Electronic Supplementary Material (ESI) for Reaction Chemistry & Engineering.

Iterative experimental design based on active machine learning reduces the experimental burden associated with reaction screening

Natalie S. Eyke, William H. Green, Klavs F. Jensen

Supporting Information (SI)

1. Data

a. Domain span

The 3-bromopyridine dataset consists of a diverse set of 1536 coupling reactions between 3-bromopyridine and a variety of different N/C/O/P/S nucleophiles in the presence of 96 different Pd precatalyst/base combinations. Each electrophile-nucleophile pair generates a unique product. The reactions were conducted at nanoscale in a well plate, at room temperature in DMSO [1].

The Suzuki dataset consists of 5760 Suzuki coupling reactions between a substituted quinoline and a substituted indazole in the presence of 384 different ligand/base/solvent combinations. The leaving groups on the quinoline and indazole are varied as well, but the cores of the coupling partners remain constant across the dataset. Therefore, in contrast to the 3-bromopyridine data, each reaction in the Suzuki dataset is not only a member of the same reaction class, but is fundamentally the same reaction and is designed to generate the same product. These reactions were conducted at nanoscale in flow, at 100°C with a residence time of one minute, with $Pd(OAc)_2$ as the catalyst [2].

The Suzuki data spans a narrower chemical space with a larger number of experiments than the 3-bromopyridine data. This makes the Suzuki data an easier modeling problem: any one of the reactions in the dataset is more useful for modeling many of the others in the set than in the 3-bromopyridine case. As a result, the Suzuki data can be modeled quite accurately, even with a small number of data points that are selected at random from the set. This allows random learning to perform quite well on the Suzuki data and helps explain why active learning doesn't begin to outperform random learning on this dataset until 1000 or so data points have been observed (Figure 4b).

To make this difference between the datasets more concrete, we computed the average Tanimoto similarity of each reaction in each dataset to the rest of the reactions in the (Figure S1); statistical details are tabulated in Table S1. These results confirm that the 3-bromopyridine reactions are substantially less similar to one another than the Suzuki reactions are to one another.

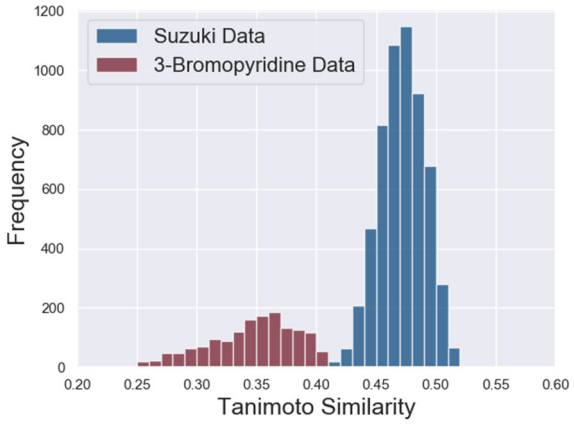

Figure S1. Distributions of average Tanimoto similarities of each reaction to all other reactions in the corresponding dataset.

Table S1. In-dataset Tanimoto similarity statistics.

Dataset	Average Similarity	Similarity Std. Dev.
3-Bromopyridine data	0.346	0.037
Suzuki data	0.471	0.019

b. Relationship between label distribution and experimental error

The experimental noise associated with a dataset places a lower bound on the test set error that can be achieved when models are trained using that data. In a separate experiment that was reported as part of the same article that describes the experimental platform used to generate the 3-bromopyridine data, the authors performed a series of 96 reactions in triplicate [1]. The results are reported as ratios of LC area counts of the desired product to that of an internal standard ("Pd/IS"). Using this data, we plotted the standard deviation in the Pd/IS ratio versus the average value of the Pd/IS ratio computed across the three replicate experiments for each reaction (Figure S2b,c). A weak linear positive correlation between the standard deviation (a proxy for the experimental error) and the average values is observed. All of the reactions for which no product was generated had an experimental error of zero; reactions that produced a nonzero amount of product had nonzero experimental error.

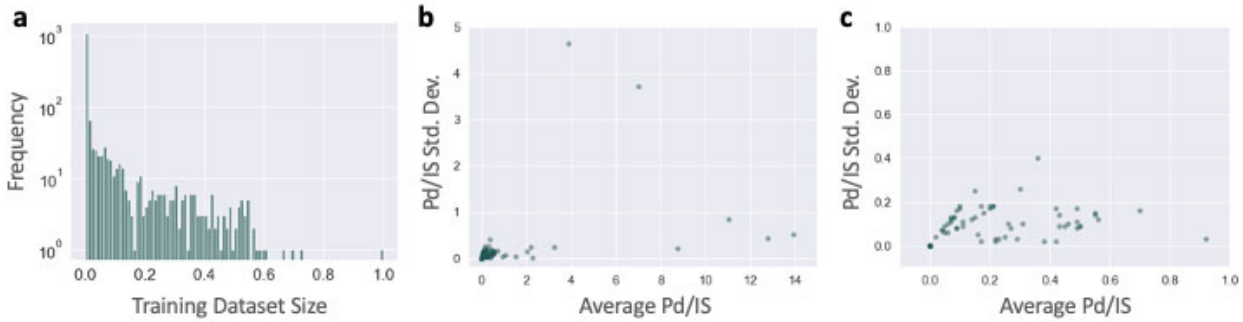

Figure S2. (a) 3-bromopyridine label distribution with bin widths of 0.01. (b) Relationship between the standard deviation in the Pd/IS area counts ratio and the average Pd/IS area counts ratio on the experimental platform used to generate the 3-bromopyridine data. (c) Same as (b), zoomed into the region near the origin.

In over 60% of the reactions in the 3-bromopyridine dataset, no product was generated (Figure S2a). This information, combined with the information gleaned from the reproducibility experiment, suggests that the average experimental error in the 3-bromopyridine dataset is very small. The small average error helps explain why active learning is able to drive to very low test set errors when applied to the 3-bromopyridine dataset.

2. Approach: input featurization, model architecture & training

The two datasets that were retrospectively analyzed in this work were generated to assess the influence of a selection of discrete reaction variables, specifically the identities of the reactants and reagents involved, on the productivity of reactions. For the purposes of machine learning, Morgan fingerprints were used to represent each of these molecules. RDKit [3] was used to generate these fingerprints from SMILES strings with a fingerprint radius of 2, which is a popular radius choice [4,5] that performed well in our early experiments, and a useFeatures setting of False.

The fingerprints of each varied molecule in a particular reaction were concatenated to produce the input representations that were fed to dense feedforward neural networks. This architecture is diagrammed in Figure S3. The reaction molecules that were varied in the 3-bromopyridine data are the nucleophile, the

precatalyst ligand, and the solvent. The molecules that were varied to generate the Suzuki dataset are the nucleophile, electrophile, ligand, solvent, and base. Continuous reaction variables (temperature, residence time, etc.) were not varied in either study and were therefore not included in the reaction featurization.

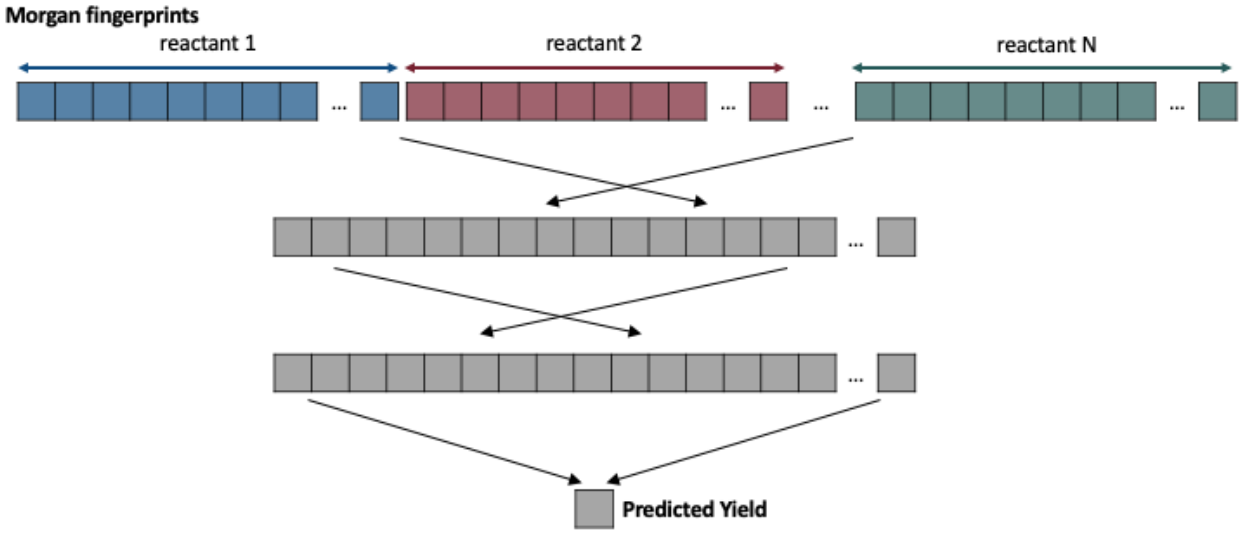

Figure S3. Neural network model architecture used for active learning.

The neural networks used for this study were implemented in PyTorch [6]. Both models were trained for a maximum of 100 epochs with early stopping. Hyperparameters were optimized using a random search. The performance of each set of hyperparameters was evaluated using a single random 80/10/10 training/validation/test split of the data. The variables examined in the hyperparameter search along with the ranges tested and settings selected for the two models are given in Table S2.

Table S2. Hyperparameter optimization results.

Network Parameter	Range Tested	3-Bromopyridine Result	Suzuki Result
Morgan Fingerprint length	128 – 2048	512	512
Number of layers	2 – 4	3	2
Hidden size	50 – 500	200	250
Layer sizes (fraction of hidden)	0.1 – 0.9	(1.0, 0.7, 0.4)	(1.0, 0.5)
Batch size	5 – 500	10	10
Patience	2 – 10	5	5
Dropout fraction	0.2 – 0.8	0.5	0.2

3. Results

a. 3-bromopyridine dataset: preferential selection of high-yielding reactions

In order to better understand how the active learning algorithm performed when operating on the 3-bromopyridine data, we plotted the test set targets (normalized LC area counts) versus predicted values (Figure S4). In this case, the algorithm preferentially selects the rare, high-yielding reactions for addition to the training data. Once only a few hundred reactions remain in the test set, all of those remaining reactions have very, very low yields (see the bottom left corner of Figure S4c). This aligns with the conclusions drawn from the progression in training/validation set target value distributions (Figure 5a-d).

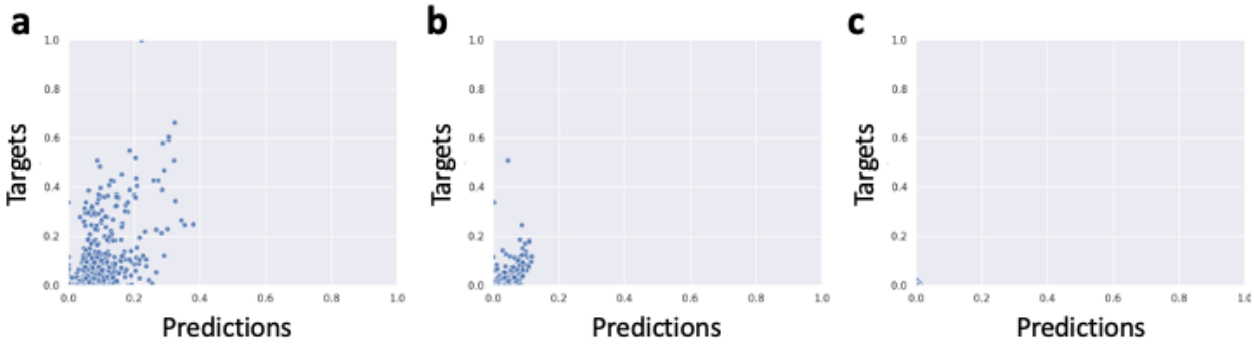

Figure S4. Active learning evolution of 3-bromopyridine test set target versus prediction plots. Units are linearly-normalized HPLC area counts ratios of product to internal standard. Number of reactions remaining in the test set: (a) 1336; (b) 1036; (c) 336.

b. Analysis of uncertainty estimation techniques

One approach to analyzing the quality of uncertainty estimates begins with the assumption that the model predictions are Gaussian-distributed. We examined the prediction distributions produced by both the ensembling technique and MC dropout, see Figure S5 for an example. Both uncertainty estimation techniques produce predictions that are roughly Gaussian-shaped. This suggests that it is appropriate to apply the features of Gaussian distributions to assess at a high level whether the uncertainty estimates are "correct." In Figure S5, the narrower spread of the predictions made by the ensembling approach is reflective of the broader trend.

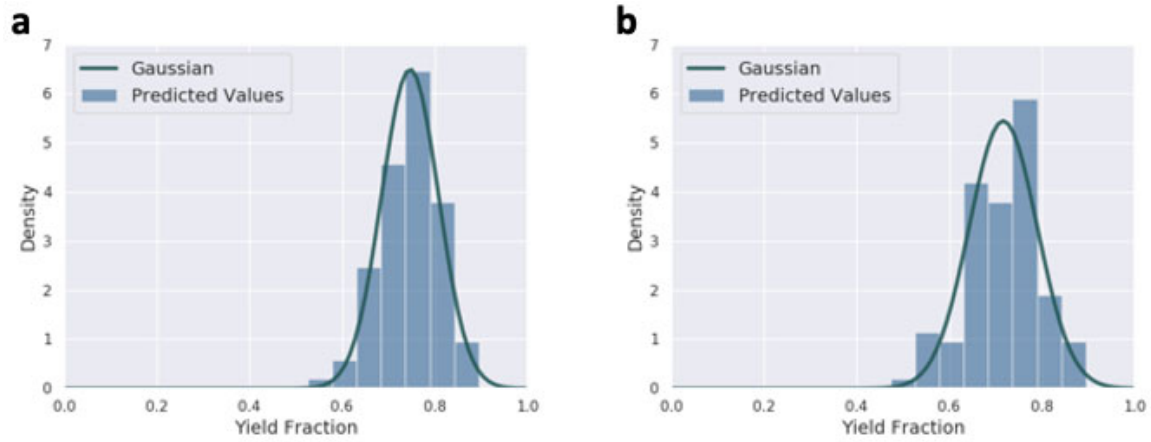

Figure S5. Prediction distributions for a randomly-selected reaction from the Suzuki dataset. (a) Ensembles; (b) MC dropout.

We analyzed how frequently both of the uncertainty estimation techniques managed to capture the true yield of a test set reaction within two standard deviations of the reaction's predicted yield. A comparison of a case when a 100-member ensemble is used to when 100 dropout masks are applied to a single trained model is shown in Figure S6. Neither technique reaches the expected value of 95%, even when the models are trained on nearly all of the data – as is often found, the uncertainty estimates are optimistic. The uncertainty estimates produced with the ensembles technique are slightly more "accurate" on average than those produced with the MC dropout technique, which is consistent with the better performance of uncertainty sampling with ensembles relative to MC dropout, although we emphasize that the most important feature of the uncertainty estimation techniques with regard to active learning performance is how they *rank* the uncertainties – in other words, we should be cautious about overinterpreting the

uncertainty estimates. Both trajectories show a slight upward trend in the quality of the uncertainty estimates with increasing dataset size, which aligns with intuition: more information about the domain leads to better estimates of the quality of our predictions.

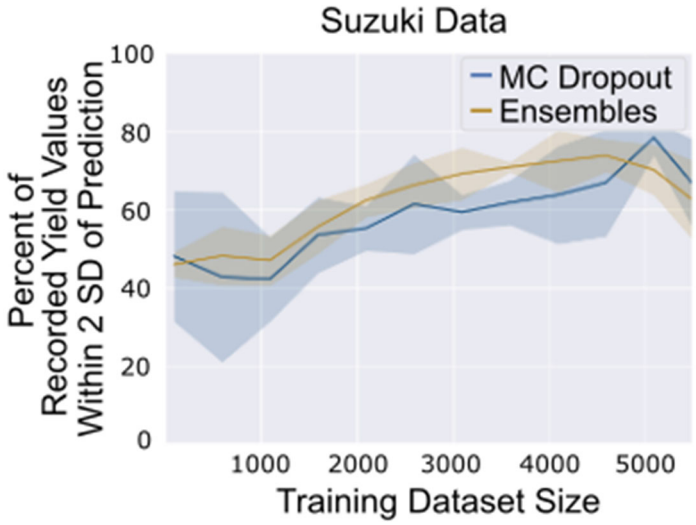

Figure S6. Frequency with which uncertainty estimates capture the true yield of a reaction within two standard deviations of the predicted value, assessed using ensembles and MC dropout masks (100 committee members in both cases, Suzuki data).

Although the MC dropout uncertainty estimation strategy performs slightly worse than ensembles, it is substantially less computationally expensive. Therefore, we sought to understand the influence of the number of dropout masks on how frequently the resulting standard deviation in the prediction effectively captured the distance between the prediction and the true yield, with a specific interest in whether increasing the number of masks would allow us to meet or exceed the performance achieved with ensembles consisting of 100 members. The quality of the uncertainty estimates produced as the number of dropout masks is varied is shown in Figure S7. We found that committees of ten masks yield better uncertainty estimates than committees of two masks, but increasing the number of masks further (to 100 and to 1000) does not further improve the quality of the estimates.

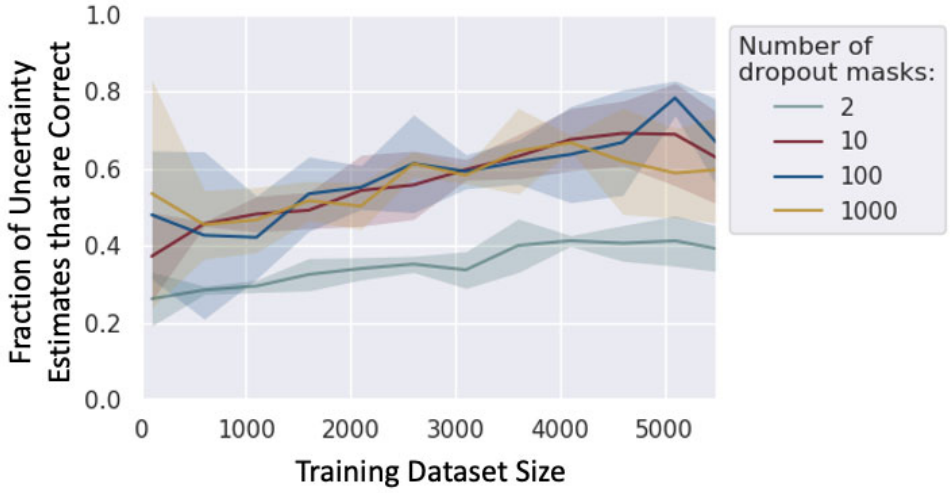

Figure S7. Frequency with which uncertainty estimates capture the true yield of a reaction within two standard deviations of the predicted value, assessed using various quantities of dropout masks, applied to the Suzuki data.

c. Random forest uncertainty sampling

We also analyzed the use of random forests for uncertainty sampling on the Suzuki data. We used SciKit Learn's RandomForestRegressor for this study. The variance in predictions across the trees was used as the uncertainty estimate. Following optimization of the random forest hyperparameters using a random search, we compared two random forest models, one with little active regularization to prevent the overfitting that is common to the decision trees that comprise the random forests, and one with greater regularization achieved by changing the the SciKit Learn RandomForestRegressor parameter "min_samples_leaf" from 1 to 10 (Figure S8a).

The less-regularized random forest-based uncertainty sampling technique outperforms random learning. It performs similarly to MC dropout over most of the dataset, but then outperforms both ensembles and MC dropout for $n > 4500$. The regularized random forest model performs worse.

Upon performing multiple ($n = 3$) repeat runs of the MC dropout, ensembles, and less-regularized random forest techniques, we also noticed that the random forest trajectory exhibited high variance and instability compared to MC dropout and ensembles for $n < 1000$ (Figure S8b).

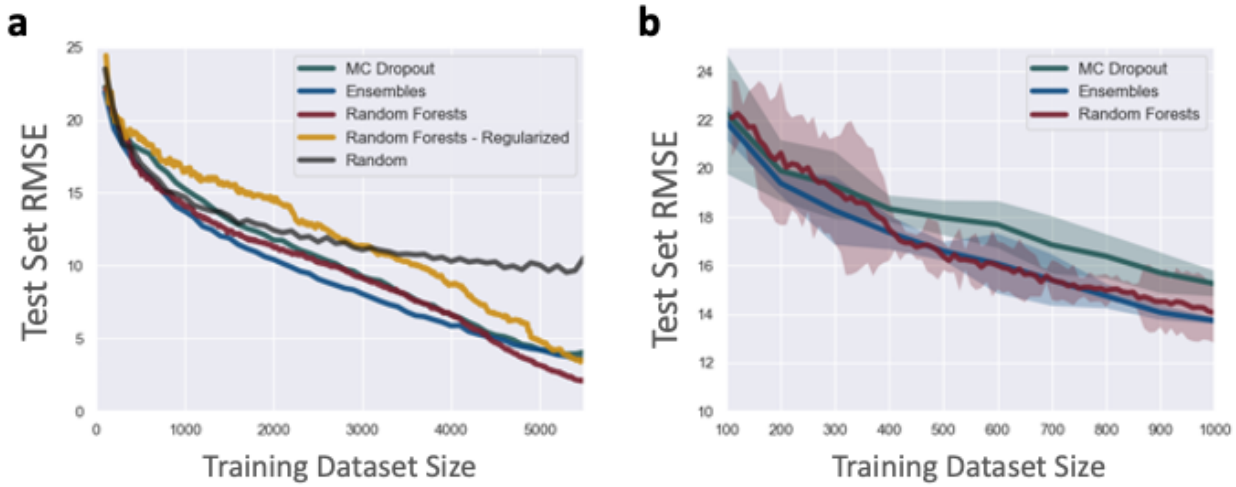

Figure S8. Comparison between random forest uncertainty sampling and neural network uncertainty sampling (with ensembles and MC dropout uncertainty estimation). (a) Full trajectory. (b) Trajectory for $n \leq 1000$.

d. Similarity-augmented sampling

When applied to the Suzuki data, active learning does not begin to outperform random learning until 1000 reactions have been added to the training set (Figure 4b). We hypothesized that this phenomenon was a result of the uncertainty rankings being insufficiently well-calibrated at this early stage. To help address this problem, we tried augmenting the selection metric with a notion of the similarity between a candidate reaction and the data already contained in the training set.

Specifically, we ranked the candidate reactions in order of decreasing uncertainty as usual, and discarded candidate reactions if a measure of their similarity to the existing training set exceeded some threshold. We studied two different similarity metrics: Tanimoto similarity and a metric based on the Euclidean distance between latent-space representations derived from the second-to-last layer of the trained neural network. We analyzed the effect of imposing cutoffs on the maximum, minimum, and average similarity values computed using these strategies. None of the combinations of similarity metric and cutoff resulted in better performance than random learning for $n < 1000$ (see Figure S9 for an example with an average Tanimoto similarity cutoff of 0.49).

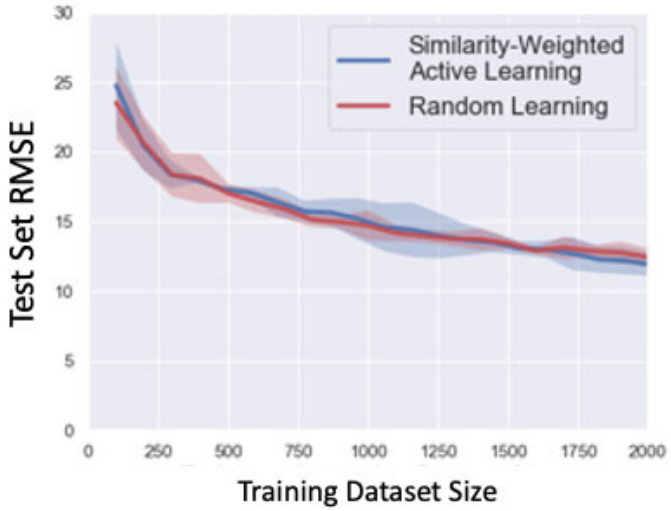

Figure S9. Comparison between random learning and active learning with an average Tanimoto similarity cutoff of 0.49.

e. Label distribution analysis

To better understand the influence that the distribution of labels within a dataset may exert on active learning performance, we subsampled the Suzuki data to create datasets with skewed label distributions. We created these datasets by dividing the Suzuki data into ten bins based on yield, and then removing reactions at random from each bin so as to create an exponentially decaying yield distribution. We created two versions of this augmented dataset: the first, which we call Suzuki Skewed, contains 1600 data points (roughly as many data points as in the 3-bromopyridine dataset), and the second ("Suzuki Very Skewed") contains 950 data points, and contains even fewer reactions with yield greater than 10% than the first (Figure S10).

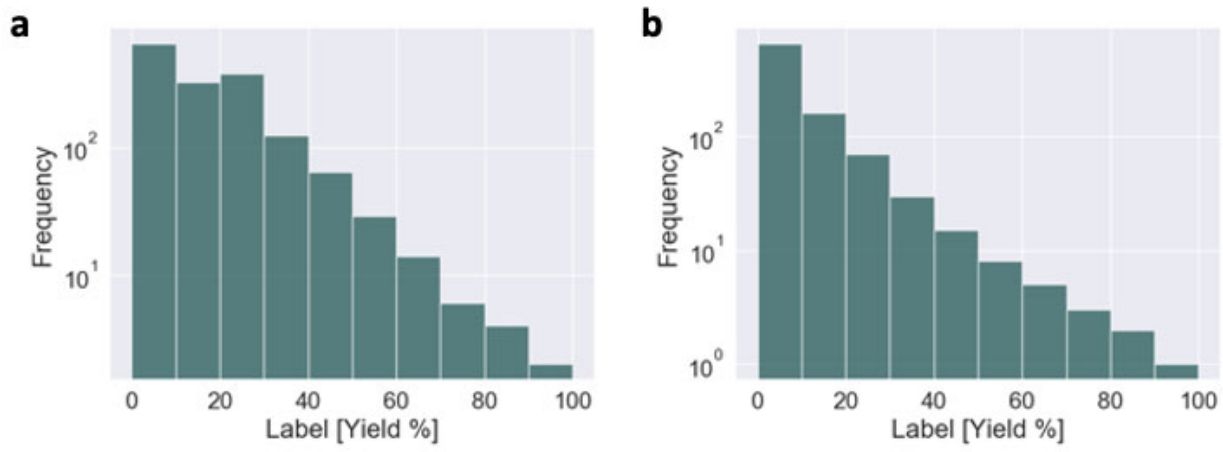

Figure S10. Yield distributions for datasets created by subsampling the Suzuki data to produce exponentially decaying yield distributions; (a) Suzuki Skewed and (b) Suzuki Very Skewed.

We applied ensemble-based uncertainty sampling to both of these datasets (Figure S11). The performance is intermediate between that of the unaugmented Suzuki dataset and the 3-bromopyridine dataset. Relative to the observations that were made with regard to the unaugmented Suzuki results, the trajectories in Figure S11 show an earlier departure of the active learning trajectory from the random learning trajectory, and greater fractional deviation between the active learning test set losses and those of the random learning trajectory.

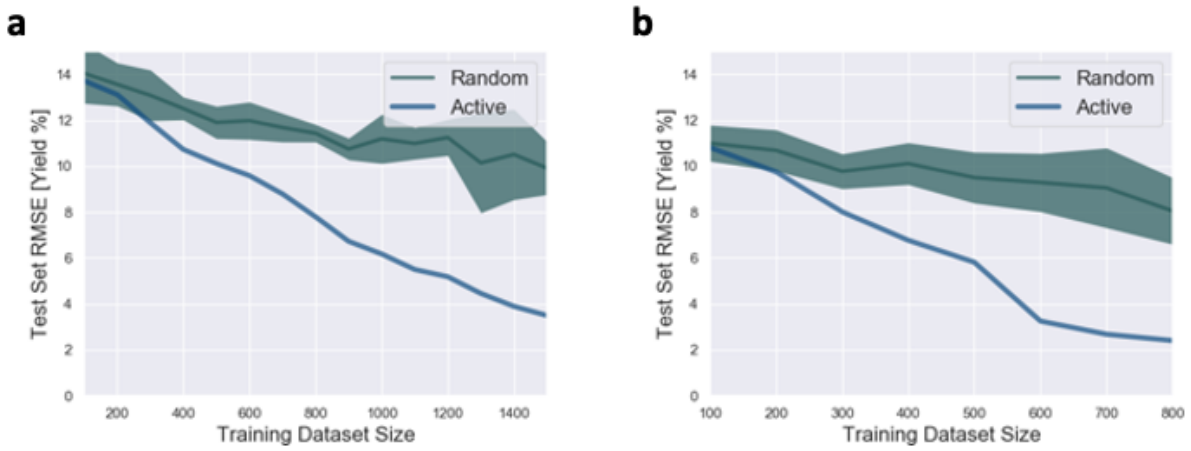

Figure S11. Active learning trajectories (ensemble-based uncertainty sampling) for datasets created by subsampling the Suzuki data to produce exponentially decaying yield distributions; (a) Suzuki Skewed and (b) Suzuki Very Skewed.

Comparison of the active learning trajectories in Figure S11 with each other and with those in Figure 3 suggest that by increasing the proportion of reactions with the same or similar yields, the minimum average test set error achievable by the model decreases. This is evident upon comparison of the initial ($n = 100$) and final loss values achieved for the Suzuki, Suzuki Skewed, and Suzuki Very Skewed datasets. This aligns with our intuition: as the dataset's labels becomes increasingly imbalanced, the model is able to make ever-better predictions for data points in or near the oversampled label region; it may make worse predictions for those data points whose value neighborhoods are underrepresented in the dataset, but since these are underrepresented, they contribute relatively little to the average error.

Further, although specific information about experimental error rates in the Suzuki data is not available, it is reasonable to expect that the same error phenomenon observed in the 3-bromopyridine data (in which the experimental error associated with a reaction that produces no product at all is lower than that for productive reactions) applies to the Suzuki data as well. Only 275 reactions in the Suzuki data were found to be zero-yielding, which represents 4.8% of the non-augmented Suzuki dataset, 17.2% of the Skewed dataset, and 28.9% of the Very Skewed dataset.

To study the influence of the low average error explicitly, we performed active learning on a version of the 3-bromopyridine dataset from which we removed all of the zero-yielding reactions, such that the remaining data possesses a higher average experimental error (Figure S12). We note that removing the zero-yielding reactions also alters the composition of the dataset, and, importantly, how it is balanced across the reactant and reagent space; since our comparison point is the non-augmented dataset, this is not a pure test of the influence of experimental error. We used 10 MC dropout masks for uncertainty estimation in this analysis. The augmentation alters the shape of the active learning trajectory: the active learning model's test set error decays more gradually. The active learning model also converges to a slightly higher final test set error.

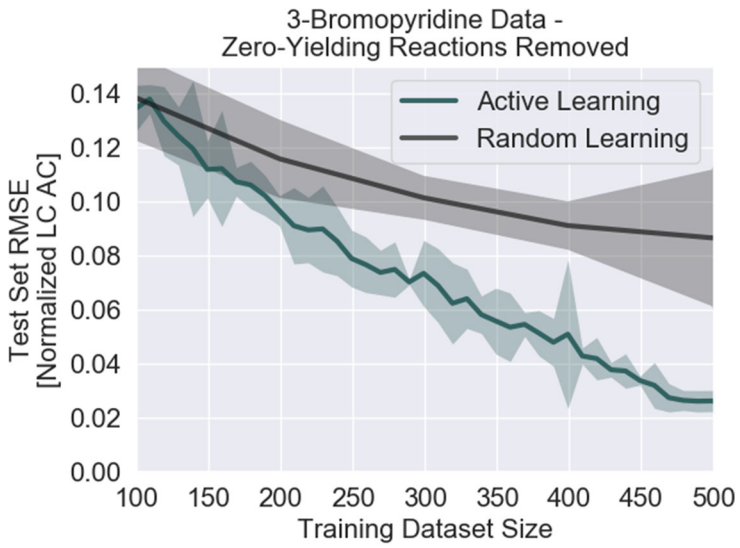

Figure S12. Active learning trajectory for the 3-bromopyridine data after removal of zero-yielding reactions.

f. Influence of added noise

To assess the influence that experimental noise exerts on the performance of the active learning algorithm, we studied the effect of adding additional Gaussian noise to both datasets. We studied two added noise conditions: 1) “low added noise,” in which random Gaussian noise with a standard deviation of 1% of the label range was added to each label, and 2) “high added noise,” in which random Gaussian noise with a standard deviation of 5% of the label range was added to each label (Figure S13).

For the Suzuki data, the most apparent effects of adding noise are higher variance in the active learning outcomes, and a delayed divergence between active and random learning. For the 3-bromopyridine data, adding noise results in higher initial and final test set losses. Compared to the non-augmented data, the high added noise case has much higher test set losses, and active learning outperforms random learning to a substantially lesser degree.

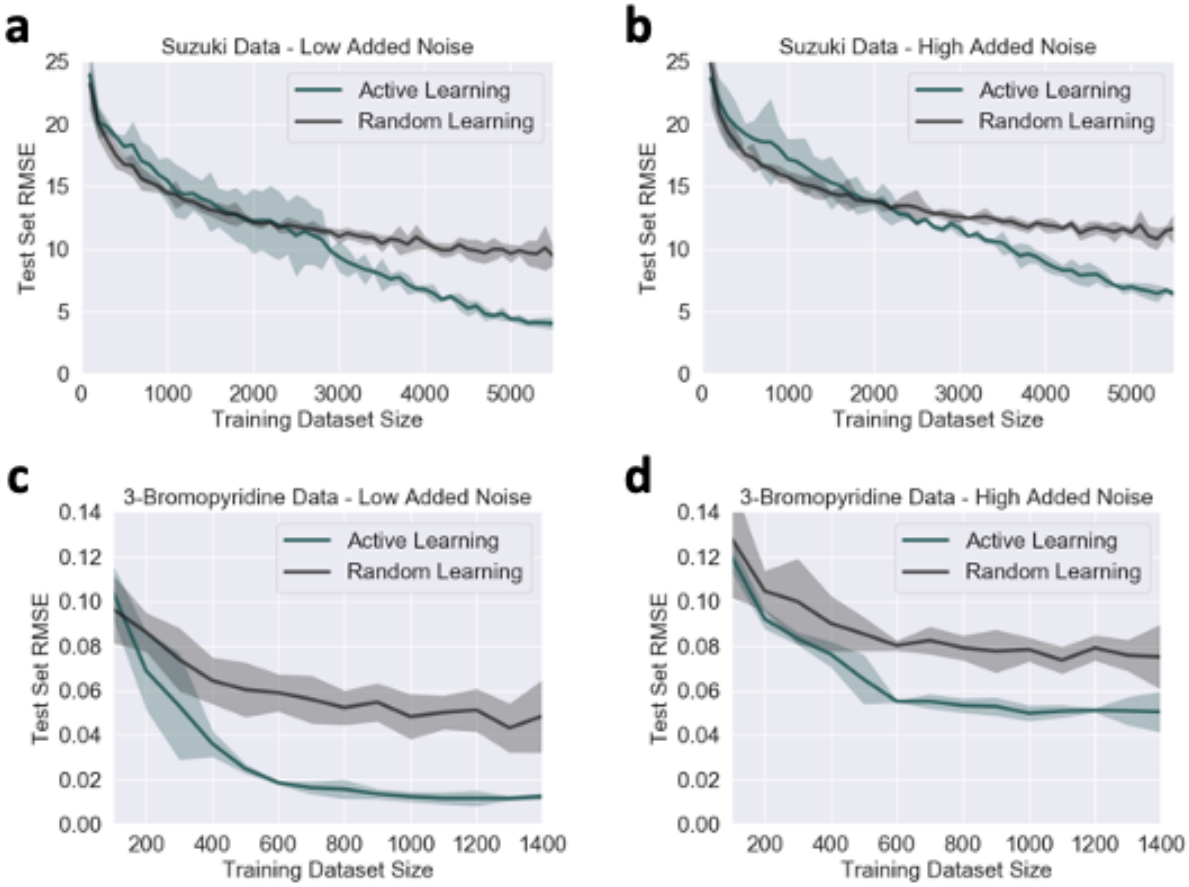

Figure S13. Active learning trajectories in the presence of added noise. (a) Suzuki data, low added noise, test set RMSE is measured in yield%; (b) Suzuki data, high added noise, test set RMSE is measured in yield%; (c) 3-bromopyridine data, low added noise, test set RMSE is measured in normalized LC area counts; (d) 3-bromopyridine data, high added noise, test set RMSE is measured in normalized LC area counts.

g. Optimization

The uncertainty sampling experiment selection criterion can be easily augmented to perform optimizations, by balancing the exploration objective that is native to the active learning approach with an exploitation objective. One strategy to achieve this balance is to combine the two objective functions into one using an exploration parameter λ (Equation S1):

$$x' = \max(\lambda\sigma(\mathbf{x}) + (1-\lambda)(\max(\mu(\mathbf{x}) - f(\mathbf{x}_{\text{best}}), 0))) \quad \text{(S1)}$$

The first term in Equation S1 represents the usual uncertainty measure used for exploration and model-building, which is employed in the bulk of the active learning analyses reported here. The second term in Equation 1 enables optimization, by assessing whether the yield of candidate reaction **x** is predicted to exceed the best value observed so far $f(\mathbf{x}_{\text{best}})$, and if so, by how much. By maximizing this new objective function, we can easily perform optimizations using essentially the same framework as before.

The Suzuki dataset presents an ideal dataset for studying the performance of optimization routines since it consists of an exhaustive combination of all of the variables explored in the domain. We were curious about how the optimization version of the algorithm would perform if it were initialized using 100 reactions all with yield less than twenty percent. We chose yield exceeding 95% as our optimization criterion, and assessed how many reactions (in addition to the initial 100) would be necessary to locate

one of the reactions with yield greater than 95% at varying values of the exploration parameter (Figure S14). Although only 1.4% of the reactions in the Suzuki dataset have yields exceeding 95%, the results indicate that it is possible to achieve optimization by performing just a few dozen reactions post-initialization. There appears to be a positive relationship between the value of the exploration parameter and the maximum optimization time observed. Higher values of lambda place greater weight on the exploration component of the objective function, so slower optimization in this case is expected.

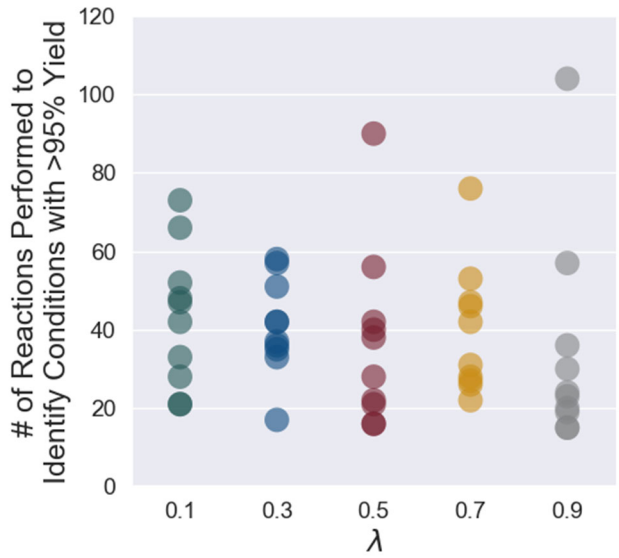

Figure S14. Number of reactions performed, post-initialization, to identify a reaction with yield exceeding 95%, versus the value of the exploration parameter. Ten independent runs of the algorithm were performed at each value of the exploration parameter.

Supplementary References

1 A. Buitrago-Santanilla, E. L. Regalado, T. Pereira, M. Shevlin, K. Bateman, L.-C. Campeau, J. Schneeweis, S. Berritt, Z.-C. Shi, P. Nantermet, Y. Liu, R. Helmy, C. J. Welch, P. Vachal, I. W. Davies, T. Cernak and S. D. Dreher, *Science*, 2015, **347**, 6217.

2 D. Perera, J. W. Tucker, S. Brahmbhatt, C. J. Helal, A. Chong, W. Farrell, P. Richardson and N. W. Sach, *Science*, 2018, **359,** 6374

3 G. Landrum, RDKit: Open-source cheminformatics. URL: http://www.rdkit.org.

4 S. Kearnes, K. McCloskey, M. Berndl, V. Pande, P. Riley, "Molecular graph convolutions: moving beyond fingerprints," *J Comput Aided Mol Des*. 2016, 30(8), 595-608. doi:10.1007/s10822-016-9938-8

5 Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, V. Pande, "MoleculeNet: a benchmark for molecular machine learning," *Chem. Sci.* 2018, 9(2), 513-30. doi: 10.1039/C7SC02664A.

6 Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A. Pytorch: An imperative style, high-performance deep learning library. InAdvances in neural information processing systems 2019 (pp. 8026-8037)