimination steps of the catalytic cycle. We observe that this core reacts more effectively with alkyl building blocks possessing higher radical LUMO energies (higher nucleophilicity) compared to other cores, suggesting that the radical capture step may be most impacted. Interestingly, the core E1 data have the most difficult time being predicted by other models. This aryl bromide is unique in that it is the only example of a di-*ortho*-substituted aryl bromide we tested. Although this aryl bromide is extremely low performing on average, likely due to steric hindrance, we do observe several substrates that afford moderate yields for this core. This is notable as the reported aryl bromides from this region of chemical space have classically been low-yielding in Ni-photoredox couplings.[40] The results of our feature-based analysis of selected aryl cores support the diversity of the chemical space from which core selections were made by providing evidence of distinct trends in alkyl bromide reactive features for the cores utilized in this study.

CONCLUSIONS

In conclusion, we have demonstrated that an active learning approach is an effective strategy toward the construction of a generalizable yield model for the Ni/photoredox catalyzed cross-electrophile coupling of aryl bromides and alkyl bromides in an HTE setting. The application of CAD quantitation allowed for semiquantitative yields to be reported for >1300 substrate pairs, providing a unique data set to the chemistry community. Using active learning, we constructed a useful yield model by sampling 1−2% of the substrate space; moreover, we demonstrated the model can be expanded to new chemical space with similar experimental coverage and employed effectively as a screening tool for potential HTE campaigns. Our future work in this area will focus on expanding the strategy to new reaction classes, leveraging mechanistic understanding to improve modeling and data gathering strategies. Furthermore, we will mine this information to identify the dark spaces in synthetic methodologies, so that reaction condition discovery/optimization efforts can be focused on the most critical challenges. We believe mapping of

https://doi.org/10.1021/jacs.5c02218
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

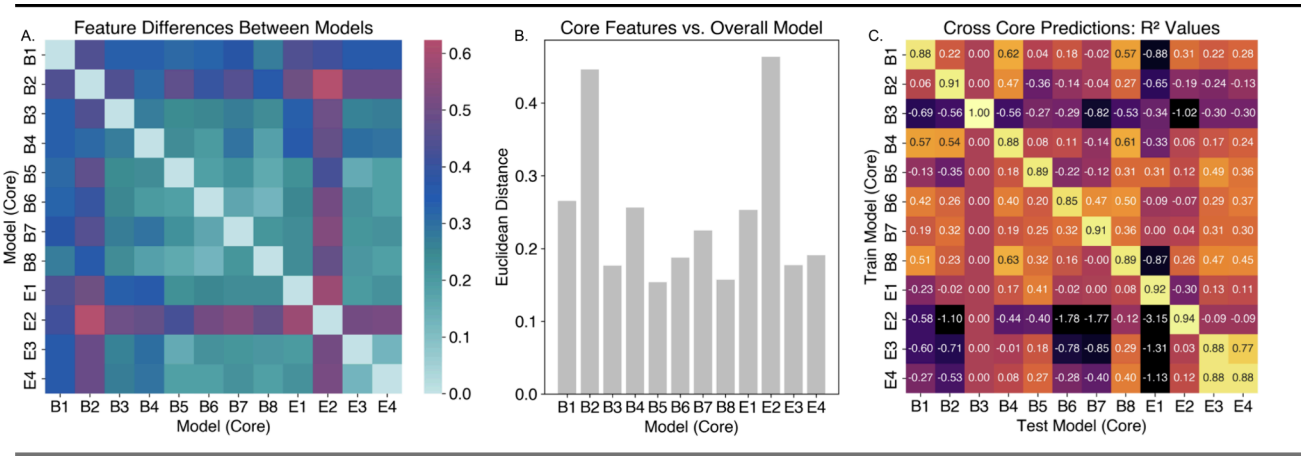


Figure 12. Comparison of models trained on individual core data. Euclidean distance of model feature importance compared across models. (B) Difference in features importance between individual models and model trained on all cores within the subset of shared building blocks (Euclidean distance). (C) Predictions of each trained core model on each other core's data, reported as R^2 values.

reaction space is a key challenge for the chemistry community, which will require continued innovation in the areas of analytics, high-throughput experimentation, data sampling strategies, and machine learning. Moreover, it will require collaborations between academia and industry to reach our highest aspirations.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacs.5c02218.

Final experimental results (XLSX)

Experimental details and characterization data (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Abigail G. Doyle − *Department of Chemistry & Biochemistry, University of California, Los Angeles, California 90095, United States;* orcid.org/0000-0002-6641-0833; Email: agdoyle@chem.ucla.edu

Kian L. Tan − *Global Discovery Chemistry, Novartis, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0001-8243-1223; Email: kian.tan@novartis.com

Authors

Lucas W. Souza − *Global Discovery Chemistry, Novartis, Cambridge, Massachusetts 02139, United States*

Nathan D. Ricke − *Global Discovery Chemistry, Novartis, Cambridge, Massachusetts 02139, United States*

Braden C. Chaffin − *Department of Chemistry & Biochemistry, University of California, Los Angeles, California 90095, United States;* orcid.org/0009-0002-8378-7453

Mike E. Fortunato − *Global Discovery Chemistry, Novartis, Cambridge, Massachusetts 02139, United States*

Shutian Jiang − *Department of Chemistry & Biochemistry, University of California, Los Angeles, California 90095, United States;* Present Address: Department of Chemistry, Princeton University, Princeton, New Jersey 08544, United States

Cihan Soylu − *Global Discovery Chemistry, Novartis, Cambridge, Massachusetts 02139, United States*

Thomas C. Caya − *Global Discovery Chemistry, Novartis, Cambridge, Massachusetts 02139, United States*

Sii Hong Lau − *Global Discovery Chemistry, Novartis, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0002-2178-0187

Katherine A. Wieser − *Global Discovery Chemistry, Novartis, Cambridge, Massachusetts 02139, United States;* Present Address: Department of Radiology, Massachusetts General Hospital, Boston, Massachusetts 02114, United States.

Complete contact information is available at: https://pubs.acs.org/10.1021/jacs.5c02218

Author Contributions

[§]N.D.R. and B.C.C. are contributed equally.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Financial support was provided by NIGMS R35 GM126986. B.C.C. supported by NSF Grant DGE-2034835. L.W.S. would like to acknowledge Dyuti Majumdar for chemistry discussion, Carol Ginsburg-Moraff for analytical contributions, J.J.L. for python support, William Ulmer for automation support, and Dr. Brooke Brown and Stefan Thibodeaux for mass spec consultation.

■ REFERENCES

(1) Jiang, Y.; Yu, Y.; Kong, M.; Mei, Y.; Yuan, L.; Huang, Z.; Kuang, K.; Wang, Z.; Yao, H.; Zou, J.; Coley, C. W.; Wei, Y. Artificial Intelligence for Retrosynthesis Prediction. *Engineering.* **2023**, *25*, 32−50.

(2) Tu, Z.; Stuyver, T.; Coley, C. W. Predictive Chemistry: Machine Learning for Reaction Deployment, Reaction Development, and Reaction Discovery. *Chem. Sci.* **2023**, *14* (2), 226−244.

(3) Fitzner, M.; Wuitschik, G.; Koller, R.; Adam, J. M.; Schindler, T. Machine Learning C-N Couplings: Obstacles for a General-Purpose Reaction Yield Prediction. *ACS Omega* **2023**, *8* (3), 3017−3025.

(4) Schleinitz, J.; Langevin, M.; Smail, Y.; Wehnert, B.; Grimaud, L.; Vuilleumier, R. Machine Learning Yield Prediction from NiCOlit, a

Small-Size Literature Data Set of Nickel Catalyzed C-O Couplings. *J. Am. Chem. Soc.* **2022**, *144* (32), 14722−14730.
(5) Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J. L. Prediction of Chemical Reaction Yields Using Deep Learning. *Mach. Learn. Sci. Technol.* **2021**, *2* (1), No. 015016.
(6) Voinarovska, V.; Kabeshov, M.; Dudenko, D.; Genheden, S.; Tetko, I. V. When Yield Prediction Does Not Yield Prediction: An Overview of the Current Challenges. *Journal of Chemical Information and Modeling.* **2024**, *64* (1), 42−56.
(7) Strieth-Kalthoff, F.; Szymkuć, S.; Molga, K.; Aspuru-Guzik, A.; Glorius, F.; Grzybowski, B. A. Artificial Intelligence for Retrosynthetic Planning Needs Both Data and Expert Knowledge. *J. Am. Chem. Soc.* **2024**, *146* (16), 11005−11017.
(8) Beker, W.; Roszak, R.; Wolos, A.; Angello, N. H.; Rathore, V.; Burke, M. D.; Grzybowski, B. A. Machine Learning May Sometimes Simply Capture Literature Popularity Trends: A Case Study of Heterocyclic Suzuki-Miyaura Coupling. *J. Am. Chem. Soc.* **2022**, *144* (11), 4819−4827.
(9) Saebi, M.; Nan, B.; Herr, J. E.; Wahlers, J.; Guo, Z.; Zurański, A. M.; Kogej, T.; Norrby, P. O.; Doyle, A. G.; Chawla, N. V.; Wiest, O. On the Use of Real-World Datasets for Reaction Yield Prediction. *Chem. Sci.* **2023**, *14*, 4997−5005.
(10) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting Reaction Performance in C-N Cross-Coupling Using Machine Learning. *Science.* **2018**, *360* (6385), 186−190.
(11) Angello, N. H.; Rathore, V.; Beker, W.; Wołos, A.; Jira, E. R.; Roszak, R.; Wu, T. C.; Schroeder, C. M.; Aspuru-Guzik, A.; Grzybowski, B. A.; Burke, M. D. Closed-loop optimization of general reaction conditions for heteroaryl Suzuki-Miyaura coupling. *Science* **2022**, *378* (6618), 399−405.
(12) Reizman, B. J.; Wang, Y. M.; Buchwald, S. L.; Jensen, K. F. Suzuki-Miyaura Cross-Coupling Optimization Enabled by Automated Feedback. *React. Chem. Eng.* **2016**, *1* (6), 658−666.
(13) Rinehart, N. I.; Saunthwal, R. K.; Wellauer, J.; Zahrt, A. F.; Schlemper, L.; Shved, A. S.; Bigler, R.; Fantasia, S.; Denmark, S. E. A Machine-Learning Tool to Predict Substrate-Adaptive Conditions for Pd-Catalyzed C-N Couplings. *Science* **2023**, *381* (6661), 965−972.
(14) Götz, J.; Jackl, M. K.; Jindakun, C.; Marziale, A. N.; André, J.; Gosling, D. J.; Springer, C.; Palmieri, M.; Reck, M.; Luneau, A.; Brocklehurst, C. E.; Bode, J. W. High-Throughput Synthesis Provides Data for Predicting Molecular Properties and Reaction Success. *Sci. Adv.* **2023**, *9* (43), No. eadj2314.
(15) Nielsen, M. K.; Ahneman, D. T.; Riera, O.; Doyle, A. G. Deoxyfluorination with Sulfonyl Fluorides: Navigating Reaction Space with Machine Learning. *J. Am. Chem. Soc.* **2018**, *140* (15), 5004−5008.
(16) Zhong, H.; Liu, Y.; Sun, H.; Liu, Y.; Zhang, R.; Li, B.; Yang, Y.; Huang, Y.; Yang, F.; Mak, F.; Foo, K.; Lin, S.; Yu, T.; Wang, P.; Wang, X. Towards Global Feasibility Prediction and Robustness Estimation of Organic Chemical Reactions with High Throughput Experimentation Data and Bayesian Deep Learning. *ChemRxiv* **2024**, gt72l.
(17) King-Smith, E.; Berritt, S.; Bernier, L.; Hou, X.; Klug-McLeod, J. L.; Mustakis, J.; Sach, N. W.; Tucker, J. W.; Yang, Q.; Howard, R. M.; Lee, A. A. Probing the Chemical 'Reactome' with High-Throughput Experimentation Data. *Nat. Chem.* **2024**, *16* (4), 633−643.
(18) Xu, Y.; Gao, Y.; Su, L.; Wu, H.; Tian, H.; Zeng, M.; Xu, C.; Zhu, X.; Liao, K. High-Throughput Experimentation and Machine Learning-Assisted Optimization of Iridium-Catalyzed Cross-Dimerization of Sulfoxonium Ylides. *Angew. Chem. Int. Ed* **2023**, *62* (48), No. e202313638.
(19) Chen, H.; Mo, Y. Accelerated Electrosynthesis Development Enabled by High-Throughput Experimentation. *Synthesis* **2023**, *55* (18), 2817−2832.
(20) Gao, Y.; Hu, K.; Rao, J.; Zhu, Q.; Liao, K. Artificial Intelligence-Driven Development of Nickel-Catalyzed Enantioselective Cross-Coupling Reactions. *ACS. Catal.* **2024**, *14* (24), 18457−18468.
(21) Lin, A.; Liu, J.; Xu, Y.; Wu, H.; Chen, Y.; Zhang, Y.; Su, L.; Zhao, X.; Liao, K. High-Throughput Experimentation and Machine Learning-Promoted Synthesis of α-Phosphoryloxy Ketones via Ru-Catalyzed P(O)O-H Insertion Reactions of Sulfoxonium Ylides. *Sci. China. Chem.* **2025**, *68*, 679.
(22) Sather, A. C.; Martinot, T. A. Data-Rich Experimentation Enables Palladium-Catalyzed Couplings of Piperidines and Five-Membered (Hetero)Aromatic Electrophiles. *Org. Proces.s Res. Dev.* **2019**, *23* (8), 1725−1739.
(23) Samha, M. H.; Karas, L. J.; Vogt, D. B.; Odogwu, E. C.; Elward, J.; Crawford, J. M.; Steves, J. E.; Sigman, M. S. Predicting Success in Cu-Catalyzed C-N Coupling Reactions Using Data Science. *Sci. Adv.* **2024**, *10* (3), No. eadn3478.
(24) Raghavan, P.; Rago, A. J.; Verma, P.; Hassan, M. M.; Goshu, G. M.; Dombrowski, A. W.; Pandey, A.; Coley, C. W.; Wang, Y. Incorporating Synthetic Accessibility in Drug Design: Predicting Reaction Yields of Suzuki Cross-Couplings by Leveraging AbbVie's 15-Year Parallel Library Data Set. *J. Am. Chem. Soc.* **2024**, *146* (22), 15070−15084.
(25) Gandhi, S. S.; Brown, G. Z.; Aikonen, S.; Compton, J. S.; Neves, P.; Martinez Alvarado, J. I.; Strambeanu, I. I.; Leonard, K. A.; Doyle, A. G. Data Science-Driven Discovery of Optimal Conditions and a Condition-Selection Model for the Chan−Lam Coupling of Primary Sulfonamides. *ACS. Catal.* **2025**, *15*, 2292−2304.
(26) Kariofillis, S. K.; Jiang, S.; Żurański, A. M.; Gandhi, S. S.; Martinez Alvarado, J. I.; Doyle, A. G. Using Data Science To Guide Aryl Bromide Substrate Scope Analysis in a Ni/Photoredox-Catalyzed Cross-Coupling with Acetals as Alcohol-Derived Radical Sources. *J. Am. Chem. Soc.* **2022**, *144* (2), 1045−1055.
(27) Sigman, M. S.; Harper, K. C.; Bess, E. N.; Milo, A. The Development of Multidimensional Analysis Tools for Asymmetric Catalysis and Beyond. *Acc. Chem. Res.* **2016**, *49* (6), 1292−1301.
(28) Rana, D.; Pflüger, P. M.; Hölter, N. P.; Tan, G.; Glorius, F. Standardizing Substrate Selection: A Strategy toward Unbiased Evaluation of Reaction Generality. *ACS. Cent Sci.* **2024**, *10* (4), 899−906.
(29) Shim, E.; Tewari, A.; Cernak, T.; Zimmerman, P. M. Machine Learning Strategies for Reaction Development: Toward the Low-Data Limit. *J. Chem. Inf. Model.* **2023**, *63* (12), 3659−3668.
(30) Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Alvarado, J. I. M.; Janey, J. M.; Adams, R. P.; Doyle, A. G. Bayesian Reaction Optimization as a Tool for Chemical Synthesis. *Nature.* **2021**, *590* (7844), 89−96.
(31) Hickman, R. J.; Aldeghi, M.; Häse, F.; Aspuru-Guzik, A. Bayesian Optimization with Known Experimental and Design Constraints for Chemistry Applications. *Digital Discovery* **2022**, *1* (5), 732−744.
(32) Smith, R. T.; Zhang, X.; Rincón, J. A.; Agejas, J.; Mateos, C.; Barberis, M.; García-Cerrada, S.; De Frutos, O.; Macmillan, D. W. C. Metallaphotoredox-Catalyzed Cross-Electrophile C Sp3 -C Sp3 Coupling of Aliphatic Bromides. *J. Am. Chem. Soc.* **2018**, *140* (50), 17433−17438.
(33) Gesmundo, N. J.; Rago, A. J.; Young, J. M.; Keess, S.; Wang, Y. At the Speed of Light: The Systematic Implementation of Photoredox Cross-Coupling Reactions for Medicinal Chemistry Research. *J. Org. Chem.* **2024**, *89* (22), 16070−16092.
(34) Dombrowski, A. W.; Gesmundo, N. J.; Aguirre, A. L.; Sarris, K. A.; Young, J. M.; Bogdan, A. R.; Martin, M. C.; Gedeon, S.; Wang, Y. Expanding the Medicinal Chemist Toolbox: Comparing Seven C(Sp2)-C(Sp3) Cross-Coupling Methods by Library Synthesis. *ACS Med. Chem. Lett.t* **2020**, *11* (4), 597−604.
(35) Chan, A. Y.; Perry, I. B.; Bissonnette, N. B.; Buksh, B. F.; Edwards, G. A.; Frye, L. I.; Garry, O. L.; Lavagnino, M. N.; Li, B. X.; Liang, Y.; Mao, E.; Millet, A.; Oakley, J. V.; Reed, N. L.; Sakai, H. A.; Seath, C. P.; MacMillan, D. W. C. Metallaphotoredox: The Merger of Photoredox and Transition Metal Catalysis. *Chem. Rev.* **2022**, *122* (2), 1485−1542.
(36) Lovering, F.; Bikker, J.; Humblet, C. Escape from Flatland: Increasing Saturation as an Approach to Improving Clinical Success. *J. Med. Chem.* **2009**, *52* (21), 6752−6756.

https://doi.org/10.1021/jacs.5c02218
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

(37) Schleinitz, J.; Carretero-Cerdán, A.; Gurajapu, A.; Harnik, Y.; Lee, G.; Pandey, A.; Milo, A.; Reisman, S. Tailoring Datasets for Regioselectivity Predictions on Complex Substrates. *ChemRxiv* **2024**, skgxb.
(38) Shim, E.; Kammeraad, J. A.; Xu, Z.; Tewari, A.; Cernak, T.; Zimmerman, P. M. Predicting Reaction Conditions from Limited Data through Active Transfer Learning. *Chem. Sci.* **2022**, *13* (22), 6655−6668.
(39) Viet Johansson, S.; Gummesson Svensson, H.; Bjerrum, E.; Schliep, A.; Haghir Chehreghani, M.; Tyrchan, C.; Engkvist, O. Using Active Learning to Develop Machine Learning Models for Reaction Yield Prediction. *Mol. Inform* **2022**, *41* (12), No. 2200043.
(40) Kariofillis, S. K.; Jiang, S.; Żurański, A. M.; Gandhi, S. S.; Martinez Alvarado, J. I.; Doyle, A. G. Using Data Science To Guide Aryl Bromide Substrate Scope Analysis in a Ni/Photoredox-Catalyzed Cross-Coupling with Acetals as Alcohol-Derived Radical Sources. *J. Am. Chem. So.c* **2022**, *144* (2), 1045−1055.
(41) Żurański, A. M.; Wang, J. Y.; Shields, B. J.; Doyle, A. G. Auto-QChem: An Automated Workflow for the Generation and Storage of DFT Calculations for Organic Molecules. *React. Chem. Eng.* **2022**, *7*, 1276−1284.
(42) CInnes, L.; Healy, J.; Saul, N.; Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Software* **2018**, *3* (29), 861.
(43) Brocklehurst, C. E.; Altmann, E.; Bon, C.; Davis, H.; Dunstan, D.; Ertl, P.; Ginsburg-Moraff, C.; Grob, J.; Gosling, D. J.; Lapointe, G.; Marziale, A. N.; Mues, H.; Palmieri, M.; Racine, S.; Robinson, R. I.; Springer, C.; Tan, K.; Ulmer, W.; Wyler, R. MicroCycle: An Integrated and Automated Platform to Accelerate Drug Discovery. *J. Med. Chem.* **2024**, *67* (3), 2118−2128.
(44) Probst, D.; Schwaller, P.; Reymond, J. L. Reaction Classification and Yield Prediction Using the Differential Reaction Fingerprint DRFP. *Digital Discovery* **2022**, *1* (2), 91−97.
(45) Landrum, G. *RDKit: Open-Source Cheminformatics*, 2006. https://www.rdkit.org (accessed February 24, 2023).
(46) Casey, S.; Perry, J. W.; Publishing Corp New York, Y. R. N.; Morgan, H. L. *Tools for Machine Literature Searching*; Interscience Publishers, Inc.: New York. N. Y, 1964; Vol. *4*. https://pubs.acs.org/sharingguidelines.
(47) Labute, P. LowModeMD - Implicit Low-Mode Velocity Filtering Applied to Conformational Search of Macrocycles and Protein Loops. *J. Chem. Inf. Model.* **2010**, *50* (5), 792−800.
(48) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-XTB - An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory. Compu.t* **2019**, *15* (3), 1652−1671.
(49) Balasubramani, S. G.; Chen, G. P.; Coriani, S.; Diedenhofen, M.; Frank, M. S.; Franzke, Y. J.; Furche, F.; Grotjahn, R.; Harding, M. E.; Hättig, C.; Hellweg, A.; Helmich-Paris, B.; Holzer, C.; Huniar, U.; Kaupp, M.; Khah, A. M.; Khani, S. K.; Müller, T.; Mack, F.; Nguyen, B. D.; Parker, S. M.; Perlt, E.; Rappoport, D.; Reiter, K.; Roy, S.; Rückert, M.; Schmitz, G.; Sierka, M.; Tapavicza, E.; Tew, D. P.; Wüllen, C. Van; Voora, V. K.; Weigend, F.; Wodyński, A.; Yu, J. M. TURBOMOLE: Modular Program Suite for Ab Initio Quantum-Chemical and Condensed-Matter Simulations. *J. Chem. Phys.* **2020**, *152* (18), 184107.
(50) Becke, A. D. Density-functional Thermochemistry. I. The Effect of the Exchange-only Gradient Correction. *J. Chem. Phys.* **1992**, *96* (3), 2155−2160.
(51) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Phys. Rev. B* **1988**, *37* (2), 785−789.
(52) Vosko, S. H.; Wilk, L.; Nusair, M. Accurate Spin-Dependent Electron Liquid Correlation Energies for Local Spin Density Calculations: A Critical Analysis. *Can. J. Phys.* **1980**, *58*, 1200−1211.
(53) Weigend, F.; Ahlrichs, R. Balanced Basis Sets of Split Valence, Triple Zeta Valence and Quadruple Zeta Valence Quality for H to Rn: Design and Assessment of Accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7* (18), 3297−3305.
(54) Klamt, A.; Schuurmann, G. COSMO: A New Approach to Dielectric Screening in Solvents with Explicit Expressions for the Screening Energy and Its Gradient. *J. Chem. Soc., Perkin Trans.* **1993**, *2*, 799−805.
(55) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(56) Ho, T. K. Random Decision Forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1995; Vol. *1*, pp 278−282.
(57) Ginsbourger, D.; Le Riche, R.; Carraro, L. Kriging Is Well-Suited to Parallelize Optimization. In Tenne, Y.; Goh, C. K., Eds., *Computational Intelligence in Expensive Optimization Problems. Adaptation Learning and Optimization*; Springer: Berlin, Heidelberg, 2010; Vol. *2*.
(58) Hughes, J. P.; Rees, S. S.; Kalindjian, S. B.; Philpott, K. L. Principles of Early Drug Discovery. *Br. J. Pharmacol.* **2011**, *162* (6), 1239−1249.

https://doi.org/10.1021/jacs.5c02218
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX