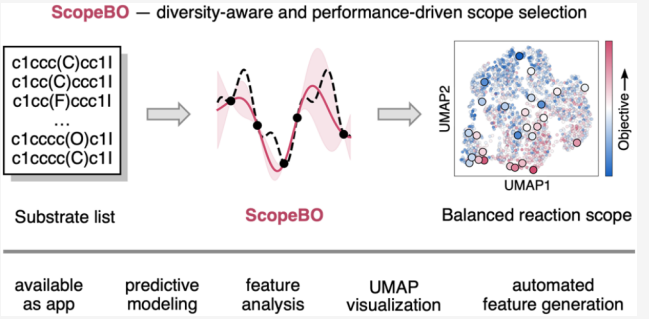
JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
pubs.acs.org/JACS
Article

Machine Learning-Guided Scope Selection to Balance Performance and Substrate Similarity

Sven Roediger, Matthew S. Sigman, and Abigail G. Doyle*

Cite This: https://doi.org/10.1021/jacs.6c10981
Read Online
ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The determination of a reaction substrate scope enables downstream users to assess whether a specific transformation is suitable for an intended application. The information content of this scope, characterized by its performance and structural diversity, is crucial to ensuring the quality of such decisions. Despite their importance, scopes have historically been evaluated subjectively, leading to well-documented selection biases that compromise the scope quality. Herein, we report a scope score to analyze scopes quantitatively within a specific chemical search space, enabling the benchmarking of common selection methods. Additionally, we developed a broadly applicable and user-friendly machine learning algorithm, ScopeBO, for selecting scopes that effectively balance substrate performance and diversity. The scope score is used to optimize the hyperparameters of ScopeBO and validate its performance across several reaction data sets. Using statistical tests, we found that ScopeBO outperforms all other selection methods considered. Consequently, ScopeBO provides an objective and standardized approach to scope selection that maximizes the information content of the evaluated substrates, addressing a long-standing challenge in reaction development.



INTRODUCTION

Organic chemistry has advanced the synthesis of medicinal compounds, materials, and polymers through the development and continuous improvement of numerous chemical transformations. Three indispensable components of establishing a new chemical transformation are the discovery, optimization, and evaluation of the reaction scope. Out of these, scope exploration provides a guide for future users by testing the reaction on a suite of substrates. Consequently, there are two aspects to a substrate scope: (1) it highlights *how well* the reaction works by featuring high-performing substrates, and (2) it demonstrates *what* the reaction can be used for by illustrating the breadth of suitable and incompatible substrates.[1,2]

Historically, reaction scopes mainly focused on showcasing substrates with high performance, evidenced by many publications mentioning the high average yield of a reaction scope in the manuscript abstract. At the same time, there has been a continuous increase of scope sizes, driven in part by a desire to highlight the range of substrates amenable to the reaction.[3] However, due to the bias toward high-performing substrates and the lack of metrics for evaluating the diversity of a scope, this "more is better" approach may result in evaluating many structurally similar examples while low-performing substrates are only sparsely reported.[3] Large scope sizes therefore may not realize their intention and often lead to an increase in labor with only marginal gains in information regarding the reaction under study. In recent years, these trends have been criticized as they lead to a misrepresentation of the reaction behavior, which makes it challenging to predict if the method can be applied to a substrate outside of the reported scope. In addition to this direct impact on future users, the selection bias in reaction scopes also limits the usefulness of literature data for machine learning purposes.[1,4−7]

To address these limitations, data-driven approaches for scope selection[8−10] and related chemical library design[11−18] have been reported that aim to maximize the diversity of the experimentally evaluated compounds. In the context of reaction scopes, these strategies are based on defining a large number of possible scope examples and quantifying their molecular structure and properties using a featurization method. Clustering techniques are employed on the resulting data sets to group similar compounds, and representatives are selected from each of these clusters to constitute the scope (Figure 1C).[8,9] However, the clustering algorithm only considers the molecular featurization during the selection process and has no insights into the reaction performance, which means that the approach is fully agnostic to reactivity. While reducing bias in selection, this strategy may deliver scope outcomes that are not

Received: May 28, 2026
Revised: August 24, 2026
Accepted: August 25, 2026

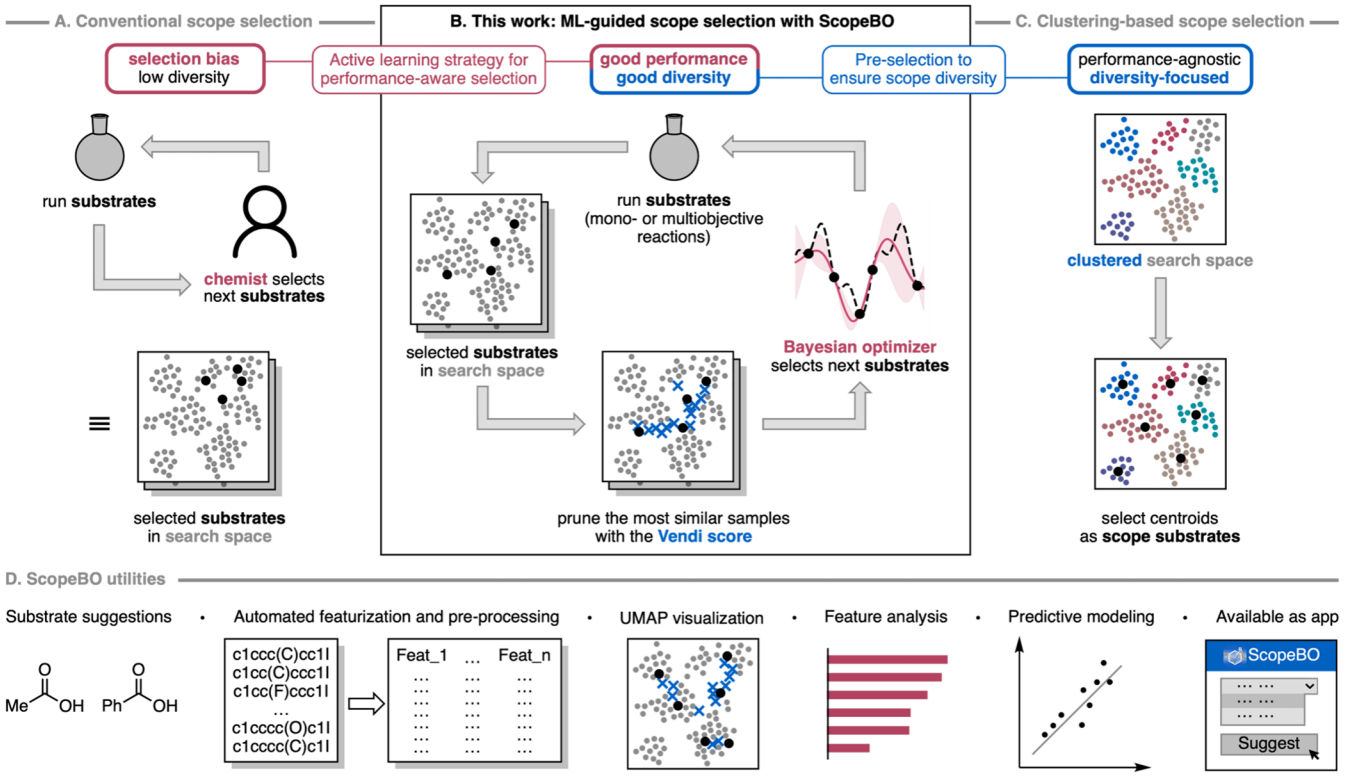


Figure 1. Overview of scope selection methods. (A): Conventional scope selection. (B): This work: ML-guided scope selection with ScopeBO. (C): Clustering-based scope selection. (D): Functionalities of the ScopeBO code package. Feat = feature. ML = machine learning. UMAP = uniform manifold approximation and projection.

representative of the reactivity found in the underlying data set even if representative of the structural variety. To summarize these contrasting approaches, human-selected reaction scopes are typically highly biased toward describing *how well* the reaction works, whereas clustering-based approaches strongly focus on *what* the substrate range is. A general approach that merges these two fundamental goals of reaction scope evaluation has not yet been reported. Additionally, quantitative evaluation metrics for scopes are required to objectively assess and compare the growing number of selection methods.

Here, we describe a scope score for the quantitative post hoc analysis of substrate scopes within a fixed chemical space. The score is leveraged to benchmark different scope selection methods across a variety of computational and experimental data sets. It is also used to optimize the hyperparameters of ScopeBO, a new user-friendly BO-based algorithm that overcomes current challenges by combining the advantages of existing approaches (Figure 1B). In particular, ScopeBO uses iterative performance-aware active learning on the substrate space with concurrent pruning of the space to remove candidate substrates that would least enhance scope diversity. ScopeBO illustrates improved performance compared with state-of-the-art approaches across different reaction classes, highlighting its broad applicability as a general tool for standardized scope selection. We provide the tool as a flexible Python code that requires only lists of possible substrate choices to provide scope suggestions. The framework further includes several utility functions such as automated feature generation and preprocessing, visualization of the proposed substrates, SHAP feature analysis[19] to gain feature importance based on the selected scope, and predictive modeling of reaction outcomes for unseen substrates (Figure 1D). All functions of ScopeBO are also available in a code-free manner via an app.

METHODS

Algorithm Development

For the development of the algorithm, we were inspired by the similarities between the way humans choose scope substrates (Figure 1A) and the manner in which experiments are selected in Bayesian optimization (BO), a machine learning technique that we,[20−22] among others,[23−27] have previously adapted for reaction optimization. In both situations, prior reaction outcomes inform the iterative selection of subsequent experimental batches. The chemist picks new scope entries based on the performance of the examples previously performed. In BO, a surrogate model is trained on existing data to estimate the underlying distribution of reactivity for untested samples, and an acquisition function uses the surrogate model predictions to determine the next batch of experiments.[28,29] These processes are then iterated until the scope or optimization target, respectively, are reached. BO is typically utilized to find the single best target sample in a predefined search space, but based on these parallels, we hypothesized that it could also be employed to find a collection of samples to define a reaction scope. While we thus envisioned that a BO acquisition function could enable the selection of high-performing samples, we posited that infusion of the BO algorithm with a preselection process, as in the clustering approaches, could facilitate the selection of diverse compounds in order to balance performance and diversity in the final scope. The clustering methods represent a one-shot preselection of substrates and are therefore not suitable to be combined with the iterative nature of the BO algorithm. Instead, we required a strategy that would allow us to infuse higher diversity in every round of the substrate selections. Here, we identified the Vendi score,[30] a diversity metric for machine learning developed by Friedman and Dieng that evaluates the similarity of data points in a set (see SI Section 2.2.4 for the score definition). The score can be understood as the effective number of fully dissimilar samples

https://doi.org/10.1021/jacs.6c10981
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

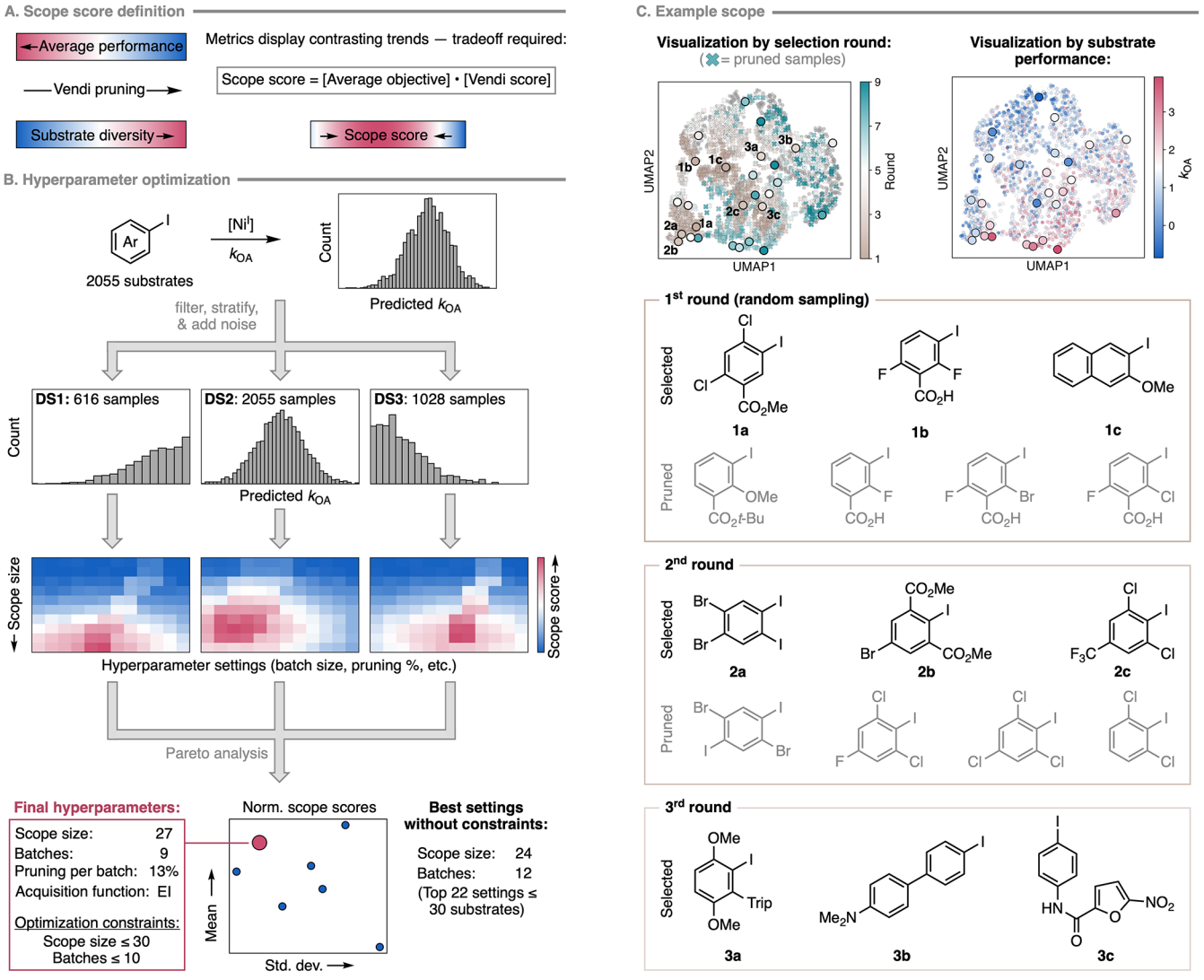


Figure 2. Investigations using aryl iodide oxidative addition data sets. (A): Scope score definition. (B): Hyperparameter optimization strategy. Plots have been simplified for illustrative purposes. Full optimization details and data are available in the SI, chapter 4. (C): UMAPs and substrates for an example scope using **DS2**. Selected samples are shown with large circles. The chosen scope is the one with a scope score closest to the average of all initiations. EI = expected improvement; OA = oxidative addition; Std. dev. = standard deviation; Trip = 2,4,6-tri-*iso*-propylphenyl.

that have the same diversity as the samples in the evaluated set. To increase the diversity of the selected samples in a BO campaign, Vendi scores could be determined for every set consisting of a single unseen sample and all prior scope entries. The samples whose sets gave the lowest Vendi score, meaning the lowest diversity, would be pruned from the search space during each BO iteration so that the remaining possible selections would be less similar to the already evaluated scope entries.[31] Through this preselection, the acquisition function is forced to explore the chemical space more broadly instead of solely focusing on high-performing regions. A proof of concept of this Vendi pruning strategy for BO, termed VendiBO, was recently reported by Dieng and Gómez-Gualdrón for the discovery of promising metal−organic framework (MOF) candidates.[32] In this study, the authors focused on finding a single high-performing MOF, whereas we envisioned defining a collection of samples to evaluate a scope. Furthermore, VendiBO has several characteristics that are challenging to directly adapt for scope selection such as the lack of human-in-the-loop capabilities, no compatibility with multiobjective optimization, and hard-coded settings that could limit use by nonexperts. We envisioned a framework that could be adapted to any reaction scope and, therefore, required easy adjustment of settings such as the data set and the objective choices (*e.g.*, yield and enantioselectivity). EDBO+, a BO reaction optimization algorithm previously developed by the Doyle group,[21] fulfills these criteria and was therefore chosen as the foundation for developing the scope selection algorithm, ScopeBO.

Scope Score Definition and Hyperparameter Optimization

After establishing the algorithm architecture by incorporating the Vendi pruning step and further functionalities, we set out to optimize the hyperparameters of ScopeBO, such as the batch and scope size as well as the level of Vendi pruning in each batch. We selected a computational data set from Sigman and Doyle on predicted oxidative addition rates for aryl iodides (ArI data set, Figure 2B) as a data set for the optimization.[33] We were drawn to this data set due to its single-reactant nature, which simplifies analysis, and its comparably large size (>2000 substrates).

To gain initial insight into the algorithm behavior, we selected scopes of 30 substrates with ScopeBO while varying the extent of the Vendi pruning. To measure the substrate performance, different metrics can be considered. We chose to use the average objective value (in this case, the oxidative addition rate) as it weighs all substrates equally and does not require additional parameters such as, *e.g.*, the analysis of the percentage of substrates performing over a certain threshold value. To determine the scope diversity, we measured the Vendi score. The result of these initial queries highlights the contrasting trends for the two metrics: increasing Vendi pruning leads to a decrease in the average

https://doi.org/10.1021/jacs.6c10981
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

oxidative addition rate but an increase of the Vendi score (Figure 2A). This illustrates that a suitable trade-off needs to be defined to determine an optimal scope with high values for both metrics.

To achieve this goal, a quantitative scope score was required that could be used for the post hoc analysis of selected scopes. We evaluated several definitions for such a score and found that the product of the performance (normalized average objective) and the diversity (normalized Vendi score[34]) provided the best results (see SI, chapter 4.4 for details). If only one of the two scope assessment values is low, then the scope score as the product will also be low. Therefore, this score definition emphasizes maximization of both factors in a scope, leading to a balanced scope design that highlights high performers but also explores reaction generality.

The Vendi score and, by extension, the scope score depend on the reactant search space, and they can therefore only be used to compare scopes that sample the same space but not those with different reactants. However, as we would only change algorithm parameters during hyperparameter optimization while keeping the search space constant, we posited that the scope score would be a suitable evaluation metric for this task.

We were cognizant that the scopes for different reactions could exhibit differences in the distribution of the objectives (*e.g.*, yield) as well as in the user-defined search space size. Indeed, systematic scope studies for varied subsets of the original ArI data set confirmed that these factors influence the algorithm performance (see SI, chapter 4.3). To determine algorithm hyperparameter settings that are universally applicable, we envisioned using parallel optimization of multiple data sets with distinct characteristics. Specifically, we curated three data sets from the original ArI one; data set **DS2** maintained the balanced objective distribution from the original data set (Figure 2B). Additional data sets were generated by partitioning the ArI data set in its top and bottom halves that resulted in bias toward low (**DS3**) or high performance (**DS1**) in the data distributions, respectively. For further differentiation, we used stratified subsampling on **DS1** so that all three data sets had different sizes (616, 1028, and 2055 substrates, respectively). As the objective values in **DS1**−**3** were predicted using a three-parameter linear regression model, we also added random noise to the objective labels to better simulate learning from noisier experimental data (see SI, chapter 3.1).

With the optimization data sets produced, we evaluated combinations of acquisition functions, batch sizes, pruning percentages and modes as well as scope sizes to explore the hyperparameter optimization (see SI, chapter 4.5 for details). Compounds were selected by ScopeBO using their reaction rate as the target objective, while the Vendi pruning step modulated the selection diversity. To ensure the convenience of employing the algorithm in future experimental campaigns, we limited the search to settings with a maximum of ten rounds of experiments and a maximum scope size of 30 samples. The scope score of the resulting scopes was determined per setting as an average across 20 random initiations of the ScopeBO. As expected from our initial experiments using different data distributions, a unique collection of hyperparameter settings resulted in the highest scope score for each data set. We therefore turned to Pareto optimization to identify the nondominated points using the three scope scores as hyperparameter optimization objectives. Among these points, we selected the hyperparameter settings that resulted in the highest mean normalized scope score and the lowest standard deviation, *i.e.*, the settings that consistently delivered high performance across the data sets. After a second round of optimization, a final Pareto analysis determined the optimal algorithm settings to involve a scope size of 27 samples with three samples selected by the expected improvement (EI) acquisition function per batch and 13% pruning per batch. Viewed across the entire scope selection, this is equivalent to the removal of almost 70% of all potential substrates, highlighting the strong influence of the pruning step on substrate selection.

Using **DS2** as an example, we then visualized one of the selected scopes on uniform manifold approximation and projections (UMAPs) (Figure 2C).[35] After the first three substrates **1a**−**c** (left UMAP, ocher circles) are determined randomly, the initial Vendi pruning is carried out, and the most similar substrates to the selected ones, in local proximity on the UMAP visualization, are removed from the search space (ocher crosses). In the second round, the performance-driven acquisition function selected two substrates (**2a**−**b**) that are similar to the top performer of the initial round (**1a**), but have not been pruned yet. The third substrate **2c** is selected in an unexplored portion of the chemical space. In the following rounds of the scope (light ocher to green), the pruned samples also align closely with the previously selected samples, enforcing the exploration of the remainder of the chemical space. In total, the samples are distributed across the majority of the search space with a greater focus on the southern section of the UMAP in which the average performance is higher than in the northern half (right UMAP). The scope selected by the algorithm therefore displays the envisioned trade-off between reaction performance and substrate diversity and performs well in both aspects. The representativeness of the scope is further highlighted by closely mirroring the objective distribution in the full search space (see SI, chapter 11). Structurally, the selected substrates contain a variety of functional groups and substitution patterns, further highlighting the diversity (see SI, chapter 9). These results also demonstrate that the algorithm can be used for different objectives besides the most typical scope objective, yield. Notably, we also established proof of principle for the compatibility with multiobjective scenarios as would be of interest in asymmetric reactions that simultaneously optimize for yield and enantioselectivity (see SI, chapters 8.7−8.9).

During optimization, we limited the scope size and the number of batches (*vide supra*) to ensure that the approach was experimentally feasible. We were curious about how the optimized settings would compare to the best possible settings that could be obtained without such constraints. In other words, if a researcher had unlimited resources and time, how many rounds of selection and total scope examples are optimal to maximize diversity and performance? Surprisingly, the top 22 settings with the highest mean normalized scope score in such an unconstrained hyperparameter optimization (see SI, chapter 4.6) all contained 30 samples or fewer with the preferred ScopeBO settings only slightly trailing the highest overall performer (24 exp.; 12 batches of 2 exp.) (scope score of 0.70 for ScopeBO versus 0.72 for the best settings). While the ideal scope size for other chemical spaces may differ slightly from the results obtained here, this result suggested that the trend toward larger scope sizes might not be the optimal strategy to achieve enhanced substrate diversity and that careful selection of the scope examples satisfies this goal while reducing the labor and material resources required for scope evaluation.

Additionally, as we optimized the algorithm by averaging different random initiations, the average objective (in this case, the oxidative addition rate) of the initial samples varied drastically between the scopes, which consequently also led to significant differences in the final scope scores. Therefore, we evaluated which initiations led to the highest scope scores and found that a high average objective in the first round of selections leads to a better worst-case performance (see SI, chapter 7). We therefore recommend evaluating three substrates that are anticipated to be high-performing as the initial scope entries and using these results to initiate the scope selection algorithm for further suggestions.

Finally, an important consideration in the algorithm performance is how the substrates are featurized using numerical descriptors. For the data presented, we used density functional theory (DFT) descriptors, obtained with Auto-QChem,[36] as prior studies have found that DFT features can lead to better results for BO in chemistry objectives.[20,21] However, since DFT calculations are computationally more expensive and less accessible than other simpler featurization approaches, we also tested the performance of RDKit descriptors,[37] Mordred descriptors,[38] Morgan fingerprints,[39] and Morfeus descriptors[40] at the GFN2-xTB[41] level of theory. This analysis indicated that DFT descriptors best describe the data (see SI, chapter 5). However, Morfeus descriptors, which are similar to those computed with DFT, although at a lower level of theory, performed comparably. In contrast to DFT calculations, Morfeus descriptors can be calculated on a personal computer within hours for a full data set and do not require access to a computational cluster. We therefore also implemented a Morfeus featurization method

https://doi.org/10.1021/jacs.6c10981
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

into ScopeBO to provide users with easy access to substrate features when DFT calculations are not available.

RESULTS AND DISCUSSION

Dependence on the Data Set Distribution

Encouraged by these results on the aryl iodide optimization data set, we tested the finalized algorithm on real-world experimental data sets using averages of 40 random initiations. First, we investigated an amide coupling data set reported by Liao and co-workers.[42] In their study, they disclosed coupling reactions for the synthesis of amides using 96 different reaction conditions. Of these, we selected three conditions that have different yield distributions with average yields of 63%, 40%, and 22% (Figure 3A). As the same products were formed using all three conditions, the Vendi scores can be compared between the data sets, and these data sets therefore present an ideal test scenario to evaluate the diversity of selections by ScopeBO when exposed to different data set distributions.

For baseline comparisons, we compared ScopeBO with Bayesian optimizers using different acquisition functions as well as the scope clustering algorithm reported by Doyle[8] (Figure 3B). In these comparisons as well as similar ones using other data sets (see SI, chapter 8 for full details), ScopeBO had a strong performance. In terms of the Vendi scores, it significantly outperformed all comparisons apart from a fully explorative algorithm. Notably, the diversity of the selected scope was also 48% greater than the diversity of the full search space. This indicates that ScopeBO highlights regions of the chemical space that are marginalized in the full data set. The average yield is only slightly below most other comparisons, resulting in the best trade-off of both metrics among the tested methods and the highest mean normalized scope score across the three data sets. While a statistical analysis did not find significant differences between the scope scores of ScopeBO and regular BO (same settings as ScopeBO apart from the absence of Vendi pruning) for these data sets, the Vendi scores obtained by ScopeBO (7.0 ± 0.2) were significantly higher and are also more consistent than the values for regular BO (6.1 ± 0.5). The Vendi pruning in ScopeBO therefore not only improves the diversity of the substrate selection but also ensures consistent scope diversity irrespective of the initial substrate choices and data set objective distribution. In contrast, the average yields of the scopes selected by ScopeBO are all slightly below the mean of the respective data set (0.3−3.4%), illustrating how this metric is strongly dictated by the underlying data set distribution.

When the selected scope entries are visualized on UMAPs (Figure 3C), the examples selected by ScopeBO encompass the chemical space more evenly than a simulation of conventional scope selection (simulated as greedy acquisition with 5% batch pruning). The examples selected by the clustering algorithm also encompass the space well, but the scope size is notably larger for this data set (37 versus 27 for ScopeBO) as it is dictated by the maximization of the silhouette score, a performance metric for clustering that is used in the scope size determination.[8] The Vendi score of the clustering scope (4.6) is however notably lower than that of ScopeBO (7.0 ± 0.2).[43] The clustering selection is based on geometric distances after a nonlinear UMAP transformation of the substrate features. This transformation is known to distort distances, and consequently similarity, between compounds, especially between clusters.[44] The UMAP/clustering analysis should therefore be regarded as a rough qualitative assessment for search space analysis, but appears less suitable as a quantitative tool for diverse selection.

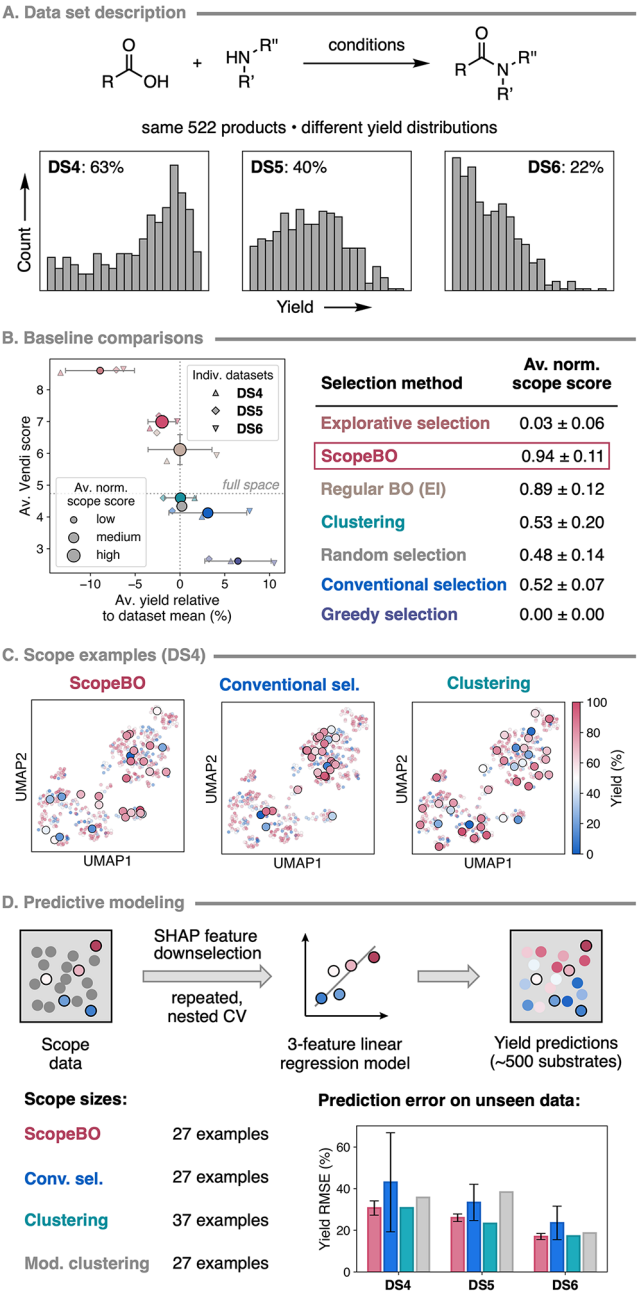


Figure 3. Investigations using amide coupling data sets. (A): Description of the data sets. (B): Baseline comparisons summarized for the three data sets. Conventional selection (conv. sel.) refers to greedy acquisition with 5% Vendi pruning. Whiskers show the standard deviation of the scores across the data sets. Some whiskers are not visible because they are smaller than the size of the data point. (C): Scope examples for the data set **DS4**. The chosen ScopeBO scope is the one with a scope score closest to the average of all initiations. The same initiation was used for the displayed conv. sel. scope. (D): Predictive modeling. Whiskers show the standard deviation across the 40 random initiations. Mod. clustering refers to clustering[8] with a fixed scope size of 27 samples. CV = cross-validation; RMSE = root mean square error; SHAP = Shapley additive explanations.

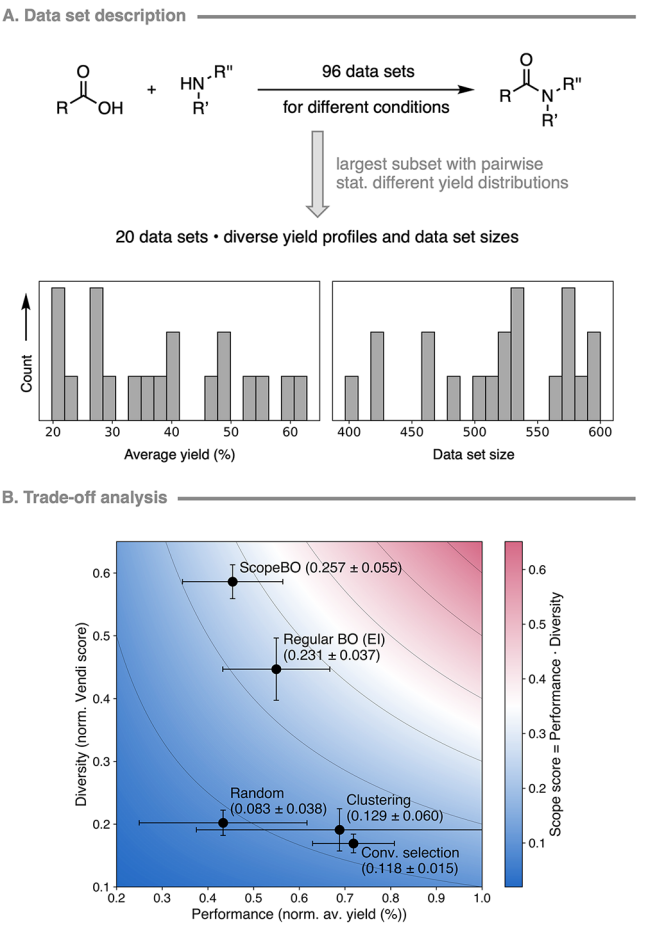
Given the good coverage of the substrate space with ScopeBO as illustrated by the Vendi score and the UMAP analysis, we wondered if we could use the small 27-sample data set of an

https://doi.org/10.1021/jacs.6c10981
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

evaluated scope to make meaningful yield predictions for the remainder of the 522-sample search space. We utilized SHAP analysis[19] to down-select the substrate features to the 20 most important descriptors and used these to train 3-feature regression models via a nested, repeated cross-validation process (Figure 3D; see SI, chapter 10). For each of the 40 initiations, we predicted the yields for the 495 unseen substrate pairs with the trained models and calculated the root-mean-square error (RMSE) of the prediction. The models trained on the conventionally selected scopes generalize significantly worse to the unseen substrates than the models trained on the ScopeBO substrates with the RMSE being 7−12% higher for the different data sets and a stronger dependence on the initiation of the scope as evidenced by the large standard deviation. This is likely a result of the lowered diversity of the selected substrates compared to ScopeBO. The models trained on the ScopeBO data had RMSEs of 17−31% for the unseen substrates in the different data sets. They performed similarly to the models for the clustering data despite the ScopeBO scopes being 27% smaller (27 *vs* 37 examples). Notably, models trained on clustering scopes with the same size as ScopeBO scopes (27 samples) had an RMSE that is on average 6% greater than that for ScopeBO (see SI, chapter 10).[45] Overall, these results imply that the diversity of the scopes selected by ScopeBO increases the predictive performance over less diverse scopes and can be leveraged to make rough qualitative predictions for a large number of unseen substrates from a small scope data set. Predictive modeling functions using either linear models or using Gaussian process regression have been implemented in the ScopeBO package.

Generality Analysis

After having established trends in how ScopeBO behaves when exposed to different data set distributions, we were curious about the generality of the workflow when tested in many different scenarios. The amide data set by Liao and co-workers[42] contains 96 subdata sets for different reaction conditions that each constitute a separate reaction scope space. From this collection, we extracted the largest subset of condition data sets that all have a pairwise statistically significantly different yield distribution (see SI Section 3.2.2). This analysis yielded 20 data sets that span a wide range of average yields and data set sizes (Figure 4A). We then simulated scopes for ScopeBO and alternative selection methods for all data sets. On average, ScopeBO achieved the highest scope score (0.257 ± 0.055), followed by regular BO using the EI acquisition function (0.231 ± 0.037). As the calculation of the scope score as the product of substrate performance and diversity enables multiple trade-offs between the two metrics to yield the same scope score, we were interested to see the position of this trade-off for the different selection methods (Figure 4B). Visualization on a scope score isosurface illustrates that ScopeBO has a higher contribution of substrate diversity than EI, which is more performance-driven. The substrate performance achieved by conventional selection is high, but the diversity is in contrast much lower. Clustering achieves a slightly higher diversity than conventional selection, but it is notably lower than that of ScopeBO, in line with the results discussed above. While the substrate performance of the clustering scope is surprisingly high, the very large standard deviation indicates the unpredictable nature of the algorithm in this regard as it is performance-agnostic in its selection. To analyze whether the observed differences are significant, a statistical analysis of the results was performed (see SI Section 8.1 for a description; Figure 4C). ScopeBO generated the best scope scores for all 20 data sets. It was the unanimously the best performer in seven data sets and shared the top performance in the other cases. Notably, the results were similar when the analysis was extended to all 32 data sets that were investigated in this manuscript (values in parentheses in Figure 4C). To quantify the improvement in performance with ScopeBO, we calculated the difference in scope scores for all data sets in which a statistical difference was found between ScopeBO and the top competitor, demonstrating that ScopeBO gave scope scores that were on average 22.2% higher (15.2% for all 32 data sets). In



C. Statistical and quantitative comparison

Rank of ScopeBO	Data set count for scope scores	Data set count for Vendi scores[a]
1st	7 (8)	20 (31)
shared-1st	13 (22)	0 (0)
2nd	0 (2)	0 (1)
Av. improvement[b]	22.2% (15.2%)	17.7% (19.0%)

Figure 4. Generality analysis. (A): Description of the data sets. (B): Trade-off analysis. Whiskers show the standard deviation of the scores across the data sets. (C): Statistical and quantitative comparison. Values in parentheses correspond to numbers across all 32 data sets that were evaluated in this manuscript. [a]The analysis excludes purely diversity-driven explorative selection methods. [b]The improvement was calculated compared to the top competitor by scope score and only includes data sets in which a statistical difference was found for the evaluated metric.

https://doi.org/10.1021/jacs.6c10981
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

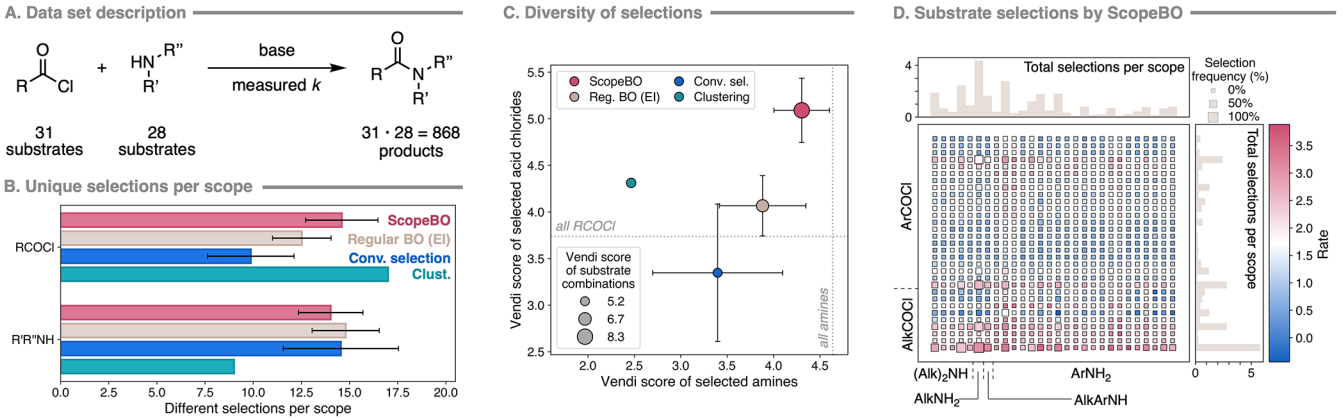


Figure 5. Analysis of selections in a combinatorial reaction space. (A): Description of the data set. (B): Unique selections per scope. (C): Diversity of selections. (D): Substrate selection by ScopeBO. Whiskers in all subfigures show the standard deviation of the scores across the initiations.

terms of the Vendi scores, ScopeBO outperformed all comparisons (excluding fully explorative methods) in all 20 amide data sets, with a Vendi score that is on average 18% higher than the top competitor.

Overall, the analysis shows that ScopeBO performs equally or better than alternative selection methods across a wide range of data sets while consistently improving the diversity of the selected substrates.

Substrate Selections in a Combinatorial Space

To analyze how ScopeBO selects scope examples in a full combinatorial search space, we explored a different amide coupling data set in which the products are formed in a Schotten-Baumann reaction (Figure 5A).[46] In this report, 31 acid chlorides are coupled with 28 different amines. An analysis of the full search space indicated that the reaction rate is mainly driven by the choice of acid chloride and that the amine identity has a smaller influence on the reaction outcome (Figure 5D). Interestingly, scopes selected by ScopeBO, regular BO (using EI), and simulated conventional selection feature approximately the same number of unique amines per scope (Figure 5B). The diversity of these amines as measured by the Vendi score is however the largest for ScopeBO (Figure 5C). In contrast, the selections of acid chlorides correlated with how performance-driven the selection method is: ScopeBO provided the highest number of unique selections and the highest diversity while both metrics decreased utilizing regular BO and conventional selection (Figure 5B,C). A clustering scope provided the highest count of unique acid chlorides that however corresponded to a lower diversity than the ScopeBO selections. The count of selected unique amines was in contrast much lower than with all other methods, illustrating that the clustering algorithm can provide lowered diversity in multisubstrate scenarios due to simply selecting the center of the substrate clusters. Viewing the combined substrates pairs, ScopeBO also outperforms these key comparisons with a Vendi score that is 27% higher than the second-best method (regular BO). Still, the selections of ScopeBO remain performance-driven as the most picked substrate combinations all have a high reaction rate (Figure 5D).

Low-Performance Scenario

Having noted that the scope performance is dependent on the underlying data set distribution, we were curious to see how ScopeBO performs in extreme cases. To test this, we used a data set for a decarboxylative coupling between a complex alkyl acid and 381 aryl bromides that was reported by MacMillan et al.[47] This data set represents a challenging case study for the algorithm as 54% of the samples gave less than 10% yield, resulting in an average yield of only 17% (Figure 6A). We found that ScopeBO performed well in baseline comparisons (see SI, chapter 8.5); however, the substrate performance in the selected scope was slightly below the average of the data set with an average yield of 14%. Furthermore, 63% of the chosen scope entries had yields of below 10%. While this low performance is a result of the data set distribution, it is arguably too low to be very informative for potential users of the reaction. To address this limitation, we hypothesized that removing substrate classes from the search space that are known to be incompatible with the reaction before scope generation would shift the objective distribution to a more suitable level. To illustrate this idea, we searched the data set for functional groups that predominantly give less than 10% yield (*e.g.*, aryl chlorides and carboxylic acids) and removed them (see SI, chapter 3.3.2). We then reran the algorithm on this curated data set that now has an average yield of 22% with 42% of samples below 10% yield. The average yield of the selected scope increased from 14% to 19%, and the number of samples below a 10% yield decreased from 63% to 51%. While these results are still modest, they underline the positive impact of removing unreactive substrate classes from the data set.

Medicinal Chemistry-Like Application

Finally, we were curious whether ScopeBO can also be applied in synthetic contexts other than traditional scope evaluation. In a typical medicinal chemistry campaign, a few lead compounds are subjected to a variety of derivatization reactions in search of highly active compounds. Such a scenario could be interpreted as a scope in which one of the reaction partners is constrained to a few examples, whereas the second reactant is varied more strongly. To test how ScopeBO would perform in such situations, we utilized another decarboxylative coupling data set published by MacMillan et al. in which three informer library aryl bromides were reacted with alkyl acids (Figure 6B).[47] The target objective in this data set is the reaction yield, but it could be replaced by a bio- or physiologically relevant property of the reaction product in a true medicinal chemistry application. First, we investigated different approaches to exploring the chemical space. One option is to conduct the coupling reactions of all informers separately. Here, all scopes could be carried out in parallel so that no information is shared between the scopes or consecutively in which case the knowledge from the previously

https://doi.org/10.1021/jacs.6c10981
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

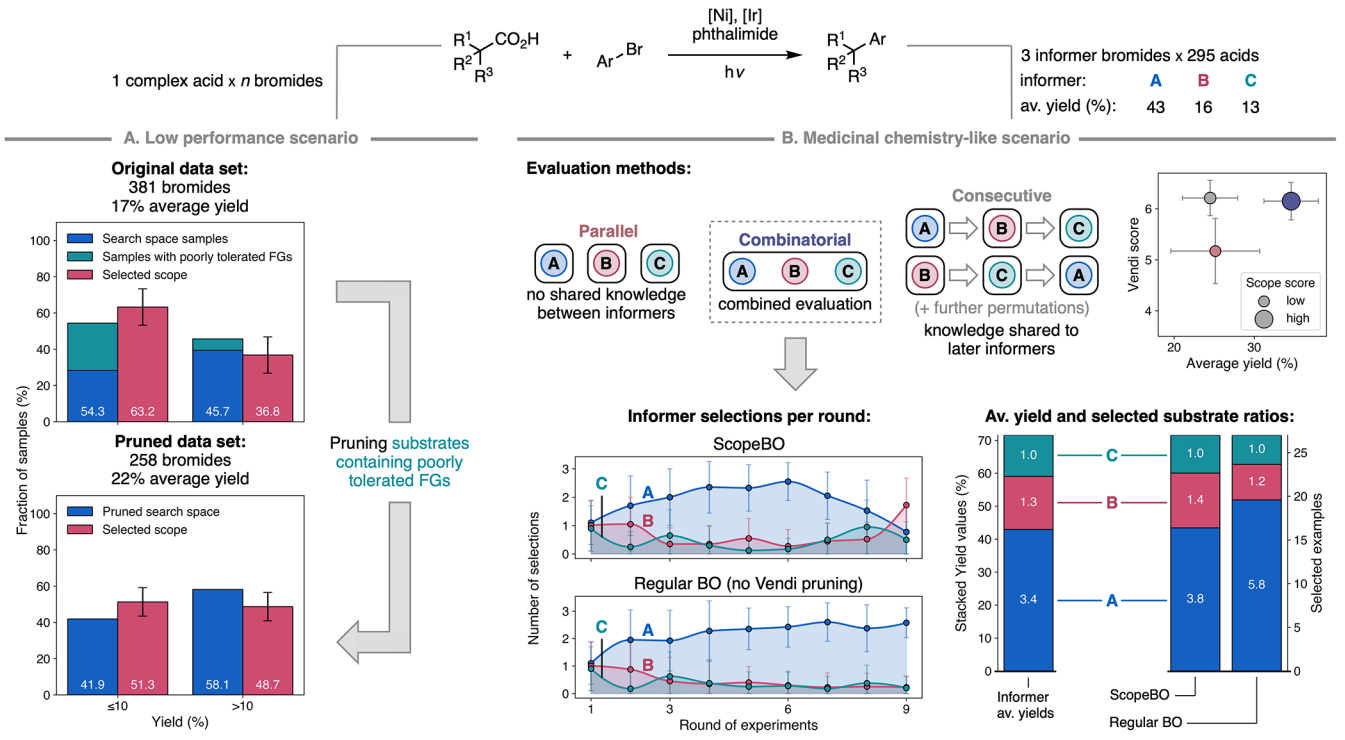


Figure 6. Investigations using decarboxylative coupling data sets. (A): Low-performance scenario. (B): Medicinal chemistry-like scenario. In the exploration of evaluation methods (top), the combined scope size was left constant at 27 samples for each method. Whiskers in all plots correspond to the standard deviation of the displayed value. FGs = functional groups.

evaluated informers is used to apprise the selections for the next one. These strategies are closely related to how reaction scopes for multicomponent reactions are typically conducted in organic chemistry where both partners are explored separately using a fixed coupling partner. Alternatively, both partners could be varied simultaneously in a combinatorial approach that, in this case, corresponds to the coupling reactions of all three informers being evaluated in the same scope. We carried out scope selections for these three strategies by using a fixed combined reaction budget. The Vendi score for the parallel evaluation is the lowest, as this strategy lacks information exchange and reduces the achievable substrate diversity, whereas the Vendi scores for the other two scenarios are higher and nearly the same. Despite the similar Vendi scores, the average yield of the combined scope of all three informers was notably higher than that in the consecutive evaluation because the combinatorial strategy permits flexible allocation of the experimental budget to the most promising informer. The combined scope scenario therefore increases the performance of the selected couplings without decreasing selection diversity.

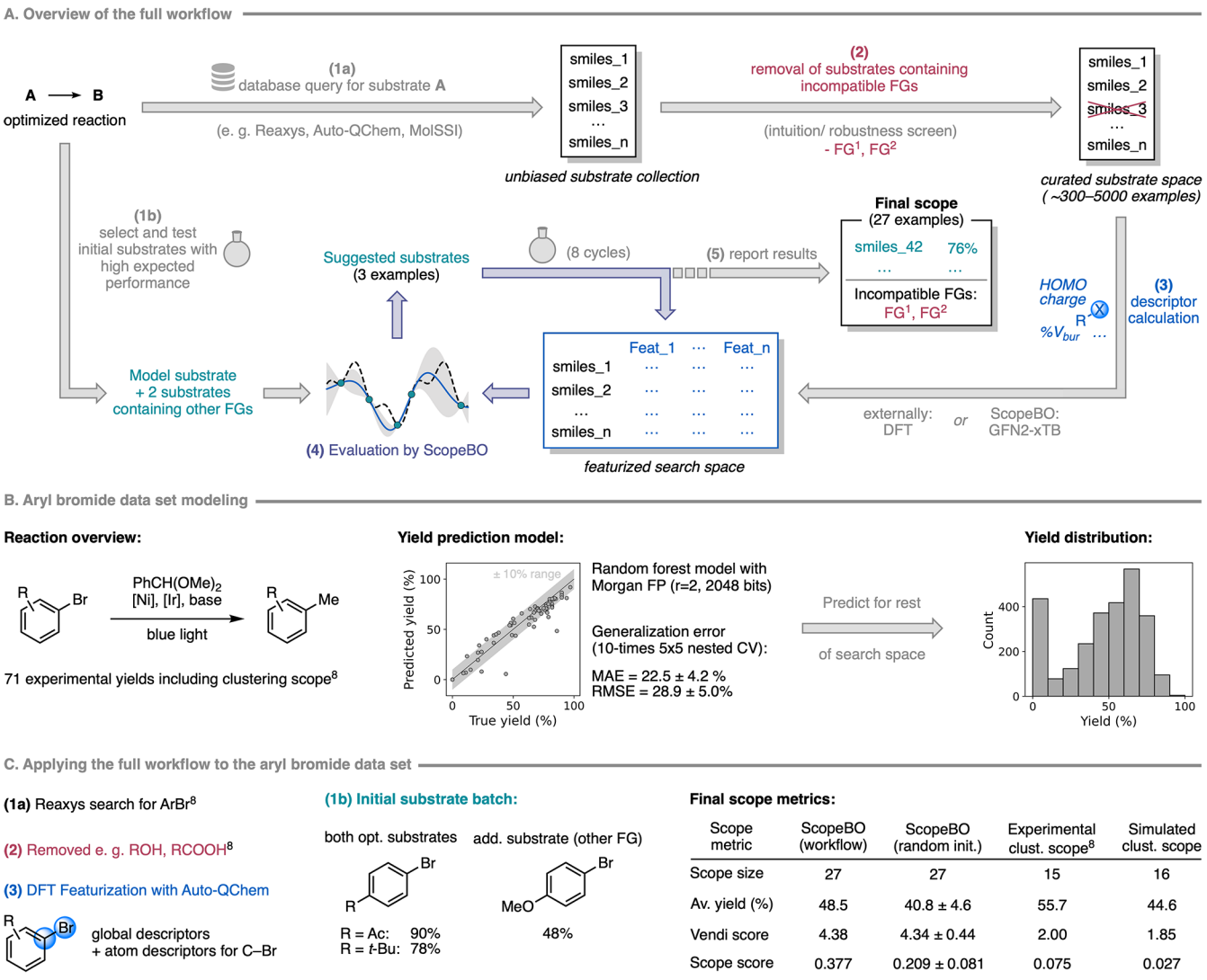
Since the three informers have notable differences in their average yield across the data set (43%, 16%, and 13%, respectively), we were interested to see how the algorithm selected informers during each round of the combined scope strategy and how the Vendi pruning influenced this step. Therefore, we compared the results of ScopeBO to those of regular BO (using the EI acquisition function). In the early rounds, both algorithms quickly committed to the highest-performing aryl bromide. Without Vendi pruning (regular BO), this commitment is maintained in the second half of the scope, and more than 2 out of 3 samples in each round are allocated to the highest-performing informer. In contrast, the number of selections of this bromide continuously decreases each round in the second half of ScopeBO runs due to the pruning of similar samples, and substrate combinations utilizing other informers are increasingly picked. Excitingly, the ratio of aryl bromide informers selected by ScopeBO closely mirrored the ratios of the average yields, illustrating that the algorithm not only promotes sample selection diversity but also captures trends in the relative sample performance. In contrast, the standard BO baseline strongly oversampled the highest-performing informer **A** with regard to the yield ratios. This behavior highlights that ScopeBO successfully focuses on the most promising lead compounds while avoiding undersampling of other candidates to enable broad chemical space coverage.

Workflow Summary

On the basis of these results, we can recommend a workflow for evaluating a substrate scope using ScopeBO (Figure 7A). To illustrate this workflow with an example, we selected the aryl bromide data set, which was used to introduce the clustering scope selection algorithm (Figure 7B).[8] The experimental data from this study were used to predict the performance for the full search space using a random forest regressor with Morgan fingerprint featurization. We then leveraged these yields to select a ScopeBO scope following our workflow recommendations (*vide infra*) and compared it to the experimentally reported clustering scope[8] in a scenario that mimics a prospective application of ScopeBO.

As a first step of the workflow, the user creates the search space that should be defined as broadly as possible to obtain a representative scope after carrying out ScopeBO. For a substrate class with broad commercial availability, the space could be generated by querying a large database like Reaxys (*e.g.*, by searching for all commercial members of the class, step 1a, see SI Section 2.5.1). In the case of a substrate class with limited direct availability, such a strategy could, however, introduce bias. As an

https://doi.org/10.1021/jacs.6c10981
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX



Scope metric	ScopeBO (workflow)	ScopeBO (random init.)	Experimental clust. scope[8]	Simulated clust. scope
Scope size	27	27	15	16
Av. yield (%)	48.5	40.8 ± 4.6	55.7	44.6
Vendi score	4.38	4.34 ± 0.44	2.00	1.85
Scope score	0.377	0.209 ± 0.081	0.075	0.027

Figure 7. Summary and example of the scope evaluation workflow. (A): Workflow overview. (B): Data set overview and yield prediction modeling for the aryl bromide data set. (C): Application of the workflow to the aryl bromide data set and comparison to an experimental clustering scope[8] as well as to a simulation of a clustering scope on the search space with the predicted yields. Ac = acetyl; CV = cross-validation; DFT = density functional theory; Feat = feature; FGs = functional groups; FP = fingerprint; MAE = mean absolute error; RMSE = root mean square error.

alternative, the space could also be defined with a surrogate substrate class that is related to the desired substrates and has broader availability (*e.g.*, carboxylic acids as a surrogate for Weinreb amides). This initial search space is then curated by removing compounds containing functional groups that are known to be incompatible with the reaction (step 2). Further functional groups could be removed based on poor performance in an optional robustness screen.[48] This step is recommended because the average performance in a ScopeBO-selected scope is dependent on the average performance across the search space (Figure 1). Removing known poor performers thus helps avoid scenarios in which most selected scope examples are unreactive. However, as this step also introduces some subjective bias in the scope diversity, it should only be carried out *before* starting scope selection, and omitted substrate categories and database search terms should be reported together with the selected scope to provide a full picture of the reaction performance across the entire chemical space.

As we noted a dependence of the algorithm performance on the search space size during our studies, we recommend that the size of the curated search space should not be drastically different from the size of the optimization data sets and therefore should ideally be between 300 and 5000 examples. However, the algorithm can, in principle, be used for any search space with at least 58 substrates. In the case of the aryl bromide data set (Figure 7C), we used the search space from the prior study[8] (*ca.* 2600 compounds) that was collected from Reaxys and refined by removing functional groups thought to be incompatible with the reaction, such as alcohols and carboxylic acids.

Next, substrate descriptors are generated (step 3). Ideally, these are calculated with DFT, for instance using Auto-QChem,[36] an automated DFT calculation submission and analysis workflow by the Doyle group, or related tools.[49−51] This strategy was used as the default in this study and also in the case of the aryl bromide data set (Figure 7C). As an alternative, Morfeus descriptors at the GFN2-xTB level, that perform similarly to DFT descriptors, can be directly calculated by ScopeBO without the need for high-performance computing. Notably, the steps 1a, 2, and 3 can be replaced by directly using existing chemical descriptor libraries such as those deposited in the Auto-QChem[52] or MolSSI[53] databases. A more detailed discussion of steps 1−3 is provided in the SI Section 2.5, and an

https://doi.org/10.1021/jacs.6c10981
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

example of the entire ScopeBO workflow is provided in the GitHub repository accompanying this publication.

In parallel with the search space generation, the user can evaluate the first batch of scope examples (step 1b). These three initial samples should be composed of the optimization model substrate and two substrates containing other functional groups that are expected to be tolerated by the reaction conditions. In the case of the aryl bromide data set, the reaction was optimized with two substrates that were thus selected for the initial batch (Figure 7C). As a third substrate, we picked 4-bromoanisole because of its complementary electronic nature compared to the two optimization substrates. The results of the experiments are recorded in the featurized search space file, and the file is submitted to the ScopeBO algorithm to iteratively build the scope over eight additional batches of three experiments (step 4). The final scope of 27 substrates should be reported together with the incompatible functional groups that were removed in the search space curation (step 5).

In the case of the aryl bromide data set, the ScopeBO scope selected following this workflow has a similar diversity but increased average yield compared to randomly initiated ScopeBO scopes (Figure 7C). In comparison to the experimental clustering scope,[8] the average yield is a bit lower, which might be partially driven by the tendency of the search space yield prediction model to underpredict yields (Figure 7B). Indeed, a simulation of a clustering scope on the same search space has a much lower average yield than the experimental example (Figure 7C). The Vendi score is, however, significantly larger for the ScopeBO scope than either clustering scope, resulting in a larger scope score.

CONCLUSION

In conclusion, we report a strategy for quantifying scope performance using the scope score, enabling benchmarking of common scope selection methods. Leveraging this score, we establish a machine learning algorithm, ScopeBO, for scope generation that merges ideas from the way humans conventionally select reaction scopes and data science-based approaches. By using ScopeBO, the practitioner can balance sample performance and diversity of the evaluated substrates. The algorithm demonstrates high scope scores across all tested data sets and against comparisons with alternative selection methods, highlighting its broad applicability for different reaction classes and mono- as well as multiobjective scenarios. We provide the ScopeBO algorithm as a user-friendly python code that also incorporates additional functions such as feature importance analysis and predictive modeling. All functions are also available in a fully code-free manner via an app, enabling easier adoption. In a broader context, we demonstrated that high scope performance can be achieved with relatively small reaction scopes by careful selection of the substrates. Our results highlight how ScopeBO reduces the prevailing selection bias in scope selection, and we expect that its adoption could help standardize this crucial aspect of reaction development in order to maximize the scope's informativeness for future users of the reaction and downstream reaction prediction tasks.

ASSOCIATED CONTENT

Data Availability Statement

The ScopeBO algorithm as well as all scripts and primary data are available on GitHub at https://github.com/doyle-lab-ucla/ScopeBO.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacs.6c10981.

Algorithm details, Python package description, installation instruction, use recommendations, data set and hyperparameter optimization details, and data analysis (PDF)

Tutorial for the app (PDF)

AUTHOR INFORMATION

Corresponding Author

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

Authors

Sven Roediger − *Department of Chemistry and Biochemistry, University of California-Los Angeles, Los Angeles, California 90095, United States*

Matthew S. Sigman − *Department of Chemistry, University of Utah, Salt Lake City, Utah 84112, United States;* orcid.org/0000-0002-5746-8830

Complete contact information is available at: https://pubs.acs.org/10.1021/jacs.6c10981

Author Contributions

All authors have given approval to the final version of the manuscript.

Funding

These studies were supported by the United States National Science Foundation (NSF) Office of Advanced Cyberinfrastructure (OAC-2118201). S.R. acknowledges the Swiss National Science Foundation (SNSF) for a Postdoc.Mobility fellowship (P500PN_222270). M.S.S. acknowledges financial support from the NSF under the CCI Center for Computer Assisted Synthesis (C-CAS) (CHE-2202693).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank the Doyle group, Dr. Richard Walroth, and Dr. Michael Bogdos for critical feedback and discussions.

ABBREVIATIONS

Ac, acetyl; ArI, aryl iodide; BO, Bayesian optimization; CV, cross-validation; DFT, density functional theory; DS, data set; EI, expected improvement; Feat, featurization; FG, function group; FP, fingerprint; MAE, mean absolute error; ML, machine learning; MOF, metal−organic framework; OA, oxidative addition; RMSE, root mean square error; SHAP, Shapley additive explanations; Std. dev., standard deviation; Trip, 2,4,6-tri-*iso*-propylphenyl; UMAP, uniform manifold approximation and projection

REFERENCES

(1) Zacate, S. B.; Dantas, J. A.; Lin, S.; Doyle, A. G.; Sigman, M. S. Considerations in Pursuing Reaction Scope Generality. *Angew. Chem., Int. Ed.* **2025**, *64*, No. e202511091.

(2) Tyler, J. L.; Trauner, D.; Glorius, F. Reaction Development: A Student's Checklist. *Chem. Soc. Rev.* **2025**, *54* (7), 3272−3292.
(3) Kozlowski, M. C. On the Topic of Substrate Scope. *Org. Lett.* **2022**, *24* (40), 7247−7249.
(4) Maloney, M. P.; Coley, C. W.; Genheden, S.; Carson, N.; Helquist, P.; Norrby, P.-O.; Wiest, O. Negative Data in Data Sets for Machine Learning Training. *J. Org. Chem.* **2023**, *88* (9), 5239−5241.
(5) Haas, B. C.; Kalyani, D.; Sigman, M. S. Applying Statistical Modeling Strategies to Sparse Datasets in Synthetic Chemistry. *Sci. Adv.* **2025**, *11* (1), No. eadt3013.
(6) Strieth-Kalthoff, F.; Sandfort, F.; Kühnemund, M.; Schäfer, F. R.; Kuchen, H.; Glorius, F. Machine Learning for Chemical Reactivity: The Importance of Failed Experiments. *Angew. Chem., Int. Ed.* **2022**, *61* (29), No. e202204647.
(7) Williams, W. L.; Zeng, L.; Gensch, T.; Sigman, M. S.; Doyle, A. G.; Anslyn, E. V. The Evolution of Data-Driven Modeling in Organic Chemistry. *ACS Cent. Sci.* **2021**, *7* (10), 1622−1637.
(8) Kariofillis, S. K.; Jiang, S.; Żurański, A. M.; Gandhi, S. S.; Alvarado, J. I. M.; 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.
(9) 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.
(10) Haas, B. C.; Goetz, A. E.; Bahamonde, A.; McWilliams, J. C.; Sigman, M. S. Predicting Relative Efficiency of Amide Bond Formation Using Multivariate Linear Regression. *Proc. Natl. Acad. Sci. U.S.A.* **2022**, *119* (16), No. e2118451119.
(11) Ruos, M. E.; Romer, N. P.; Deichert, J. A.; Alabanza, L. M.; Gandhi, S. S.; Brown, G. Z.; Walroth, R. C.; Cruz, K.; Gosselin, F.; Hong, A. Y.; Sigman, M. S.; Doyle, A. G. Data Science-Guided Development of Deoxyfluorination Reagents with Enhanced Reactivity, Practicality, and Safety. *J. Am. Chem. Soc.* **2025**, *147* (29), 25815−25824.
(12) Gandhi, S. S.; Brown, G. Z.; Aikonen, S.; Compton, J. S.; Neves, P.; Alvarado, J. I. M.; 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* (3), 2292−2304.
(13) Gensch, T.; Gomes, G. d. P.; Friederich, P.; Peters, E.; Gaudin, T.; Pollice, R.; Jorner, K.; Nigam, A.; Lindner-D'Addario, M.; Sigman, M. S.; Aspuru-Guzik, A. A Comprehensive Discovery Platform for Organophosphorus Ligands for Catalysis. *J. Am. Chem. Soc.* **2022**, *144* (3), 1205−1217.
(14) Gensch, T.; Smith, S. R.; Colacot, T. J.; Timsina, Y. N.; Xu, G.; Glasspoole, B. W.; Sigman, M. S. Design and Application of a Screening Set for Monophosphine Ligands in Cross-Coupling. *ACS Catal.* **2022**, *12* (13), 7773−7780.
(15) 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.
(16) Henle, J. J.; Zahrt, A. F.; Rose, B. T.; Darrow, W. T.; Wang, Y.; Denmark, S. E. Development of a Computer-Guided Workflow for Catalyst Optimization. Descriptor Validation, Subset Selection, and Training Set Analysis. *J. Am. Chem. Soc.* **2020**, *142* (26), 11578−11592.
(17) Hueffel, J. A.; Sperger, T.; Funes-Ardoiz, I.; Ward, J. S.; Rissanen, K.; Schoenebeck, F. Accelerated Dinuclear Palladium Catalyst Identification through Unsupervised Machine Learning. *Science* **2021**, *374* (6571), 1134−1140.
(18) Reymond, J.-L. The Chemical Space Project. *Acc. Chem. Res.* **2015**, *48* (3), 722−730.
(19) Lundberg, S. M.; Lee, S.-I. *A Unified Approach to Interpreting Model Predictions*, Proceedings of the 31st International Conference on Neural Information Processing Systems; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp 4768−4777.
(20) 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.
(21) Torres, J. A. G.; Lau, S. H.; Anchuri, P.; Stevens, J. M.; Tabora, J. E.; Li, J.; Borovika, A.; Adams, R. P.; Doyle, A. G. A Multi-Objective Active Learning Platform and Web App for Reaction Optimization. *J. Am. Chem. Soc.* **2022**, *144* (43), 19999−20007.
(22) Romer, N. P.; Min, D. S.; Wang, J. Y.; Walroth, R. C.; Mack, K. A.; Sirois, L. E.; Gosselin, F.; Zell, D.; Doyle, A. G.; Sigman, M. S. Data Science Guided Multiobjective Optimization of a Stereoconvergent Nickel-Catalyzed Reduction of Enol Tosylates to Access Trisubstituted Alkenes. *ACS Catal.* **2024**, *14* (7), 4699−4708.
(23) Häse, F.; Roch, L. M.; Aspuru-Guzik, A. Chimera: Enabling Hierarchy Based Multi-Objective Optimization for Self-Driving Laboratories. *Chem. Sci.* **2018**, *9* (39), 7642−7655.
(24) Häse, F.; Aldeghi, M.; Hickman, R. J.; Roch, L. M.; Aspuru-Guzik, A. Gryffin: An Algorithm for Bayesian Optimization of Categorical Variables Informed by Expert Knowledge. *Appl. Phys. Rev.* **2021**, *8* (3), No. 031406.
(25) Dalton, D. M.; Walroth, R. C.; Rouget-Virbel, C.; Mack, K. A.; Toste, F. D. Utopia Point Bayesian Optimization Finds Condition-Dependent Selectivity for N-Methyl Pyrazole Condensation. *J. Am. Chem. Soc.* **2024**, *146* (23), 15779−15786.
(26) Fitzner, M.; Šošić, A.; Hopp, A. V.; Müller, M.; Rihana, R.; Hrovatin, K.; Liebig, F.; Winkel, M.; Halter, W.; Brandenburg, J. G. BayBE: A Bayesian Back End for Experimental Planning in the Low-to-No-Data Regime. *Digital Discovery* **2025**, *4* (8), 1991−2000.
(27) Sin, J. W.; Chau, S. L.; Burwood, R. P.; Püntener, K.; Bigler, R.; Schwaller, P. Highly Parallel Optimisation of Chemical Reactions through Automation and Machine Intelligence. *Nat. Commun.* **2025**, *16* (1), No. 6464.
(28) Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R. P.; de Freitas, N. Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proc. IEEE* **2016**, *104* (1), 148−175.
(29) Guo, J.; Ranković, B.; Schwaller, P. Bayesian Optimization for Chemical Reactions. *Chimia* **2023**, *77* (1−2), 31−38.
(30) Friedman, D.; Dieng, A. B. The Vendi Score: A Diversity Evaluation Metric for Machine Learning, 2023. arXiv:2210.02410. arXiv.org e-Print archive. https://arxiv.org/abs/2210.02410.
(31) Miranda-Quintana, R. A.; Rácz, A.; Bajusz, D.; Héberger, K. Extended Similarity Indices: The Benefits of Comparing More than Two Objects Simultaneously. Part 2: Speed, Consistency, Diversity Selection. *J. Cheminf.* **2021**, *13* (1), No. 33.
(32) Liu, T.-W.; Nguyen, Q.; Dieng, A. B.; Gómez-Gualdrón, D. A. Diversity-Driven, Efficient Exploration of a MOF Design Space to Optimize MOF Properties. *Chem. Sci.* **2024**, *15* (45), 18903−18919.
(33) Tang, T.; Hazra, A.; Min, D. S.; Williams, W. L.; Jones, E.; Doyle, A. G.; Sigman, M. S. Interrogating the Mechanistic Features of Ni(I)-Mediated Aryl Iodide Oxidative Addition Using Electroanalytical and Statistical Modeling Techniques. *J. Am. Chem. Soc.* **2023**, *145* (15), 8689−8699.
(34) We compared the Vendi score to other diversity metrics and observed a good alignment with most tested methods (see SI, chapter 6). In principle, the score could also be defined using alternative methods.
(35) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, 2020. arXiv:1802.03426. arXiv.org e-Print archive. https://arxiv.org/abs/1802.03426.
(36) 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* (6), 1276−1284.
(37) Landrum, G.; Tosco, P.; Kelley, B.; Ric; Sriniker; Cosgrove, D.; Gedeck; Vianello, R.; NadineSchneider; Kawashima, E.; N, D.; Jones, G.; Dalke, A.; Cole, B.; Swain, M.; Turk, S.; AlexanderSavelyev; Vaucher, A.; Wójcikowski, M.; Take, I.; Probst, D.; Ujihara, K.; Scalfani, V. F.; Guillaume, G.; Lehtivarjo, J.; Walker, R.; Pahl, A.; Berenger, F.;

Jasondbiggsstrets123. *Rdkit/Rdkit: 2023_03_3 (Q1 2023) Release*; Zenodo, 2023. DOI: 10.5281/zenodo.8254217.
(38) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: A Molecular Descriptor Calculator. *J. Cheminf.* **2018**, *10* (1), No. 4.
(39) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5* (2), 107−113.
(40) Jorner, K.; Turcani, L.*Kjelljorner/Morfeus: V0.7.2*; Zenodo, 2022 DOI: 10.5281/zenodo.7017599.
(41) 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 Comput.* **2019**, *15* (3), 1652−1671.
(42) Zhang, C.; Lin, Q.; Yang, C.; Kong, Y.; Yu, Z.; Liao, K. Intermediate Knowledge Enhanced the Performance of the Amide Coupling Yield Prediction Model. *Chem. Sci.* **2025**, *16* (26), 11809−11822.
(43) Due to the deterministic nature of the clustering selection (just one scope), we could not statistically compare ScopeBO results with clustering results.
(44) Jeon, H.; Park, J.; Shin, S.; Seo, J. Stop Misusing T-SNE and UMAP for Visual Analytics, 2025. arXiv:2506.08725. arXiv.org e-Print archive. https://arxiv.org/abs/2506.08725.
(45) Constraining the clustering algorithm to a fixed size can decrease the quality of the clustering. In this case, the silhouette score, a metric for this, however remained almost unchanged (0.784 vs 0.791), indicating that the quality loss should be minimal.
(46) Baumer, M.; Gallegos, L.; Anstine, D.; Kubaney, A.; Regio, J.; Isayev, O.; Bernhard, S.; Gomes, G. Democratizing Reaction Kinetics through Machine Vision and Learning *ChemRxiv* 2025 DOI: 10.26434/chemrxiv-2025-4tk40.
(47) Kullmer, C. N. P.; Kautzky, J. A.; Krska, S. W.; Nowak, T.; Dreher, S. D.; MacMillan, D. W. C. Accelerating Reaction Generality and Mechanistic Insight through Additive Mapping. *Science* **2022**, *376* (6592), 532−539.
(48) Collins, K. D.; Glorius, F. A Robustness Screen for the Rapid Assessment of Chemical Reactions. *Nat. Chem.* **2013**, *5* (7), 597−601.
(49) Alegre-Requena, J. V.; Sowndarya, S. V. S.; Pérez-Soto, R.; Alturaifi, T. M.; Paton, R. S. AQME: Automated Quantum Mechanical Environments for Researchers and Educators. *WIREs Comput. Mol. Sci.* **2023**, *13* (5), No. e1663.
(50) Shved, A. S.; Ocampo, B. E.; Burlova, E. S.; Olen, C. L.; Rinehart, N. I.; Denmark, S. E. Molli: A General Purpose Python Toolkit for Combinatorial Small Molecule Library Generation, Manipulation, and Feature Extraction. *J. Chem. Inf. Model.* **2024**, *64* (21), 8083−8090.
(51) Haas, B. C.; Hardy, M. A.; V, S. S. S.; Adams, K.; Coley, C. W.; Paton, R. S.; Sigman, M. S. Rapid Prediction of Conformationally-Dependent DFT-Level Descriptors Using Graph Neural Networks for Carboxylic Acids and Alkyl Amines. *Digital Discovery* **2025**, *4* (1), 222−233.
(52) Żurański, A. M.; Wang, J. Y.; Shields, B. J.; Doyle, A. G. Auto-QChem. database. https://autoqchem.org/ (accessed November 20, 2025).
(53) MolSSI Descriptor Libraries. https://descriptor-libraries.molssi.org/ (accessed November 20, 2025).

https://doi.org/10.1021/jacs.6c10981
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX