

Evaluating Predictive Accuracy in Asymmetric Catalysis: A Machine Learning Perspective on Local Reaction Space

Isaiah O. Betinol, Aleksandra Demchenko, and Jolene P. Reid*



Cite This: ACS Catal. 2025, 15, 6067–6077



Read Online

ACCESS |

Metrics & More

Article Recommendations

Supporting Information

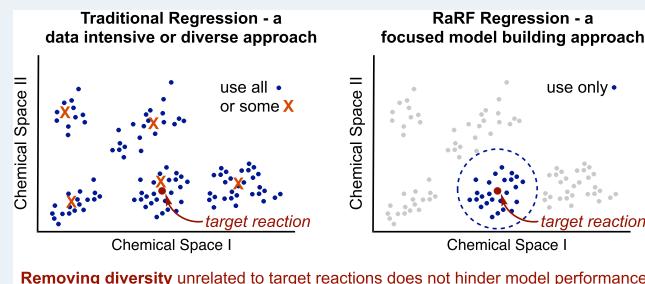
ABSTRACT: Machine learning (ML) models are increasingly being employed in asymmetric catalysis to predict reaction outcomes and optimize enantioselective processes. Despite the trend of expanding data set sizes to improve model performance, asymmetric catalysis presents unique challenges, including the difficulty of acquiring high-quality experimental data and the often-limited availability of structurally diverse examples. Consequently, rational data set design requires the practitioner to choose whether to collect data that maximizes diversity in the training set or data that maximizes representation around a target prediction. A key challenge in these studies is understanding the role of local reaction space—specifically, how much predictive accuracy is driven by nearest neighbors (structurally and electronically similar data points) and the next-nearest neighbors? This study investigates the predictive power of ML models trained with varying levels of local representation in the reaction space. We provide a framework, a radius-based random forest (RaRF) algorithm, to systematically probe the effects of including diverse reactions dissimilar to a target prediction. We show that when the training set is representative of the target reaction, the gains from increasing data set diversity are modest—typically less than 0.1 kcal/mol in predictive error—and increasing to only 0.5 kcal/mol for extrapolative tests, highlighting the need for targeted data set design. Furthermore, these findings hold even for complex architectures and features. Finally, we demonstrate that a targeted, neighborhood-oriented strategy greatly accelerates the identification of predictive models compared to diversity-driven approaches.

KEYWORDS: machine learning, data set design, asymmetric catalysis, enantioselectivity, prediction

INTRODUCTION

The small molecule universe (SMU) comprising more than 10^{60} synthetically feasible molecules, represents an immense frontier for chemists.^{1,2} While it is impossible to experimentally study the properties and reactivity of each molecule, machine learning (ML) and artificial intelligence (AI) offer powerful tools for predicting molecular behaviors. These technologies are increasingly enabling advancements in reaction prediction, mechanistic understanding, and synthesis planning, accelerating exploration within this vast chemical space.

Data collection, however, remains one of the most resource-intensive phases in developing and applying predictive models (Figure 1A). Building high-quality data sets often requires hundreds to thousands of data points, which demand significant experimental effort. While data mining and careful curation of existing data sets can reduce the need for new experiments, these approaches are constrained by the limited availability of well-curated, machine-readable reaction databases, reporting biases, and inconsistencies in the experimental literature. This challenge is particularly acute in asymmetric catalysis given challenges in high-throughput analytical techniques which hinder the rapid accumulation of new data.^{3–6} Furthermore, while common databases such as



SciFinder and the Open Reaction Database⁷ have provided organized and consistent data for yields and transformations, the same does not hold for enantioselectivity. For these reasons, even the largest curated databases in asymmetric catalysis (~12,000 reactions to the best of our knowledge⁸) pale in comparison to the available data for general yield or reactivity predictions (~1,000,000 reactions in USPTO-full for example⁹) and molecular property predictions (134,000 molecules in QM9 for example¹⁰).

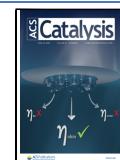
Despite these limitations, domain-specific databases have enabled important advances in predicting reaction outcomes and properties (Figure 1B). A prominent example is the use of linear free energy relationships (LFER), where constraining the reaction space to a single transformation enables insights into how minute changes relate structure to selectivity.^{11,12} In other words, designing a data set to address focused mechanistic

Received: February 10, 2025

Revised: March 21, 2025

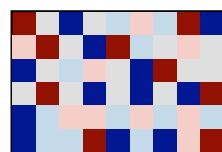
Accepted: March 25, 2025

Published: March 31, 2025



ACS Publications

© 2025 American Chemical Society

A. Challenges in creating asymmetric catalysis databases

exhaustive screening and slow analysis

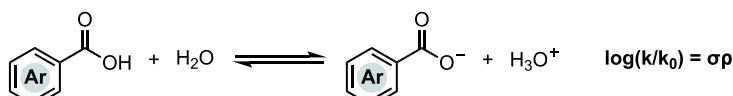
REACTANTS
 PRODUCT
 CONDITIONS

YIELD EE ?

inconsistent formatting in current databases

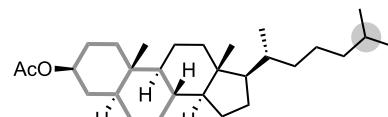
B. Constrained training set approaches

Linear Free Energy Relationships (LFER): Constrain structural variation

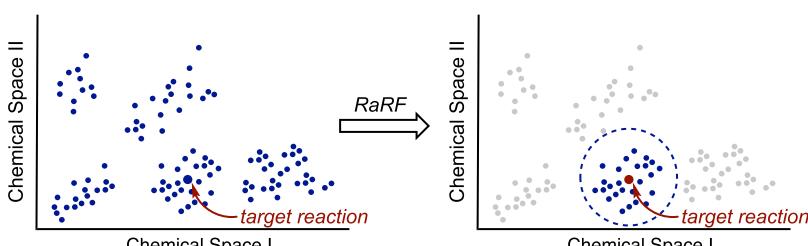


Active learning driven target-specific datasets for CH activation

look for similar reactive sites



look for similar scaffolds

C. This work

Is structurally diverse data necessary for accurate predictions of a target reaction?

Figure 1. (A) Challenges in creating large databases for asymmetric catalysis. (B) Constrained training set approaches to model chemical data. (C) Illustration of this work.

questions can maximize information gain with minimal data. These reaction-specific approaches have led to widespread adoption of LFER-related techniques (e.g., MLR) in asymmetric catalysis,^{13–15} and tailored featurization can enable accurate predictions in these small-data regimes with more complex regression algorithms.¹⁶ Indeed, the importance of local representation has also been emphasized in several reports for predicting yields of metal-catalyzed cross-coupling reactions. For example, ML models for C–O couplings trained exclusively on reactions with common coupling partners achieved performance comparable to models trained on unrestricted data sets nearly ten times larger.¹⁷ Similarly, transfer-learning strategies, where only mechanistically relevant reactions are used to train a targeted model, have proven effective in predicting palladium-catalyzed cross-coupling reactions with limited training data.¹⁸ Alternative “target-aware” methods to generate relevant training sets have also been proposed, though these approaches have not yet been widely vetted or adopted for asymmetric catalysis.^{19–24} One notable example in asymmetric catalysis was reported by

Reisman, Milo and co-workers who used active learning to identify optimal targeted training sets for predicting complex molecules in regioselective C–H functionalization reactions.²⁵

These important advances notwithstanding, broader sentiment by the scientific community posits that larger, more diverse data sets are required for accurate and generalizable models.^{26–32} Recognizing the challenges in meeting these data demands, we instead considered whether a targeted training set design could enable accurate predictions of a target reaction while using only a fraction of the data. To achieve this, key questions remain about the optimal balance between data set diversity and local representation. Specifically, how much predictive accuracy can be gained from including reactions that are structurally or electronically similar to the target (Figure 1C)? Additionally, how does incorporating chemically diverse reactions influence the predictive power of models trained on asymmetric catalysis data sets?

Herein, we provide a comprehensive analysis of the effects of diversity vs locality in asymmetric catalysis. To answer these questions, we developed a radius-based random forest

regression (RaRFR) algorithm. This method dynamically selects training data based on proximity within a defined cutoff radius in feature space, allowing us to systematically examine how model performance changes with the inclusion of neighbors at varying levels of similarity (Figure 2). By

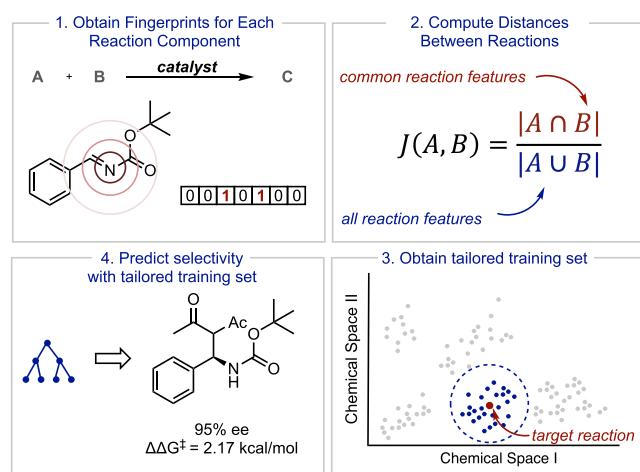


Figure 2. RaRFR workflow.

methodically evaluating this approach across multiple data sets of varying complexity and diversity, we show that focusing on localized reaction spaces offers a more efficient and effective alternative to traditional diversity-driven strategies. Our findings suggest that the inclusion of diverse reactions dissimilar to a target prediction often provides marginal improvements in predictive accuracy—between 0.1 and 0.5 kcal/mol—even with hundreds to thousands of additional data points. In other words, models trained on localized reaction spaces achieve comparable, if not superior, performance with significantly fewer data points, highlighting the practical benefits of a localized, target-driven approach to ML-guided asymmetric catalysis.

METHODS

In our laboratories, we have previously employed domain-specific approaches to elucidate important structural characteristics of BINOL-derived catalysts.³³ More recently, we developed ML protocols to evaluate catalyst generality, culminating in a computer-driven workflow for identifying widely applicable enantioselective catalysts.³⁴ In these studies, reaction space coverage was chosen based on an attempt to balance relevance with data set size, producing ML models that were well-validated and effective for downstream applications. While the mechanistic basis of the models led to high extrapolative ability,³⁵ the reliance on expert knowledge for reaction selection and database construction presented a significant barrier to adoption for nonspecialists.

Building on this foundation, we questioned how a refined workflow could rigorously investigate the impact of training set reactions unrelated to a target prediction on model performance. This ML framework should incorporate systematic validation to establish a more rigorous basis for training set selection. The workflow is distinguished by its combination of critical design features: (1) a systematic method for training set selection that is adaptable to any reaction and agnostic to specific mechanisms; (2) dynamic tailoring of the training set to emphasize the most relevant data for each prediction by

leveraging proximity in the feature space; (3) potentially robust model performance for predicting outcomes on novel catalyst and substrate structures, including those outside the selectivity range of the training data.

We introduce RaRFR (Figure 2) as a robust method of evaluating the impacts of nonlocal data set diversity. This process is outlined in Figure 2, which serves as an example of a general procedure that can be adopted across unique reaction examples. To initiate, a comprehensive data set of reaction examples is curated (see the section on selected case studies), and molecular fingerprints (ECFP4,³⁶ nbits = 2048) are calculated for each reaction component to define the chemical space. These fingerprints are then concatenated to create an overall unique reaction fingerprint encoding all available structural information (e.g., nucleophile, electrophile, catalyst, solvent). We choose 2D molecular fingerprints as they are straightforward and computationally inexpensive to compute, while also providing similar performance to alternative descriptors in many cases.³⁷ We note, however, that their inability to inherently capture 3D conformations can limit their applicability when conformational effects play a significant role. The binary nature of fingerprints enables us to calculate a physically meaningful distance metric, the Jaccard distance, between a pair of reactions that can be consistently applied over a variety of reactions.^{38,39} This distance is defined by the proportion of features common to both reaction fingerprints as formulated in eq 1 by

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

Where $|A \cap B|$ represents the intersection of reaction fingerprints A and B (i.e., the number of shared bits), and $|A \cup B|$ is the union of reaction fingerprints (i.e., the sum of bits across both reaction fingerprints). This metric is commonly referred to as the Tanimoto similarity⁴⁰ when applied to molecular similarity comparisons in the literature.

Following reaction featurization, the radius-based random forest (RaRF) algorithm is employed. Here, a subset of reactions is dynamically selected based on their proximity in feature space. Specifically, for each reaction to be predicted, the training set is reduced to include only those reactions within a user-defined radius of the target prediction, thus excluding reactions dissimilar to the target. Following this training set reduction, a random forest model is then trained on the reduced data set and used to predict the target reaction. This algorithm is summarized in Supporting Information (Algorithm S1).

While RaRF is fundamentally a neighborhood regression approach, it differs significantly from more conventional methods like k -nearest neighbors (k -NN) regression or standard radius regression. Unlike k -NN regression, which selects a fixed number of neighbors regardless of their similarity to the target reaction, RaRF ensures a consistent level of similarity by using a radius-based cutoff. This distinction is particularly important in sparse data scenarios, where the nearest neighbors in k -NN may vary widely in similarity, making it difficult to draw robust conclusions about the effect of local representation on predictive accuracy. Additionally, both k -NN regression and radius regression rely on simple averaging of neighbor outcomes, a relatively naïve strategy compared to non-neighborhood regression methods. As the neighborhood size increases in these methods,

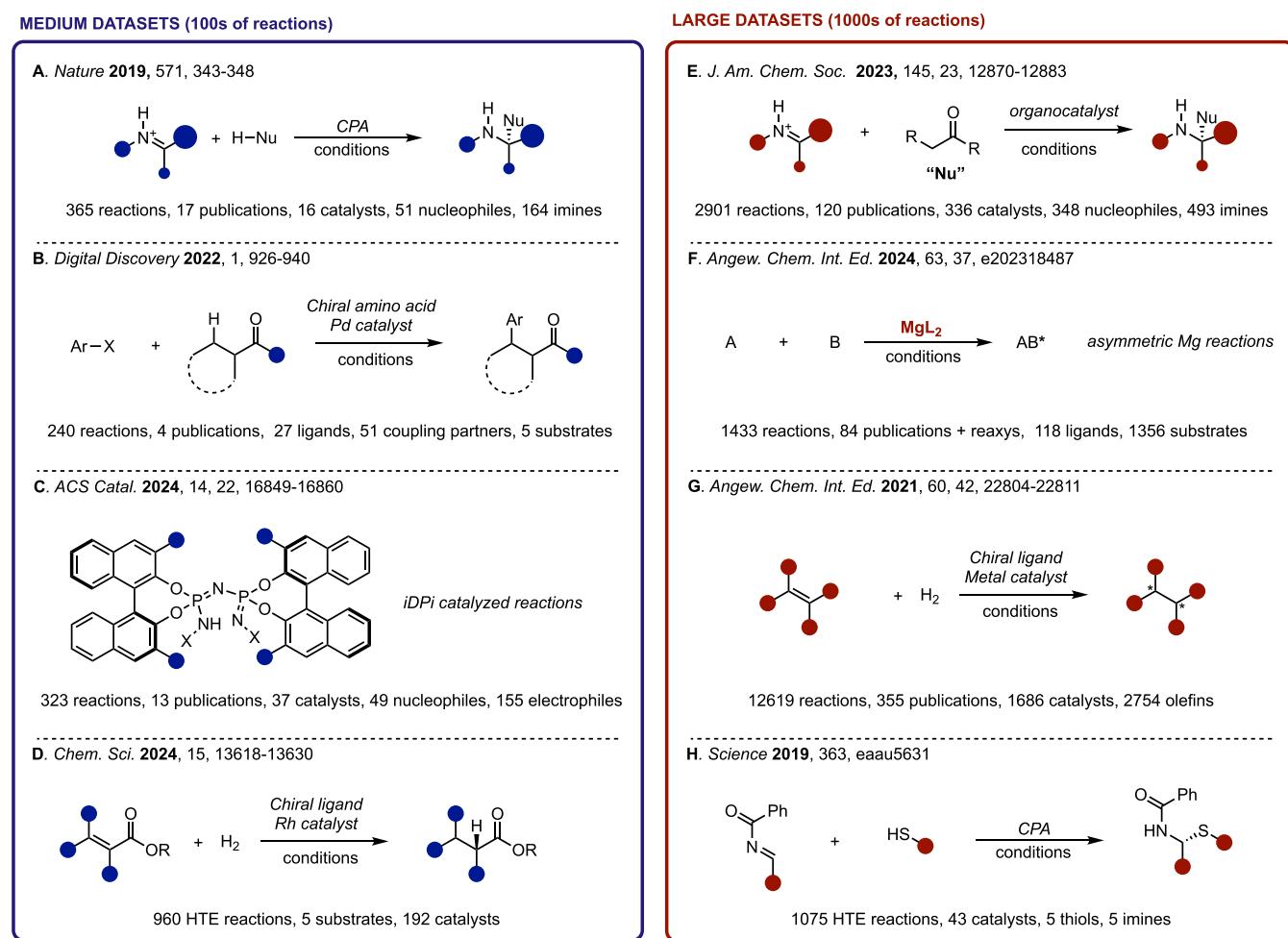


Figure 3. Data sets interrogated in this study.^{8,33,34,46–50}

predictions often degrade as unrelated reactions are incorporated into the training set.

RESULTS AND DISCUSSION

Case Study Selection. With the workflow defined, we benchmarked the approach across eight diverse asymmetric catalysis case studies to examine the effects of locality across data sets with varying heterogeneity, size, structural diversity, and chemical reactivity (Figure 3). The selected data sets encompass a rich body of literature and feature excellent organocatalysts, including chiral phosphoric acids, imidodiphosphorimidates, and secondary amines, along with transition metal catalysts based on palladium, rhodium, or magnesium for example. While numerous other data sets exist, our selection was deliberately tailored to systematically investigate how data set composition influences ML performance.

For example, homogeneous data sets, derived from a single reaction type, provide precise and focused insights but can lack the chemical diversity needed for broader applicability. Conversely, heterogeneous data sets, spanning multiple reaction types, capture greater chemical diversity but are more challenging to model due to increased dimensionality. Even within a single reaction class, significant variability can arise if diversity in catalyst or reaction structures is introduced. For example, in Case Study E, focused on organocatalytic Mannich reactions, the data set includes covalent catalysts (e.g., secondary amines), hydrogen-bonding catalysts (e.g.,

thioureas), and Brønsted acid catalysts (e.g., chiral phosphoric acids), each of which activates substrates and induces selectivity through fundamentally different mechanisms, despite the conserved overall transformation. Taken these ideas into consideration, we included six homogeneous reaction data sets where reactant transformations are consistent (A, B, D, E, G, H), alongside two heterogeneous data sets that span multiple reaction types (C, F).

Similar challenges emerge when comparing data sets derived from literature reports versus those generated using high-throughput experimentation (HTE). Literature data sets often suffer from reporting biases and higher dimensionality, which can hinder the development of truly predictive models.^{17,41–43} HTE data sets, while fully defined in reaction space, can introduce biases arising from combinatorial patterns.^{44,45} As such, we incorporated two HTE-derived data sets (D, H) alongside six literature-based data sets, which vary significantly in their levels of diversity.

Last, we examined the impact of data set size on ML performance by selecting case studies ranging from 240 to 12,619 reactions. These were categorized into medium-sized data sets (between 100 and 1000 reactions, A–D) and large data sets (greater than 1000 reactions, E–H), with four data sets in each category. We reasoned that small data sets containing only tens of reactions (such as the substrate scope from a single report) already constrain the diversity of reaction components and thus are not explored further.

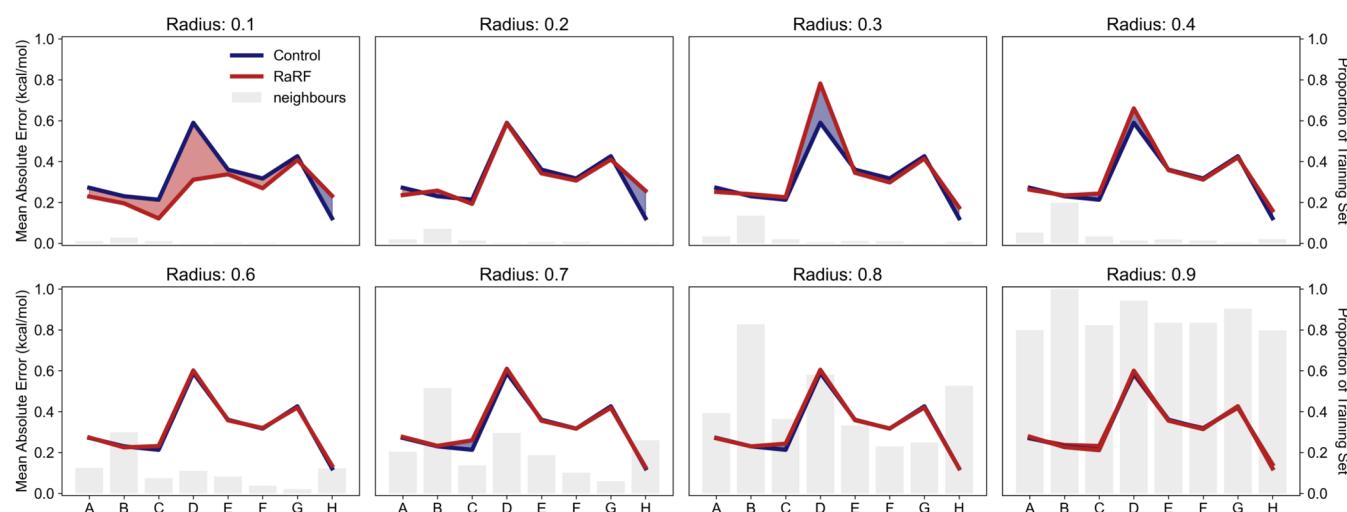


Figure 4. Mean absolute error for RaRFRegression vs full training set controls. Red regions indicate better performance for RaRFRegression, blue regions indicate better performance for full training set control, gray bars indicate proportion of training set used by RaRFRegression. Test sets obtained using random sampling.

Through this comprehensive evaluation, we aim to elucidate how data set composition and characteristics influence ML performance, providing insights for optimizing data generation campaigns to achieve predictive accuracy in asymmetric catalysis. Using these case studies, we applied the RaRF algorithm to address two key questions: (1) how much predictive accuracy is driven by a reaction's nearest neighbors, and (2) whether neighborhood-based approaches to data set collection can outperform diversity-oriented strategies in developing predictive models.

1. Almost All Predictive Performance Stems from a Reaction's Nearest Neighbors. Our first test in evaluating the efficacy of models trained on a restricted data set emulated a scenario where training and test set reactions have a similar distribution of properties. Each data set was split into an 80:20 training/test set through random sampling before being predicted using RaRF, with varying radii cutoffs, or a baseline models trained on the complete data set. More specifically, radii for RaRFRegression were systematically varied between 0.1 and 0.9 with smaller radii representing stricter inclusion criteria, focusing on the most similar reactions, and larger radii encompass broader data sets with decreasing similarity. This approach enables us to evaluate how the degree of similarity in the training data influences model performance. Control models were developed using random forest (RF), neural network (NN), and *k*-nearest neighbors (*k*NN), with hyperparameters optimized via Bayesian optimization. These models were selected for their established utility in predicting organic reaction outcomes. Specifically, RF was chosen for its robustness in avoiding overfitting by averaging outputs across many decision trees and its demonstrated superior performance over other ML techniques in data sets of varying size, as highlighted in recent studies.^{51,52} NN were included for their capacity to model nonlinear relationships and capture complex patterns in high-dimensional data sets. The *k*NN baseline was specifically designed to mimic a chemist's approach, where the closest reaction in the database with a known outcome serves as a direct analog for the prediction. This baseline assesses the utility of extrapolating predictions based solely on proximity in chemical space and is distinctive from RaRF. All models are deployed to predict the difference in transition state energies

leading to competing enantiomeric products, $\Delta\Delta G^\ddagger$, which is related to the enantiomeric ratio according to

$$\Delta\Delta G^\ddagger = -RT \ln(e.r.)$$

Finally, all models are evaluated with mean absolute error (MAE) as it is unbiased by sample size or data distribution.

Figure 4 visualizes this data in a line diagram, where peaks and valleys represent the MAE of the data sets. Blue regions represent cases where the control models outperformed RaRFRegression, red regions indicate superior performance by RaRFRegression, and the gray bars indicate the average proportion of the training set used by the RaRFRegressor. To complement this, eight panels are presented to demonstrate variations with increasing radii. This visualization facilitates identifying relative model performance differences as measured by radius. In each case, the control MAEs remain constant, while RaRF values and the training set sizes will vary. Alternatively, the best results obtained for each case study using either RaRF or the control are tabulated in Table 1.

As demonstrated by these visualizations, similar predictive accuracy is achieved whether the ML model has access to the full training set or is restricted to reactions within a defined structural neighborhood. In most cases, the MAE differences

Table 1. Best Mean Absolute Error Obtained for Each Case Study

case study	best RaRF(kcal/mol)	control (kcal/mol)
A	0.26 (0.4) ^a	0.27
B	0.22 (0.6) ^a	0.23
C	0.21 (0.9) ^a	0.21
D	0.58 (0.5) ^a	0.59
E	0.36 (0.5) ^a	0.36
F	0.27 (0.1) ^a	0.32
G	0.42 (0.7) ^a	0.43
H	0.12 (0.8) ^a	0.12

^aBracketed numbers denote the radius for each respective RaRF model. Best RaRF models are restricted to those where all test set reactions have at least one neighbor. Where multiple models have the same performance (to two significant figures), the lowest radius is shown.

between approaches are minimal, often within 0.1 kcal/mol, even though the RaRFRregression model uses only a fraction of the training set. At the smallest radii (e.g., 0.1), where only tens of reactions are typically used for training, RaRFRregression outperformed full training set approaches across all case studies except for Denmark's thiol addition data set (Case Study H). This superior performance can be attributed to the presence of highly representative examples within the training set, however we note that, at low radii (typically <0.3), this prediction boost can somewhat be attributed to the removal of difficult test reactions which have no training set reactions within the set radii. Notably, these results are consistent across multiple train/test splits, though some variation is observed especially at the smallest radii (Figure S9).

The performance differences between RaRFRregression and full training set models narrows significantly as the radii increase, with differences between the blue and red lines becoming virtually negligible. For example, at a radius of 0.6, where less than 20% of the data is typically used, the models demonstrate nearly indistinguishable performance and continuing until radius = 0.9. The most intriguing variations arise at intermediate radii, where no consistent trend links performance to radius size. Importantly, profound contrasts emerge between high-throughput experimentation (HTE) data sets (e.g., D, H) and literature-derived data sets. For instance, while full training set controls generally outperformed RaRFRregression for HTE data sets, the rhodium-catalyzed hydrogenation data set displayed substantial fluctuations. RaRFRregression excelled at low radii (0.1, 0.2), delivering superior predictions, but faltered at intermediate radii (0.3, 0.4), illustrating the complexity of balancing neighborhood size and prediction accuracy. Interestingly, the overall differences remained modest, with the largest variations hovering around ~0.2 kcal/mol, and most cases showing even smaller discrepancies.

The Denmark data set focused on CPA-catalyzed thiol additions stands out as a particularly challenging case.⁵⁰ This data set encompasses 43 CPA catalysts arranged in a 5×5 matrix of nucleophiles and imines, resulting in a total of 1075 reactions. In this data set, the catalysts were intentionally selected to maximize structural diversity, posing a significant obstacle for neighborhood-based models. As expected, this data set was the sole case where full training set models consistently outperformed RaRFRregression across all radii. Even so, RaRFRregression demonstrated its efficiency by achieving comparable performance to the full training set models. Remarkably, at a radius of 0.5, RaRFRregression required an average of just 41 out of the ~700 available training reactions yet delivered predictions that were statistically indistinguishable from those of the full training set (Figure 5).

With these results in hand, we next questioned if similarly small gains in model accuracy would be observed in more difficult prediction scenarios. In this regard, perhaps the most difficult tasks involve predictions for reactions with no explicit overlap with the training set. To explore such cases, we employed leave-one-reaction-out (LORO) analysis. In this approach, a single reaction (based on a specific publication) is excluded from the data set and treated as the validation set. Next, the model is retrained on the remaining reactions and used to predict the excluded one. For HTE-like data sets not sourced from literature, we adapted this methodology to a leave-one-catalyst-out analysis, where individual catalysts or

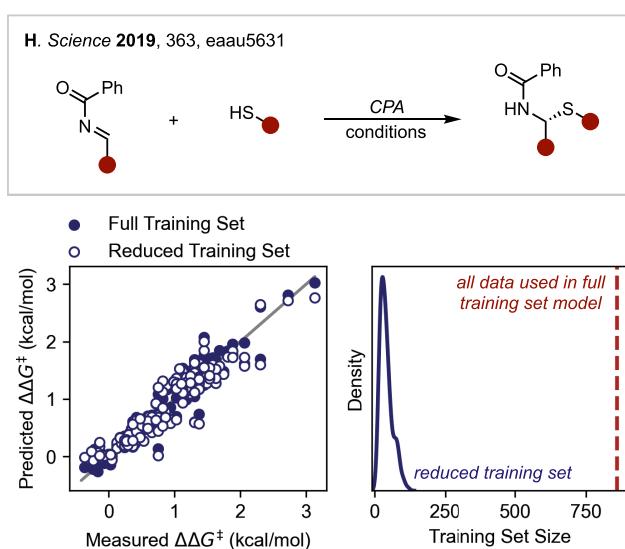


Figure 5. Comparison of RaRFRregression predictions at radius = 0.5 versus the full training set control (left) and kernel density estimate plot of the average training set size used for each prediction.

chiral ligands were excluded and subsequently predicted. These results are visualized in Figure 6 and the best results where all test set reactions have at least one neighbor tabulated in Table 2.

As expected, the neighborhood around prediction points was noticeably sparser in LORO analyses compared to stratified sampling tests. In certain case studies, no suitable neighbors were available for training until moderate radius thresholds (>0.3), and radii exceeding 0.7 were often necessary to ensure all test set reactions had at least one neighbor in the training set. Consequently, much larger variations are observed between case studies at low radii. An intriguing example arose in case study C (iDPI reactions), where the control model outperformed RaRFRregression by nearly 1 kcal/mol at a radius of 0.4. Upon further analysis, we suggest that this disparity likely stems from limited training data—only five reactions had neighbors within this radius. This is supported by the near indistinguishable performance between RaRFRregression and controls for this case study at a radius of 0.7, where ~5% of the data is used for each prediction. In general, while the neighborhood-based approach occasionally surpassed full training set models at very small radii, the full training set controls generally provided better predictions for most radii. Nevertheless, RaRFRregression largely matched the performance of full training set controls across all case studies above radii = 0.6. These findings highlight that adding diversity to training sets enhances predictive accuracy by a maximum of ~0.25–0.5 kcal/mol when predictions lack direct representation in the training set.

Across all cases, we consistently observed that prediction accuracy within 0.5 kcal/mol could be achieved using a fraction of the data typically required by traditional models like RF, k-NN, or shallow neural networks when utilizing molecular fingerprints. To further test our model's capability, we compared it against state-of-the-art deep neural networks (DNNs) for predicting selectivity. Here, DNNs refer to neural networks containing more than one hidden layer. This benchmark, provided by Sunoj and co-workers,⁴⁶ involved palladium-catalyzed C–H activation reactions (case study B) modeled using a DNN architecture with five hidden layers and

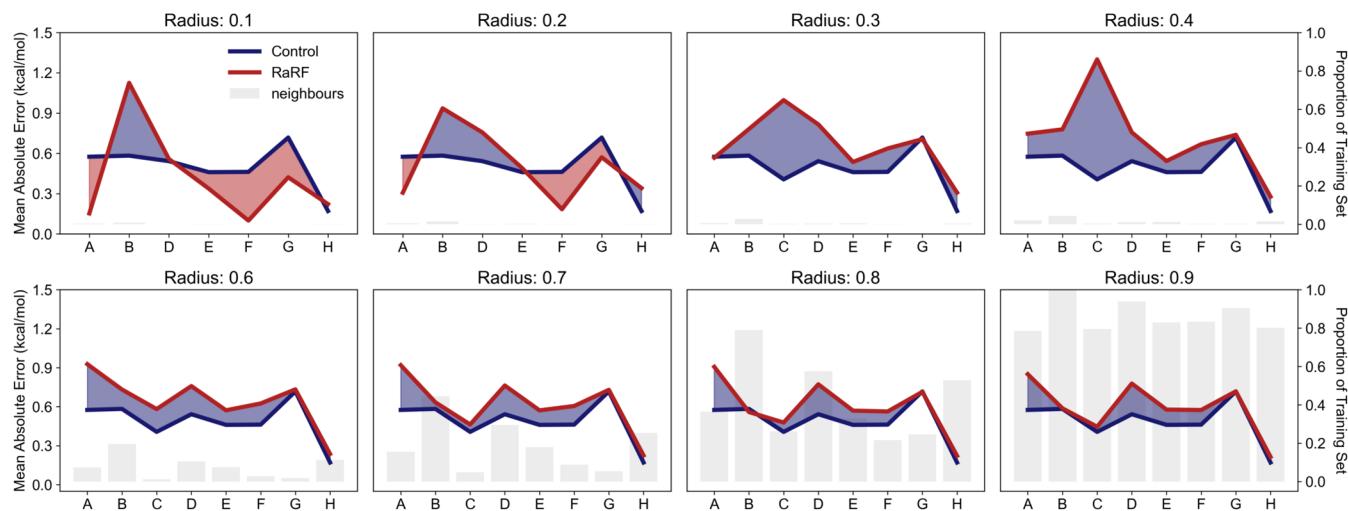


Figure 6. Mean absolute error for leave-one-reaction-out (LORO) or leave-one-catalyst-out analysis. Red regions indicate better performance for RaRFRregression, blue regions indicate better performance for full training set control, gray bars indicate proportion of training set used by RaRFRregression.

Table 2. Best Mean Absolute Error Obtained for Each LORO/LOCO Case Study

case study	best RaRF(kcal/mol)	control(kcal/mol)
A	0.85 (0.9) ^a	0.58
B	0.56 (0.8) ^a	0.58
C	0.45 (0.9) ^a	0.41
D	0.75 (0.5) ^a	0.54
E	0.57 (0.8) ^a	0.46
F	0.57 (0.9) ^a	0.46
G	0.71 (0.8) ^a	0.72
H	0.21 (0.9) ^a	0.17

^aBracketed numbers denote the radius for each respective RaRF model. Best RaRF models are restricted to those where all test set reactions have at least one neighbor. Where multiple models have the same performance (to two significant figures), the lowest radius is shown.

up to 400 neurons. Furthermore, their approach relied on tailored features derived from DFT-calculated structures, thus enabling us to determine the efficacy of fingerprint-based neighborhoods combined with complex DFT-based features for prediction.

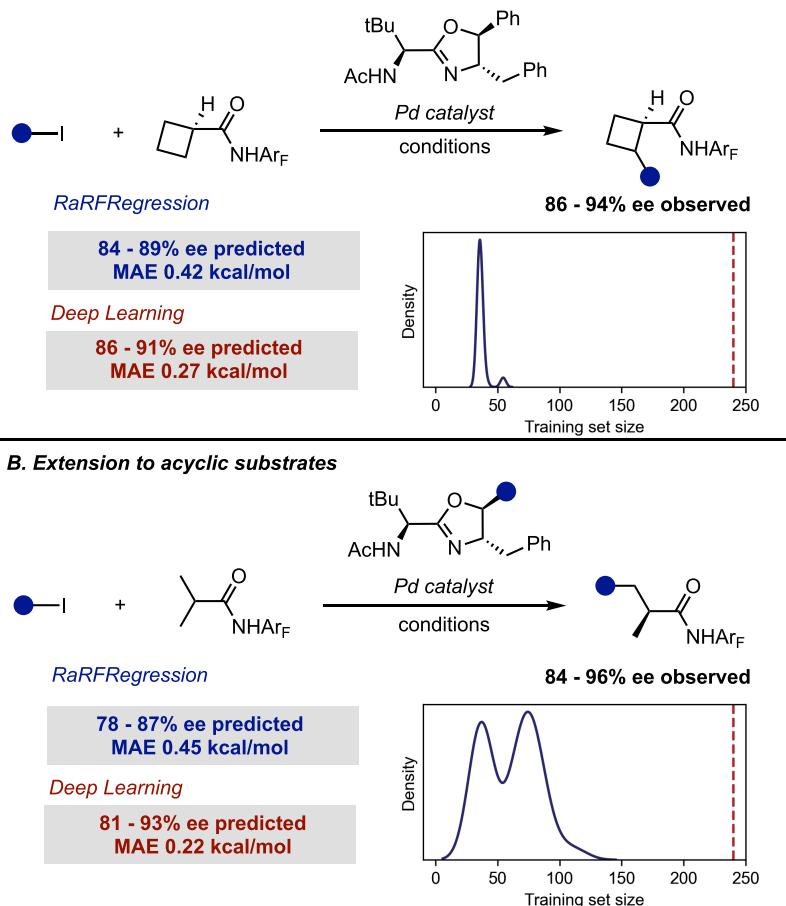
As part of this effort, we applied RaRFRregression to three out-of-sample reactions involving new ligands, substrates, or both, using the smallest radius (0.8) that ensures all test set reactions have at least one neighbor in the training set. Here, reaction-specific DFT features, such as NBO charges, IR frequencies, and buried volume, were incorporated to predict selectivity following neighborhood generation. Subsequently, the results were compared against the optimal DNN architecture. Remarkably, despite relying on simpler methods and using only a fraction of the training data, RaRFRregression underperformed the DNN by just 0.2 kcal/mol for the first two case studies (Figure 7A7B). Moreover, in a case study involving entirely new ligands and substrate types, RaRFRregression exceeded DNN performance (Figure 7C). These results suggest that increasing model complexity offers diminishing returns in predictive accuracy. Additionally, they emphasize the limited value of unrelated training set diversity, with such additions still improving accuracy by less than 0.5 kcal/mol.

#2. Neighborhood Approaches Outperform Diversity-Oriented Approaches. Having established that adding diversity unrelated to the target prediction only marginally improves model performance, we next explored how this insight could be applied at the outset of experimental campaigns in asymmetric catalysis. Researchers often have a clear goal for the types of reactions they aim to model but face uncertainty about the optimal data to collect for effective modeling. While many studies advocate for maximizing data set diversity as a primary objective when designing training sets for cheminformatics applications, recent investigations indicate that even extremely large and diverse data sets encompassing numerous catalytic reactions often fail to produce accurate models for predicting yields or reaction conditions.^{41,53} This challenge likely persists when predicting enantioselectivities, where small energy differences in transition states—on the order of kcal/mol—can result in significant variations in observed selectivities. To simulate the initial stages of a cheminformatics-driven exploration, we aimed to answer a critical question: what is more important for prediction accuracy—local neighborhood representation or data set diversity?

To address this, we developed the *k*Reduction method, designed to simulate a diversity-oriented approach to data set curation. In brief, *k*Reduction employs *k*-medoids clustering to group similar reactions, and retains only the medoids while removing all other reactions. This ensures that the curated data set maximizes diversity while minimizing redundancy. Using the reduced training set, a random forest model is then applied to predict target reactions in the same manner as the RaRFRregressor. While both RaRFRregression and *k*Reduction approaches incorporate some degree of local representation, their underlying strategies differ significantly (Figure 8). The diversity-oriented *k*Reduction approach emphasizes broad coverage of chemical space, ensuring representation across diverse reaction types. In contrast, the neighborhood-based RaRFRregression approach prioritizes dense coverage of local chemical space, focusing on reactions most similar to the target.

We applied both RaRFRregressor and *k*Reduction methods across multiple case studies using radii ranging from 0 to 1

A. Extension to new ligand and coupling partners



C. Extension to new ligand and substrates

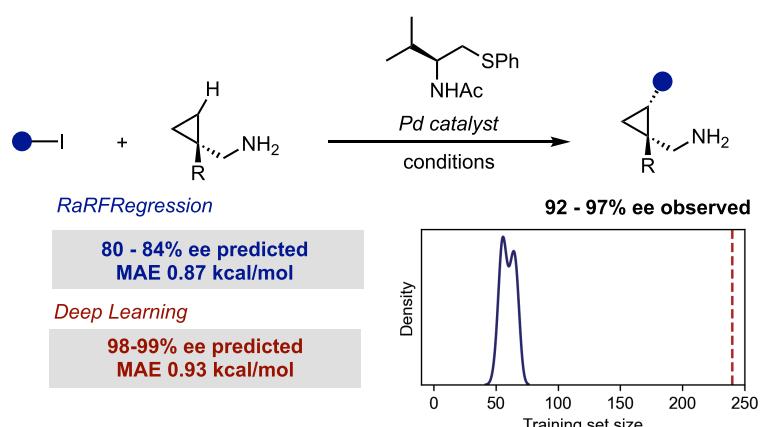


Figure 7. Extrapolation to out-of-sample reactions.

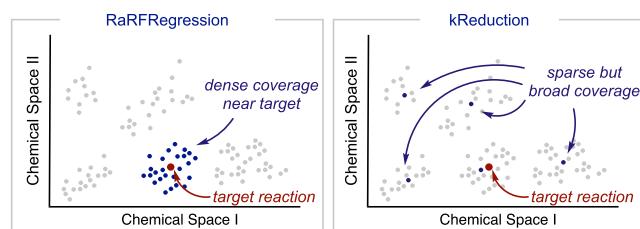


Figure 8. Comparison of neighborhood-oriented models (RaRF, left) and diversity-oriented models (right, k Reduction).

with a step size of 0.01. To ensure a fair comparison, the training set sizes for *k*Reduction were matched to the average training set size used by RaRFRRegressor at each radius. Specifically, the number of clusters in *k*Reduction corresponded to the average number of neighbors employed by RaRFRRegressor for predictions at the same radius. Recognizing that training set composition can significantly influence model performance, we repeated each simulation 10 times with different training-test splits. For case study G, we reduced the number of tests to 3 training-test splits and a step size of 0.1 due to computational constraints given the large data set size (12,619 reactions). As a benchmark, control random forest

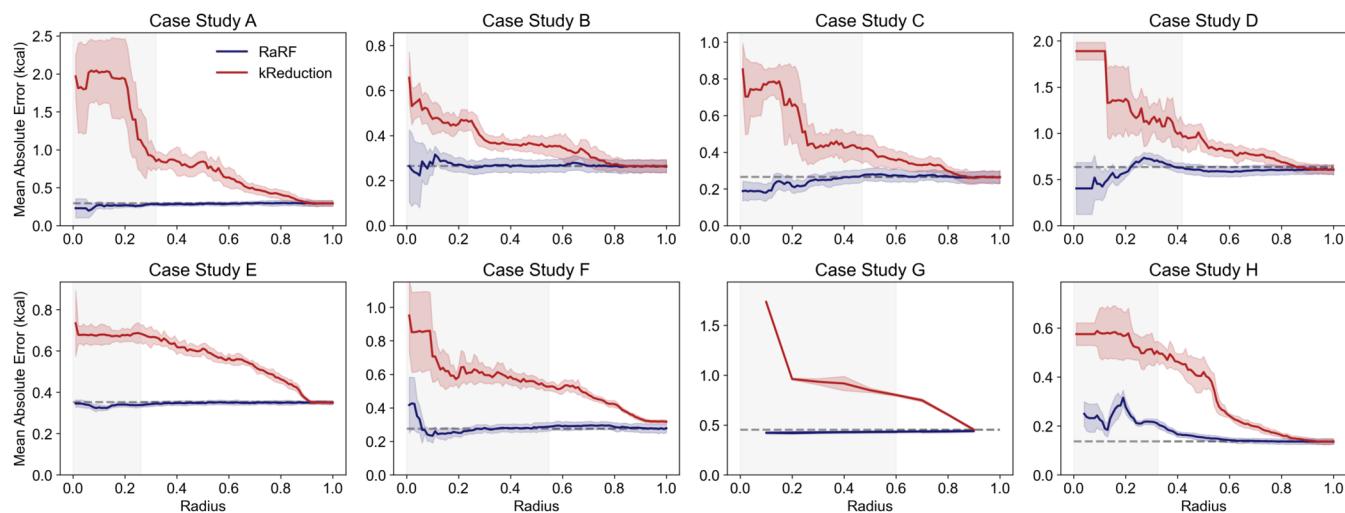


Figure 9. Mean absolute error for RaRFRegression (neighborhood-oriented, blue) or *k*Reduction (diversity-oriented, red) approaches at varying radii. Training set sizes for *k*Reduction are set such that they are equal to the average training set size used by the RaRFRegressor at each respective radii. Gray regions denote at radii where at least one test set reaction has no training set neighbors.

models with no training set reduction were also evaluated. The results, visualized in Figure 9, are summarized with bold lines representing the average performance across runs, and the shaded regions representing one standard deviation from the mean. Last, the gray shaded regions denote radii where at least one test set reaction cannot be predicted.

Strikingly, RaRFRegression consistently outperformed *k*Reduction across all case studies at small to medium radius values. While RaRFRegression closely mirrored the performance of full training set controls, the *k*Reduction approach yielded predictions that were up to 2 kcal/mol less accurate with small training sets. Although some variability was observed between literature-derived data sets versus HTE data sets, as well as between small and large reaction sets, the overarching trend was clear: prioritizing target-aware data sets significantly accelerates the development of robust models compared to a untargeted focus on maximizing diversity. Notably, the diversity-oriented method only began to approach the performance of the neighborhood-based tactic when at the highest radii (>0.8). Given these observations, we argue that the data set diversity is only critical to the extent that the target prediction's neighborhood is sufficiently populated. For practitioners aiming to develop models for specific reaction predictions, targeted training set curation should take precedence over a broad-diversity approach.

CONCLUSIONS

This study provides critical insights into the relationship between data set diversity and predictive performance in cheminformatics. While the conventional approach often advocates for maximizing data set diversity, our findings reveal that such efforts yield limited benefits unless they directly enhance the representation of the local neighborhood around the target predictions. Instead, we highlight the pivotal role of target-focused data set curation, where training sets are thoughtfully constructed to include data points that closely resemble the specific reactions or outcomes being modeled. Given the immense dimensionality of chemical space and difficulties in obtaining high-quality, consistent, and diverse data, the development of a universal “Oracle-like” predictive model in asymmetric catalysis will be extremely challenging.

Thus, it is by addressing these fundamental data limitations and tailoring data sets to specific target reactions—rather than solely focusing on algorithmic complexity—that predictive models in asymmetric catalysis can be obtained.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acscatal.5c01051>.

Computational modeling details and all tabulated results (PDF)

Repository of code and data (ZIP)

AUTHOR INFORMATION

Corresponding Author

Jolene P. Reid – Department of Chemistry, University of British Columbia, Vancouver, British Columbia V6T 1Z1, Canada; orcid.org/0000-0003-2397-0053; Email: jreid@chem.ubc.ca

Authors

Isaiah O. Betinol – Department of Chemistry, University of British Columbia, Vancouver, British Columbia V6T 1Z1, Canada; orcid.org/0000-0001-9953-7761

Aleksandra Demchenko – Department of Chemistry, University of British Columbia, Vancouver, British Columbia V6T 1Z1, Canada

Complete contact information is available at: <https://pubs.acs.org/10.1021/acscatal.5c01051>

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Financial support to J.P.R. was provided by the University of British Columbia, the Natural Sciences and Engineering Research Council of Canada (NSERC) and the CFI John R. Evans Leaders Fund. I.O.B. acknowledges NSERC for a PGSD Research Fellowship. A.D. acknowledges the University of British Columbia for a 4YF fellowship. Computational

resources were provided by the Digital Research Alliance of Canada and the Advanced Research Computing (ARC) center at the University of British Columbia.

■ REFERENCES

- (1) Bohacek, R. S.; McMinn, C.; Guida, W. C. The Art and Practice of Structure-Based Drug Design: A Molecular Modeling Perspective. *Med. Res. Rev.* **1996**, *16* (1), 3–50.
- (2) Tielens, A. G. G. M. The Molecular Universe. *Rev. Mod. Phys.* **2013**, *85* (3), No. 1021.
- (3) Wagen, C. C.; McMinn, S. E.; Kwan, E. E.; Jacobsen, E. N. Screening for Generality in Asymmetric Catalysis. *Nature* **2022**, *610*, 680–686.
- (4) Howard, J. R.; Shuluk, J. R.; Bhakare, A.; Anslyn, E. V. Data-Science-Guided Calibration Curve Prediction of an MLCT-Based Ee Determination Assay for Chiral Amines. *Chem.* **2024**, *10* (7), 2074–2088.
- (5) Nie, W.; Wan, Q.; Sun, J.; Chen, M.; Gao, M.; Chen, S. Ultra-High-Throughput Mapping of the Chemical Space of Asymmetric Catalysis Enables Accelerated Reaction Discovery. *Nat. Commun.* **2023**, *14* (1), No. 6671.
- (6) Korch, K. M.; Hayes, J. C.; Kim, R. S.; Sampson, J.; Kelly, A. T.; Watson, D. A. Selected Ion Monitoring Using Low-Cost Mass Spectrum Detectors Provides a Rapid, General, and Accurate Method for Enantiomeric Excess Determination in High-Throughput Experimentation. *ACS Catal.* **2022**, *12* (11), 6737–6745.
- (7) Kearnes, S. M.; Maser, M. R.; Wleklinski, M.; Kast, A.; Doyle, A. G.; Dreher, S. D.; Hawkins, J. M.; Jensen, K. F.; Coley, C. W. The Open Reaction Database. *J. Am. Chem. Soc.* **2021**, *143* (45), 18820–18826.
- (8) Xu, L.; Zhang, S.; Li, X.; Tang, M.; Xie, P.; Hong, X. Towards Data-Driven Design of Asymmetric Hydrogenation of Olefins: Database and Hierarchical Learning. *Angew. Chem., Int. Ed.* **2021**, *60* (42), 22804–22811.
- (9) Dai, H.; Li, C.; Coley, C.; Dai, B.; Song, L. *Retrosynthesis Prediction with Conditional Graph Logic Network*; Wallach, H.; Larochelle, H.; Beygelzimer, A.; Alché-Buc, F. D.; Fox, E.; Garnett, R., Eds.; Proceedings of the 33rd International Conference on Neural Information Processing Systems; Curran Associates, Inc., 2019; pp 8872–8882.
- (10) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Sci. Data* **2014**, *1* (1), No. 140022.
- (11) Wells, P. R. Linear Free Energy Relationships. *Chem. Rev.* **1963**, *63* (2), 171–219.
- (12) Hammett, L. P. The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives. *J. Am. Chem. Soc.* **1937**, *59* (1), 96–103.
- (13) Crawford, J. M.; Kingston, C.; Toste, F. D.; Sigman, M. S. Data Science Meets Physical Organic Chemistry. *Acc. Chem. Res.* **2021**, *54* (16), 3136–3148.
- (14) Sigman, M. S.; Harper, K. C.; Bess, E. N.; Milo, A. The Development of Multidimensional Analysis Tools for Asymmetric Catalysis and Beyond. *Acc. Chem. Res.* **2016**, *49* (6), 1292–1301.
- (15) Reid, J. P.; Betinol, I. O.; Kuang, Y. Mechanism to Model: A Physical Organic Chemistry Approach to Reaction Prediction. *Chem. Commun.* **2023**, *59* (72), 10711–10721.
- (16) Cuomo, A. E.; Ibarraran, S.; Sreekumar, S.; Li, H.; Eun, J.; Menzel, J. P.; Zhang, P.; Buono, F.; Song, J. J.; Crabtree, R. H.; Batista, V. S.; Newhouse, T. R. Feed-Forward Neural Network for Predicting Enantioselectivity of the Asymmetric Negishi Reaction. *ACS Cent. Sci.* **2023**, *9* (9), 1768–1774.
- (17) Schleinitz, J.; Langevin, M.; Smail, Y.; Wehnert, B.; Grimaud, L.; Vuilleumier, R. Machine Learning Yield Prediction from NiCOLit, a Small-Size Literature Data Set of Nickel Catalyzed C–O Couplings. *J. Am. Chem. Soc.* **2022**, *144* (32), 14722–14730.
- (18) Shim, E.; Kammeraad, J. A.; Xu, Z.; Tewari, A.; Cernak, T.; Zimmerman, P. M. Predicting Reaction Conditions from Limited Data through Active Transfer Learning. *Chem. Sci.* **2022**, *13* (22), 6655–6668.
- (19) Kim, J. Y.; Khan, S. A.; Vlachos, D. G. Similarity-Based Machine Learning for Small Data Sets: Predicting Biolubricant Base Oil Viscosities. *J. Phys. Chem. B* **2024**, *128* (48), 11963–11970.
- (20) Lemm, D.; von Rudorff, G. F.; von Lilienfeld, O. A. Improved Decision Making with Similarity Based Machine Learning: Applications in Chemistry. *Mach. Learn. Sci. Technol.* **2023**, *4* (4), No. 045043.
- (21) Yang, S.; Sun, M.; Shi, C.; Liu, Y.; Guo, Y.; Liu, Y.; Lu, Z.; Huang, Y.; Pu, X. Data-Quality-Navigated Machine Learning Strategy with Chemical Intuition to Improve Generalization. *J. Chem. Theory Comput.* **2024**, *20* (23), 10633–10648.
- (22) Yu, J.; Zhu, L.; Qin, R.; Zhang, Z.; Li, L.; Huang, T. Combining K-Means Clustering and Random Forest to Evaluate the Gas Content of Coalbed Bed Methane Reservoirs. *Geofluids* **2021**, *2021*, No. 9321565.
- (23) Xie, T.; Wittreich, G. R.; Curnan, M. T.; Gu, G. H.; Seals, K. N.; Tolbert, J. S. Machine-Learning-Enabled Thermochemistry Estimator. *J. Chem. Inf. Model.* **2025**, *65* (1), 214–222.
- (24) Xerxa, E.; Vogt, M.; Bajorath, J. Influence of Data Curation and Confidence Levels on Compound Predictions Using Machine Learning Models. *J. Chem. Inf. Model.* **2024**, *64* (24), 9341–9349.
- (25) Schleinitz, J.; Carretero-Cerdán, A.; Gurajapu, A.; Harnik, Y.; Lee, G.; Pandey, A.; Milo, A.; Reisman, S. E. Designing Target-Specific Data Sets for Regioselectivity Predictions on Complex Substrates. *J. Am. Chem. Soc.* **2025**, *147* (9), 7476–7484.
- (26) Park, S.; Han, H.; Kim, H.; Choi, S. Machine Learning Applications for Chemical Reactions. *Chem. - Asian J.* **2022**, *17* (14), No. e202200203.
- (27) Toyao, T.; Maeno, Z.; Takakusagi, S.; Kamachi, T.; Takigawa, I.; Shimizu, K. Machine Learning for Catalysis Informatics: Recent Applications and Prospects. *ACS Catal.* **2020**, *10* (3), 2260–2297.
- (28) Strieth-Kalthoff, F.; Sandfort, F.; Segler, M. H. S.; Glorius, F. Machine Learning the Ropes: Principles, Applications and Directions in Synthetic Chemistry. *Chem. Soc. Rev.* **2020**, *49* (17), 6154–6168.
- (29) Pinus, S.; Genzling, J.; Burai-Patrascu, M.; Moitessier, N. Computational Methods for Asymmetric Catalysis. *Nat. Catal.* **2024**, *7* (12), 1272–1287.
- (30) Abraham, B. M.; Jyothirmai, M. V.; Sinha, P.; Viñes, F.; Singh, J. K.; Illas, F. Catalysis in the Digital Age: Unlocking the Power of Data with Machine Learning. *WIREs Comput. Mol. Sci.* **2024**, *14* (5), No. e1730.
- (31) Shi, Y.-F.; Yang, Z.-X.; Ma, S.; Kang, P.-L.; Shang, C.; Hu, P.; Liu, Z.-P. Machine Learning for Chemistry: Basics and Applications. *Engineering* **2023**, *27*, 70–83.
- (32) Raghavan, P.; Haas, B. C.; Ruos, M. E.; Schleinitz, J.; Doyle, A. G.; Reisman, S. E.; Sigman, M. S.; Coley, C. W. Dataset Design for Building Models of Chemical Reactivity. *ACS Cent. Sci.* **2023**, *9* (12), 2196–2204.
- (33) Reid, J. P.; Sigman, M. S. Holistic Prediction of Enantioselectivity in Asymmetric Catalysis. *Nature* **2019**, *571* (7765), 343–348.
- (34) Betinol, I. O.; Lai, J.; Thakur, S.; Reid, J. P. A Data-Driven Workflow for Assigning and Predicting Generality in Asymmetric Catalysis. *J. Am. Chem. Soc.* **2023**, *145* (23), 12870–12883.
- (35) Betinol, I. O.; Kuang, Y.; Reid, J. P. Guiding Target Synthesis with Statistical Modeling Tools: A Case Study in Organocatalysis. *Org. Lett.* **2022**, *24* (7), 1429–1433.
- (36) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742–754.
- (37) Sidorov, P.; Tsuji, N. A Primer on 2D Descriptors in Selectivity Modeling for Asymmetric Catalysis. *Chem. - Eur. J.* **2024**, *30* (10), No. e202302837.
- (38) Bajusz, D.; Rácz, A.; Héberger, K. Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *J. Cheminf.* **2015**, *7* (1), No. 20.
- (39) Todeschini, R.; Consonni, V.; Xiang, H.; Holliday, J.; Buscema, M.; Willett, P. Similarity Coefficients for Binary Chemoinformatics

- Data: Overview and Extended Comparison Using Simulated and Real Data Sets. *J. Chem. Inf. Model.* **2012**, *52* (11), 2884–2901.
- (40) Tanimoto, T. T. IBM Internal Report 17th November, 1957.
- (41) Beker, W.; Roszak, R.; Wolos, A.; Angello, N. H.; Rathore, V.; Burke, M. D.; Grzybowski, B. A. Machine Learning May Sometimes Simply Capture Literature Popularity Trends: A Case Study of Heterocyclic Suzuki–Miyaura Coupling. *J. Am. Chem. Soc.* **2022**, *144* (11), 4819–4827.
- (42) 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.
- (43) Li, D.-Z.; Gong, X.-Q. Challenges with Literature-Derived Data in Machine Learning for Yield Prediction: A Case Study on Pd-Catalyzed Carbonylation Reactions. *J. Phys. Chem. A* **2024**, *128* (48), 10423–10430.
- (44) Chuang, K. V.; Keiser, M. J. Comment on “Predicting Reaction Performance in C–N Cross-Coupling Using Machine Learning.” *Science* **2018**, *362* (6416), No. eaat8603.
- (45) Zahrt, A. F.; Henle, J. J.; Denmark, S. E. Cautionary Guidelines for Machine Learning Studies with Combinatorial Datasets. *ACS Comb. Sci.* **2020**, *22* (11), 586–591.
- (46) Hoque, A.; Sunoj, R. B. Deep Learning for Enantioselectivity Predictions in Catalytic Asymmetric β -C–H Bond Activation Reactions. *Digital Discovery* **2022**, *1* (6), 926–940.
- (47) Sanocki, M.; Russell, H. C.; Handjaya, J.; Reid, J. P. Relative Generality and Risk: Quantitative Measures for Broad Catalyst Success. *ACS Catal.* **2024**, *14* (22), 16849–16860.
- (48) Kalikadien, A. V.; Valsecchi, C.; van Putten, R.; Maes, T.; Muuronen, M.; Dyubankova, N.; Lefort, L.; Pidko, E. A. Probing Machine Learning Models Based on High Throughput Experimentation Data for the Discovery of Asymmetric Hydrogenation Catalysts. *Chem. Sci.* **2024**, *15* (34), 13618–13630.
- (49) Baczevska, P.; Kulczykowski, M.; Zambroń, B.; Jaszczebska-Adamczak, J.; Pakulski, Z.; Roszak, R.; Grzybowski, B. A.; Mlynarski, J. Machine Learning Algorithm Guides Catalyst Choices for Magnesium-Catalyzed Asymmetric Reactions. *Angew. Chem., Int. Ed.* **2024**, *63* (37), No. e202318487.
- (50) Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. Prediction of Higher-Selectivity Catalysts by Computer-Driven Workflow and Machine Learning. *Science* **2019**, *363* (6424), No. eaau5631.
- (51) Singh, S.; Pareek, M.; Changotra, A.; Banerjee, S.; Bhaskararao, B.; Balamurugan, P.; Sunoj, R. B. A Unified Machine-Learning Protocol for Asymmetric Catalysis as a Proof of Concept Demonstration Using Asymmetric Hydrogenation. *Proc. Natl. Acad. Sci. U.S.A.* **2020**, *117* (3), 1339–1345.
- (52) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting Reaction Performance in C–N Cross-Coupling Using Machine Learning. *Science* **2018**, *360* (6385), 186–190.
- (53) Saebi, M.; Nan, B.; Herr, J. E.; Wahlers, J.; Guo, Z.; Zurański, A. M.; Kogej, T.; Norrby, P.-O.; Doyle, A. G.; Chawla, N. V.; Wiest, O. On the Use of Real-World Datasets for Reaction Yield Prediction. *Chem. Sci.* **2023**, *14* (19), 4997–5005.

CAS BIOFINDER DISCOVERY PLATFORM™

ELIMINATE DATA SILOS. FIND WHAT YOU NEED, WHEN YOU NEED IT.

A single platform for relevant, high-quality biological and toxicology research

Streamline your R&D

CAS
A division of the
American Chemical Society