Supporting Information for

# Predicting Reaction Conditions from Limited Data through Active Transfer Learning

Eunjae Shim[1], Joshua A. Kammeraad[1,2], Ziping Xu[2], Ambuj Tewari[2], Tim Cernak[1,3,*] and Paul M. Zimmerman[1,*]

1. Department of Chemistry, University of Michigan, Ann Arbor, MI
2. Department of Statistics, University of Michigan, Ann Arbor, MI
3. Department of Medicinal Chemistry, University of Michigan, Ann Arbor, MI

Email: tcernak@med.umich.edu, paulzim@umich.edu

**Table of Contents**

## Structure of the full dataset: yield label distribution, substrate structures and descriptors

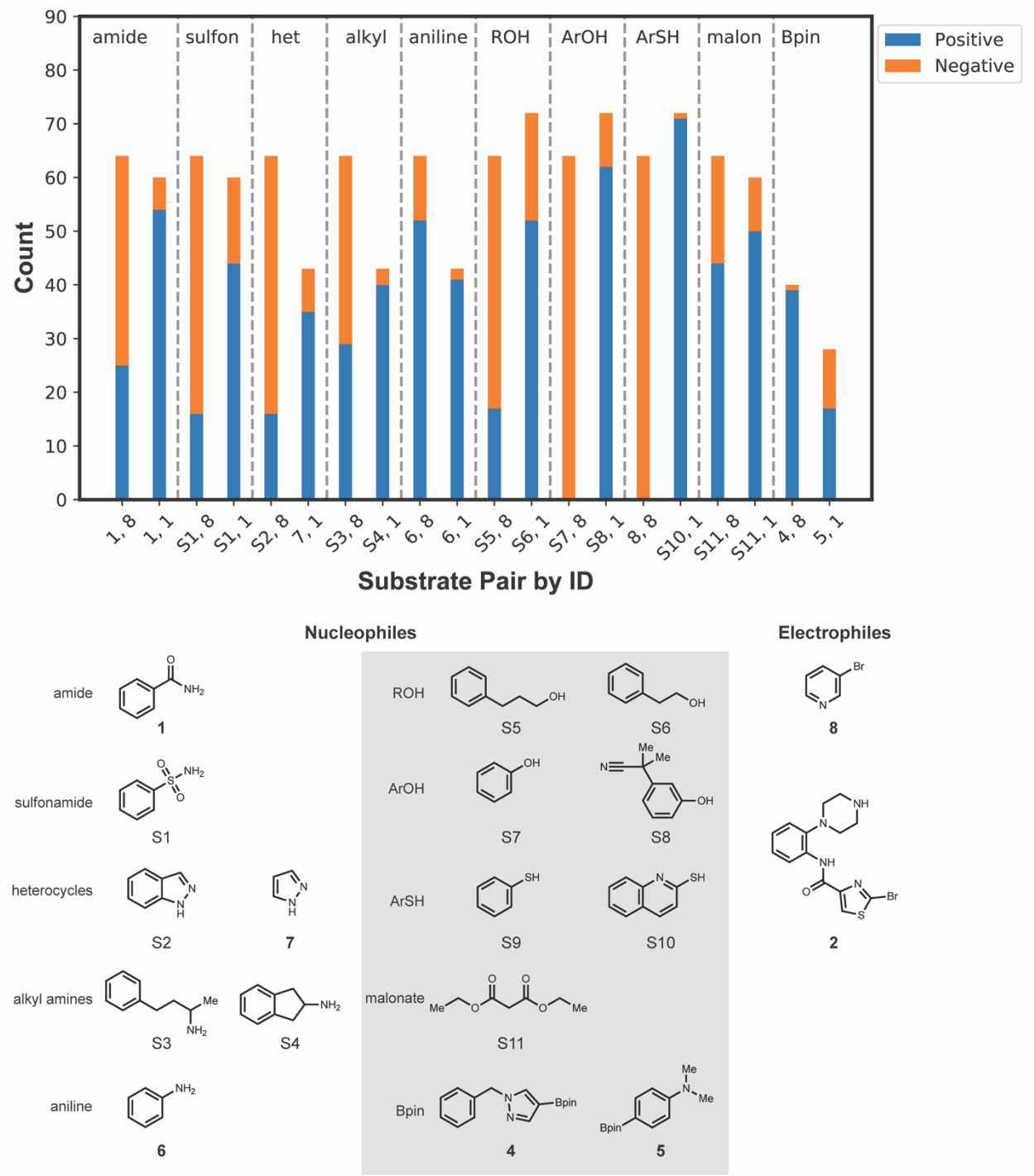



*Figure S1. (Top) Binary yield distribution for every nucleophile type. The x-axis labels denote the ids of substrate pairs in the order of nucleophile-electrophile. (Bottom) Structures of all substrates with their IDs.*

The catalyst, base and solvents used within the dataset can be accessed through the structure query language database file provided on github[1]. Below is the list of descriptors used for each reaction component.

• Nucleophile (6 descriptors): highest occupied molecular orbital energy (HOMO), lowest unoccupied molecular orbital energy (LUMO), volume, area, natural bond order (NBO) of the atom forming the new bond (N for nitrogen nucleophiles and C for C-C coupling reactions), NBO of a hydrogen atom connected to the reacting atom (for all nucleophiles except pinacol boronates) or boron (pinacol boronate).

• Electrophile (27 descriptors): HOMO, LUMO, Volume, Area, $^{13}$C-NMR chemical shift of the carbon atom forming the new bond, NBO of the same carbon, NBO of the Br atom, NBO values of the two adjacent atoms, 9 pairs of vibration frequency and intensity values around the reacting carbon atom.

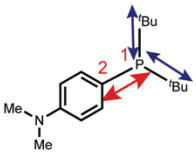
• Catalyst (26 descriptors): buried volume, Sterimol L, B1 and B5 parameters, NBO values of Pd and P in the pre-catalyst complex, HOMO, LUMO, 9 pairs of vibration frequency and intensity values around P calculated for the ligand.

• Base (7 descriptors): HOMO, LUMO, volume, area, NBO at the basic atom, dipole moment, proton affinity

• Solvent (7 descriptors): dielectric constant, molecular weight, density, Hansen D, P and H parameters, dipole moment.

Vibrational descriptors were used for electrophiles and ligands. For electrophiles, the first six pairs are from modes that involve the bonds between C and its neighboring bond that is not Br. The following three pairs of frequency and intensity are of modes that involve the C-Br bond. For ligands, the first six pairs come from modes that involve P and C of the two identical substituents while the last three pairs are from modes that involve the bond between P and C of the di-aryl group. Within each group (first six pairs and last three pairs), the modes were sorted in ascending order of frequency values.

Frequency and intensity values were extracted from modes where displacement is largest along bonds of interest. First, the bond vector of interest is computed from the optimized geometry. Then, for all vibrational modes with frequency above 500cm$^{-1}$, frequency and intensity values, along with the displacement vector of the bond were extracted from the output file. The inner products of the bond vector and displacement vectors were computed. Three modes with highest absolute values were used for the frequency and intensity descriptors. This process is explained in Figure S2.




APhos

**For each bond of interest :**

**1. Compute bond vector from optimized geometry**

```
Step 13
  P       1.04519984     -2.10870206     -0.67353716
  C       2.89489871     -1.56603096     -0.79601820
```

vector : -1.84969887 -0.54267110 0.12248104

**2. For each vibrational mode, get displacement vector**

| Mode: | 37 | | | 38 | | | 39 | | |
|---|---|---|---|---|---|---|---|---|---|
| Frequency: | 588.80 | | | 619.78 | | | 654.08 | | |
| Force Cnst: | 0.7293 | | | 1.0980 | | | 1.8092 | | |
| Red. Mass: | 3.5703 | | | 4.8515 | | | 7.1773 | | |
| IR Active: | YES | | | YES | | | YES | | |
| IR Intens: | 12.233 | | | 9.922 | | | 0.212 | | |
| Raman Active: | YES | | | YES | | | YES | | |
| | X | Y | Z | X | Y | Z | X | Y | Z |
| P | 0.006 | 0.185 | 0.009 | 0.167 | 0.009 | 0.015 | -0.030 | -0.013 | 0.039 |
| C | -0.053 | -0.147 | 0.044 | -0.031 | -0.033 | 0.008 | 0.004 | 0.011 | -0.007 |
| vector : | -0.059 | 0.332 | -0.035 | 0.196 | 0.042 | 0.007 | -0.034 | -0.024 | 0.046 |

**3. Compute absolute values of inner products of the bond vector and displacement vectors**

••• Mode 37 : 0.0667, Mode 38 : 0.5896, Mode 39 : 0.0815 •••

**4. Get three modes of highest values**

*Figure S2. Procedure for extracting vibrational frequency and intensity values.*

## Analysis of transfer from Bpin to phenyl sulfonamide

Figure 3B shows the transfer of Bpin models to predict phenyl sulfonamide reactions resulted in an ROC-AUC score of 0.04. While the low score may seem to be unsatisfactory, it indicates that the models can almost perfectly distinguish between positive and negative reactions. A deeper analysis was conducted to explore the reason behind this observation.

Table S1 shows the distribution of labels in each dataset (i.e. number of reactions that gave positive and negative yields, respectively). For electrophile 8, there is a near-perfect inverse relationship between their labels when comparing the two nucleophile types, indicating that the opposite outcome occurs when switching nucleophile type.

| Source (Bpin) | | | | Target (Sulfonamide) | | | |
|---|---|---|---|---|---|---|---|
| Nuc. Id | Elec. Id | #Positives | #Negatives | Nuc. Id | Elec. Id | #Positives | #Negatives |
| 4 | 8 | **39** | **1** | S1 | 8 | **0** | **40** |
| 5 | 2 | 17 | 11 | S1 | 2 | 4 | 6 |

*Table S1. Label distributions for source (Bpin) and target (sulfonamide) by electrophile that explains the low ROC-AUC for electrophile 1 (i.e. inverse prediction).*

To demonstrate how the label distribution impacts model's decisions on target reactions, decision paths (i.e. which descriptor values were evaluated at each node) applied to the 50 sulfonamide target reactions were analyzed for one Bpin random forest model (25 trees of depth two). Since the electrophile is common between the source and target datasets, if the decision path involves any electrophile descriptor, the decision tree will correctly recognize the target electrophile and predictions are based on this

information. On the other hand, the source dataset involves two nucleophiles 4 and 5, each reacting with electrophiles 8 and 2, respectively. If a decision path involves a descriptor of a nucleophile, the decision tree may recognize nucleophile S1 as either nucleophile 4 or 5, which leads to predictions as if the reaction involved electrophile 8 or 2, respectively. When only descriptors of catalyst, base and solvent are used throughout the decision path, the decision tree does not explicitly utilize electrophile information for making a prediction. Predicted probabilities of the correct label were collected from each decision tree, on every sulfonamide reaction (Figure S3 A and B shows results on reactions of electrophile 8 and 2, respectively). Higher values would lead to higher 'overall' ROC-AUC values.

In agreement to what is expected from the near-opposite labels, when the reacting electrophile is correctly recognized (~65% of all evaluations) as electrophile-8, decision trees almost always make a wrong prediction (Figure S3A, left column). Only when the reacting electrophile is incorrectly recognized as 2 is when a reaction between sulfonamide-S1 and electrophile-8 could be correctly predicted with high confidence, while the average value is still below 0.4 (Figure S3A, middle column). The predicted probability values on target reactions with electrophile-2 were undecisive, ranging throughout 0 and 1 (Figure S3B). Highly confident incorrect predictions of electrophile-8 reactions and the predictions that slightly favor the incorrect prediction of reactions that involve electrophile-2, combined, explains the ROC-AUC close to 0.

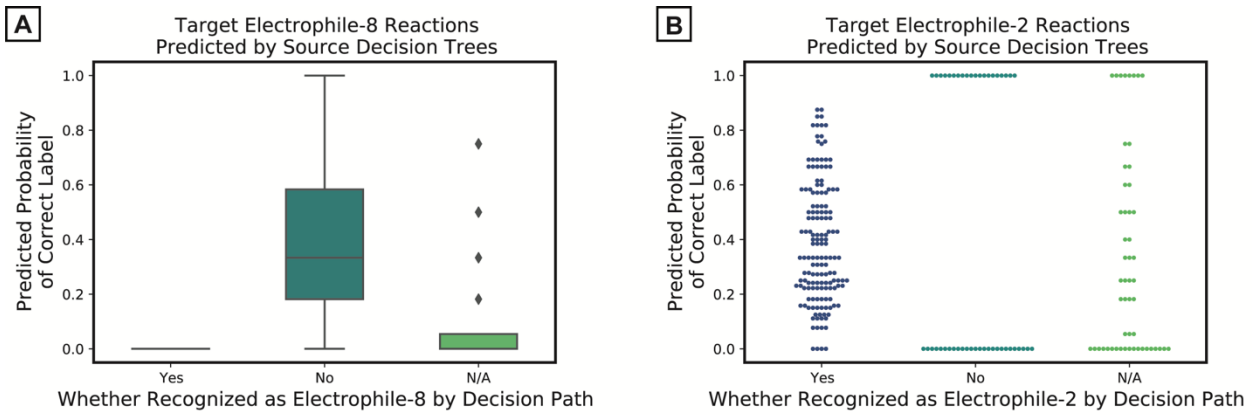



*Figure S3. Decision trees, which compose random forests, make decisions through a series of evaluations of descriptor values. Due to the structure of the dataset, evaluation of nucleophile descriptors implicitly infers the identity of the electrophile of the reaction. (Left) Predicted probability values of 40 electrophile-8 reactions produced by decision trees in a Bpin model. When the electrophile is correctly inferred, as 8, the wrong prediction is made with high confidence (left column). Predictions that favor the correct label is made when the electrophile identity was assumed incorrectly (middle column). (Right) Predicted probability values of 10 electrophile-2 reactions produced by decision trees of the same Bpin model. The values are spread throughout 0 and 1.*

Lastly, Bpin models were evaluated against amide models on sulfonamide reactions with electrophile-2 only, to remove the effect of inverse label distribution of reactions of sulfonamide and electrophile-8. The ROC-AUC scores on the 10 electrophile-2 sulfonamide reactions (Table S2) were compared between the 10 Bpin and amide models used in Figure 3B, which were trained on different random seeds. An independent two-sample t-test showed that the ROC-AUC scores for amide models are significantly higher than those for Bpin models (p-value 4.7e-4), supporting our claim that a model

built on reactions of similar mechanism is likely to transfer better than others. As such, balanced datasets with representation from positive and negative examples, in contrast to most publicly available ones, may play an important role for the investigation of how different substrate types relate to each other.

| Model | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Bpin | 0.083 | 0.375 | 0.250 | 0.042 | 0.292 | 0.250 | 0.375 | 0.250 | 0.396 | 0.042 |
| Amide | 0.396 | 0.750 | 0.667 | 0.667 | 0.396 | 0.667 | 0.625 | 0.208 | 0.771 | 0.375 |

*Table S2. ROC-AUC scores of models trained on two different nucleophile types predicting 10 target reactions between sulfonamide and electrophile-2.*

As ROC-AUC values below 0.1 continued to be observed in Table S2, even after limiting the predictions to reactions of electrophile 2, the labels of the reactions were further inspected. 5 out of 10 common reaction conditions resulted in different yield labels for Bpin and sulfonamide reacting with electrophile 2. This, along with the fact that only 10 reactions are being evaluated and the high variance across models (amide models show ROC-AUC score as low as 0.208, as shown in model 8 of Table S2), could all be reasons beneath the low ROC-AUC. However, the possibility of the mechanistic and structural differences between the Bpin and sulfonamide nucleophiles could also be playing a role.

## Structure of datasets for Figure 4A

The dataset has been organized to study model generalizability and active transfer learning as explained in Figure 4A. As shown in Figure S1, 3-Bromopyridine and 2-bromothiazole correspond to electrophile 8 and 2, respectively. In Figure S4A, 'Elec2 Train' corresponds to the 18 reactions that are common across the source and target datasets. 'Elec2 Test' describes additional reactions between the target nucleophile and 2-bromothiazole (25 for heterocycle, alkyl amine and aniline, as stated in the main text, 22 for amide and sulfonamide, and 30 for ROH). In other words, when a nucleophile dataset is used as the source, the reactions would consist of the left and middle columns, while the target dataset is composed of the middle and right columns, within each section in Figure S4A.

Reactions with 3-bromopyridine (electrophile 8), of which most of the source data consists, are well balanced in outcome. In contrast, for 2-bromothiazole (electrophile 2), the portion of negative reactions is significantly smaller. Therefore, *defining negative reactions as desired outcomes* represents a harder problem, as described in the Computational Details section. Particularly, as in Figure S4A, heterocycle (pyrazole), alkyl amine and aniline do not show any desired outcomes for 'Elec2 Train', making the problem even more challenging.

The reagents (catalysts, bases and solvents) used for target reactions are those that were each individually employed in the source data. This is shown in Figure S4B, where the clusters on the right side are formed by overlaps between source and target reactions. This 'familiarity of reagents' is deliberately analogous to how chemists would start an exploration, before considering totally new reagents or combinations.

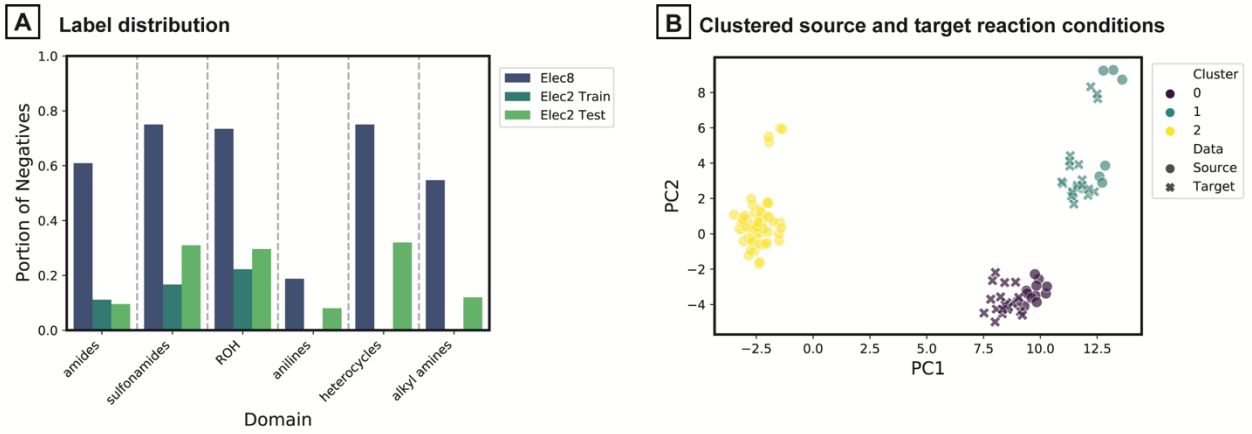



*Figure S4. (A) Yield label distributions of nitrogen nucleophile reaction data for generalizability and active transfer learning studies. (B) PCA of clustered source and target reaction conditions that are jittered horizontally to reduce overlap between points. Yellow datapoints on the left side correspond to source reactions with electrophile 8. Overlap even after jittering shows that the source and target reaction conditions used for electrophile 2 are highly similar.*

## Cross validation within source datasets

Models of various combinations of maximum depth and number of trees were evaluated through 5-fold cross-validation (CV). Results presented in the heatmaps are an average of ROC-AUC over models with 25 different random seeds, with standard deviations provided in parentheses. Random seeds impact the bootstrapping of the samples used when building trees and the sampling of features to consider when looking for the best split at each node[2].

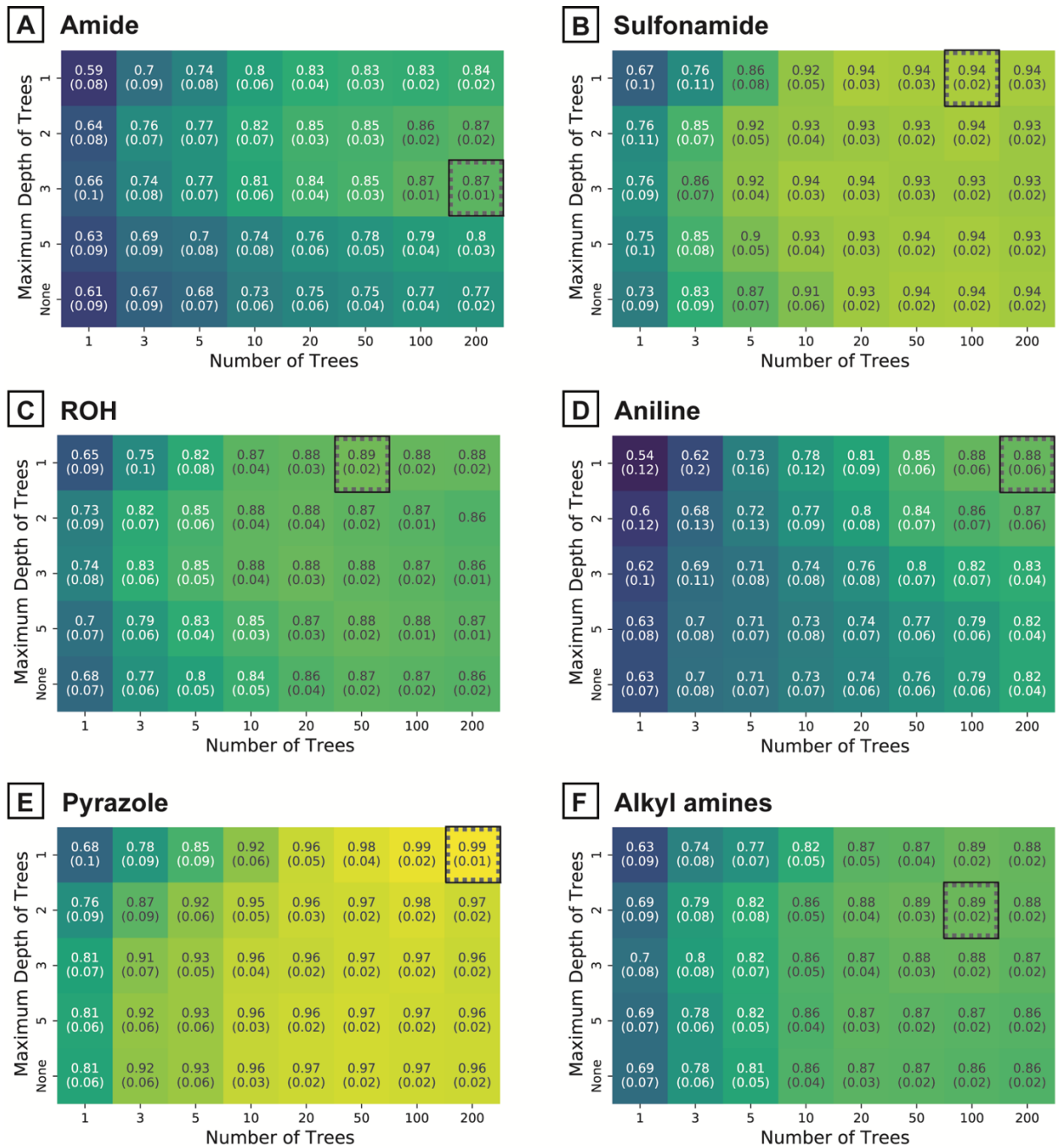



*Figure S5. In-domain CV results for each source nucleophile. Black boxes indicate the hyperparameter combination that results in the highest average CV score.*

**Model transfer with different source and target domain combinations**

Generalizability of source models to the target reactions that include unseen reagent combinations was analyzed in Figures 4B and 4C. Here, further details on the comparison between CV-determined and simpler models in transfer between all source-target pairs are provided. Our goal is to show that there may exist simpler models with comparable (or even better) transfer performance.

The simpler models used for the comparisons were determined after applying the following two filters. First, in-domain CV scores (used to generate Figure S5) of models with different hyperparameter combinations were compared to that of the models that give maximum score, through independent t-tests (black boxes in Figure S5). Independent t-tests were conducted over CV scores of 25 independent runs (from different random seeds) for the two compared models. The models that have statistically different CV scores were subject to the next filter. Note that this contrasts with the common practice which selects models that show comparable performance to the best model and thus evaluates models for transfer that would not be conventionally considered. Among these inferior in-domain models, we further limit the models to those with a smaller number of maximum possible evaluation nodes (computed as num_trees * $2^{max\_depth}$, strongly limiting the maximum depth values that can be considered) than that of the CV-determined model. These filters were inspired by regularization, which may provide sub-optimal training scores, but better generalizability. Then, the model that provides highest transfer ROC-AUC was selected to compare transfer performance with the CV-determined model.

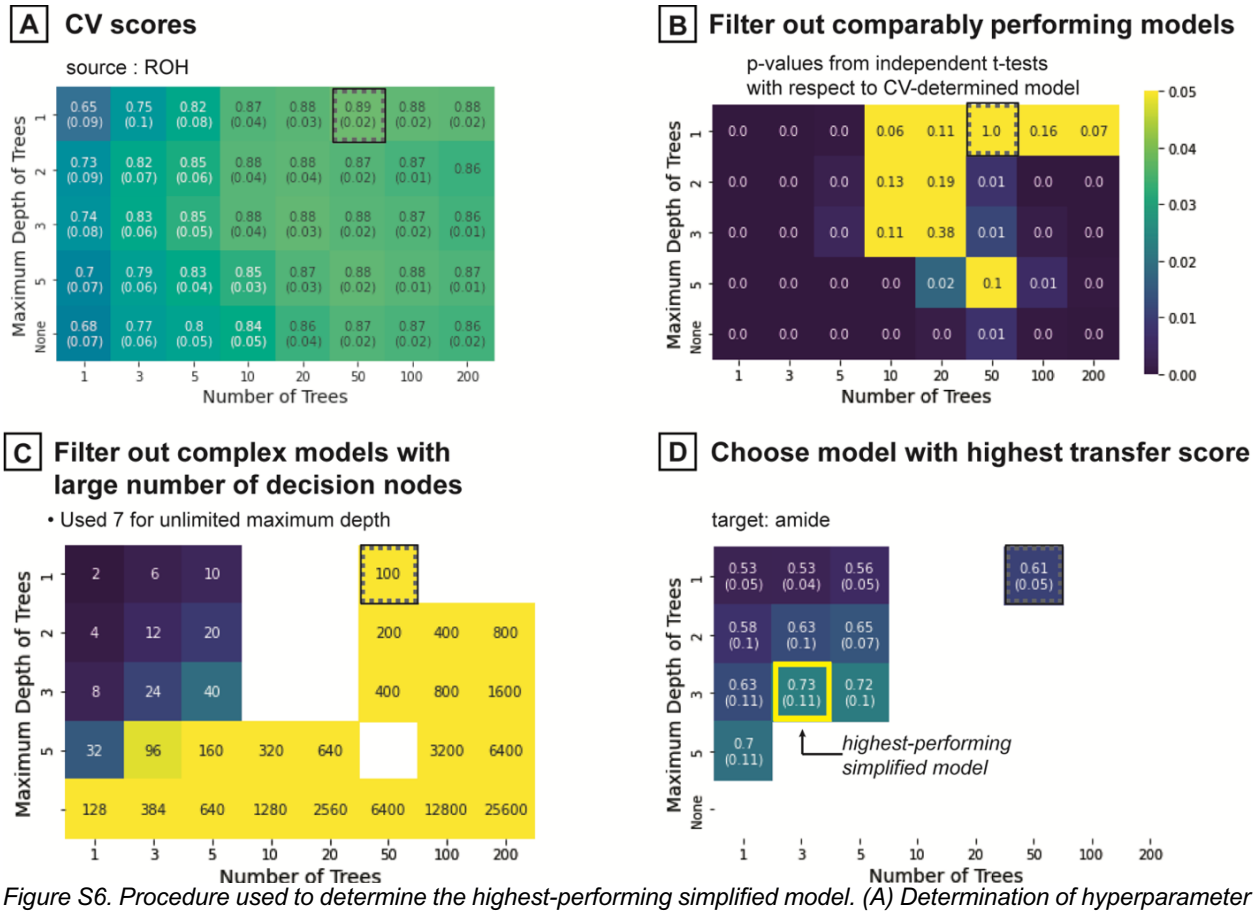



*Figure S6. Procedure used to determine the highest-performing simplified model. (A) Determination of hyperparameter combination with highest average CV score. (B) CV scores of other hyperparameter combinations are compared to the CV-determined model through independent t-tests. Only those that show significantly worse performance ($p<0.05$) are further considered. (C) Then, the maximum number of decision nodes are compared. Models that reduce the number by at least 50% are considered. (D) Finally, we compare the highest transfer ROC-AUC (yellow box) with the CV-determined model's ROC-AUC.*

The full results are provided in Table S3, which is summarized in Figure S7. In some cases, a simplified model significantly outperforms (determined by independent t-tests) CV-determined models, though the reverse is also possible. In 14 of 30 cases, there is no statistical significance to the differences in hyperparameter choice (as shown with X markers in Figure S7). The simplified models with better performance outnumber the CV-determined models (9 green circles vs 7 purple circles). The biggest contributors to CV-determined models statistically outperforming simplified models are transferring with aniline source models, where the aniline source data is significantly different in structure than others (Figure S4A).

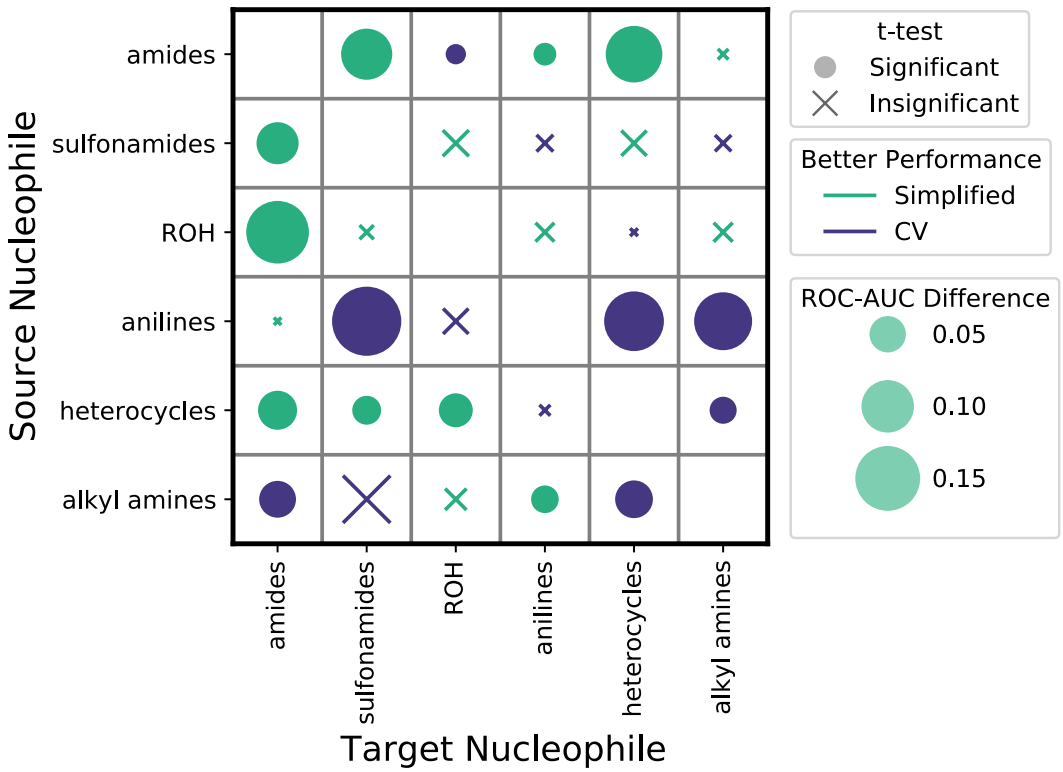



*Figure S7. Summary of the comparison between CV-determined models and simpler models. The marker shape shows whether the p-value resulting from independent t-test is below 0.05. Marker color and size shows which model scheme resulted in a higher average ROC-AUC and to which extent, respectively.*

| Source | Target | CV-#Trees | CV-Depth | Simpler-#Trees | Simpler-Depth | CV Avg ROC-AUC (std) | Simpler Avg ROC-AUC (std) |
|---|---|---|---|---|---|---|---|
| Amide | Sulfon | 200 | 3 | 50 | 1 | 0.69 (0.05) | **0.81 (0.09)** |
| | ROH | | | 50 | 2 | **0.83 (0.01)** | 0.82 (0.02) |
| | Aniline | | | 5 | 1 | 0.91 (0.01) | **0.93 (0.01)** |
| | Het | | | 10 | 1 | 0.52 (0.02) | **0.62 (0.06)** |
| | Alkyl | | | 5 | 1 | 0.93 (0.01) | **0.94 (0.01)** |
| Sulfon | Amide | 100 | 1 | 3 | 2 | 0.59 (0.04) | **0.65 (0.07)** |
| | ROH | | | 3 | 3 | 0.77 (0.03) | **0.79 (0.04)** |
| | Aniline | | | 5 | 1 | **0.90 (0.02)** | 0.89 (0.06) |
| | Het | | | 5 | 1 | 0.60 (0.03) | **0.62 (0.04)** |
| | Alkyl | | | 5 | 1 | **0.91 (0.02)** | 0.90 (0.06) |
| ROH | Amide | 50 | 1 | 3 | 3 | 0.61 (0.05) | **0.73 (0.11)** |
| | Sulfon | | | 5 | 3 | 0.84 (0.03) | 0.84 (0.07) |
| | Aniline | | | 5 | 1 | 0.90 (0.02) | **0.92 (0.03)** |
| | Het | | | 5 | 1 | 0.62 (0.04) | 0.62 (0.06) |
| | Alkyl | | | 5 | 1 | 0.91 (0.02) | **0.93 (0.03)** |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Aniline | Amide | 200 | 1 | 20 | 2 | 0.59 (0.04) | 0.59 (0.08) |
| | Sulfon | | | 20 | 1 | **0.82 (0.03)** | 0.68 (0.18) |
| | ROH | | | 20 | 1 | **0.79 (0.04)** | 0.77 (0.06) |
| | Het | | | 20 | 1 | **0.74 (0.04)** | 0.63 (0.09) |
| | Alkyl | | | 20 | 1 | **0.82 (0.08)** | 0.72 (0.21) |
| Het | Amide | 200 | 1 | 5 | 2 | 0.56 (0.02) | **0.60 (0.05)** |
| | Sulfon | | | 10 | 2 | 0.77 (0.02) | **0.80 (0.04)** |
| | ROH | | | 10 | 2 | 0.71 (0.03) | **0.75 (0.05)** |
| | Aniline | | | 20 | 1 | **0.93 (<0.01)** | 0.92 (0.02) |
| | Alkyl | | | 10 | 1 | **0.94 (<0.01)** | 0.92 (0.04) |
| Alkyl | Amide | 100 | 2 | 5 | 2 | **0.56 (0.07)** | 0.53 (0.06) |
| | Sulfon | | | 5 | 2 | **0.63 (0.13)** | 0.56 (0.17) |
| | ROH | | | 5 | 2 | 0.75 (0.04) | **0.76 (0.06)** |
| | Aniline | | | 5 | 1 | 0.91 (0.01) | **0.93 (<0.01)** |
| | Het | | | 5 | 2 | **0.62 (0.06)** | 0.57 (0.07) |

*Table S3. Full analysis of CV-determined models vs best-performing simpler models. The latter was chosen from hyperparameter combinations that result in significantly worse CV scores and with smaller number of decision nodes. The hyperparameters of both models and corresponding average transfer ROC-AUC scores and standard deviation are shown. ROC-AUC values that are higher than the other by at least 0.01 is highlighted bold.*

## Adversarial controls for model transfer

Table S3 shows that models transferred amongst N-nucleophiles make meaningful predictions (ROC-AUC values > 0.5). To show that this is not due to a spurious correlation between the datasets, 25 source models were trained on shuffled yield labels through 5-fold cross-validation (yield labels were shuffled differently each time). The transfer performance of y-shuffled models was compared to that of unshuffled models for all source-target pairs. The heatmap below shows the p-values from independent t-tests for each scenario. Except one case (alkyl amines → amide) where transfer failed in the original model, all comparisons show significant difference. Therefore, models trained with proper data labels (i.e. unshuffled y) are crucial for making meaningful transfer.

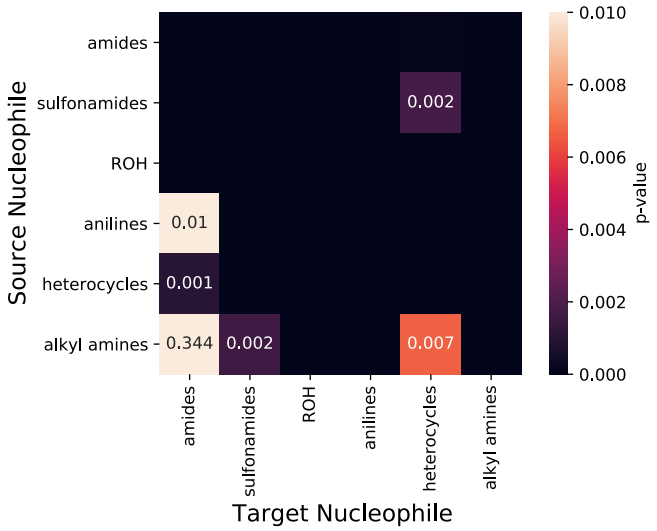



*Figure S8. Heatmap of p-values resulting from independent t-tests between y-shuffled models and CV-determined models.*

Models built upon DFT descriptors were compared to models using fingerprints and one-hot labels. Each reaction was represented with fingerprints by concatenating each component's Morgan fingerprint of length 1,024 and radius of 2 (resulting vector is length 5,120). Each column of a source one-hot label array corresponds to a compound used within the entire source dataset. Each reaction is then represented with 1's at the columns of the compounds used and 0's at all other elements, resulting in five 1's. Importantly, for target reactions, since the target nucleophile was never used in the source dataset, there all nucleophile columns have the value of 0.

Figure S9 shows the results of all transfer scenarios, sorted by representation. The diagonal plots show average in-domain CV scores. For four of the six source nucleophiles, models based on fingerprints showed a significantly higher in-domain CV score. However, when transfer performance is compared, independent t-tests after Bonferonni correction show DFT descriptor models significantly outperform models trained with other two representations in 16/30 transfers. Fingerprints perform better in 6/30 scenarios than other representations, while only 2 scenarios favor one-hot encoding. For remaining 6 cases, the three representations did not result in statistically better models. Particularly, for transfers where performance differs significantly, the ROC-AUC benefit of descriptors over fingerprints is larger than fingerprints over descriptors, as shown in Figure S10. This result suggests that the use of DFT descriptors can be beneficial over fingerprints or one-hot encoding. This might be due to the better representation of mechanisms with DFT descriptors than one-hot encoding, which is equivalent to distinguishing compounds by their names, and fingerprints, which are purely structural.

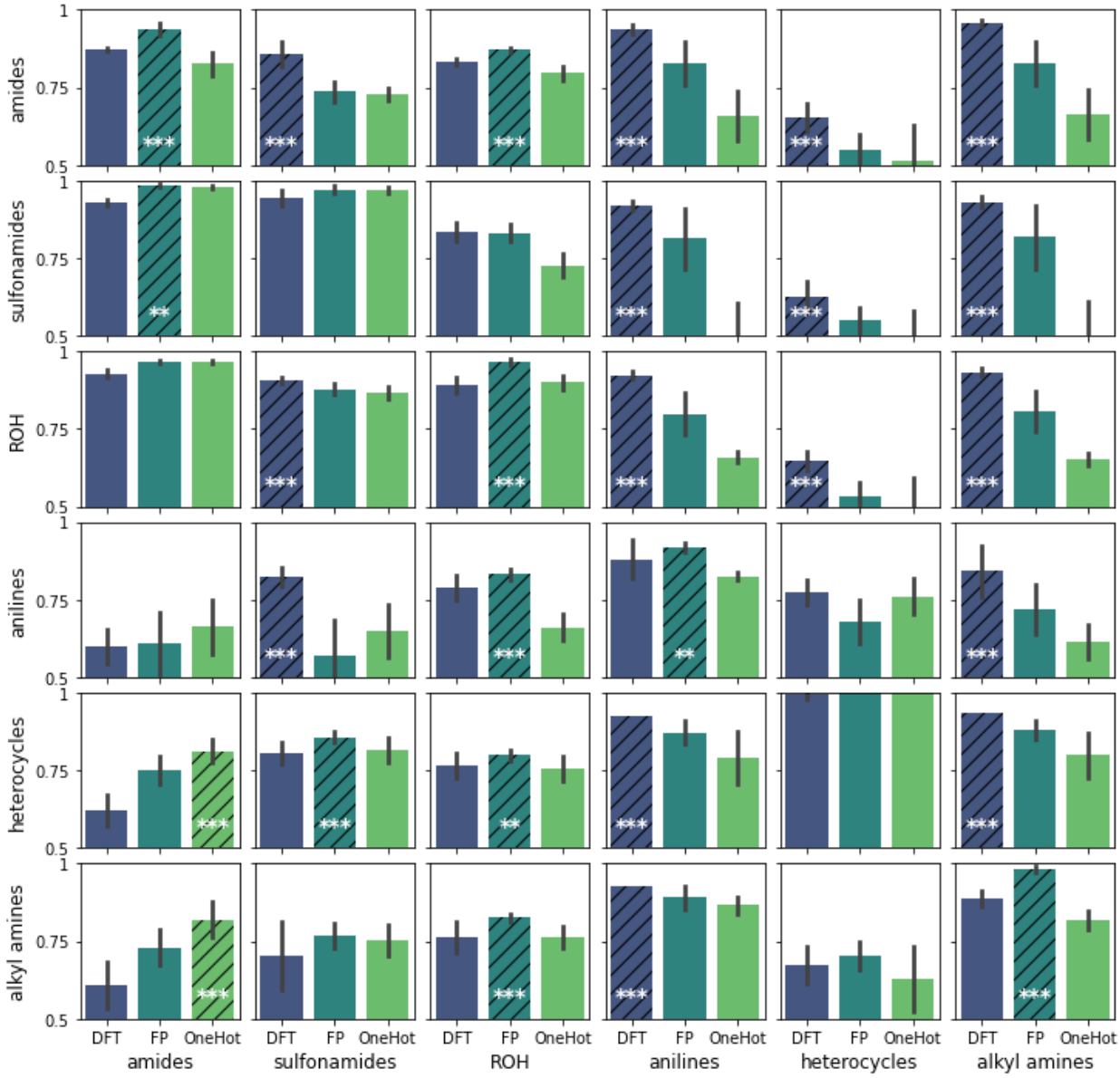



*Figure S9. Adversarial control with fingerprints and one-hot labels. Diagonals correspond to in-domain CV performance. Row and column labels correspond to source and target nucleophiles, respectively. Black diagonal hatches of each bar denote that the corresponding representation's models perform statistically better than the other two, after Bonferonni corrected p-values of independent t-tests. ** denotes p<0.01. *** denotes p < 0.001.*

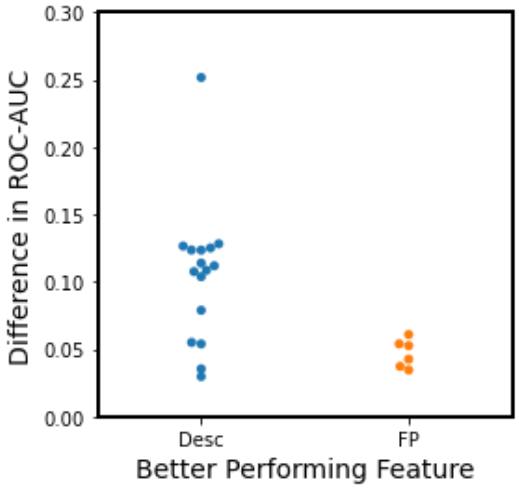



*Figure S10. ROC-AUC benefit of the statistically better performing representation.*

## ATL results of 'target tree growth' strategy with other source datasets

From this section onwards, the target nucleophile is pyrazole unless stated otherwise. With the other two source datasets and pyrazole as target, the target tree growth model is compared to other baselines. The descriptors selected by models at different iterations were also analyzed. Similar observations to that in the main text (Figure 5) can be made for the other two source nucleophiles.

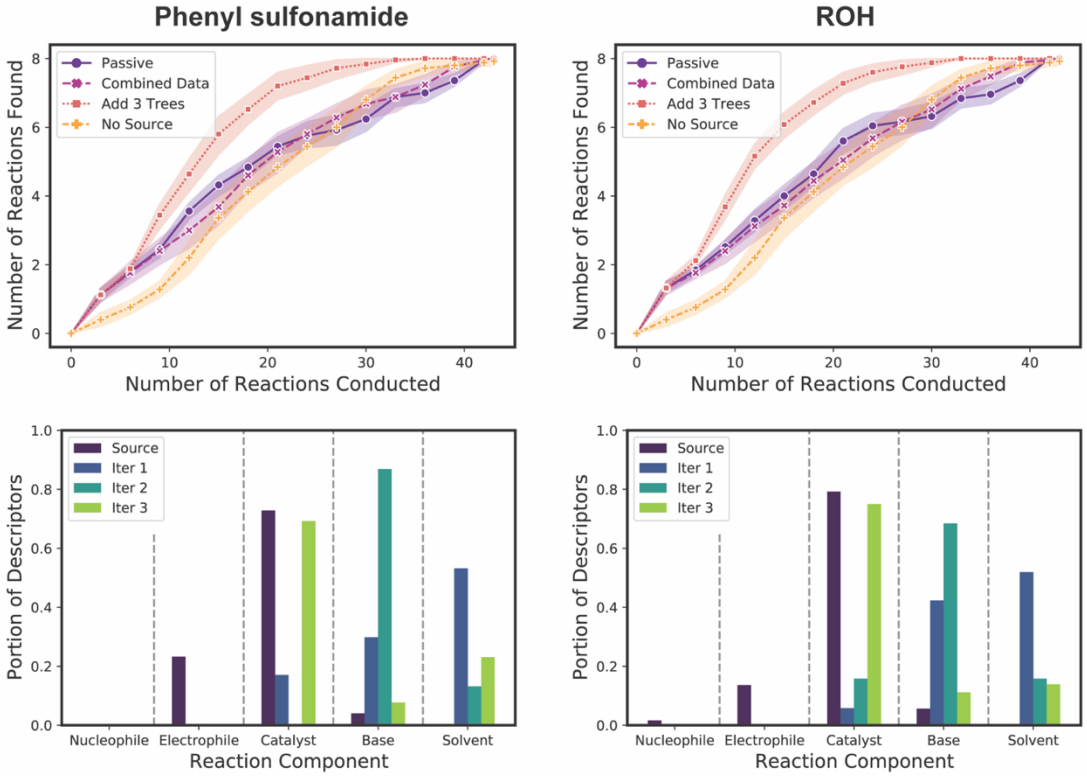



*Figure S11. (Top row) Results of the target tree growth ATL strategy with different source datasets. (Bottom row) Portions of descriptors selected by models at each iteration, with different source datasets.*

## Reaction selection behavior of different strategies

The analysis of selected reactions provided insight on how the adaptation of target tree growth ATL proceeded along iterations. Therefore, the behaviors of the passive model, which does not update knowledge from collected target reactions, and AL on combined source and collected target data is shown in Figure S12.

One notable aspect of the target tree growth ATL was its choice to move onto a different catalyst in the third iteration after confirming that the use of the bottom two bases yield negative reactions in the second iteration (two bottom reactions are left unlabeled in the bottom right cluster in Figure S12A). In contrast, the passive source model exhausts the negative reactions in the right top cluster at the third iteration. This is continued at latter iterations at the clusters located at PC1 near 1 (Figure S12B). Similarly, models updated with combined data also exhausts the reactions that involve rightmost catalyst in the first three iterations. However, this is changed in the fourth iteration, where reactions are sampled from clusters with different catalysts, which implies the newly combined target data may have started to make an influence in the selection process. However, unlike the target tree growth ATL, it continues to sample the bottom two reactions of a cluster, even though it has seen numerous negative examples, leading to inefficient exploration. This demonstrates the difference in what the models learn, following different strategies.

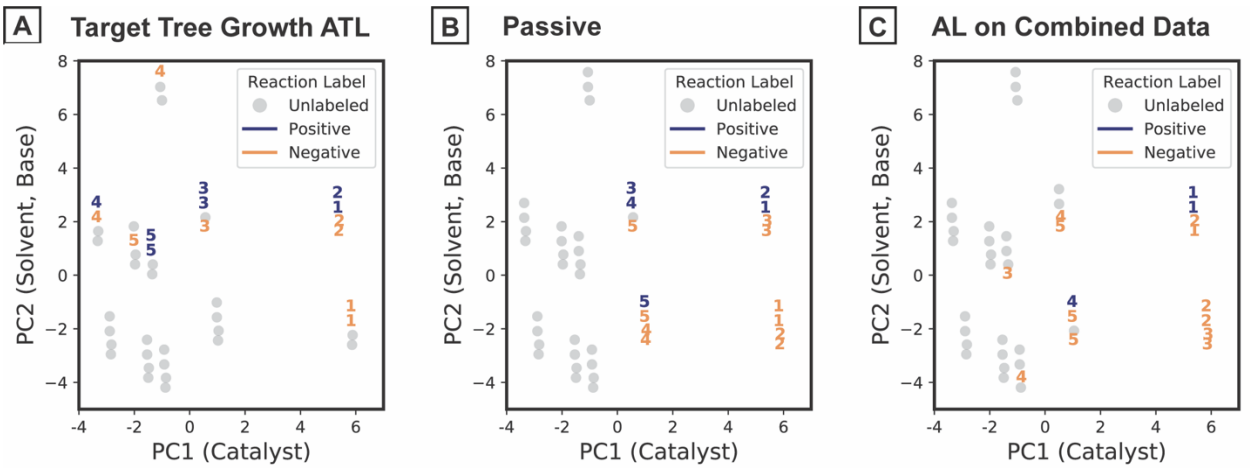



*Figure S12. Examples of which reactions were selected at each batch (up to the 5th iteration), for different strategies. Grey circles are unlabeled reactions. The number markers denote the iteration number, and their colors identify the labels of the reaction.*

**Comparison of model performance on source and collected target data to understand the adaptation of models in the target domain**

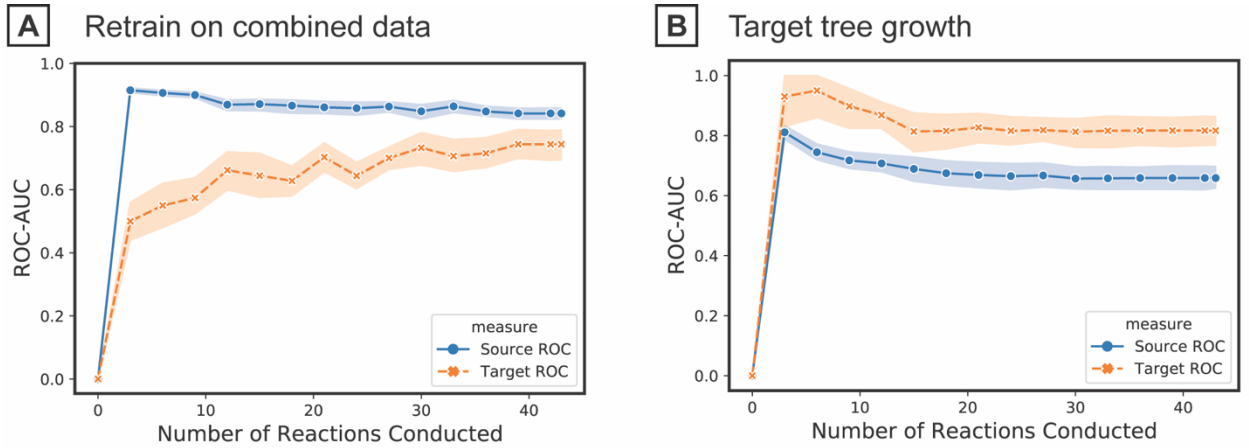



*Figure S13. Evaluations of the adaptability of two different ATL strategies. (A) and (B) correspond to results of retraining on combined data with weight of 5 on target datapoints and target tree growth, respectively. The plots show how models at every iteration perform on the source data (blue curves) and collected target data up to that point (orange curves).*

To evaluate the adaptability of each ATL strategy in the target reaction space, the performance of models at each iteration were evaluated on the target data that has been collected up to that point and compared to their performance on source data which is also in hand. Figure S13A shows the results on models that were trained on the combined source and target data. Even with sample weight of 5 on target datapoints, models fail to show meaningful performance on the target data they have been trained on (Figure S13A, orange curve), while performance on the source data is remains high (Figure S13A, blue curve). In contrast, the target tree growth strategy, which adds decision trees trained only on newly collected target data, performs better on the collected target data than the source (Figure S13B) even after the first iteration. These results indicate higher degree of adaptation in the target domain for the target tree growth strategy compared to models retrained on the combined data.

**ATL results when models are trained on combined source and target data with importance weights favoring the target**

For this target, the first ATL strategy retrains a random forest on the dataset that combines the source data with all collected target reactions every iteration. Random forests of five trees with maximum depth of one were used. Since the source data outnumbers the target data, especially in early iterations, and our goal is to model the target data, weights greater than one for the target were considered. Weight of zero is equivalent to using the source model as is (i.e. passive), while the weight of one considers each source and target reaction data point equally. For clarity, only the average number of reactions found across 25 model instances are shown.

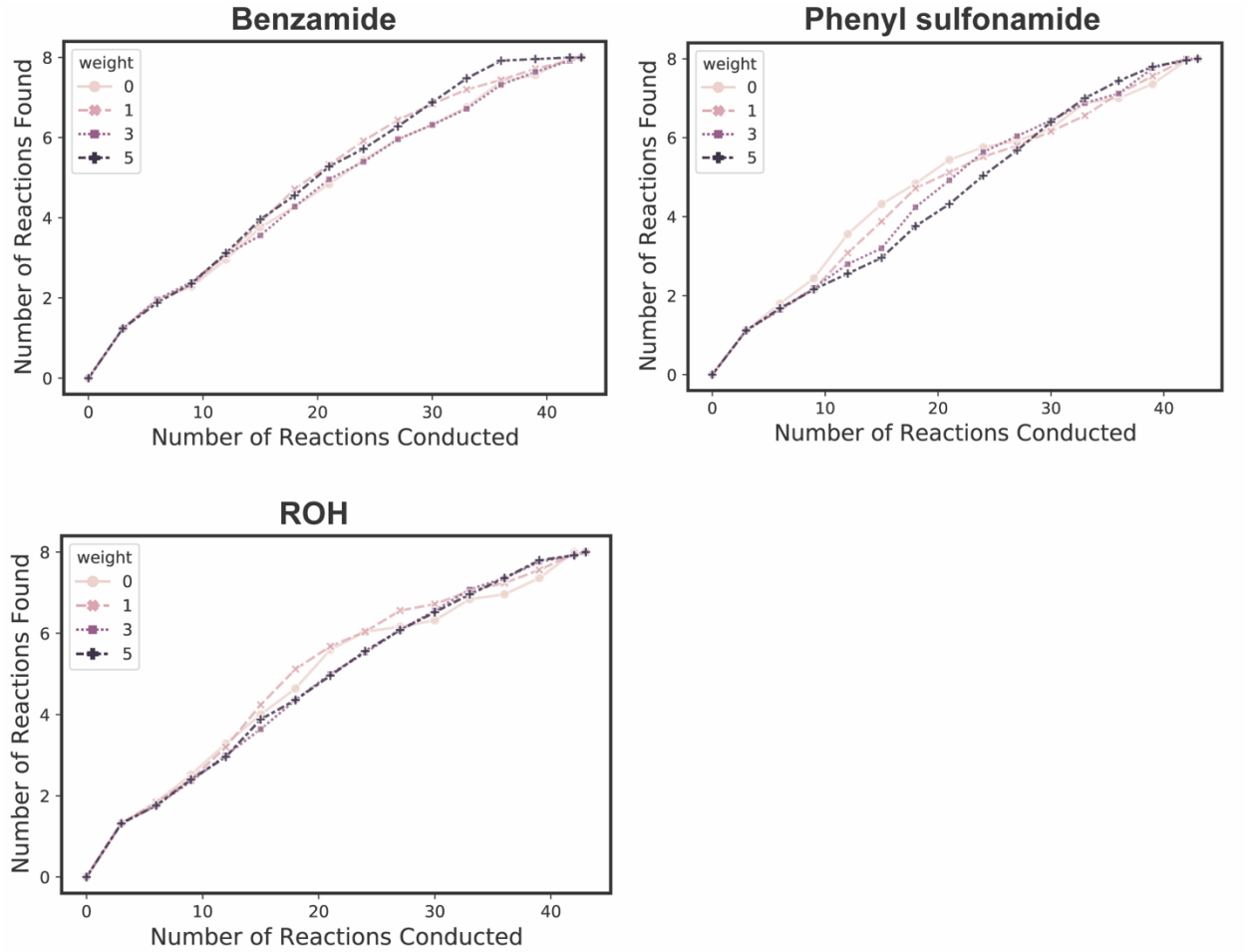



*Figure S14. Performance of different importance weight values on collected target reactions when models are updated every iteration on the combined data. Source nucleophile is written as the title and the target is pyrazole.*

For most cases in Figure S14, the performance curves lie near the diagonal, which corresponds to random selection. This suggests that model updates based on combined datasets do not provide significant benefit over random selection or passive learning (weight=0), probably due to the low adaptability in the target reaction space as demonstrated in Figure S14A.

**ATL results when different number of 'target trees' are added to the model every iteration**

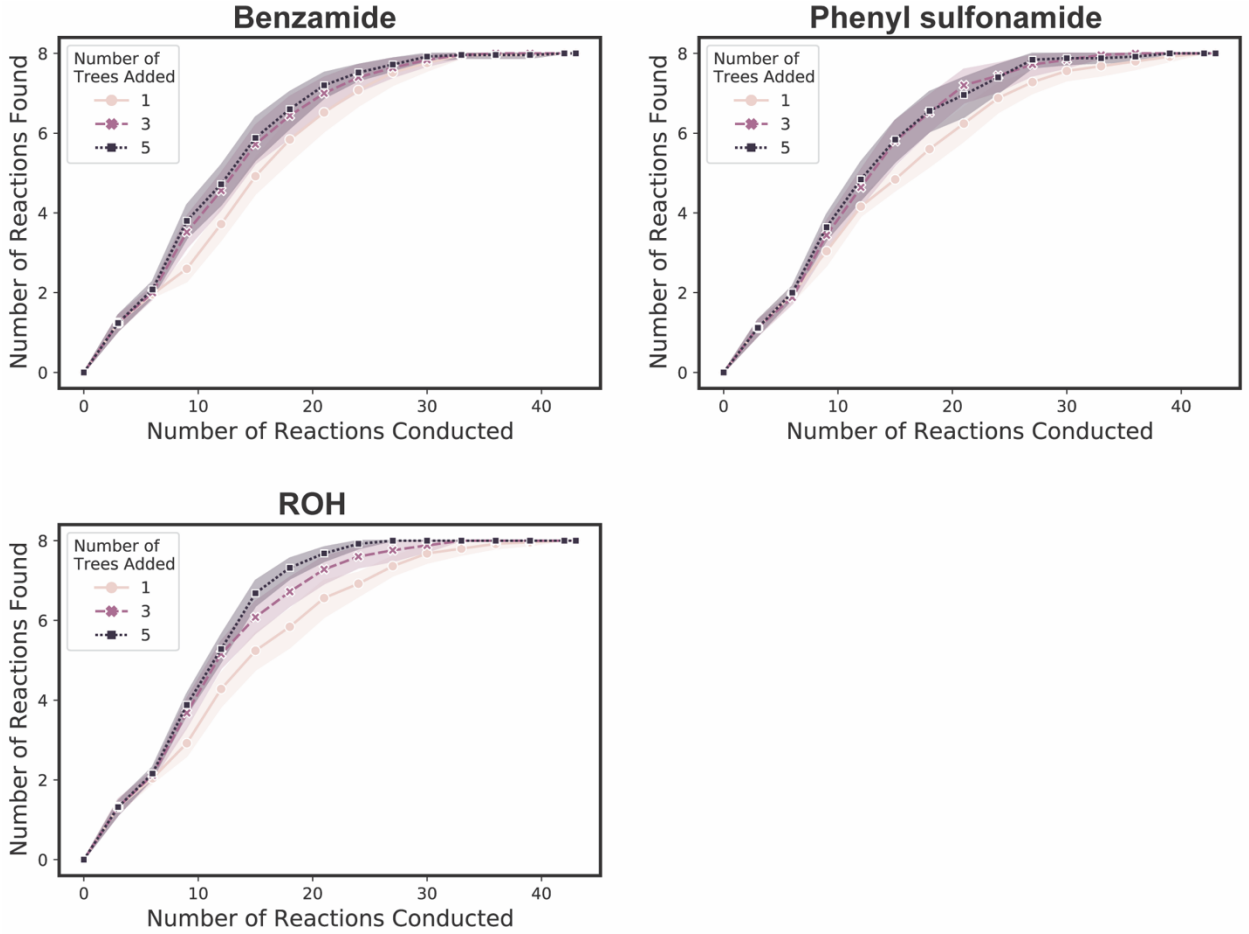



*Figure S15. Target tree growth results when different number of decision trees are added every iteration.*

Next, the addition of different numbers of decision trees were evaluated. Adding one target tree every iteration (pink curves) does not show as good performance as adding three (magenta curves). This is attributed to the insufficiency of using a single decision node to extract useful reactivity information. Adding five trees (black curves) does not give significant benefit compared to three, probably because only three reactions are collected in each iteration.

**ATL results of 'target tree growth' strategy with various reaction selection criteria**

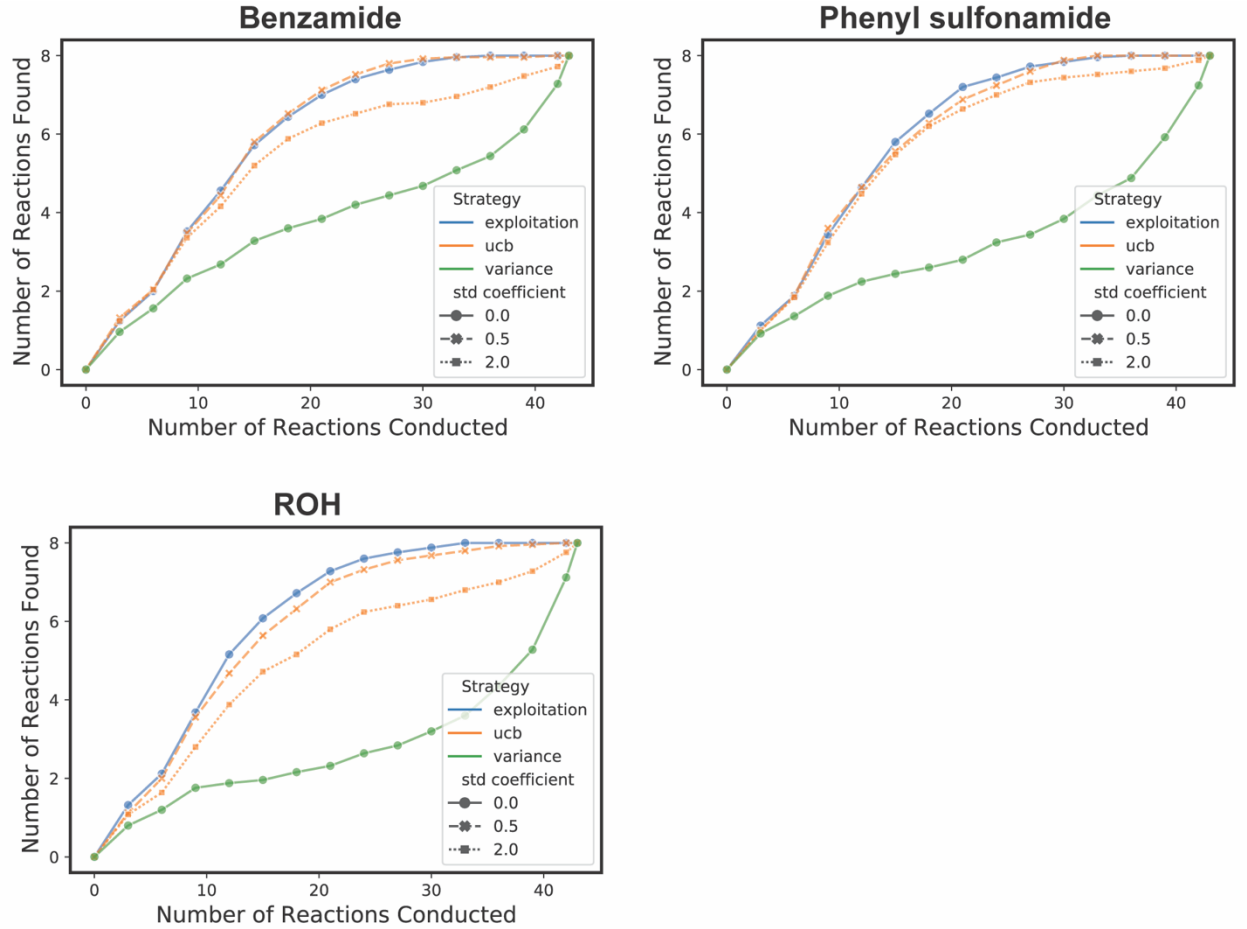



*Figure S16. Evaluation of various reaction selection strategies under the target tree growth strategy. Std coefficient of 0 for upper confidence bound (UCB) is equivalent to exploitation. Confidence intervals were not shown for clarity between exploitation and UCB.*

Next, various reaction selection strategies were compared. See Computational Details for a description of the strategies. As the objective is to find datapoints with a certain label rather than reducing the error, selecting uncertain data points (Figure S16, green curves) in the active learning iteration is not beneficial. Upper confidence bound (UCB), which combines exploitation and uncertainty of the model, seems to perform (Figure S16, orange curves) on par with pure exploitation (Figure S16, blue curves) when the coefficient on variance is 0.5. This is partially due to the small dataset that the model can explore. Higher coefficient of 2 was not as beneficial probably due to the same reason as the ineffectiveness of selecting uncertain datapoints.

**Effect of using source random forest models with more or deeper trees on target tree growth ATL**

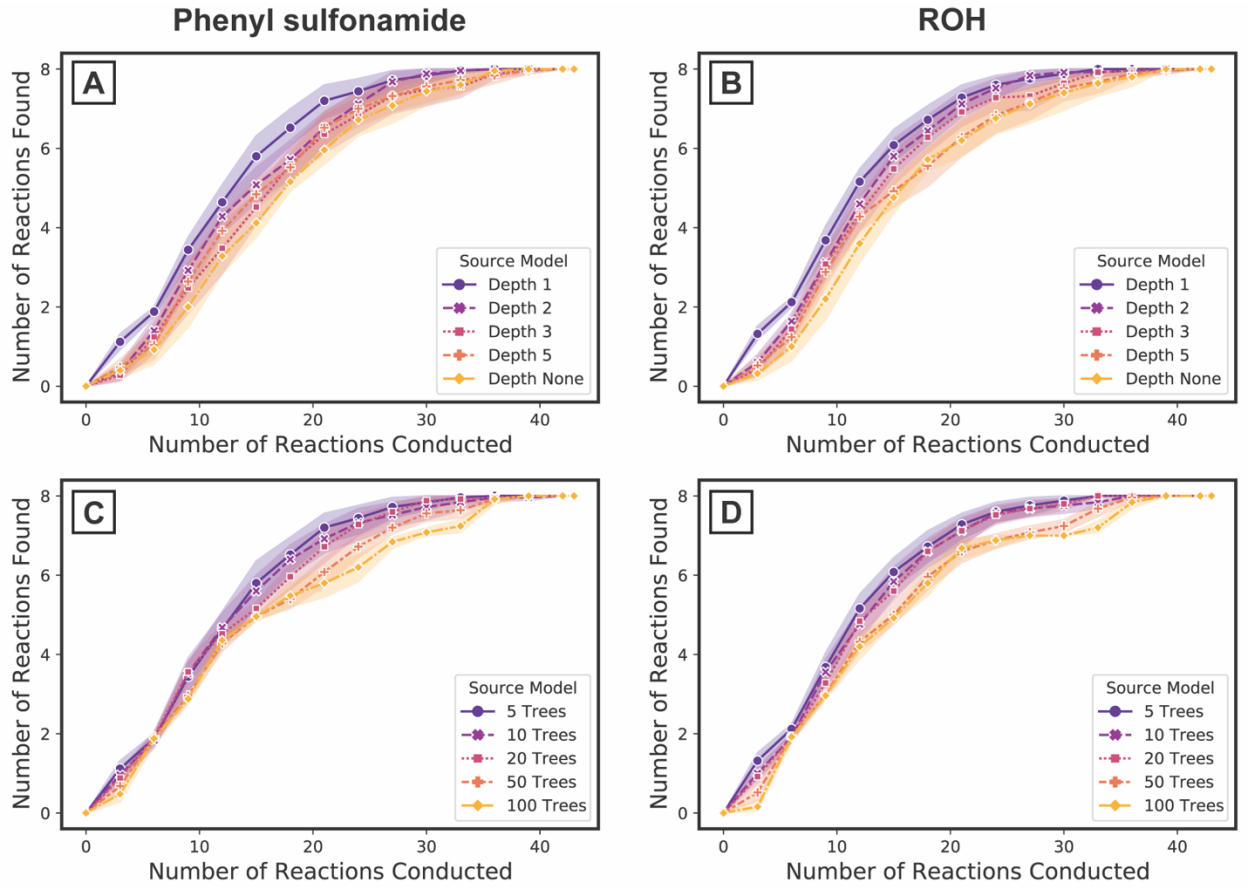



*Figure S17. Evaluation of ATL performance with different source model hyperparameters. Left column corresponds to phenyl sulfonamide as source, while the right column corresponds to ROH as source. (A) and (B) starts with the source model with five decision trees, where their maximum depth differs. The decision trees in the source models of (C) and (D) are limited to maximum depth of one, differing in the number.*

As in the Discussion section (Figure 6), the importance of source model simplicity is evaluated for phenyl sulfonamide and aliphatic alcohols. Similarly, source models that are more complex than five trees of maximum depth one does not show better target tree growth performance. While the extent to which the number of trees affects performance is greater for phenyl sulfonamide (Figure S17C, after 5$^{th}$ batch) and ROH (Figure S17D, after 5$^{th}$ batch) than benzamide as source, it is not as profound as having different maximum depth values.

The greater impact of the maximum depth of the trees can be explained by the way predictions are made. When random forests make a prediction on a datapoint, the average of probability values of the leaf nodes that the datapoint arrives in each decision tree is computed[2]. These probability values are the ratio of the labels of the *training data* that arrived at leaf nodes. Therefore, a deeper tree, which is more likely to have a purer composition of labels at leaf nodes due to higher probability of overfitting, will have higher probability values. Ultimately, the contribution of the unaltered source model to the overall

probability value is larger, making it harder to adapt in the target reaction space, as the source model has no knowledge of the target reactivity.

### Effect of combined source data on target tree growth ATL

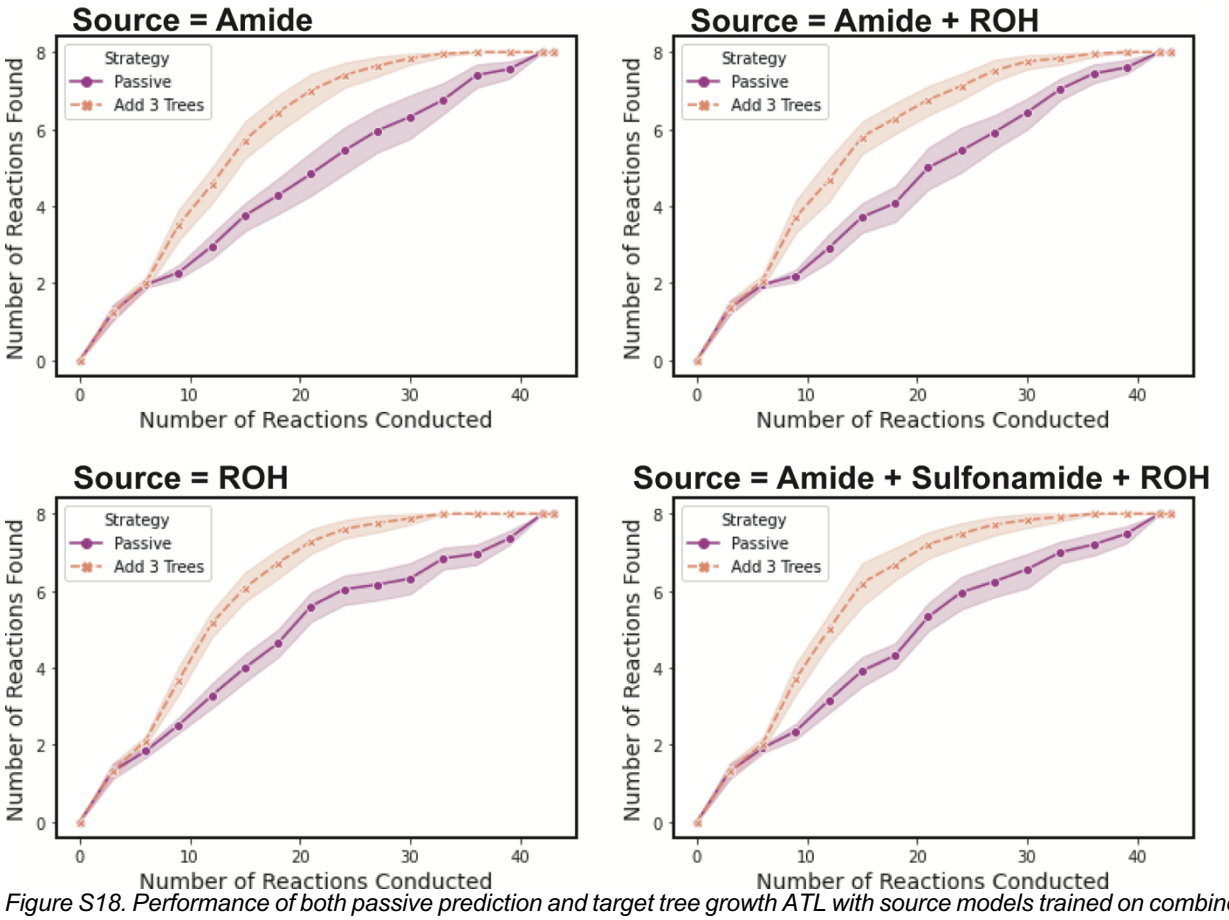



*Figure S18. Performance of both passive prediction and target tree growth ATL with source models trained on combined datasets shown in the title of each plot.*

When there are multiple source datasets with similar screened reaction conditions, a better predictivity can be expected. Accordingly, target tree growth ATL was conducted on source models trained on combined datasets of amide, sulfonamide and ROH. However, a simple concatenation of datasets does not seem to result in better performance than the best performing single model (ROH).

### Comparison of target tree growth ATL with AL strategies

In the main text and the Supporting Information up to this point, analyses were centered around pyrazole as the target nucleophile. To understand the overall benefit of the target tree growth ATL in various scenarios, it was compared with the two AL baselines between all source-target pairs of the six nitrogen-based nucleophiles. In contrast to trellis of plots up to this point (e.g. Figure 3B or Figure S9), rows and columns

of Figure S19 correspond to the target and source, respectively. 100 randomly initiated models (5 trees of depth 1) were trained and transferred.

In all tests, the target tree growth ATL outperforms AL with no transfer, while the degree of benefit differs between target nucleophiles. When compared to AL on combined source and target data, the number of reactions found is nearly the same at the first iteration. This is explained by the dominance of the source data at the early stage for both strategies. From the second iteration, two trends are observed. When over 1/3 of the target candidates are positive (i.e. one can expect one positive reaction just by sampling three random reactions, when sulfonamide and ROH are targets), the performance of the target tree growth ATL slightly trails the latter baseline. More importantly, however, for challenging cases where less than 20% of the candidates are positive and applying the positive reaction conditions from the source data would be unsuccessful (except amide, where 2/5 positive reaction conditions are also positive in source), target tree growth ATL outperforms the AL baseline. Additionally, even though the aniline source models used for ATL, without any updates, transfer poorly compared to CV-determined models (worse than that listed in Table S3), ATL allows efficient exploration for these challenging cases (Figure S20, leftmost column).

Based on this observation, an overall summary of the target tree growth ATL's benefit over each AL baseline is provided in Figure S21. The portion of reactions found by each strategy was compared to the maximum number of positive reactions that could be found up to that iteration. At each iteration, a Friedman test, followed by a Holm-Dunn post-hoc test, was performed to evaluate the statistical difference in performance, of which the p-values are presented in Tables S4 and S5. These results demonstrate how effective the target tree growth method could be in various scenarios, especially being significantly more efficient in finding good reaction conditions for the most challenging reaction types.

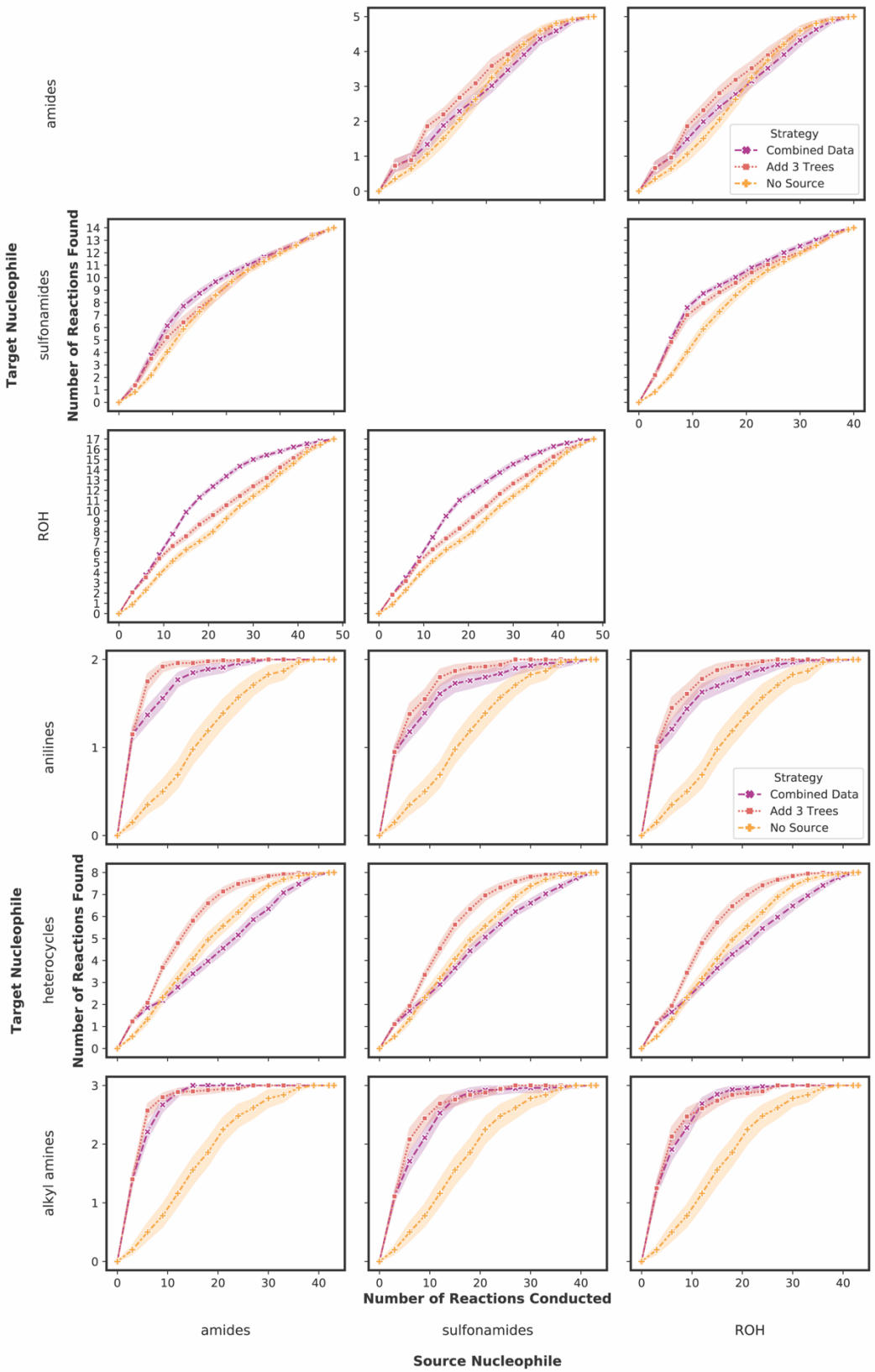



*Figure S19. Trellis of performance plots of ATL vs AL baselines. Source nucleophiles are amide, sulfonamide and ROH, from left to right. Target nucleophiles are amide, sulfonamide, ROH, aniline, pyrazole and alkyl amines from top to bottom.*

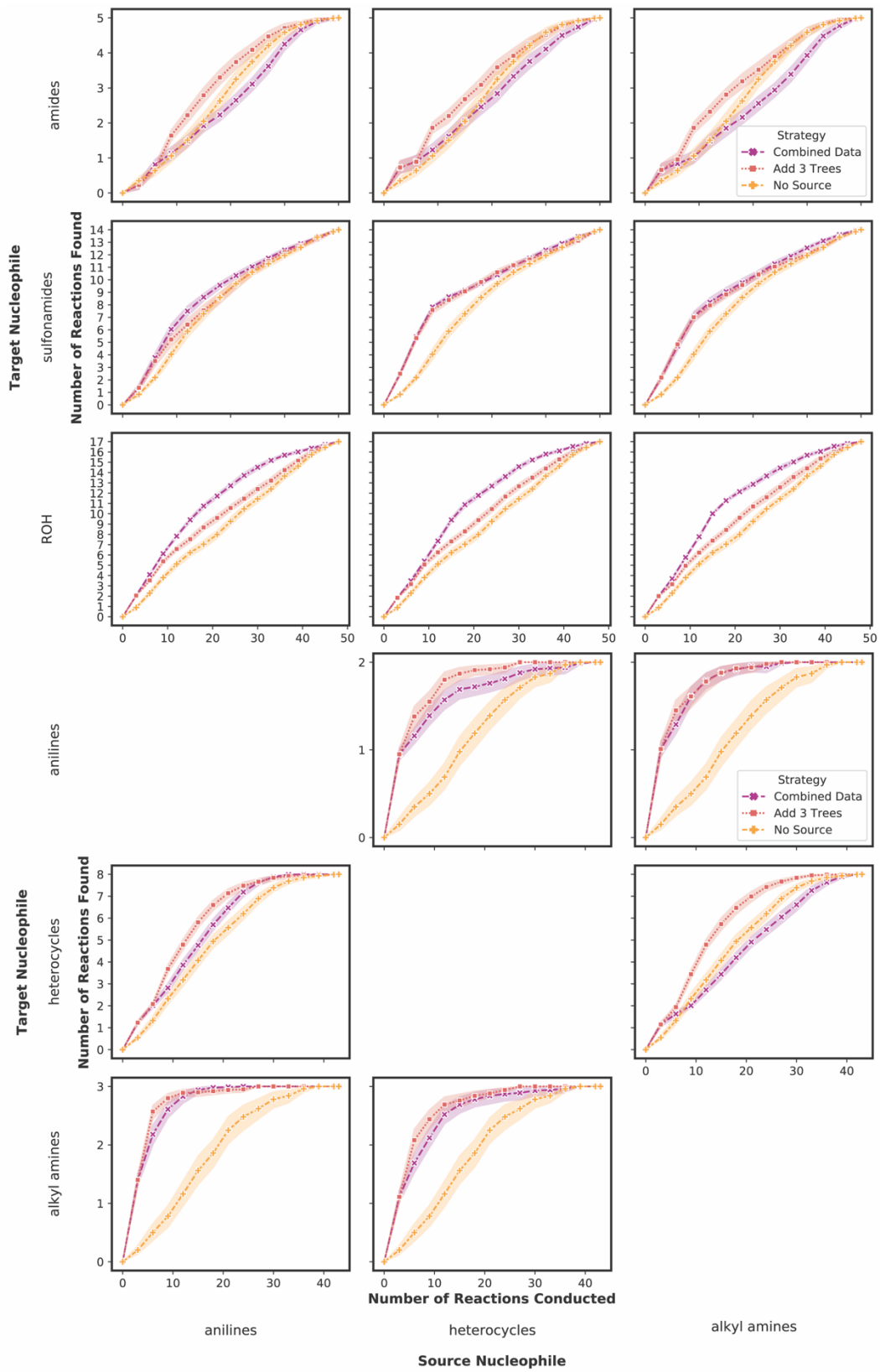



*Figure S20. Trellis of performance plots of ATL vs AL baselines. Source nucleophiles are aniline, pyrazole and alkyl amines, from left to right. Target nucleophiles are amide, sulfonamide, ROH, aniline, pyrazole and alkyl amines from top to bottom.*

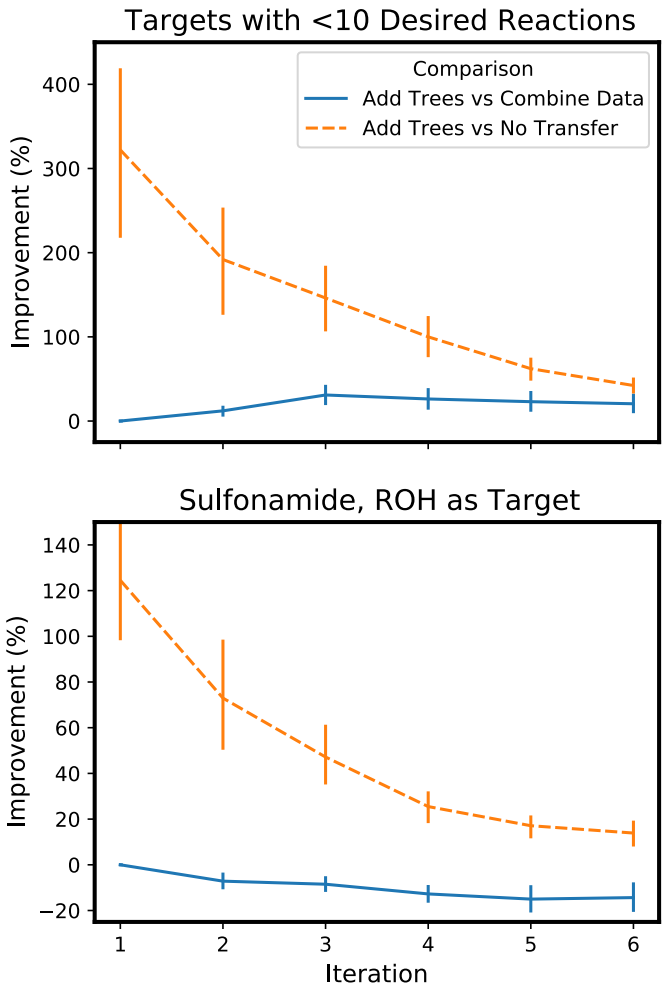



*Figure S21. Overall improvement on the portion of number of reactions found by a strategy, compared to the maximum number of reactions that can be found up to each iteration. (Top) Analysis conducted on four target nucleophiles where the portion of positives < 20%. (Bottom) Analysis conducted on sulfonamide and ROH targets, where the portion of positives > 33%.*

| Strategy | Iteration | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| Combined Data | 1.00 | 0.390 | 0.030 | 0.020 | 0.039 | 0.035 |
| No Transfer | 6.87e-7 | 4e-6 | 1.60e-7 | 9.96e-7 | 1.1e-5 | 4.8e-5 |

*Table S4. P-values resulting from the Holm-Dunn post-hoc test on the portion of number of reactions found by a strategy, compared to the maximum number of reactions that can be found up to each iteration, from the four targets with eight or less desired reactions. The values are compared to target tree growth ATL, which performs best as shown in Figure S21 (top).*

| Strategy | Iteration | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| Combined Data | 1.00 | 0.358 | 0.153 | 0.153 | 0.027 | 0.035 |
| No Transfer | 3.61e-4 | 1.50e-3 | 3.69e-3 | 3.69e-3 | 0.102 | 0.193 |

*Table S5. P-values resulting from the Holm-Dunn post-hoc test on the portion of number of reactions found by a strategy, compared to the maximum number of reactions that can be found up to each iteration, from sulfonamide and ROH targets with >33% positives. The values are compared to target tree growth ATL, which trails combined data but outperforms no transfer (Figure S21 bottom).*

## Additional examples of target tree growth ATL – 1

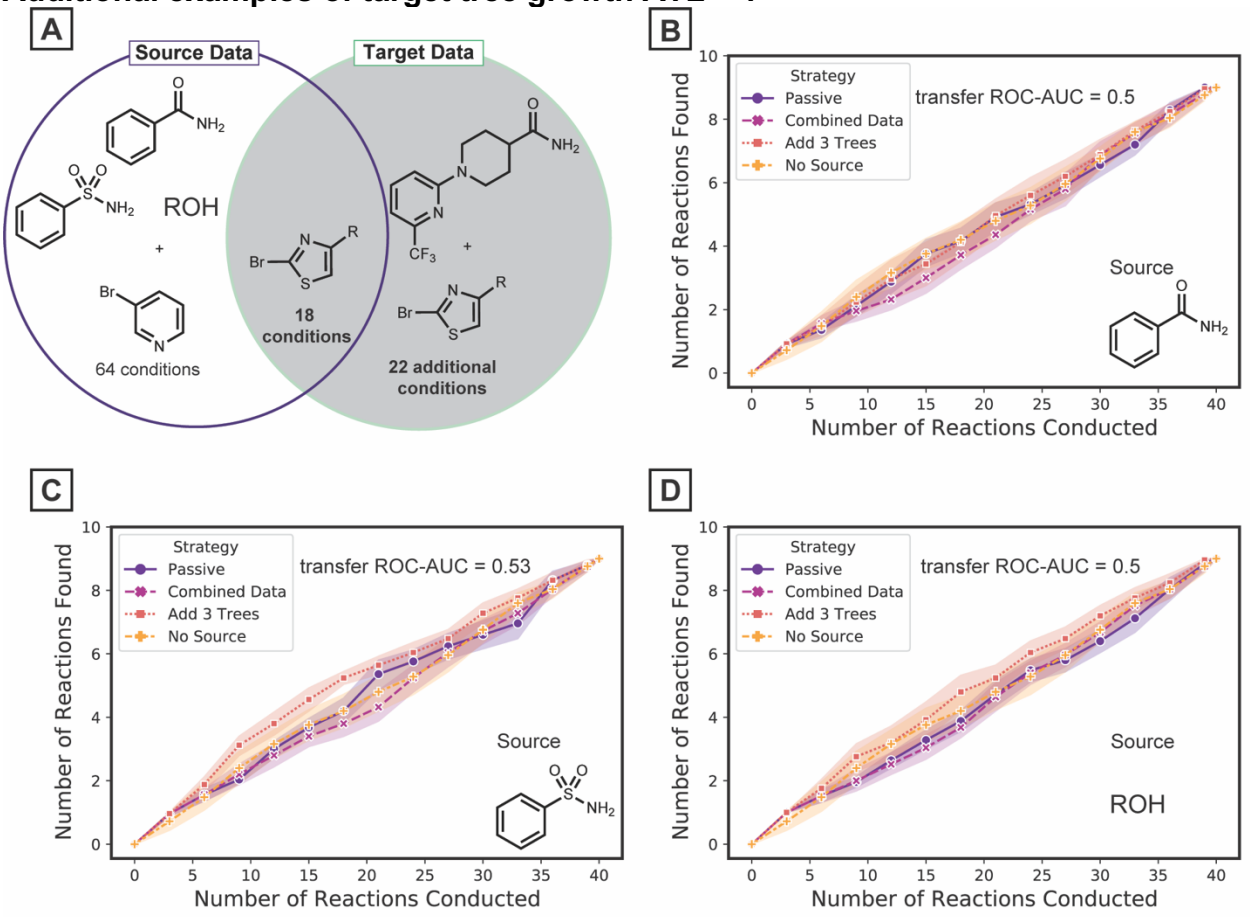



*Figure S22. (A) Problem setting for ATL. Three nitrogen source nucleophiles are all considered, while the target is now a complex aliphatic primary amide with 40 reaction condition candidates of which 9 have desired outcomes. (B), (C) and (D) correspond to ATL results with respective source data. The source model has five trees of maximum depths of one.*

Additional source and target pairs are examined to further investigate the applicability of our target tree growth strategy. Here, the same three source nucleophiles as the main text and 40 reactions of a complex aliphatic primary amide are considered as target. As in the main text, three reactions were labeled every iteration, and the shades in each plot show the 95% confidence interval from 25 different initial source models.

For benzamide, the target tree growth strategy did not show significant benefit over other baselines. Compared to when source data was not utilized, target tree growth showed a slightly inferior rate of desired reaction condition identification (Figure S22B, red vs. orange curves). For sulfonamide, there is a benefit of finding one more reaction at the mid-stage of the exploration ($3^{rd}$~$6^{th}$ iteration), compared to other baselines. Lastly, for ROH, the advantage of the target tree growth model is minimal.

Two explanations can be made for the reduced benefit of the target tree growth ATL compared to having pyrazole as target. First, the target nucleophile here is more complex in structure compared to the source nucleophiles. The source model therefore cannot address the deviation of reactivity that may arise from the structural differences, such as the nitrogen atoms that are not part of the amide functional group. Pyrazole, in contrast, is among the simplest structures within their classes, letting the source model

adapt relatively easily. Second, the reactivity of the target nucleophile is hard to model, as maximum cross-validation ROC-AUC that can be achieved with models trained on the 40 target reactions is 0.64 (not shown, available in jupyter notebook). This means that adding decision trees trained on three reactions may not be sufficient to make effective predictions on the remaining reactions.

### Additional example of model transfer and target tree growth ATL - 2

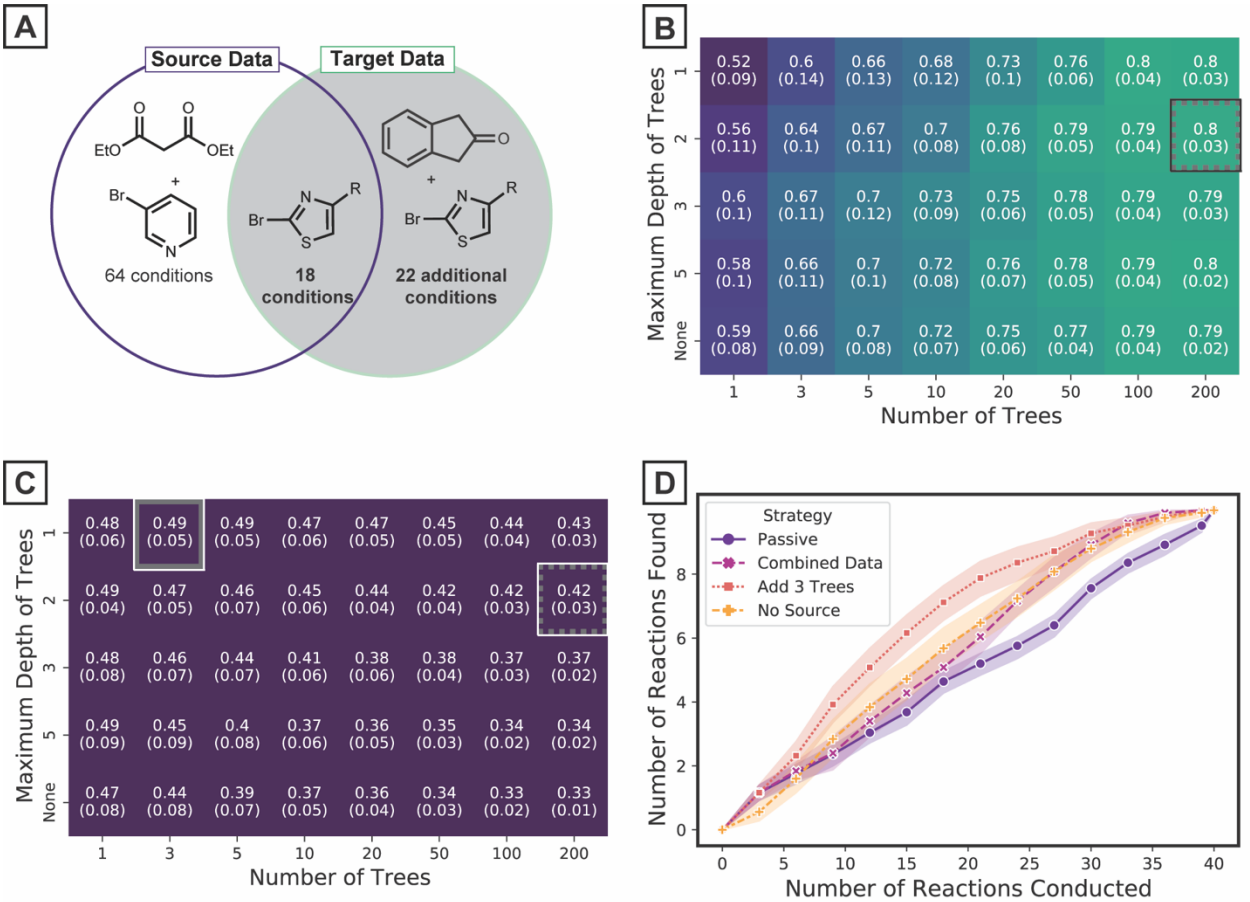



*Figure S23. (A) Problem setting for transfer from diethyl malonate to 2-indanone. 10 out of 40 reactions give positive yields. (B) Cross-validation results for modeling with malonate data. The optimal performance is highlighted with the dashed red square. (C) Model transfer performance of malonate models on the 40 reactions of 2-indanone. Dashed red square corresponds to the combination of hyperparameters determined by cross-validation. Solid square denotes simpler hyperparameters yet achieving comparable transfer performance. (D) ATL results with source models of three trees of depth one.*

Both model transfer generalizability and ATL were evaluated for C-C coupling reactions. The datasets were prepared in a similar manner to what was considered in the main text (Figure 4A), where the diethyl malonate source data was composed of the same 82 reaction conditions as the three nitrogen nucleophile source datasets. For candidate target reactions, 40 reaction conditions for the coupling between 2-indanone and 2-bromothiazole was used (Figure S23A). As 10 out of 40 reactions showed >0% yield for the target, *we look for positive reactions*, unlike the case where aniline, alkyl amine or pyrazole was considered as target.

Cross-validation determines random forests of 200 trees of maximum depth of two as optimal source models (Figure S23B, dashed grey square). Simpler models, with three trees of depth one, perform better when transferred (Figure S23C, solid grey square). For consistency, we use the source models of five trees of depth one for the ATL experiment.

From the third iteration, using no source data outperforms both the passive use of the source model and ATL on combined source and target data (Figure S23D orange vs. purple and magenta curves), suggesting the importance of proper target information update. Like the results observed in the main text (Figure 5A), the target tree growth strategy (red curve) accelerates the rate of hits compared to the other two baselines. This result is surprising because the transfer ROC-AUC of the source model is 0.49, slightly worse than random guessing. It contrasts with the previous examples, where source models with transfer ROC-AUC of 0.5 and 0.53 did not achieve a similar boost in rate (Figures S22B and D) through target tree growth. From the molecular structure perspective, the target nucleophile does not have functional groups other than the reacting center that might undergo side reactions. Also, the benzylic carbon, which is the reaction center, is stabilized by the carbonyl and phenyl rings, which makes it relatively similar to the source nucleophile. Thus, we conclude that the chemical and structural relevance is important for the ATL strategy to succeed.

## Additional example of model transfer and target tree growth ATL - 3

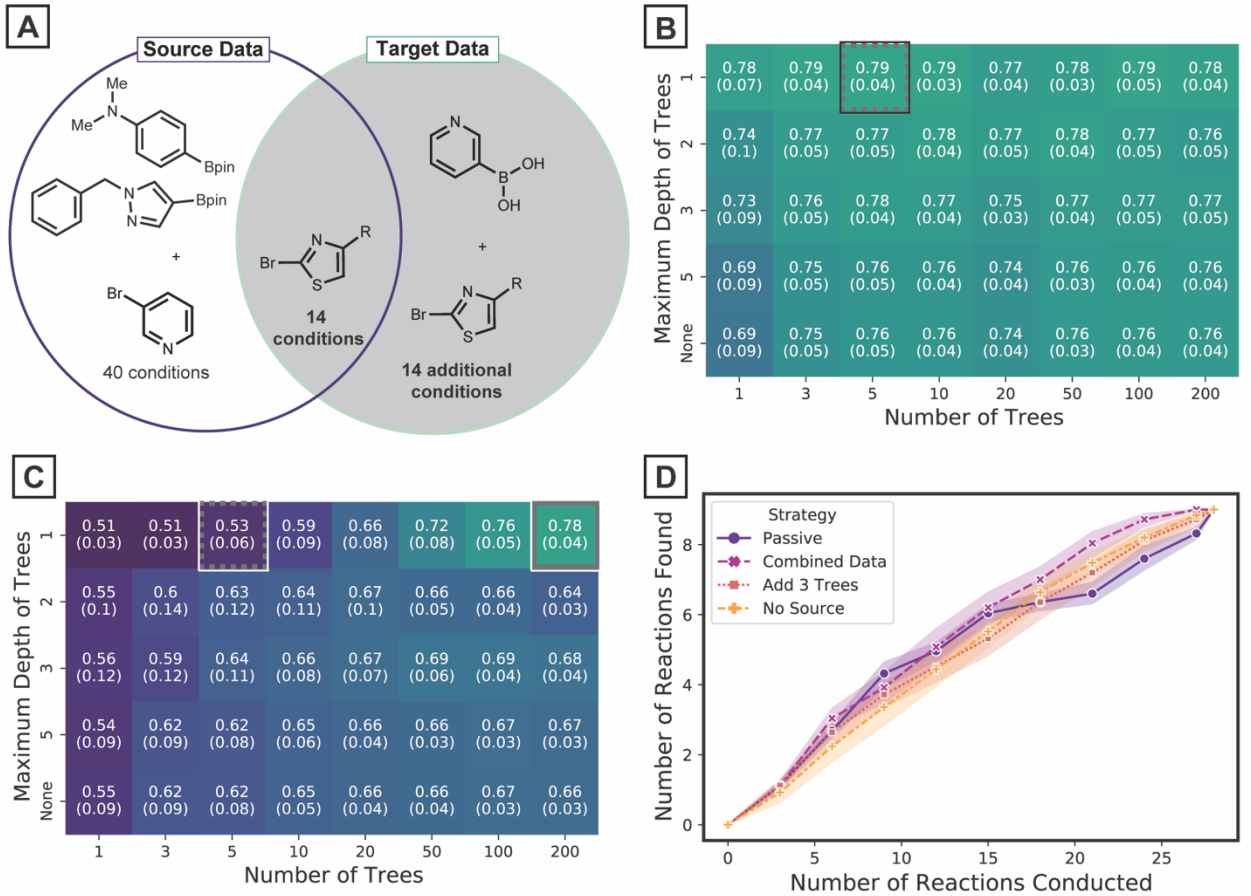



*Figure S24. (A) Problem setting for transfer from pinacol boronates to 3-pyridine boronic acid. (B) Cross-validation results for modeling with pinacol boronate data. The optimal performance is highlighted with the dashed grey square. (C) Model transfer performance of pinacol boronate models on the 28 reactions of 3-pyridine boronic acid. Dashed grey square corresponds to the combination of hyperparameters determined by cross-validation. Solid grey square (top right) denotes hyperparameters that achieve optimal transfer performance. (D) ATL results with source models of 5 trees of maximum depth of one.*

Lastly, we consider the transfer scenario between two nucleophile types that undergo Suzuki coupling. The source data consists of 40 reactions between 4-(N,N-dimethylamino)phenylboronic acid pinacol ester (top nucleophile under source data in Figure S24A) and 3-bromopyridine and 14 reactions between 1-benzylpyrazol-4-boronic acid pinacol ester (bottom nucleophile under source data in Figure S24A) and 2-bromothiazole. Notably, 39 of 40 reactions are successful for the former substrate pair. Unlike previous cases, there are 28 target reactions between pyridin-3-ylboronic acid and 2-bromothiazole, of which 9 show positive yields. Therefore, we look for positive yielding reactions.

Cross-validation with 54 source pinacol boronate data points reveals source models with five trees of maximum depth one to be optimal for the source data (Figure S24B, dashed grey square). Cross-validation scores do not differ between models with different hyperparameters, probably due to the extreme yield label distribution in one substrate pair of the source data. This also results in source models failing to extract

information on reactivity that is transferable. As such, highest transfer performance was achieved with a large number of trees (Figure S24C).

Nevertheless, ATL experiments were conducted with source models of 5 trees of maximum depth one. This case, various strategies show similar performance. Specifically, strategies that use source models are slightly more efficient than no transfer, up to the 4$^{th}$ iteration, probably due to the information from the reactions of source nucleophile with 2-bromothiazole. The target tree growth model (Figure S24D, red curve) does not exceed other baselines that involve model transfer. Two aspects may be playing a role. First is the misleading guidance of the source model as it does not include meaningful information on reactivity from biased data. Second is the difficulty of modeling the target data (as can be seen in jupyter notebook 'complexity_vs_transfer', cross-validation scores when training with the target dataset is below 0.75). Also, the target dataset considered here consist of ~33% positives, reminiscent of what was observed in Figures S19 and S20 when sulfonamide or ROH was target - the portion of positives were ~35% and target tree growth strategy could not outperform AL on combined data.

In summary, the additional examples imply the importance of structural similarity between the source and target nucleophiles for the target tree growth model to be beneficial. While not definitive, given mechanistic similarity and structural similarity, ATL seems to perform better than the baselines considered in this study on more challenging cases where the ratio of positives are smaller. Lastly, the use of ATL with biased source data should be applied with caution, as the source model may fail to learn meaningful information.

The numpy arrays of descriptors and corresponding yield labels of the three target nucleophiles used in the Supporting Information are available on github[1] as separate joblib files.

## References

1. https://github.com/ZimmermanGroup/ActiveTransfer.

2. https://scikit-learn.org/stable/modules/ensemble.html#forest (accessed October 27, 2021).