JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JACS

Article

Incorporating Synthetic Accessibility in Drug Design: Predicting Reaction Yields of Suzuki Cross-Couplings by Leveraging AbbVie's 15-Year Parallel Library Data Set

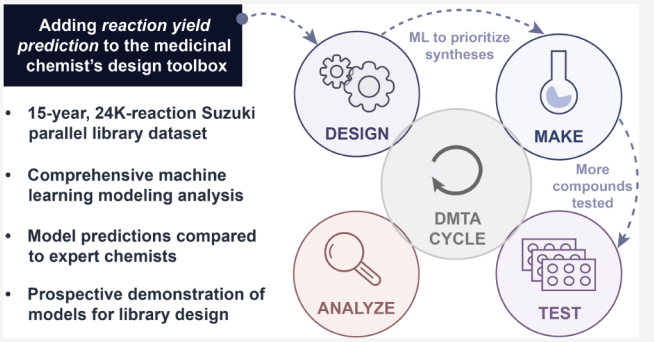
Priyanka Raghavan,[‖] Alexander J. Rago,[‖] Pritha Verma, Majdi M. Hassan, Gashaw M. Goshu, Amanda W. Dombrowski, Abhishek Pandey, Connor W. Coley,* and Ying Wang*

Cite This: *J. Am. Chem. Soc.* 2024, 146, 15070−15084

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Despite the increased use of computational tools to supplement medicinal chemists' expertise and intuition in drug design, predicting synthetic yields in medicinal chemistry endeavors remains an unsolved challenge. Existing design workflows could profoundly benefit from reaction yield prediction, as precious material waste could be reduced, and a greater number of relevant compounds could be delivered to advance the design, make, test, analyze (DMTA) cycle. In this work, we detail the evaluation of AbbVie's medicinal chemistry library data set to build machine learning models for the prediction of Suzuki coupling reaction yields. The combination of density functional theory (DFT)-derived features and Morgan fingerprints was identified to perform better than one-hot encoded baseline modeling, furnishing encouraging results. Overall, we observe modest generalization to unseen reactant structures within the 15-year retrospective library data set. Additionally, we compare predictions made by the model to those made by expert medicinal chemists, finding that the model can often predict both reaction success and reaction yields with greater accuracy. Finally, we demonstrate the application of this approach to suggest structurally and electronically similar building blocks to replace those predicted or observed to be unsuccessful prior to or after synthesis, respectively. The yield prediction model was used to select similar monomers predicted to have higher yields, resulting in greater synthesis efficiency of relevant drug-like molecules.

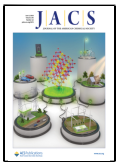


INTRODUCTION

Hundreds to thousands of molecules need to be designed, made, tested, and analyzed (DMTA) in small-molecule drug discovery programs before one compound is chosen as the clinical candidate molecule.[1] Often, the synthesis of target compounds is the bottleneck of this process, most significantly contributing to overall DMTA cycle times.[2] In this context, the translation of designed molecules from paper to experimentally accessible molecules is of critical importance to the efficiency of the DMTA cycles. The design phase in the cycle generally involves careful triaging of potential synthetic targets of interest through a combination of techniques. These often include calculation of physiochemical properties,[3] prediction of activity[4] and/or ADME properties,[5] molecular diversity analysis,[6] or using protein−ligand docking scores[7] to estimate how well the compounds might bind to the target of interest. The extent to which these tools are used varies from program-to-program, but ultimately these techniques guide synthetic efforts toward the next best molecules for projects.

Chemical libraries synthesized in parallel are particularly useful for enabling efficient and rapid exploration of structure−activity relationships (SAR) in medicinal chemistry programs. Unlike singleton synthesis, the synthesis of many compound analogues in parallel is often accomplished using a robust transformation that enables the simultaneous evaluation of multiple SAR hypotheses and can therefore help reduce the bottleneck that synthesis can impose on the DMTA cycle. While libraries have become routine practices in most modern medicinal chemistry programs, the selection of library analogues does not explicitly consider the likelihood of a successful synthesis. The need for improving such considerations has been demonstrated with the analysis of AbbVie's internal library synthesis data set. The study highlighted that synthesis success rates[8] of the most prevalent synthetic transformations used by medicinal chemists in pursuing drug candidates hovered only around 60−70% in most cases, even

Received: January 3, 2024
Revised: April 24, 2024
Accepted: April 25, 2024
Published: May 20, 2024



https://doi.org/10.1021/jacs.4c00098
J. Am. Chem. Soc. 2024, 146, 15070−15084

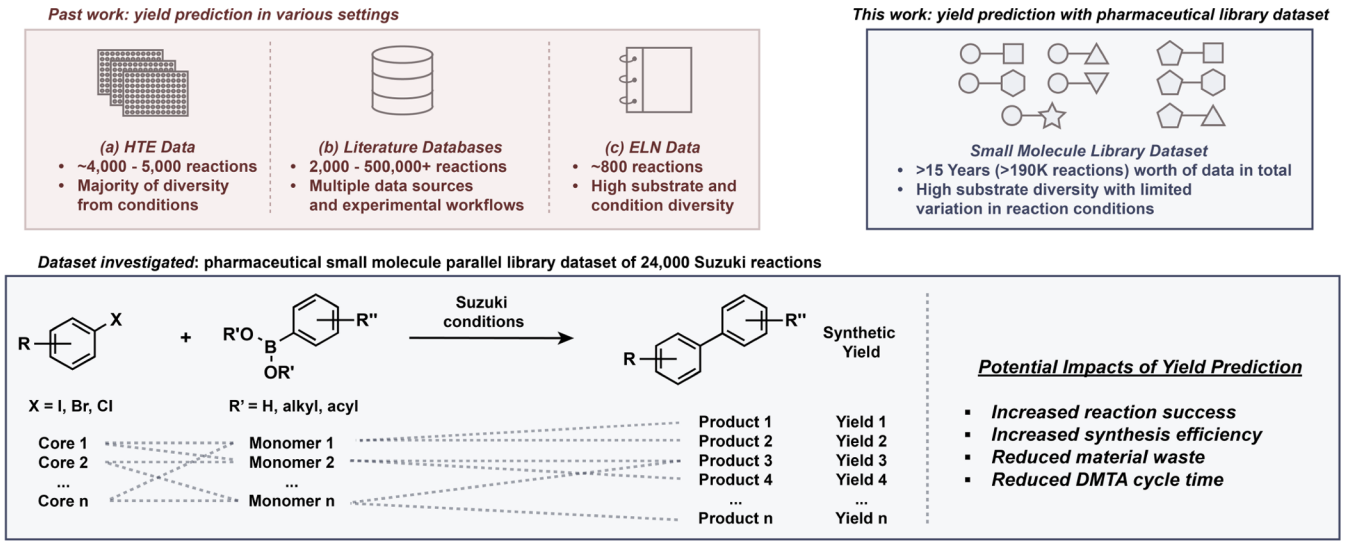


Figure 1. Prior endeavors into reaction yield prediction: (a) HTE data modeling (Doyle 2018, Zheng 2021), (b) literature database modeling (Schwaller 2021, Vuilleumier 2022, Neves 2023), (c) ELN data set modeling (Wiest 2023), and this work: incorporating yield prediction in the synthetic chemists' design toolbox using AbbVie's parallel library data set.

for the robust, well-established Suzuki−Miyaura cross-coupling.[9] As such, predictive models that would allow medicinal chemists to gauge the proposed synthesis efficiency of the transformation (e.g., reaction yields) will significantly improve the overall DMTA cycle time, in addition to reducing the amounts of precious advanced intermediates lost due to unsuccessful reactions.

To enable the development of such a tool, it is important to curate a relevant data set for model development and, equally important, to identify and attempt to control factors that complicate yield prediction. A myriad of data sets of different sources and sizes, reaction scales, and structural diversity have been reported in the literature to investigate yield prediction (Figure 1). These data sets include those derived from high-throughput experimentation (HTE),[10] literature-derived reaction databases,[11] and proprietary internal data sets.[12] While HTE-derived data sets have successfully been used for the accurate prediction of reaction yields for Buchwald−Hartwig amination[10a,12a] and Suzuki cross-coupling reactions,[12a] these represent highly exhaustive exploration of very narrow, combinatorial reaction spaces and thus would not be expected to enable generalization beyond their scopes. Although HTE is an effective technology for generating large reaction data sets quickly and efficiently,[13] the outputs are not isolated yields in most cases, regardless of whether the data is used for ML. Instead, assay yields are often reported, such as UV area percents, percent conversions, or product/internal standard ratios.[10,14] This may pose questions regarding the relevance of using HTE readouts to predict synthetic yields at larger scales due to the confounding factors that isolation may introduce, and as a result, the prediction from a model trained exclusively on HTE data may not necessarily translate into material delivery to assays. Public data sets of varying sizes, sourced from the USPTO,[11] scientific literature,[15] and beyond[16] have also been used to build models for yield prediction and other tasks such as condition recommendation.[17] These data sets, while large, exhibit significant procedural variation among different data sources, causing yield prediction models to exhibit low performance.[11] The underrepresentation of negative results in literature data further contributes to this low performance,[18] although Glorius and co-workers have found that enriching literature data sets with negative data can lead to improved model performance.[19]

For these reasons, several pharmaceutical companies have turned to their internal chemistry electronic laboratory notebook (ChemELN) data or other proprietary data sets, which in principle cover large chemical spaces with pharmaceutically relevant structures and transformations. Unfortunately, ChemELN data cleanup and processing can be a very daunting task, which generally entails sizable human and capital investments.[20] These data sets also exhibit variations in reaction conditions, scale, and reactant properties (among other factors that may or may not be recorded), all of which increase the challenge of building ML models.[12a] Meanwhile, other studies have built classification models using targeted reactions to generate data sets for ML to address these limitations,[21] although this approach could potentially struggle to reach the chemical space desired for broader deployment in compound design workflows for drug discovery. We hypothesized that a data set composed exclusively of medicinal chemistry synthesis efforts might be the most relevant data set to use to build a yield prediction model for practical use in drug discovery.

At AbbVie, our centralized chemistry technology group was established in the early 2000s to enable the rapid generation of chemical libraries in parallel to facilitate medicinal chemistry SAR studies. The libraries are experimentally conducted by reacting a common intermediate (core) with a collection of monomers chosen by a medicinal chemist under the same reaction conditions for each to achieve the synthesis of a collection of similar but differentiated products. Nearly 9,000 libraries, spanning over 100 small-molecule drug discovery programs, were completed by our group, resulting in the synthesis and registration of ∼160,000 compound lots. Historically, compounds registered via library synthesis accounted for up to 50% of all of the compounds registered by AbbVie medicinal chemists annually. The data associated with the synthesis of this vast compound collection has been captured in our laboratory information management system (LIMS) database, which is maintained separately from

https://doi.org/10.1021/jacs.4c00098
J. Am. Chem. Soc. 2024, 146, 15070−15084

AbbVie's ChemELN data. Recorded details include library submission information, parallel library experimental planning, reagent calculations, final synthesis results, and product characterization data. The main workflow for library synthesis has remained consistent over the years, where crude reactions are analyzed by UPLC-MS and then purified via reverse-phase mass-directed preparative HPLC. Moreover, automation has been introduced at several points over the years, including the processing of crude reactions for purification, mass-directed purification, transferring of purified compounds into vials, and final weighing of the compounds, all of which decreased the likelihood of variation in results introduced due to experimental error. Additionally, this data was generated exclusively in support of AbbVie's global medicinal chemistry efforts across a diverse array of therapeutic areas and functions over the last two decades. These factors have enabled us to be intimately familiar with both how the data was generated and how the data sets are formatted. AbbVie's chemical library data set contains a large amount of negative data, as well. This addresses one limitation of using data sets sourced from scientific literature and patents for reaction yield modeling. The data also exhibits a much greater number of reactions, unique reactant structures, and molecular diversity than previously reported HTE data sets while containing a smaller number of unique conditions compared to other HTE and literature data sets of a single reaction type (see the Supporting Information, Figure S15). Due to these factors, we hypothesized that a model trained using this data might generalize better to more diverse unseen reactant structures than models trained on dense combinatorial data[10a,11] Unlike many other data sets evaluated for ML yield prediction, this data set solely reports the isolated yields of compounds, as opposed to assay yields (e.g., NMR yields or LCMS area percent values), which is an important consideration for differentiating whether compounds could be obtained in pure form. While purification itself can introduce variation that the model is not explicitly privy to into the final measured yields, our reaction workflow aimed to be as consistent as possible by using mass-directed reverse-phase preparative HPLC for all compound purification. Using crude LCMS end points for model training may alleviate this variation when the data is calibrated using internal standards, or when there is a small number of the substrate pairs, and can therefore be useful for tasks involving reactivity-only end points (e.g., reaction condition prediction[21b]). However, a model trained using isolated yields would directly inform on a critical metric for medicinal chemistry—the ability to deliver the reaction products into biological assays with sufficient purity. Therefore, the accumulation of this past library reaction data presented an exciting opportunity to build reaction yield prediction models (Figure 1).

In this work, we detail our efforts to build and evaluate such models for Suzuki cross-couplings using AbbVie's extensive chemical library data set and demonstrate the utility of the model for pharmaceutical compound design. We illustrate the idea of using such models as aides to chemists in library design by showcasing a comparison of the model's predictions to those made by expert medicinal chemists, as well as case studies demonstrating model deployment.

RESULTS AND DISCUSSION

Data Set Selection and Preparation. We limited our study to a single library transformation. Consistent with literature reports on the prevalence of reactions used in medicinal chemistry,[22] Suzuki cross-couplings represent the second most-conducted library transformation in AbbVie's library data set, with an overall success rate of ∼68% for reaction product registration (i.e., desired compounds were isolated with sufficient purity for assays). Suzuki reactions are also well known to have more general/robust reaction conditions than other cross-coupling transformations[21,23] and are ubiquitous in medicinal chemistry.[24] Therefore, we anticipated the impact of reaction condition variation on the results of the library reactions to be reduced (i.e., that yields are more influenced by cross-coupling partners than reaction conditions). Additionally, more than 85% of the Suzuki library reactions were completed by just three AbbVie chemists. Therefore, we chose to evaluate this subdata set as an exemplary parallel library data set for using historic library data to build yield prediction models.

As stated previously, intimate knowledge of the data itself is essential to prepare a data set suitable for machine learning. Using this knowledge, we were able to manually correct mislabeled reactants, erroneously entered data, and artifacts of data storage. The initial Suzuki library reaction data set contained >40,000 individual data points; due to the nature of how the data was stored, not every data point corresponded to an individual valid reaction. Therefore, an extensive data set processing workflow was developed to furnish a data set suitable for ML. First, the names of several catalysts, bases, and solvents were found to vary throughout the data set, as these were often entered as free text by the chemists conducting the libraries. For instance, a single palladium catalyst contained two dozen unique labels in the data set (Figure 2a). Such entities were identified and assigned to a common name using the molecular weight entered by the chemist during experimental design to assist manual review. Next, reactions that were separated into multiple product lots during purification were combined into one data point, as these were often left separate when registering the reaction products (Figure 2b). However, reactions were also removed from the data set in their entirety under some circumstances. For instance, reactions that exhibited transformations beyond a typical Suzuki reaction were identified and removed from the data set (e.g., a subsequent deprotection or ester hydrolysis after the cross-coupling; Figure 2c), as were compounds that underwent two or more rounds of purification. These additional reactions and purification measures can introduce unpredictable amounts of variation into the final measured isolated yield due to the efficiency of the subsequent synthetic transformation(s) and the potential challenge of repurifying the compounds, respectively. Furthermore, reactions that did not have any reaction conditions associated with them were removed in their entirety. Combined, these efforts furnished a cleaned data set of just over 24,000 individual Suzuki reactions (Figure 2d).[25]

Processed Suzuki Library Reaction Data Set. The processed data set contains 23,236 unique Suzuki reaction products from 24,203 individual reactions,[26] which were synthesized across 629 libraries. Each library consists of a series of reactions that share common reaction conditions and a common core structure with a set of differentiated monomer structures. Typically, the common core is an aryl halide, and the monomer structures are organoboranes (both acids and esters) in this data set; the reverse is true for a small subset of the data set. Each reaction in the data set is labeled by the

https://doi.org/10.1021/jacs.4c00098
J. Am. Chem. Soc. 2024, 146, 15070−15084

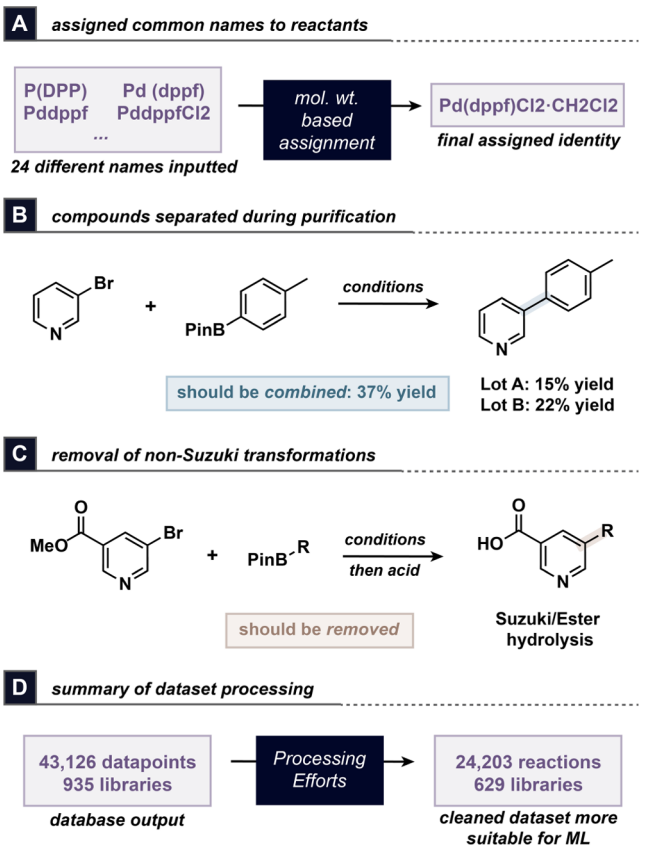


Figure 2. Scenarios encountered and addressed during data set processing, including (A) multiple names for the same reagent, (B) combining of multiple product lots postpurification, (C) removal of transformations beyond a typical Suzuki coupling, and (D) the overall summary of these processing efforts. Generic example structures are shown for panels (B) and (C).

isolated percent yield of the reaction (Figure 3a). Visualizing the chemical space of the core and monomer structures with t-SNE[27] confirms this, as the majority of the core structures cluster around each other with the aryl halide monomers. This analysis also revealed that the diversity present in the boronic acid/ester monomers was greater than that which was present in the MIDA boronate monomers (Figure 3b). Analysis of the functional handles used for core and monomer compounds reveals that this Suzuki data set reaches ~3% of the possible core/monomer pairings.[28] Overall, the data set exhibits a large range of diversity across several drug-relevant physiochemical properties (see the Supporting Information, Figure S14). Furthermore, the reaction conditions chosen by the chemist conducting the library are often highly general and robust.[9] While 118 unique reaction condition combinations are present in the data set, >80% of the reactions were conducted with one of five palladium catalysts; similarly, all but a small number of reactions were conducted with four inorganic bases commonly employed in Suzuki couplings (Figure 3c). This rather low diversity in reaction condition components speaks to the convenience of Suzuki cross-coupling reactions, where a small number of unique reagents can facilitate the synthesis of diverse compound libraries. This is in stark contrast to, for instance, C−N couplings, where the nature of the amine coupling partner can dramatically influence the catalyst choice for a reaction.[29] It also indicates that the data set diversity primarily stems from the combination of core and monomer structures. Furthermore, the distribution of isolated yields in the data set shows that most successful library reactions furnished the desired products in 10−40% yield (Figure 3d). Therefore, the yields can be differentiated into four classes: unsuccessful (0%), low (>0−10% yield), medium (>10−30% yield), and high (>30% yield). In this data set, the unsuccessful reactions include those where the reaction simply did not work, in addition to those where the compound could not be

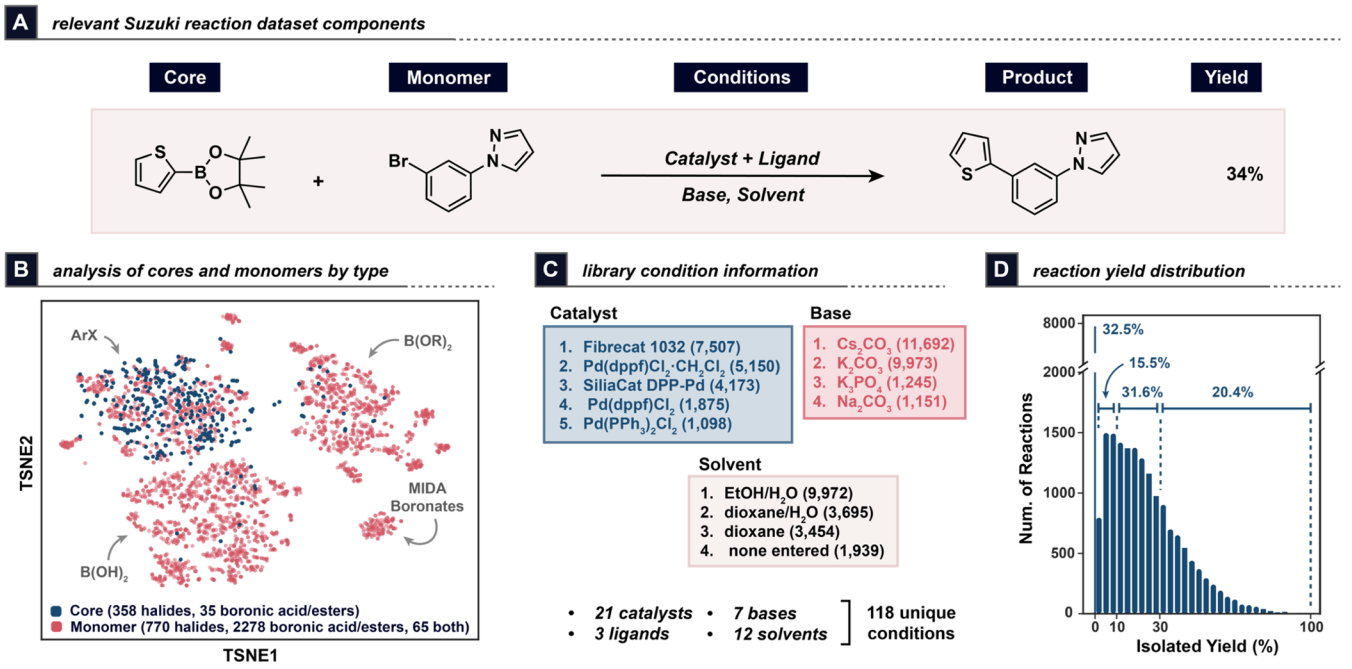


Figure 3. Processed Suzuki library data set, showing (A) the general format of the Suzuki medicinal chemistry library data set, (B) the types and distribution of core and monomer structures present in the data set (via t-SNE clustering of the PCA decomposition of structural fingerprints), (C) commonly used catalysts, bases, and solvents used throughout the data set, and (D) the distribution of isolated yields in the data set.

https://doi.org/10.1021/jacs.4c00098
J. Am. Chem. Soc. 2024, 146, 15070−15084

successfully purified or otherwise isolated; any result where the desired product was registered (i.e., the compound exhibited sufficient purity for evaluation in biological assays) was considered successful, regardless of the final yield. Furthermore, we estimated inherent data set noise due to uncaptured reaction factors by analyzing sets of repeated reactions within the data set, finding that over half of these 201 reaction sets exhibited a standard deviation in the isolated yield of <5% (see the Supporting Information, Figure S19).

In full, the data set consists of 393 unique core and 3,113 unique monomers. About 2/3 of the unique cores present in the data set were used in a single library, typically making them present in 20−70 reactions. For comparison, roughly half of the unique monomer structures were used in a single library, indicating that these monomers appear just once throughout the data set (Figure 4a). Many of the remaining monomers

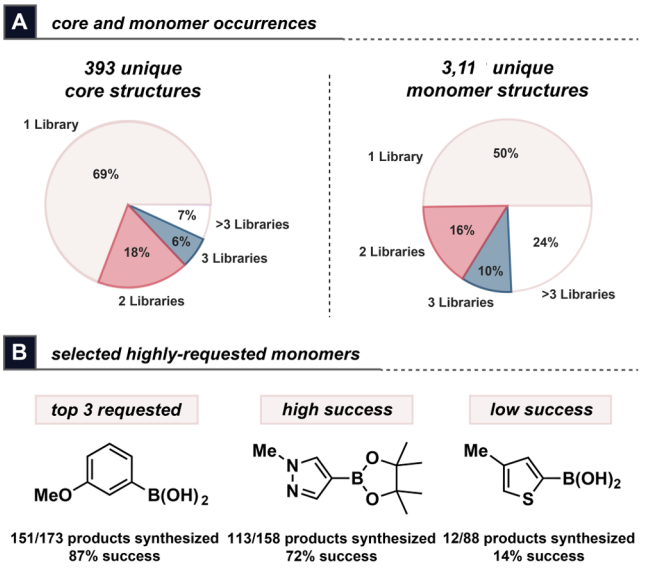


Figure 4. Suzuki data set reactant statistics showing (A) the number of library occurrences of core and monomer structures and (B) selected examples of highly requested monomers in the retrospective data set and their success rates.

were requested multiple times, with some being included in >100 libraries. As described in our group's prior analysis, highly requested monomers are not always highly successful monomers.[9,30] For example, pyrazoles and anisoles are highly requested functional motifs that are often successfully cross-coupled in Suzuki libraries, while 2-thiazoles are highly requested/desirable motifs in medicinal chemistry, but poor substrates for the transformation (Figure 4b).

Retrospective Modeling. We focused our evaluation of the processed 15-year data set on 3 molecular featurization approaches (Figure 5a). As a baseline, one-hot encoded (OHE) features were assigned depending on the identity of the core and monomer in each reaction, since this featurization approach has exhibited good performance in interpolative prediction tasks[21,31] and can be used to tell whether the feature-based models are learning beyond data set statistics. Morgan fingerprints (FP) generated using RDKit were selected, as they are widely used, fast to obtain, and computationally inexpensive to generate. Furthermore, they have demonstrated good performance on chemistry-related prediction tasks.[32] Alongside these, ab initio quantum chemical features obtained with density functional theory (DFT) for the reactive site atoms and molecules were calculated, although these features are more computationally intensive compared to FPs. DFT featurization was accomplished using an automated pipeline (built with modifications to Auto-QChem[33]) involving Gaussian 16 (G16)[34] calculations to obtain features for each atom in the core/monomer molecules, in addition to whole molecule features (see the Supporting Information, Figure S1). All reactions in the data set were first atom-mapped, and the 3,473 unique aryl halides and organoboranes were extracted. For each molecule, 1−20 conformers were generated using RDKit;[35] then, to reduce computational overhead, the lowest-energy conformer was selected using semiempirical (GFN2-xTB)[36] energy calculations. The selected lowest-energy conformer was then used to conduct geometry optimization and frequency calculations (APFD/6-31G*) using G16.[34] This generated 25 molecule-level descriptors per molecule, and 19 atom-level descriptors per atom per molecule. For each reaction, atom-level features were extracted for the four atoms corresponding to the reactive site (C−B bond on the organoborane, C-X bond on the aryl halide). Computed descriptors that exhibited varied distributions across the data set were selected for modeling (see the Supporting Information, Figures S2 and S3). This resulted in a final DFT feature vector of length 48 for each reaction, composed of 8 molecule features per molecule and 8 atom-level features per reactive site atom for the 4 reactive site atoms (see the Supporting Information, Section 2d for details regarding the DFT featurization pipeline).

Beyond feature selection, 2 "axes" of modeling were chosen for this study, namely, the data set splitting method and the prediction tasks (Figure 5b). The splits and prediction tasks were chosen to be in alignment with those that would be conducive for developing useful prospective tools for the parallel library design workflow. For the splits, random, monomer-based, and core-based splits were used to evaluate the model's performance. The random split provides no user bias in the distribution of structures into the train/test sets, and therefore it is likely that identical or very similar structures will appear in both train and test. Meanwhile, monomer- and core-based splits ensure that the model is evaluated on unseen monomer and core structures, respectively. These splits are more challenging and allow for analysis of how well the models extrapolate in terms of potential use cases for the prediction of new library reaction yields (core split), or the prediction of follow-up library yields on a core that has previously undergone library submission with unseen monomers (monomer split). For the core-based split, in particular, the test reactions may also contain unseen monomer structures due to the nature of how parallel libraries are conducted, meaning the model is predicting yields for both unseen and seen monomers and therefore represents the evaluation of the greatest challenge−extrapolation to entirely unseen reactions. In practice, follow-up libraries may employ cores or monomers that were already seen by the model. Indeed, due to the nature of libraries, there are 33 structures that are used as both a core and a monomer in different libraries. Regarding the task axis of modeling, binary classification, multiclass classification, and regression tasks were chosen, as they are each relevant to tasks in medicinal chemistry. In this data set, a 0% isolated yield indicates an inability to obtain enough product postpurification (or insufficient purity, potentially without enough material for repurification) for downstream assays and testing, so even

https://doi.org/10.1021/jacs.4c00098
J. Am. Chem. Soc. 2024, 146, 15070−15084

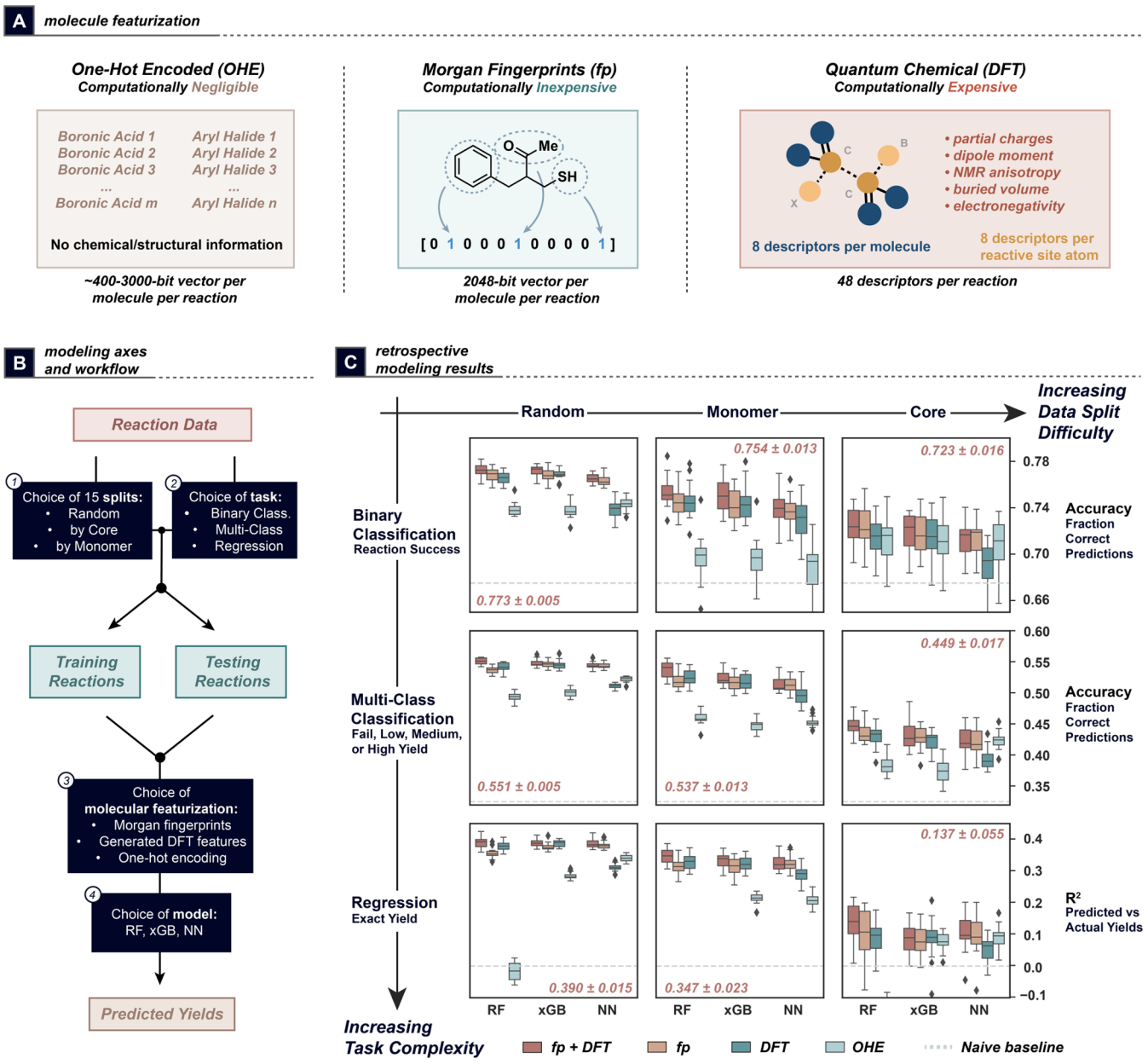


Figure 5. (A) Data set featurization approaches, (B) workflow for reaction yield prediction, (C) results for the 108 models built for retrospective predictions of Suzuki library yields using random forest (RF), extreme gradient boost (xGB), or feed-forward neural network (NN) models with fingerprint (fp), quantum chemical (DFT), or one-hot encoded (OHE) features. The naïve baseline is shown as a horizontal dashed line, and results from the 108 yield prediction models are depicted as box plots of aggregate results from 15 data shuffles. The best result from the feature-based models is written in each plot, colored by the featurization type that yielded this result. The RF OHE-based regression models for the monomer- and core-based splits exhibited average R^2 values that were less than −0.1 and therefore do not appear in the plots. Outliers within the box plots represented those past 1.5 times the interquartile range.

good model performance on a binary classification (0 or non-0 yield) task is useful to medicinal chemistry programs. For multiclass yield classification, four yield bins of 0%, >0−10%, >10%−30%, and >30%−100% yield provide additional nuance for the prediction of successful reactions and would allow a medicinal chemist to prioritize compounds predicted to be in higher yield bins. Finally, regression, as the most challenging prediction task, was investigated to determine whether it would be possible to rank-order the monomers for a given library with more discrimination than a classification model. Although past yield prediction efforts have largely focused on regression, a trustworthy classification model is still incredibly valuable for design in medicinal chemistry, as these models would directly evaluate whether a product could be synthesized with sufficient purity for evaluation in biological assays.

For the model types, we selected random forest (RF), extreme gradient boosting (xGB), and feed-forward neural network (NN) machine learning models (Figure 5b).[37] We limited our analysis to relatively "simple" model types, as past yield prediction efforts have consistently seen little to no improvement, or even detriment, from using more complex deep learning architectures on real-world data.[11,12] Despite this, we also evaluated using a graph neural network and features from a language-based transformer model, finding that

https://doi.org/10.1021/jacs.4c00098
J. Am. Chem. Soc. 2024, 146, 15070−15084

neither could provide a significant advantage over the best results obtained from these three models (see the Supporting Information, Tables S6 and S7). Given the 3 choices of data set splits, prediction tasks, and models, we trained and evaluated 108 models in total using OHE, FP, DFT, and a combination of FP and DFT features. In all cases, the aforementioned features derived from the aryl halide and boronic acid were used. The reaction conditions were represented in all models using OHE for the catalyst + ligand system and base, and multihot encoding (MHE) for the solvent(s). This representation was then concatenated to the molecule feature vectors to obtain the final input vectors for model training. In each of the 108 modeling scenarios evaluated, 15 data shuffles were modeled to obtain the final results.

The combination of FPs and DFT features generally afforded the highest performance observed across all prediction tasks and data set splits (Figure 5c). A potential explanation for this observation lies in the overall complexity of modeling chemical reactivity, where structural and quantum chemical features can both provide useful information regarding a structure's reactivity that the other cannot. For example, the presence of proximal functional groups in the molecules can be informed with FPs, while electronic features of the reactive site can be informed with DFT features. Furthermore, the OHE-based models generally exhibited poorer accuracies and larger variances compared to the feature-based models, suggesting that the models are indeed extrapolating to unseen structures to some degree. Only minor differences between the ML models can be seen, with the RF model generally slightly outperforming the xGB and NN models for feature-based modeling. As expected, random splits afford the most optimistic predictions across all prediction tasks evaluated. The monomer-based split exhibited a slight performance drop in comparison; however, considering the model is being evaluated on new monomer structures, we were pleased to observe similar performance. Meanwhile, the core-based split exhibited a larger drop in performance and is clearly the most challenging way to split the data set for the evaluation of unseen chemical matter. This performance drop is likely observed due to the lower number of cores present in the data set (393) as compared to monomers (3113), along with the potential for unseen monomer structures to also appear in the test split. Furthermore, the core molecules are usually more structurally complex than the monomer structures and therefore pose a greater generalization challenge. The core- and monomer-based splits thus exhibited higher standard deviations than the random split models, indicating that performance is more sensitive to the distribution of reactions between train/test for these more complex splitting methods. We reiterate that the different split types mimic different scenarios of how these models would be used prospectively and should not be seen as a design choice for the model itself.

A direct comparison of the prediction tasks should be avoided, as modeling naturally becomes more challenging when more specific reaction outcomes are being predicted. As such, a better comparison would be of each task/split combination to that of the naïve baseline (the accuracy that would be achieved if the classification models predicted only the highest-occupied class in the data set, or predicting the mean yield for regression) and to the OHE baseline models. While modest generalizability is observed across the board, it is particularly apparent for the monomer-based split, where all three of the feature-based approaches could afford significantly higher accuracies and R^2 values for the classification and regression tasks, respectively, as compared to the naïve and OHE baselines. Meanwhile, comparing the results from the core-based split shows that it is more challenging for the model to extrapolate to unseen core structures (see the Supporting Information, Sections 2e−2h for selected results, confusion matrices, and parity plots).

In summary, the combination of the RF model with combined FP + DFT featurization consistently furnished the highest performance across all 9 data set splits and prediction tasks evaluated. Binary accuracies ranging from ∼72 to 78% and multiclass accuracies ranging from ∼45 to 55% were achieved with the classification models depending on the data set splitting strategy chosen. For the regression models, R^2 values of ∼0.35−0.39 were achieved for the random and monomer-based splits, while R^2 values of ∼0.13 for the more challenging core split were obtained. As the OHE-based models perform consistently worse than the feature-based models, this suggests that the feature-based models are indeed generalizing to unseen structures in this retrospective study to some extent. These results suggest that the classification models might be sufficient for prospective use, while regression modeling could be beneficial for the prediction of new monomer yields. Given the results of this study, we chose to use combined FP + DFT features for all studies that follow.

Comparing the Retrospective Model to Expert Medicinal Chemist Predictions. In the overall workflow of library design, the de facto baseline is the human expert. We therefore decided to compare the model's predictions to those made by 11 expert medicinal chemists at AbbVie. The chemists we surveyed support a wide array of therapeutic areas across different AbbVie sites, exhibit a variety of synthetic chemistry backgrounds, and have varying amounts of experience. Some of the chemists surveyed had been medicinal chemists for as few as 6 months or as tenured for over 15 years in the field. Two scenarios were evaluated: (1) whole-library predictions and (2) partial library predictions. In the latter, the monomers used in the training split were provided to the chemists, along with their corresponding isolated yields. Three libraries were selected for each scenario with varying numbers of reactions and model performance. Chemists were provided with a brief background of this project, along with the distribution of reaction yields found in the entire retrospective data set, and the chemists were asked to evaluate whether they thought each reaction would give zero/low/medium/high yields (as defined by the yield bins for the multiclass classification models). In total, each chemist provided 117 predicted categorical yield values, which were used to compare their predictions of reaction success and reaction yield bins against the binary and multiclass classification models' predictions, respectively.[38,39]

Gratifyingly, the model could predict both library reaction success (binary classification) and binned yields (multiclass classification) with greater accuracy than the median achieved by the medicinal chemists for most of the libraries evaluated (Figure 6a). The model performs particularly poorly for libraries 2 and 6, for both classification prediction tasks, although its performance is still within the range of AbbVie expert medicinal chemists' accuracies. The range of prediction accuracies achieved by the chemists was also much wider for binned yield predictions than it was for reaction success predictions, with no single medicinal chemist achieving the highest accuracy for every library evaluated. Furthermore, we

https://doi.org/10.1021/jacs.4c00098
J. Am. Chem. Soc. 2024, 146, 15070−15084

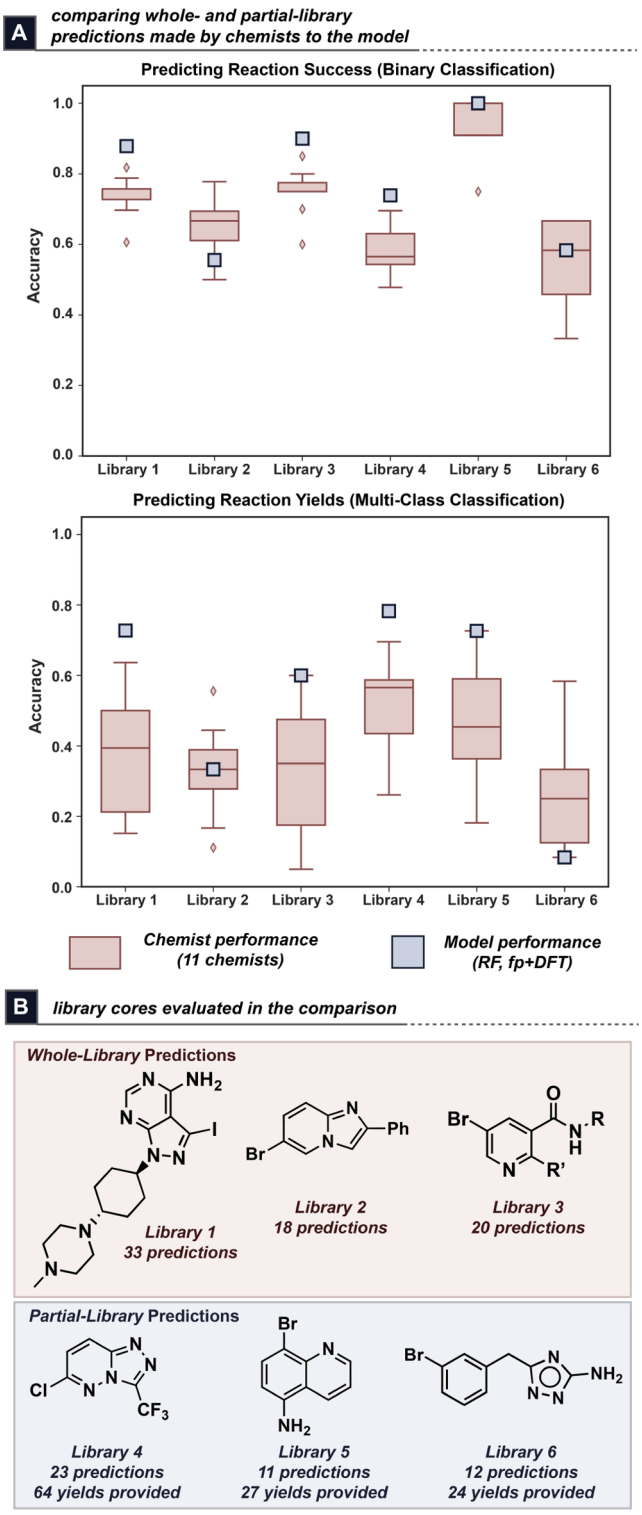


Figure 6. Comparison of the model's predictions to those made by 11 expert medicinal chemists. (A) Performance of the chemists (depicted as box plots) vs the model. The results from predicting reaction success (binary classification) are depicted in the top plot, while the results obtained from predicting binned reaction yields (multiclass classification) are shown in the bottom plot. (B) Core structures and number of reactions in the libraries selected to evaluate the models' performance against medicinal chemists. The whole-library predictions were compared to the core-split models' predictions, while the partial-library predictions were compared to the monomer-split models' predictions.

did not notice increased chemist performance with a larger number of reaction yields provided for the partial-library predictions, and the chemist predictions were thus found, overall, to be more subjective than the models' (see the Supporting Information, Section 4). Overall, the model can furnish reaction success and yield predictions with 10−25% greater accuracy than the average of the medicinal chemists surveyed. The results obtained from this comparison are in alignment with those obtained when comparing other ML models to chemists' intuition.[40] As is often discussed, predictive tools should augment expert chemists and their intuition and need not be seen as replacements. This study showcases the potential of this model to serve as a more reliable tool to enhance existing library design tools by assisting chemists in the prioritization of desirable monomers that are more likely to be successful in synthesis.

Pseudo-Prospective Evaluation. Having conducted extensive retrospective modeling on historic library reaction data and demonstrating that the models can often outperform expert medicinal chemists' predictions, we now sought to evaluate how well our models could perform when faced with newer data and potentially more novel structures. For this analysis, we used a held-out data set of 574 library reactions (18 libraries) from 16 cores and 475 monomers conducted by our group more recently than the data set used for retrospective analysis (post-mid-2021, the cutoff date of the retrospective data set).[41] This task represents an extreme time-based split, as 90% of the training data set was acquired before 2019 (Figure 7a). There are 11 new cores and 207 new monomers in this data; therefore, 44% of the unique reactant molecules were unseen by the models, in addition to new combinations of previously seen molecules. Applying the previously trained models (all using FP + DFT features and trained on the full retrospective data set) to this data, we found that overall performance metrics dropped for all 3 prediction tasks (see the Supporting Information, Section 2j). To investigate this decrease in performance, we divided this held-out data set into 4 subdata sets: (1) reactions involving cores and monomers both seen in the retrospective data set (110 reactions), (2) reactions involving seen cores but unseen monomers (140 reactions), (3) reactions involving unseen cores but seen monomers (237 reactions), and (4) reactions involving unseen cores and monomers (87 reactions).

The models generally outperformed the naïve baselines for multiclass classification for each data subset but performed similarly for binary classification (Figure 7b). It is important to note that the distribution of yields has shifted relative to the retrospective data, with a higher reaction success rate and higher yields observed on average (Figure 7c). For example, in a multiclassification sense, the most populous class in the retrospective data is the unsuccessful reaction class (0 yielding); however, in the held-out data, it is the highest yield class (>30% yielding). As such, the naïve baseline is still considered to be class 0 for the multiclass predictions, as this was the highest-occupied binned yield in the retrospective data set, and the model is not aware of this distributional yield shift.

Aligning with expectations, the subdata set exhibiting the highest overall improvement compared to the naïve baseline was composed of the reactions with seen cores and seen monomers. Confounding variables and the drift in experimental procedures over time (e.g., new chemists, the introduction of improvements into the workflow, etc.) can explain the decrease in interpolative performance from the

https://doi.org/10.1021/jacs.4c00098
J. Am. Chem. Soc. 2024, 146, 15070−15084

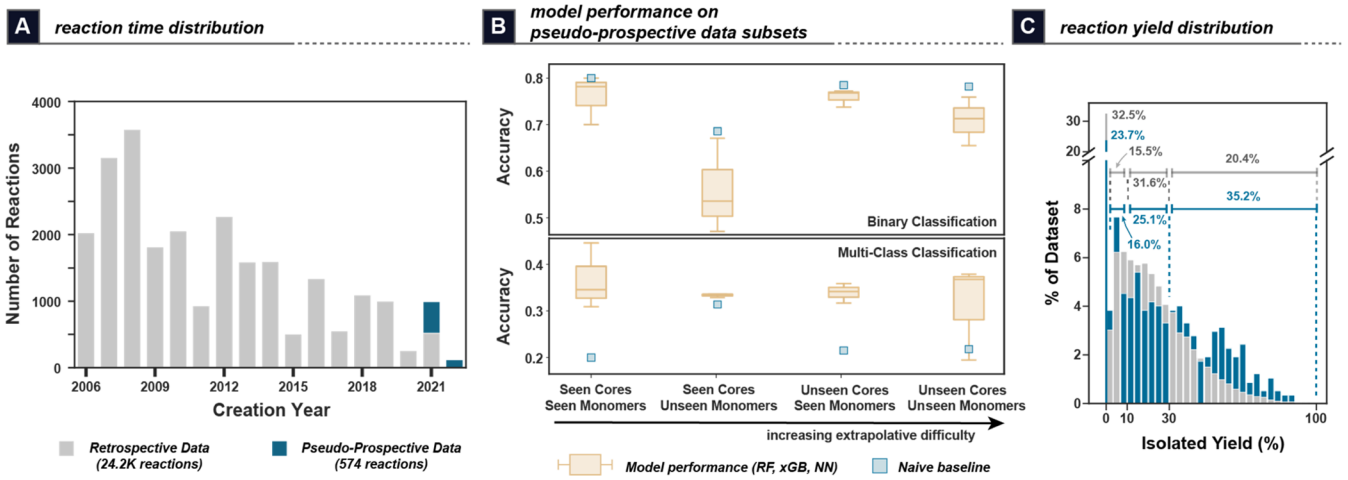


Figure 7. Pseudo-prospective evaluation on a held-out data set, showing the (A) distribution of the reactions over time compared to the retrospective data set's, (B) accuracies obtained from the binary (top) and multiclass (bottom) predictions of the held-out data set, which was separated into subdata sets depending on the extrapolative difficulty of the predictions, and (C) distribution of the held-out data set's yields compared to the retrospective data set's relative to the overall percentage of each data set (gray represents the retrospective data, while color represents the post-2021 data). In panel (b), the naïve baselines are defined as the accuracy achieved if the model predicted the highest-occupied class from the retrospective data for each subset of the held-out data set. The pseudo-prospective data set contains reactions that were experimentally conducted between May 2021 and February 2022.

retrospective evaluation to this prospective evaluation. For example, there has been a shift of chemists in recent years ensuring that Suzuki reactions are conducted under inert atmosphere, which will generally increase yields. These results underscore the importance of controlling for confounding variables that can affect yield prediction tasks, such as reaction environment, temperature, concentration, etc. The post-mid-2021 core structures also exhibited, on average, higher molecular weights, more heteroatoms, more rotatable bonds, and lower QED (see the Supporting Information, Figure S20). This suggests increased complexity of the post-mid-2021 cores as compared to the retrospective data set's cores, and can help explain the observed performance drop. Therefore, it would be prudent to frequently retrain the models as new data is collected for long-term deployment, as this could increase the models' robustness to better handle experimental changes and better adjust to drifts in the reaction yields and molecular complexities.

Prospective Application toward Library Design. The most fundamental application of yield prediction to compound design is the selection (or replacement) of monomers in the library design workflow. Having fewer library members fail synthesis would decrease precious material waste, save personnel time, and deliver a greater number of design-relevant compounds to the DMTA iterative cycle. We identified and explored two specific opportunities for model incorporation in this context (Figure 8). The first setting is monomer prioritization, where replacement monomers can be proposed for those that are predicted to be unsuccessful by the binary classification model. This scenario is analogous to using the model for compound design prior to conducting a library, and as such, the model used to make predictions for this application has not yet seen the target library's core.[42] The second setting is monomer rescue, where replacement monomers are proposed for those that were observed to be unsuccessful during synthesis, using knowledge of the initial experimental results during model training to better inform selection of the replacement monomers. In this scenario, the

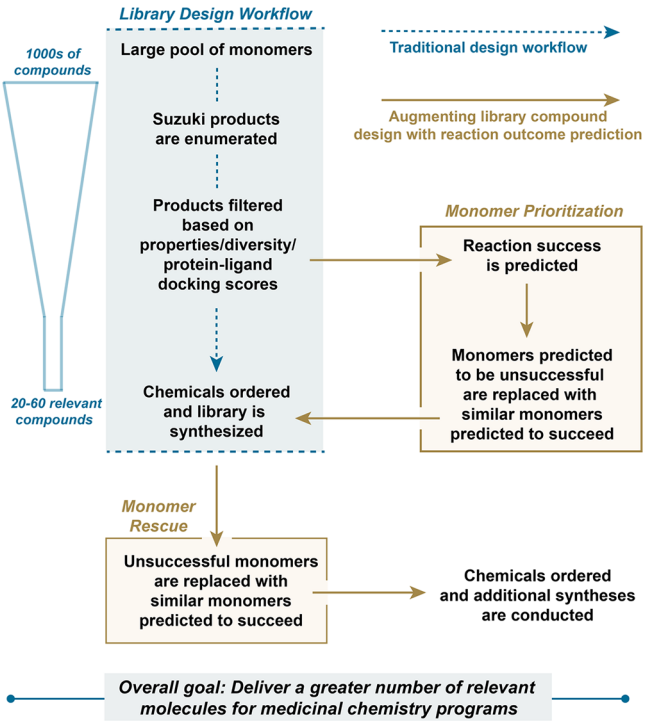


Figure 8. Supplementing the drug design workflow using reaction outcome prediction via Monomer Prioritization (i.e., prior to conducting any syntheses) and Monomer Rescue (i.e., after the library has been synthesized). Similar monomers were suggested using 3D shape and electrostatic potentials to calculate and quantify similarity.

library is being "rescued" by replacing the unsuccessful monomers with similar monomers that were predicted to be successful. For both scenarios, to preserve the query monomers' design goals in terms of structural and electronic likeness, an open-source package, *espsim*, was used to determine molecular similarity.[43] Similarity scores for 3D shape and electrostatic potentials were calculated using this

https://doi.org/10.1021/jacs.4c00098
J. Am. Chem. Soc. 2024, 146, 15070−15084

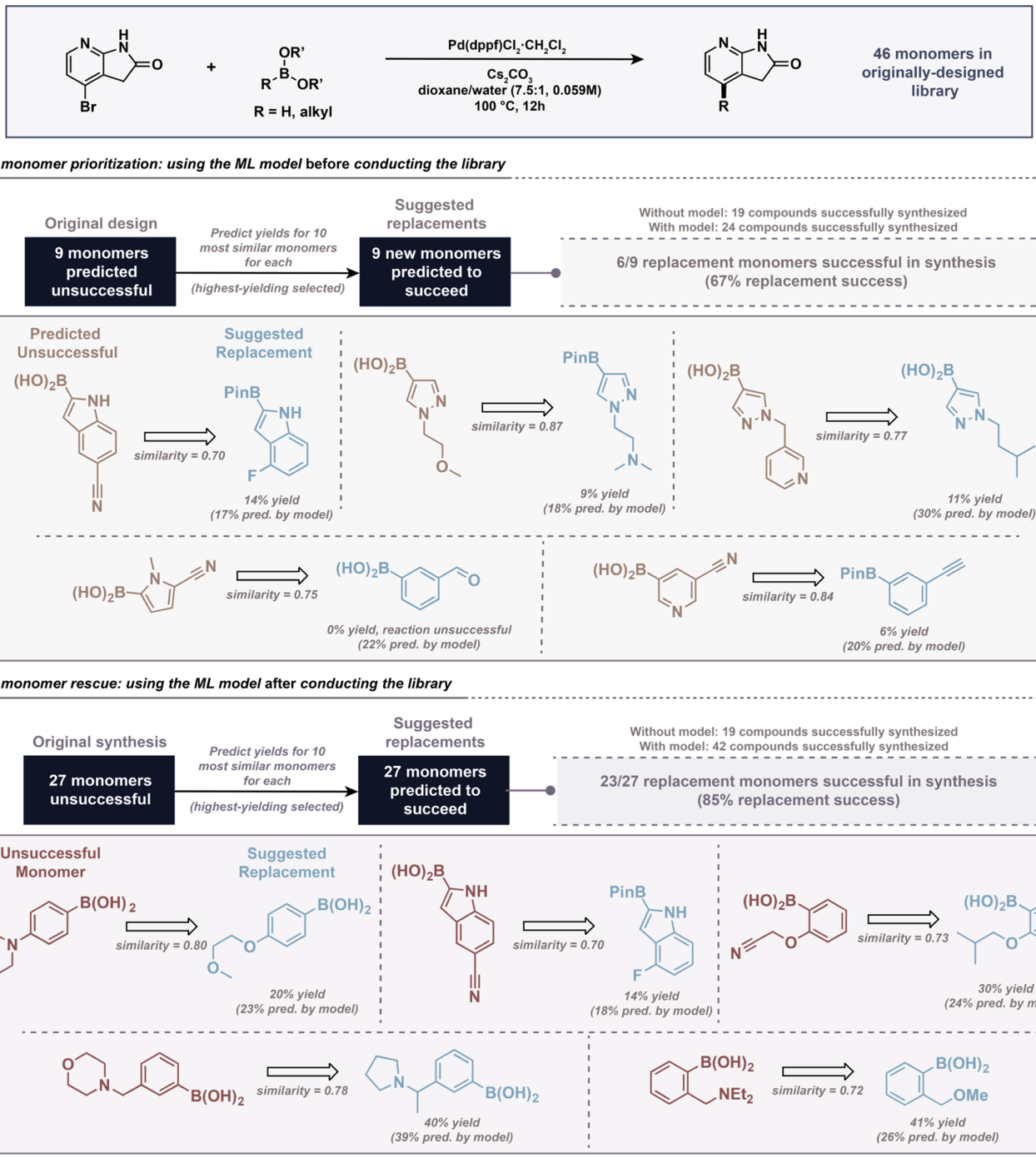


Figure 9. Using the yield prediction model in drug design via (A) replacing monomers predicted to be unsuccessful when the library has not been seen by the model and (B) replacing monomers that were unsuccessful during the prior library synthesis when the model has seen the library. The query monomers are shown with their suggested replacement monomers, along with the similarity scores, predicted yields, and isolated yields.

package and multiplied by each other to obtain the final similarity score between the query monomer and its potential replacements. The 10 most similar monomers were determined for each query monomer using this approach, and the similar monomer with the highest predicted regression yield was selected as the replacement for synthesis (see the Supporting Information, Figure S30, for an example). For the studies presented in this manuscript, we limited the available pool of replacement monomers to the 1,934 unique boronic acid/ester monomers present in the retrospective data set, although the list of potential replacement monomers could be expanded if needed (see the Supporting Information, Section 6). In accordance with our findings from the retrospective modeling, the RF models with FP + DFT features were used for all predictions in this study.[44]

The specific evaluation of both monomer replacement scenarios used a 46-compound library sourced from the retrospective data set that used an aza-oxindole core (Figure 9), in which 19 of the chosen monomers were successful and 27 were unsuccessful. In the monomer prioritization scenario for this library, 9 of the originally designed 46 monomers were predicted by the binary model to be unsuccessful.[45] Using the

https://doi.org/10.1021/jacs.4c00098
J. Am. Chem. Soc. 2024, 146, 15070−15084

workflow described previously, 6 of these were successfully replaced with design-relevant compounds (Figure 9a). Therefore, the prospective model-guided design led to the successful synthesis of 24 total compounds if the model was used at the time of the original synthesis.[46] Meanwhile, in the monomer rescue scenario, the model is better informed on the performance of monomers with the core of interest. Carrying out the *rescue* approach on the same library enabled the successful replacement of 23 of the 27 previously unsuccessful monomers (Figure 9b). This amounts to the successful synthesis of 42 total compounds using this core when combined with the successful reactions from the original library synthesis. As a particular example, the model anticipated that monomers bearing tertiary aniline groups would not lead to the highest yields of the potential replacement monomers (see the Supporting Information, Figure S30). We repeated this monomer rescue workflow for a second library using a different core with a different set of reaction conditions and again confirmed its viability (see the Supporting Information, Figure S23 for selected examples). This further demonstrates the broad potential of the model to suggest the successful replacement of monomers with new reactants that adhere to the original design as closely as possible (see the Supporting Information, Tables S14 and S15 for the full list of query monomers, replacement monomers, similarity scores, predicted yields, and isolated yields). It should also be emphasized that the monomer replacement approach can be flexible. While the replacement monomer with the highest predicted yield was chosen for this study, in practical use, a medicinal chemist can choose to prioritize monomer similarity score over the predicted reaction yield depending on the specific needs of the medicinal chemistry program they support.

The results of this study clearly demonstrate how the yield prediction models could be incorporated into the DMTA cycle for medicinal chemistry campaigns to achieve increased design and synthesis efficiency. When included in the design phase, medicinal chemists will be able to assess the synthesis outcome via predicted yields, which represents a large void in the current design process. The implications could be significant, as the complex core is the limiting reactant and often requires lengthy and tedious synthesis that may sometimes occupy a chemist's efforts for more than one month. Thus, judicious use of the advanced material to yield the maximum number of library analogues becomes critically important in shortening the DMTA cycle. Our results give us confidence that, by including synthetic yield prediction into the design funnel as showcased by the example given (monomer prioritization, Figure 9a), higher synthesis efficiency could be achieved. Additionally, by combining the yield prediction model with molecular similarity search for key analogues that were unsuccessful in synthesis, one could quickly identify similar monomers that are predicted to exhibit higher synthetic yields. These alternative analogues can be efficiently synthesized with speed to funnel more relevant compounds through the DMTA cycle (monomer rescue, Figure 9b), in parallel to identifying workable synthesis methods/conditions for the specific analogues that were unsuccessful in synthesis, if deemed necessary.

CONCLUSIONS

Among the various applications of ML in chemistry, reaction yield prediction remains to be a very challenging task. This is primarily due to the intrinsic complexity of the parameters that may influence the reaction yields, the lack of high-fidelity data sets that are suitable for ML, and our currently limited experience and knowledge on general representations well suited for reactivity prediction tasks. In theory, our 15-year parallel library data set offers the ideal real-world data set for reaction yield prediction in the context of medicinal chemistry research. Compared to data sets extracted from corporate ChemELNs, data sets extracted from libraries bear unique advantages, including semiautomated reaction processes, high consistency of the data points due to the nature of the parallel synthesis format, and the minimal variation in parameters that may influence yields, such as operators and reaction scales. Using isolated library yields as the end point is also much more pertinent for reaction yield prediction, contrary to most HTE outputs. The balance of positive and negative data points in our data set is another advantage over existing public data sets, let alone the high relevance of the chemical structures in our data set for pharmaceutically relevant reaction yield prediction. Furthermore, parallel library synthesis is executed and managed by a centralized lab within AbbVie using our highly customized LIMS. This ensured the consistency and comprehensiveness of the data capture, curation, and storage, which is generally not the case for most large and diverse reaction data sets. Nevertheless, significant efforts were spent on our data set cleanup to make the data set suitable for ML. For example, as mentioned previously, the variation in reagent nomenclature required manual efforts to consolidate and standardize the catalysts, ligands, bases, and solvents used. In addition, there were variations in the ways different operators handled certain library synthesis scenarios. These cases were identified, consulted with the original operators, and dealt with individually. The extent of this effort is reflected by the final 24,203 reactions used for this study versus the original >40,000 Suzuki coupling data points in our library database, as a significant portion of the data points were eliminated to achieve high data quality and integrity. Taken together, our experience highlights the critical need to involve data scientists much earlier in the data set generation and storage discussions. Internally, we have strategically decided to involve data scientists at the very beginning of the discussion of our next-generation LIMS to ensure the data is captured in more ML-friendly ways, foreseeing the desire to utilize the data set downstream for modeling purposes without the need for data set processing. One can deduce that such a strategy could also benefit corporate ChemELN or other chemistry-related software development, an aspect that was often overlooked in the past.

The 24,203 Suzuki reactions curated over the course of 15 years only represented 3% of all of the possible combinations of cores and monomers, even without including other factors such as reaction conditions. Some condition information was not represented in this data set (e.g., concentrations and temperatures), as they were not captured by our LIMS. For example, information related to purification and product stability was also not captured, which represents factors that are challenging to account for when modeling data sets derived from isolated yields. Even though we expect that Suzuki reactions should be less sensitive to reaction conditions than other reaction types, as detailed earlier, these factors exemplify a huge challenge in the field of reaction yield prediction, even for our high-quality library synthesis data set. The potential parameters that may influence the reaction yields are vast and

https://doi.org/10.1021/jacs.4c00098
J. Am. Chem. Soc. 2024, 146, 15070−15084

may not be easily identified or quantified. As such, questions may be raised regarding how many and which reaction data points one would need to predict yields on even one transformation with confidence. Furthermore, obtaining isolated yields is also much more labor-intensive and time-consuming than, for instance, reaction assessment via UV analysis of crude reaction mixtures against an internal standard. The necessary time and financial investments of building such data sets can be considerable, even though many companies have initiated efforts to build and centralize their internal reaction data sets.[16a,47] Taken together, if we as a community are to make strides in this field, two important factors to consider would be (1) how to harmonize the data into an ML-ready format from the start[47,48] and (2) how to obtain a critical mass of data points to make sufficiently accurate predictions.

Despite these existing challenges, our evaluation of AbbVie's 15-year parallel library data set to build yield prediction models has demonstrated promising results for prospective use of reaction yield prediction in the design stage of the DMTA cycle. The upkeep of a persistent database since 2006 was essential to furnish a data set suitable for machine learning in this study. Retrospective model performance exceeds naive baselines by a substantial margin, and FP and/or DFT representations enable at least partial generalization to unseen structures, as evidenced by superiority over OHE-based models. Various featurization methods and ML models were investigated in this study. Although DFT featurization improved the overall model performance, for incorporating synthetic yield prediction effectively in the routine drug design for SAR studies, computationally inexpensive FP representations alone may be sufficient. Although there is room for improvement in model performance in this study, gratifyingly, we have demonstrated that the best RF model, using a combination of FP and DFT features, could often outcompete the prediction accuracy of expert medicinal chemists. This comparison provided strong evidence that this model could increase reaction success rates if deployed in the medicinal chemistry design workflow. Taken one step further, we showed a practical, prospective application of the models toward increasing parallel library reaction success, both prior to and after conducting a library, by suggesting structurally and electronically similar monomers predicted to achieve higher synthetic yields. To our knowledge, this signifies for the first time that a synthesis outcome prediction tool could be incorporated into the design cycle along with other design tools used routinely by medicinal chemists, with the great potential to shorten DMTA cycle time and improve overall synthesis efficiency for SAR endeavors. We envision that this model will be used broadly internally in AbbVie as one of the essential design tools in directing SAR studies, and we are internally conducting case studies of the models' use in the DMTA cycle.

Our current efforts are directed toward designing experiments to supplement our existing models to better generalize to unseen reactant structures, along with investigating other parallel library data sets for reaction yield prediction using the comprehensive approach detailed in this manuscript. We hope this work inspires others to investigate and work together to accelerate the field of reaction yield prediction, whether from the aspects of reaction data set curation/design, molecular featurization, or the development of new ML architectures.

■ ASSOCIATED CONTENT

Data Availability Statement

The data underlying this study are available in the published article and its online Supporting Information. One-hot encoded versions of the proprietary data sets have been made available in the GitHub repository for this project (shown below).

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacs.4c00098.

Data set information, Supporting Figures, and new compound characterization data (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Connor W. Coley − *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0002-8271-8723; Email: ccoley@mit.edu

Ying Wang − *Advanced Chemistry Technologies Group, AbbVie, Inc., North Chicago, Illinois 60064, United States;* orcid.org/0000-0001-9516-3483; Email: wang.ying@abbvie.com

Authors

Priyanka Raghavan − *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States;* orcid.org/0009-0002-5949-4570

Alexander J. Rago − *Advanced Chemistry Technologies Group, AbbVie, Inc., North Chicago, Illinois 60064, United States;* orcid.org/0000-0002-9621-1974

Pritha Verma − *Advanced Chemistry Technologies Group, AbbVie, Inc., North Chicago, Illinois 60064, United States*

Majdi M. Hassan − *RAIDERS Group, AbbVie, Inc., North Chicago, Illinois 60064, United States*

Gashaw M. Goshu − *Advanced Chemistry Technologies Group, AbbVie, Inc., North Chicago, Illinois 60064, United States*

Amanda W. Dombrowski − *Advanced Chemistry Technologies Group, AbbVie, Inc., North Chicago, Illinois 60064, United States;* orcid.org/0000-0001-9596-263X

Abhishek Pandey − *RAIDERS Group, AbbVie, Inc., North Chicago, Illinois 60064, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/jacs.4c00098

Author Contributions

‖P.R. and A.J.R. contributed equally to this work. The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding

P.R. acknowledges support from the MIT-Takeda fellowship. P.R. and C.W.C. acknowledge additional support from the NSF Center for Computer Assisted Synthesis (C-CAS) under Grant CHE-2202693.

Notes

The authors declare the following competing financial interest(s): A.J.R. (postdoctoral researcher), P.V., A.W.D., A.P., and Y.W. are employees of AbbVie, Inc. G.M.G. is a

contract employee of AbbVie, Inc. M.M.H. is a former employee of AbbVie, Inc. The design, study conduct, and financial support for this research were provided by AbbVie, Inc. AbbVie, Inc., participated in the interpretation of data, review, and approval of this publication. P.R. and C.W.C. declare no competing financial interests.

Code Availability The underlying code and one-hot encoded versions of the data sets can be found at the following link: https://github.com/priyanka-rag/suzuki_yield_predict_external.

■ ACKNOWLEDGMENTS

The authors acknowledge the following AbbVie, Inc., scientists for their support: Ana Aguirre, Leena Bhatt, Andrew Bogdan, Emma Clay-Barbour, Marlon Cowart, Ethan Hoff, Keith Galyan, Cheng Ji, Renhe Li, Timothy Montavon, Augustine Osuma, Aaron Roth, Katerina Sarris, Phillip Searle, Anurupa Shrestha, Aristidis Vasilopoulos, Frank Wagenaar, Jan Waters, Rick Yarbrough, and Jonathon Young. The authors also acknowledge John Bradshaw of MIT for generating the language-based features investigated during the retrospective modeling.

■ REFERENCES

(1) Wesolowski, S. S.; Brown, D. G. The Strategies and Politics of Successful Design, Make, Test, and Analyze (DMTA) Cycles in Lead Generation. *Lead Gener.* **2016**, 487−512.

(2) Blakemore, D. C.; Castro, L.; Churcher, I.; Rees, D. C.; Thomas, A. W.; Wilson, D. M.; Wood, A. Organic synthesis provides opportunities to transform drug discovery. *Nat. Chem.* **2018**, *10*, 383−394.

(3) (a) Singh, N.; Sun, H.; Chaudhury, S.; AbdulHameed, M. D. M.; Wallqvist, A.; Tawa, G. A physicochemical descriptor-based scoring scheme for effective and rapid filtering of kinase-like chemical space. *J. Cheminf.* **2012**, *4* (1), No. 4, DOI: 10.1186/1758-2946-4-4. (b) Rossi Sebastiano, M.; Garcia Jimenez, D.; Vallaro, M.; Caron, G.; Ermondi, G. Refinement of Computational Access to Molecular Physicochemical Properties: From Ro5 to bRo5. *J. Med. Chem.* **2022**, *65*, 12068−12083.

(4) Alberga, D.; Trisciuzzi, D.; Montaruli, M.; Leonetti, F.; Mangiatordi, G. F.; Nicolotti, O. A New Approach for Drug Target and Bioactivity Prediction: The Multifingerprint Similarity Search Algorithm (MuSSeL). *J. Chem. Inf. Model.* **2019**, *59*, 586−596.

(5) Di Lascio, E.; Gerebtzoff, G.; Rodríguez-Pérez, R. Systematic Evaluation of Local and Global Machine Learning Models for the Prediction of ADME Properties. *Mol. Pharmaceutics* **2023**, *20*, 1758−1767.

(6) (a) Janssen, A. P. A.; Grimm, S. H.; Wijdeven, R. H. M.; Lenselink, E. B.; Neefjes, J.; van Boeckel, C. A. A.; van Westen, G. J. P.; van der Stelt, M. Drug Discovery Maps, a Machine Learning Model That Visualizes and Predicts Kinome−Inhibitor Interaction Landscapes. *J. Chem. Inf. Model.* **2019**, *59*, 1221−1229. (b) Saldívar-González, F. I.; Medina-Franco, J. L. Approaches for enhancing the analysis of chemical space for drug discovery. *Expert Opin. Drug Discovery* **2022**, *17*, 789−798, DOI: 10.1080/17460441.2022.2084608.

(7) Torres, P. H. M.; Sodero, A. C. R.; Jofily, P.; Silva-Jr, F. P. Key Topics in Molecular Docking for Drug Design. *Int. J. Mol. Sci.* **2019**, *20*, 4574.

(8) For AbbVie's library synthesis dataset, the synthesis success rate is defined as the fraction of reaction products that were successfully registered for downstream evaluation in biological assays.

(9) Dombrowski, A. W.; Aguirre, A. L.; Shrestha, A.; Sarris, K. A.; Wang, Y. The Chosen Few: Parallel Library Reaction Methodologies for Drug Discovery. *J. Org. Chem.* **2022**, *87*, 1880−1897.

(10) (a) 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*, 186−190. (b) Fu, Z.; Li, X.; Wang, Z.; Li, Z.; Liu, X.; Wu, X.; Zhao, J.; Ding, X.; Wan, X.; Zhong, F.; Wang, D.; Luo, X.; Chen, K.; Liu, H.; Wang, J.; Jiang, H.; Zheng, M. Optimizing chemical reaction conditions using deep learning: a case study for the Suzuki−Miyaura cross-coupling reaction. *Org. Chem. Front.* **2020**, *7*, 2269−2277. (c) Götz, J.; Jackl, M. K.; Jindakun, C.; Marziale, A. N.; André, J.; Gosling, D. J.; Springer, C.; Palmieri, M.; Reck, M.; Luneau, A.; Brocklehurst, C. E.; Bode, J. W. High-throughput synthesis provides data for predicting molecular properties and reaction success. *Sci. Adv.* **2023**, *9*, No. eadj2314, DOI: 10.1126/sciadv.adj2314. (d) Fitzner, M.; Wuitschik, G.; Koller, R.; Adam, J.-M.; Schindler, T. Machine Learning C−N Couplings: Obstacles for a General-Purpose Reaction Yield Prediction. *ACS Omega* **2023**, *8*, 3017−3025.

(11) Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J.-L. Prediction of chemical reaction yields using deep learning. *Mach. Learn.: Sci. Technol.* **2021**, *2*, No. 015016.

(12) (a) Saebi, M.; Nan, B.; Herr, J. E.; Wahlers, J.; Guo, Z.; Zurański, A. M.; Kogej, T.; Norrby, P.-O.; Doyle, A. G.; Chawla, N. V.; Wiest, O. On the use of real-world datasets for reaction yield prediction. *Chem. Sci.* **2023**, *14*, 4997−5005. (b) Yarish, D.; Garkot, S.; Grygorenko, O. O.; Radchenko, D. S.; Moroz, Y. S.; Gurbych, O. Advancing molecular graphs with descriptors for the prediction of chemical reaction yields. *J. Comput. Chem.* **2023**, *44*, 76−92.

(13) (a) Shevlin, M. Practical High-Throughput Experimentation for Chemists. *ACS Med. Chem. Lett.* **2017**, *8*, 601−607. (b) Mennen, S. M.; Alhambra, C.; Allen, C. L.; Barberis, M.; Berritt, S.; Brandt, T. A.; Campbell, A. D.; Castañón, J.; Cherney, A. H.; Christensen, M.; Damon, D. B.; Eugenio De Diego, J.; García-Cerrada, S.; García-Losada, P.; Haro, R.; Janey, J.; Leitch, D. C.; Li, L.; Liu, F.; Lobben, P. C.; Macmillan, D. W. C.; Magano, J.; McInturff, E.; Monfette, S.; Post, R. J.; Schultz, D.; Sitter, B. J.; Stevens, J. M.; Strambeanu, I. I.; Twilton, J.; Wang, K.; Zajac, M. A. The Evolution of High-Throughput Experimentation in Pharmaceutical Development and Perspectives on the Future. *Org. Process Res. Dev.* **2019**, *23*, 1213−1242.

(14) (a) Cook, A.; Clément, R.; Newman, S. G. Reaction screening in multiwell plates: high-throughput optimization of a Buchwald−Hartwig amination. *Nat. Protoc.* **2021**, *16*, 1152−1169. (b) Christensen, M.; Yunker, L. P. E.; Adedeji, F.; Häse, F.; Roch, L. M.; Gensch, T.; dos Passos Gomes, G.; Zepel, T.; Sigman, M. S.; Aspuru-Guzik, A.; Hein, J. E. Data-science driven autonomous process optimization. *Commun. Chem.* **2021**, *4*, No. 112, DOI: 10.1038/s42004-021-00550-x. (c) Rago, A. J.; Vasilopoulos, A.; Dombrowski, A. W.; Wang, Y. Di(2-picolyl)amines as Modular and Robust Ligands for Nickel-Catalyzed C(sp2)−C(sp3) Cross-Electrophile Coupling. *Org. Lett.* **2022**, *24*, 8487−8492. (d) Gesmundo, N. J.; Tu, N. P.; Sarris, K. A.; Wang, Y. ChemBeads-Enabled Photoredox High-Throughput Experimentation Platform to Improve C(sp2)−C(sp3) Decarboxylative Couplings. *ACS Med. Chem. Lett.* **2023**, *14*, 521−529.

(15) 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*, 14722−14730.

(16) (a) Neves, P.; McClure, K.; Verhoeven, J.; Dyubankova, N.; Nugmanov, R.; Gedich, A.; Menon, S.; Shi, Z.; Wegner, J. K. Global reactivity models are impactful in industrial synthesis applications. *J. Cheminf.* **2023**, *15*, No. 20, DOI: 10.1186/s13321-023-00685-0. (b) Liu, Z.; Moroz, Y. S.; Isayev, O. The challenge of balancing model sensitivity and robustness in predicting yields: a benchmarking study of amide coupling reactions. *Chem. Sci.* **2023**, *14*, 10835−10846.

(17) (a) Fitzner, M.; Wuitschik, G.; Koller, R. J.; Adam, J.-M.; Schindler, T.; Reymond, J.-L. What can reaction databases teach us about Buchwald−Hartwig cross-couplings? *Chem. Sci.* **2020**, *11* (48), 13085−13093. (b) Beker, W.; Roszak, R.; Wołos, 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*, 4819−4827.
(18) 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. *Org. Lett.* **2023**, *25*, 2945−2947.
(19) 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*, No. e202204647, DOI: 10.1002/anie.202204647.
(20) (a) Christ, C. D.; Zentgraf, M.; Kriegl, J. M. Mining Electronic Laboratory Notebooks: Analysis, Retrosynthesis, and Reaction Based Enumeration. *J. Chem. Inf. Model.* **2012**, *52*, 1745−1756. (b) Ghiandoni, G. M.; Bodkin, M. J.; Chen, B.; Hristozov, D.; Wallace, J. E. A.; Webster, J.; Gillet, V. J. Development and Application of a Data-Driven Reaction Classification Model: Comparison of an Electronic Lab Notebook and Medicinal Chemistry Literature. *J. Chem. Inf. Model.* **2019**, *59*, 4167−4187.
(21) (a) Xu, J.; Kalyani, D.; Struble, T.; Dreher, S. D.; Krska, S.; Buchwald, S. L.; Jensen, K. F. Roadmap to Pharmaceutically Relevant Reactivity Models Leveraging High-Throughput Experimentation *ChemRxiv* 2022 DOI: 10.26434/chemrxiv-2022-x694w. (b) 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*, 965−972.
(22) Brown, D. G.; Boström, J. Analysis of Past and Present Synthetic Methodologies on Medicinal Chemistry: Where Have All the New Reactions Gone? *J. Med. Chem.* **2016**, *59*, 4443−4458.
(23) Angello, N. H.; Rathore, V.; Beker, W.; Wołos, A.; Jira, E. R.; Roszak, R.; Wu, T. C.; Schroeder, C. M.; Aspuru-Guzik, A.; Grzybowski, B. A.; Burke, M. D. Closed-loop optimization of general reaction conditions for heteroaryl Suzuki-Miyaura coupling. *Science* **2022**, *378*, 399−405.
(24) Roughley, S. D.; Jordan, A. M. The Medicinal Chemist's Toolbox: An Analysis of Reactions Used in the Pursuit of Drug Candidates. *J. Med. Chem.* **2011**, *54*, 3451−3479.
(25) The structures contained in this dataset are proprietary and cannot be shared. The one-hot encoded version of the dataset has been made available as a separate dataset.
(26) The difference between the total number of reactions and number of unique reaction products originates from (1) duplicate reactions conducted in different libraries and (2) the use of different functional handles (e.g. B(OH)2 vs Bpin) to install the same R-group
(27) van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579−2605.
(28) Note: There are 358 × 2343 = 838,794 possible core/monomer pairs for the aryl halide cores and 35 × 835 = 29,225 possible core/monomer pairs for the organoborane cores. Therefore, our dataset is reaching 23,837/868,019 of the possible core/monomer pairs. This number is greater than the number of unique products due to multiple monomers in the dataset (e.g., boronic acid vs boronic ester) leading to the same unique product.
(29) Ingoglia, B. T.; Wagen, C. C.; Buchwald, S. L. Biaryl monophosphine ligands in palladium-catalyzed C−N coupling: An updated User's guide. *Tetrahedron* **2019**, *75*, 4199−4211.
(30) Wang, Y.; Haight, I.; Gupta, R.; Vasudevan, A. What is in Our Kit? An Analysis of Building Blocks Used in Medicinal Chemistry Parallel Libraries. *J. Med. Chem.* **2021**, *64*, 17115−17122.
(31) Chuang, K. V.; Keiser, M. J. Comment on "Predicting reaction performance in C−N cross-coupling using machine learning. *Science* **2018**, *362*, No. eaat8603, DOI: 10.1126/science.aat8603.
(32) Haywood, A. L.; Redshaw, J.; Hanson-Heine, M. W. D.; Taylor, A.; Brown, A.; Mason, A. M.; Gärtner, T.; Hirst, J. D. Kernel Methods for Predicting Yields of Chemical Reactions. *J. Chem. Inf. Model.* **2022**, *62*, 2077−2092.
(33) Żurański, A. M.; Wang, J. Y.; Shields, B. J.; Doyle, A. G. Auto-QChem: an automated workflow for the generation and storage of DFT calculations for organic molecules. *React. Chem. Eng.* **2022**, *7*, 1276−1284.
(34) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 16*, Rev. C.01; Wallingford, CT, 2016.
(35) Landrum, G.et al. *rdkit/rdkit: 2022_03_5 (Q1 2022) Release*, Zenodo 2022.
(36) 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*, 1652−1671.
(37) Note: The RF model was obtained from the scikit-learn package, the xGB model was provided by Kaggle, and the NN model was implemented using the pytorch package-.
(38) The random forest model used in this exercise used default hyperparameters.
(39) The one-hot encoded reactions with the true results, the model's predictions, and each chemists' predictions are available as a separate file. As with the rest of the dataset, these structures are proprietary and cannot be shared.
(40) (a) 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*, 89−96. (b) Caldeweyher, E.; Elkin, M.; Gheibi, G.; Johansson, M.; Sköld, C.; Norrby, P.-O.; Hartwig, J. F. Hybrid Machine Learning Approach to Predict the Site Selectivity of Iridium-Catalyzed Arene Borylation. *J. Am. Chem. Soc.* **2023**, *145*, 17367−17376.
(41) The structures contained in this pseudo-prospective dataset are proprietary and cannot be shared. A one-hot encoded version of the dataset has been made available.
(42) The model used for the monomer prioritization study was trained on the entire retrospective dataset, minus 11 libraries identified as potential choices for the replacement study. The core used in the monomer prioritization part of this study also appeared as a monomer once in the dataset. This reaction was identified and removed from the training data used to obtain the monomer prioritization predictions to ensure that the library/core was unseen to the model. For all monomer rescue predictions, the model was trained using the entirety of the retrospective dataset.
(43) Bolcato, G.; Heid, E.; Boström, J. On the Value of Using 3D Shape and Electrostatic Similarities in Deep Generative Methods. *J. Chem. Inf. Model.* **2022**, *62*, 1388−1398.
(44) The binary random forest model for predicting whether an originally selected monomer might be unsuccessful in the monomer prioritization study used the optimized hyperparameters from the random forest core-split binary model. The regression random forest models for selecting the replacement monomers used in this study used the optimized hyperparameters from the core-split and monomer-split regression models for monomer prioritization and monomer rescue, respectively.
(45) 8/9 of these monomers were unsuccessful during the original synthesis.
(46) Since one of the monomers that would have been replaced was successful in the original synthesis, the total number of compounds would have been 24 instead of 25.

(47) 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*, 18820−18826.
(48) 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*, 2196−2204.

https://doi.org/10.1021/jacs.4c00098
J. Am. Chem. Soc. 2024, 146, 15070−15084