ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

Machine Learning C−N Couplings: Obstacles for a General-Purpose Reaction Yield Prediction

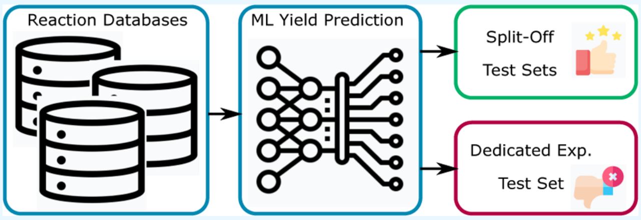
Martin Fitzner, Georg Wuitschik,* Raffael Koller, Jean-Michel Adam, and Torsten Schindler

Cite This: *ACS Omega* 2023, 8, 3017−3025

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Pd-catalyzed C−N couplings are commonplace in academia and industry. Despite their significance, finding suitable reaction conditions leading to a high yield, for instance, remains a challenging and time-consuming task which usually requires screening over many sets of conditions. To help select promising reaction conditions in the vast space of reagent combinations, machine learning is an emerging technique with a lot of promise. In this work, we assess whether the reaction yield of C−N couplings can be predicted from databases of chemical reactions. We test the generalizability of models both on challenging data splits and on a dedicated experimental test set. We find that, provided the chemical space represented by the training set is not left, the models perform well. However, the applicability domain is quickly left even for simple reactions of the same type, as, for instance, present in our plate test set. The results show that yield prediction for new reactions is possible from the algorithmic side but in practice is hindered by the available data. Most importantly, more data that cover the diversity in reagents are needed for a general-purpose prediction of reaction yields. Our findings also expose a challenge to this field in that it appears to be extremely deceiving to judge models based on literature data with test sets which are split off the same literature data, even when challenging splits are considered.



1. INTRODUCTION

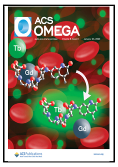
Chemistry is in the midst of a data-driven shift.[1−3] Many areas of the broader field experience a surge in development aimed at utilizing data, from machine learning (ML) models replacing aspects of computational chemistry and property calculations[4] to retrosynthetic planning[5,6] to optimizing reaction conditions.[7,8] The latter example is a problem particularly suitable for these new approaches due to the quickly exploding space of possibilities. The optimization of reaction conditions is a frequently encountered problem in academia and industry,[9] often posing a bottleneck to efficient, fast, and productive retrosynthesis. Screening over many sets of reaction conditions is still required to find, for example, acceptable yields. With all the knowledge stored in reaction databases populated over the last decades, ML is a promising route to improve the state of the art for reaction condition optimizations.

Several concepts can be designed to harness data-driven technology for the purpose of optimizing reaction conditions. Work by Gao et al.,[10] Genheden et al.,[11] as well as Li and Eastgate[12] used neural networks that predict the success probability for a set of substances in each relevant reagent class. In these approaches, often the accuracy of the top-N most probable predictions is evaluated, which rates if the proposed set of N conditions contains the one that achieved the highest yield. This is a rigorous approach when the outcome for all possible reagent combinations that the model could predict is known. However, the metric can also be computed if the predictions contain combinations for which the reaction outcome is not measured, assuming the best conditions are the best known conditions. This assumption is dangerous and could distort our ability to judge whether the model can truly rank reaction conditions. In particular, for a study like this one in which we will use literature databases where the outcome of most conceivable combinations is not known, this approach is unsuitable.

Recently, we have also seen great progress in studies using Bayesian optimization as a foundation,[13−15] where a new model is iteratively built for each new reactant pair based on the performed measurements. This ansatz is very suited to the optimization problem in the lab, and promising results have been reported. However, it is unclear whether the knowledge of previously performed optimizations can be included to make use of all available data.

Another route to help select promising reaction conditions is simply regression of the reaction yield or other desired properties. Since the model evaluation is generally cheap compared to training, for most use-cases it should be feasible to simply iterate over all combinations of reagents and rank them. This could be followed by grouping and sorting to accommodate experimental design, which is often limited in

Received: September 13, 2022
Accepted: October 25, 2022
Published: January 11, 2023



terms of which combinations can be effectively screened together. For a model to be useful, it is only necessary to achieve the correct ranking of conditions and not the exact numerical target. This approach by default can only be evaluated if all the considered conditions have actually been measured, which is more restrictive but then leads to a rigorous evaluation of the models' capacity to optimize. Another advantage of the regression approach is that, provided the physical descriptors have been used to characterize reagents, new reagents can be evaluated without having to redesign/retrain the model since the output is the target and not a set of pre-fixed categorical probabilities.

The regression approach has been tested on a variety of data sets, predominantly focused on reactions from high-throughput experimentation (HTE), typically on the order of a few thousand data points. Ahnemann et al.[7] created a set of around 4000 HTE reactions spanning a C−N coupling space. Data splits with their random forest model could achieve an excellent performance of R^2 of around 0.92. When additives were left out for a test set, the performance dropped to R^2 of around 0.83, which suggests reasonable generalizability, albeit the other reagents remained the same between training and test sets. Dong et al.[16] also performed a work on this C−N data set, suggesting that good model performances could also be achieved with substantially smaller training data set size. Probst et al.[17] introduced a novel reaction fingerprint and tested it on these data, which achieved a similar or better performance. Overall, for this dataset, both the training performance and extrapolation performance to unseen additives have been tested with a variety of different methods, all suggesting that ML can push the boundaries of HTE yield prediction problems. Similar HTE data sets have been investigated, for example, by Nielsen et al.,[18] who could show good predictive performance for a deoxyfluorination data set treated with the modeling approach of ref 7. With a set of novel steric descriptors, Zahrt et al.[19] were able to demonstrate the prediction of enantioselectivity with promising performance on several test cases, such as unseen substrates and unseen catalysts. Sandfort et al.[20] tested their model variant utilizing concatenated fingerprints on this selectivity as well as the yield prediction problem of Ahnemann et al.,[7] suggesting their flexible model could easily be used for different chemical prediction problems.

Less attention has been dedicated to creating a general-purpose model aimed at covering larger chemical spaces than generated in a typical HTE data set. Skoraczyński et al.[21] created a model predicting whether the yield is above a certain threshold, training on a large set of reactions not restricted to any specific sub-type. They conclude that the yield prediction is not possible and attribute this to the descriptors available. Schwaller et al.[22] use the chemical transformer[23,24] for predicting the reaction yields of different splits of the C−N data from ref 7, as well as random splits of US Patent and Trademark Office (USPTO) data filtered for C−N cross couplings. While excellent regression performances could be achieved for the HTE data, they find that the same approach gives unsatisfactory results for the USPTO database, which covers a large chemical space. Probst et al. could improve on this performance with their differential reaction fingerprints,[17] albeit with an overall performance of $R^2 < 0.2$ the prediction on the USPTO set remains difficult. The results of these works suggests that a general-purpose yield prediction is a substantial challenge and might be hindered by either the digital description of the reaction or the available data.

In this work, we tested whether an ML model trained on Buchwald−Hartwig C−N coupling[25−29] data comprising three of the largest available reaction databases can achieve a general-purpose yield prediction. To this end, we investigated a variety of modeling approaches and data splits, including a dedicated test set mimicking industrial settings created specifically for this work. The size of our training data set is roughly one order of magnitude larger than what has been used in previous works. While a promising performance could be reached for even the harder self-consistent splits, the model was not able to make good predictions on the external set. We attribute this to strong biases in yield and reagent diversity in the databases that display opposite tendencies to what is typically found in the lab.

In Section 2, we describe the different modeling approaches investigated. This is followed by discussing the prediction performance for data splits (Section 3.1) as well as an analysis on whether anything can be gained from dividing the C−N coupling reaction space further into reasonable sub-types in Section 3.2. The performance on the test plates is presented in Section 3.3. We study various factors potentially explaining the outcome in Section 3.4. We conclude in Section 4.

2. METHODS

As the foundation for this work, we use the cleaned reaction data set analyzed in ref 30. These reactions comprise the Reaxys[31,32] database from Elsevier, the Scifinder[33] database from Chemical Abstract Services (CAS) as well as patents from the USPTO.[34,35] These have been cleaned regarding several aspects, for instance, yield availability, availability of chemical structures of all involved molecules, and an unambiguous assignment of the role of each involved reagent (i.e., solvent, base, and ligand). As in ref 30, we do not consider the type of pre-catalyst in this work. Furthermore, all duplicated entries have been removed. For a detailed workflow and technical details, the interested reader is referred to ref 30. In total, we have around 62 000 unique reactions.

From these data, we train several ML models utilizing different features:

1. The multiple fingerprint features (MFF) model from ref 20. This one uses a large set of fingerprints of different types to characterize all the involved molecules.
2. AttentiveFP: A message-passing graph-neural network (GNN) with an attention mechanism.[36] Originally developed for molecular problems, we expand the model by using one graph per involved molecule so it can be applied to reactions.
3. Our custom model, using four types of features together with the gradient booster XGBoost.[37] The features we include are: (i) atom-pair fingerprints of length 1024 for reactants and ligand; (ii) RDKit features,[38] describing various topological aspects for reactants and ligand; (iii) tabular features for solvents[39] and bases (pKA in water and DMSO as well as base charge); and (iv) a locally adapted form of topological features for reactants, only considering a certain radius around the reactive site. This can be, for instance, the number of atoms, bonds, or H-bond donors within this radius. For some of the models, we also tried to refrain from introducing features

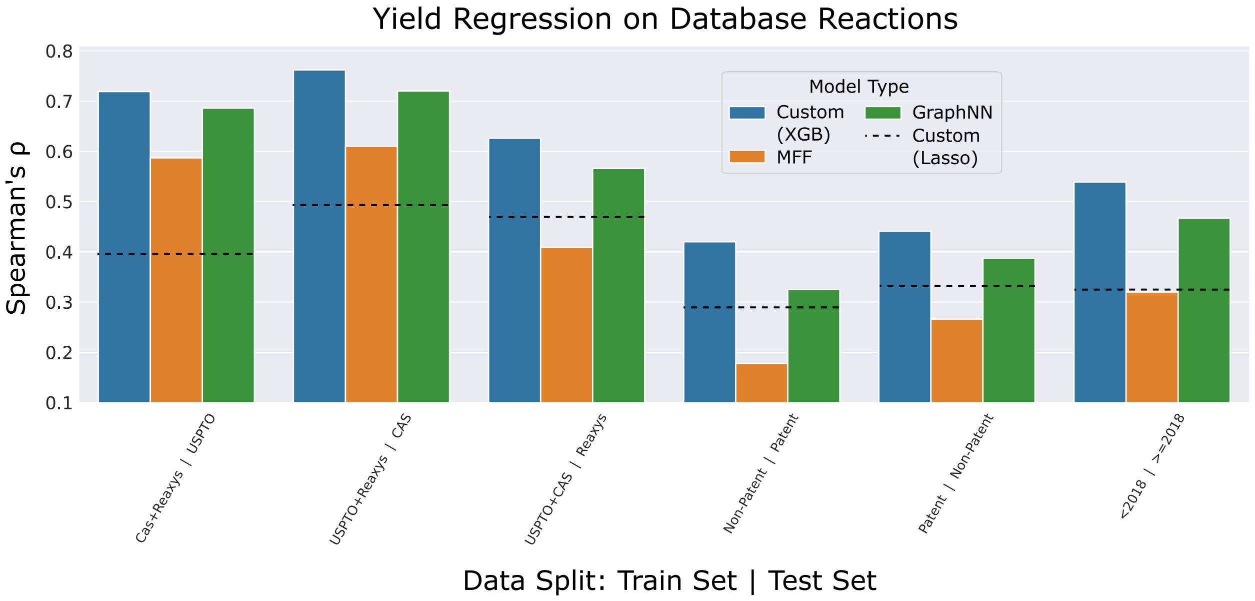


Figure 1. Yield regression results based on training on reaction databases: The *y*-axis shows Spearman's ranking correlation ρ for different data splits and models. The data used for these results was the cleaned reaction set from ref 30, comprising one public and two commercial data sets. The *x*-label denotes the training/test part before/after the "|". The different model types are explained in the Methods section.

for the reagents and instead treat them categorically with CatBoost.[40,41]

In this work, we approach the problem of optimizing reaction conditions from the regression perspective. While alternative approaches can be conceived (as mentioned in the Introduction), the regression of the yield can easily be utilized for optimization. Provided a model can predict the yield well, it would be straightforward to perform a virtual screening of reaction conditions through iteration, subsequently ranking the results by predicted yield. This actually does not require a perfect quantitative prediction of the yield, typically rated with metrics such as the coefficient of determination R^2, mean absolute error (MAE), or root-mean-square error (RMSE), but only needs the correct order. For this reason, we choose to evaluate results, unless indicated otherwise, with the Spearman's ranking correlation ρ. ρ takes values between 1 (in perfect order of the predicted yields) and -1 (predicted yields are in perfect opposite order to the actual yields). A random ordering would have $\rho \approx 0$.

3. RESULTS AND DISCUSSION

3.1. Database Performance. With the main aim to predict the reaction yield of C−N coupling reactions, we first focus on the database reactions. In Figure 1 we show the Spearman's ranking correlation ρ for different data splits (indicated on the *x*-axis). Different colors represent the different models outlined earlier.

First, we can recognize that our custom model performs best in all cases. The GNN achieves a similar but consistently worse performance, which could be explained by missing the tabular features to describe solvents and bases, which are not well encoded in their graphs. Attempts to include these features as simple feed forward layers in the neural network did not result in better performance, which could also be due to the resulting mixed architecture requiring a more sophisticated design. The MFF model, in this case, trails behind in performance. We explain this as due to it having the largest amount of features and memory requirement. Our problem requires the characterization of five molecules (both reactants, solvent, base, and ligand), resulting in about 90, 000 features per reaction for 1024 bit fingerprints. The search for hyperparameters was difficult due to this excessive memory requirement. Most of the time, a simple linear LASSO baseline (using our custom features) is also outperformed by either approach.

Second, the performance depends a lot on the data split. Separating training and test sets by databases leads to good performances with ρ values ∼0.6 to 0.75. This is a promising result since all reactions that were duplicates in different databases were removed. We note that after de-duplication of the commercial databases, the training and test sets were still of comparable size, indicating that all databases provide a large number of unique reactions to the overall literature data.

Splitting by patents and non-patents, however, leads to a considerably worse performance. The focus of patent and non-patent publications can often be very different, with the former often not having optimized reactions or even sets of screened reactions as the focus. Also, we anticipate that patent publications contain many more complex molecules from, for example, the pharmaceutical industry, while non-patent reactions more often contain simpler substrates that are used in publications presenting new methodology. Other regression metrics for this figure can be found in the Supporting Information.

As we move from splits by databases to splits by time and document type, we expect to have more out-of-sample reactant pairs in the corresponding test set. In line with this, we see a performance decrease that suggests despite the large training set size, the models are challenged by out-of-sample reactant pairs.

Similar works by Schwaller et al.[22] and Probst et al.[17] gauged the performance across splits of the USPTO dataset irrespective of reaction type and reached R^2 values ∼0.2. For Buchwald−Hartwig couplings, we report R^2 values of 0.531, 0.604, and 0.386 for the first three splits in Figure 1, that is, where the USPTO, CAS, and Reaxys databases were the test sets.

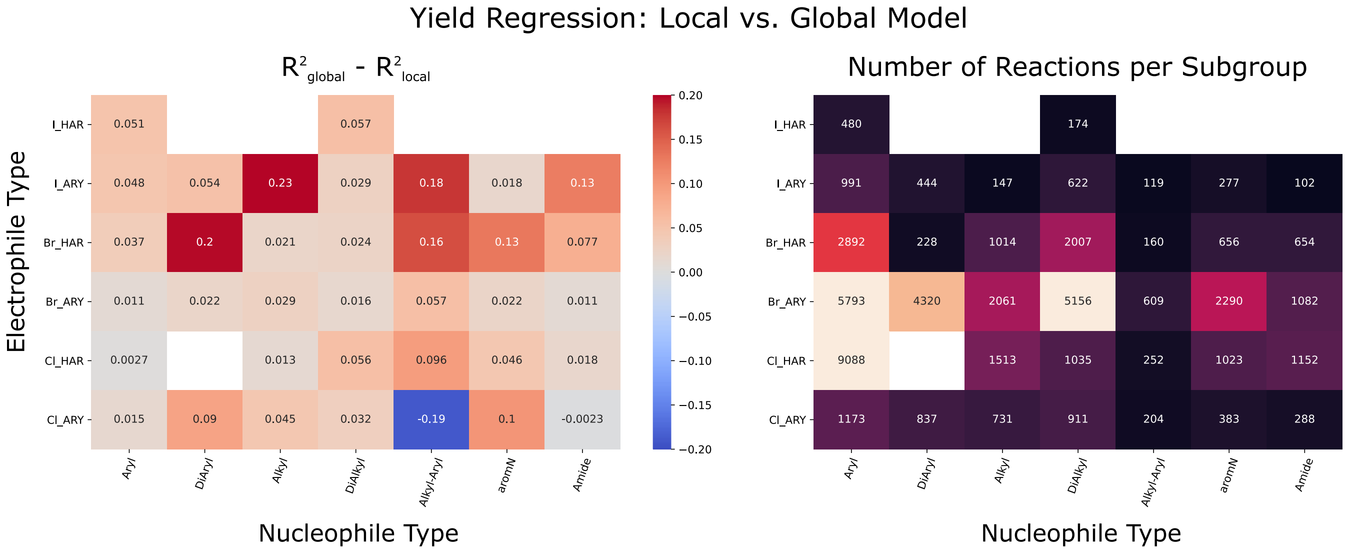


Figure 2. Yield regression performance of global and local models: (a) the data was split based on the reaction's nucleo- and electrophile type (*x*- and *y*-axes) classified in the same manner as in ref 30. Specifically, the electrophile (*y*-axis) has been classified by the leaving group connected to either aryl (ARY) or heteroaryl (HAR), while the nucleophile (*x*-axis) was classified based on the surrounding of the nitrogen. We take 80% of each subgroup for training and the rest for testing. The global model was trained on all trainable parts, while for each subgroup we also created a local model only trained on reactions corresponding to that tile. The number displayed is the R^2 of the global model of that tile minus the R^2 of the local model. (b) Number of reactions for each subgroup.

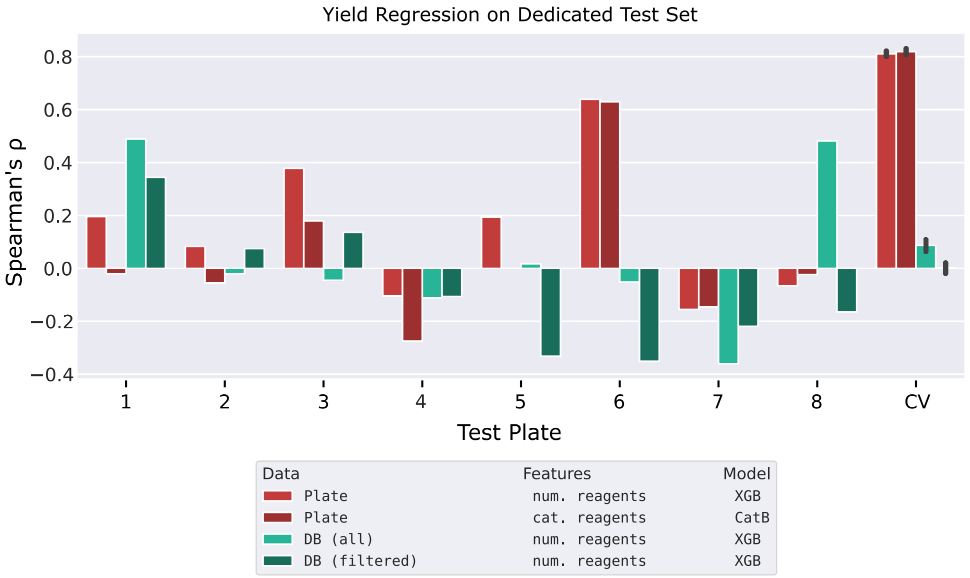


Figure 3. Prediction results on experimental test plates. The *x*-label indicates the test plate number or "CV" for 20 times repeated five-fold cross-validation across all plates. Different colors indicate different models used: (i) an XGBoost[37] (XGB) model using our custom features trained on the remaining non-test plates; (ii) same as (i) but using categorical encoding with CatBoost[40] (CatB) for reagents; (iii) an XGB model using our custom features trained on the database reactions discussed above (DB all) and (iv) same as (iii) but trained only on reactions with an identified scale >500 mg (DB filtered).

We also took reactions published before 2018 and tested them on reactions coming thereafter. This time-based split is perhaps the most realistic estimate of the use-case, as a model deployed in the lab would likely encounter new chemistry over time. Our custom model reached a ρ of about 0.5. This is not perfect, but in our internal assessments of yield rankings, we find empirically that with a ρ of 0.6 or higher, we can already perform a helpful prioritization of conditions. With some more work on the model (such as oversampling to counter the lack of diversity discussed in a later section), this indicates a useful model for ranking reaction conditions could be achieved.

3.2. Local versus Global Models.

Before moving on to the results on experimental plate data, we investigated one more aspect of the database model. In our previous work,[30] we found it helpful to further distinguish different types of substrates for the reactions. For instance, one can distinguish the nucleophile by aryls, alkyls, di-aryls, and so forth. The electrophile can be distinguished by the leaving group and whether it is connected to an aryl or heteroaryl. This opens the question of whether there is any benefit to making this distinction also to the yield predicting model.

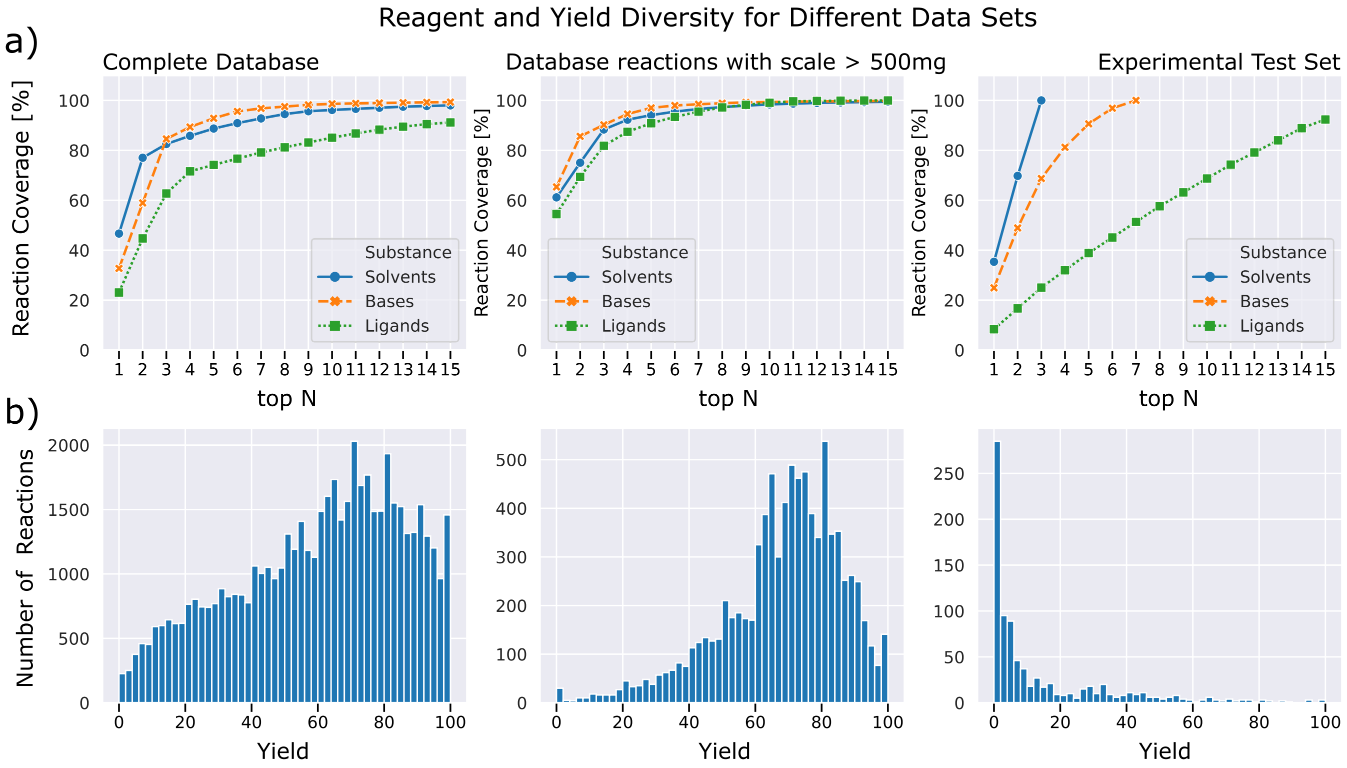


Figure 4. Diversity mismatch between the database reactions and the test plates: (a) Cumulative amount of reactions covered by considering the top N solvents/bases/ligands and (b) histograms of the yields. The first column is for the entire database, the second for the subset of reactions with a scale of >500 mg, and the third is for our experimental test set (not included in other columns).

In Figure 2 we assess the R^2 of models fitted only on a certain sub-group of reactants (termed local models) versus a model fitted on all the data. Stratification is applied to ensure equal parts of all subgroups in the training data of the global model. The result is very clear as, except for one group, the performance of the global model is always equivalent to or better than the local models. This has two positive implications: (i) it makes handling the model deployment easier since only one model has to be considered and (ii) the fact that there seems to be synergy from combining different sub-groups of reactions means that our model was able to learn those synergies, that is, it has learned to some extent the underlying chemistry.

3.3. Performance on Test Plates. The data splits from Figure 1 were aimed at testing the generalizability of the trained models. The variability of the performance as well as its absolute value (e.g., for the time-based split) allows no final conclusion as to whether the model could actually serve as a starting point for a development that could be deployed in a lab. For this reason, we also created a set of 768 reactions from 8 screening plates that serve solely as a test set for our model. For each plate, we use a different pair of reactants, which were selected with the commercial availability in mind. We verified that none of the reactant pairs were present in our reaction database, so they can serve as entirely unseen reactions. Each plate was screened with a different set of reaction conditions from common industrial choices for solvents, bases, and ligands. These data are provided as part of this publication. We hope that the community can use this data set for their own modeling approaches as either an external training or test set.

We use these data in two ways for assessing the model as well as the training data (database or plate data) performance. Figure 3 shows the results from a leave-one-plate-out approach, where seven of the eight plates were used for training and the remaining for testing. The red bars correspond to models using our custom features trained on the seven corresponding training plates. We tried both using numerical features for reagents (with XGBoost[37]) or treating them categorically (with CatBoost[40]). The green bars are the results from the previously discussed model trained on all database reactions. In addition to this, a version that was trained only on reactions that had an identifiable scale of >500 mg was also performed. The last group of bars refers to the plate data not split by plate but with 20 times repeated five-fold cross-validation of the entire plate data.

The results vary and seem rather inconsistent. Overall, the performance is poor, sometimes even achieving a negative ranking correlation. The leave-one-plate-out results are only good for plate 7, with no obvious chemical or other reason behind that. The database model has reasonable performance only for plates 1 and 8, and we also note that the scale-filtered model performs worse overall. The cross-validation split achieves an excellent performance when the model is trained on plate data, but this is the only split where the training and test sets are very similar by construction. When the literature data is used to train the model (green bars for CV split), the ranking correlation drops to 0.

These results are in contrast to the results seen in the database splits. While the poor performance of the models built with plate data can be explained by a rather small training set that is not able to cover chemical diversity at all, it is particularly disappointing to see that the data from reaction databases cannot improve on this. To test whether a dataset starting from our plate reactions and using active learning to

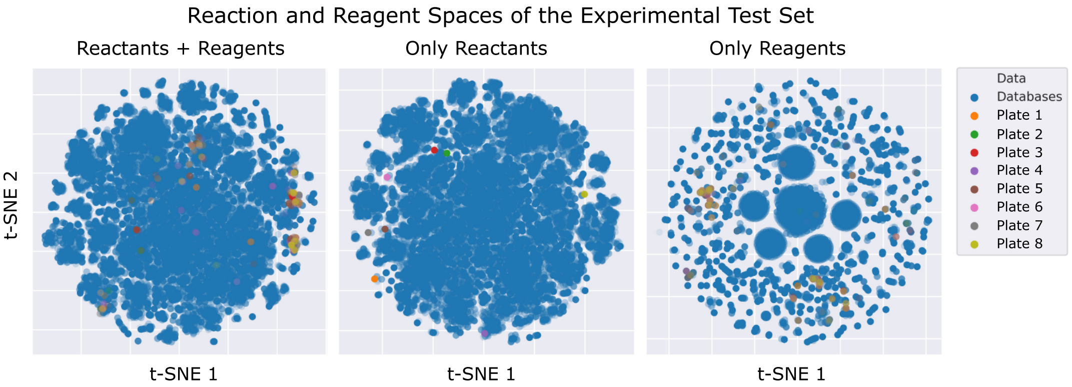


Figure 5. Visualization of the feature space of reactions. Shown are the first two components of a t-SNE dimensionality reduction with a perplexity of 50. The first picture includes features for both reactants and reagents, while the other two are only considering features for either reactants or reagents.

selectively include portions of the literature can improve the predictive performance (see the Supporting Information), but no consistent upgrade could be achieved. In what follows, we try to rationalize this and identify the most likely issue.

3.4. Diversity Mismatch. The first possible explanation for the poor generalizability of the database model is to question the featurization of the reactions and the model choice. However, we have tested three sets of different molecular features, including the graph description via neural networks. While we cannot rule out the existence of a better treatment of reactions from the ML aspect, it seems unlikely that the poor performance is due to this. Furthermore, we have tested our custom feature set with the amination data from Doyle and coworkers.[7] Despite the fact that our model has no capability to treat generic reagents (such as the additive used in that work), we achieve a similar performance (see the Supporting Information) to the published metrics from several groups.[20,22] While it would be interesting to see modeling approaches that are entirely different (such as SMILES-based transformers[23,24]), we do not believe that the choice of the ML model is the main reason for the poor performance we observe in Figure 3.

Instead, we see the main culprit in the data itself. The most striking evidence for this can be found in Figure 4a. The curves show how many solvents, bases, or ligands are needed to cover a certain fraction of reactions in the data. A steep curve indicates a stark bias where only very few reagents are needed to cover most reactions. In contrast, if the reagents would cover the data points uniformly, the line would be straight and less steep. We observe a strong mismatch between our database (first two columns) and the plate reactions (third column), the latter displaying a much better diversity. Since our plates were designed explicitly to mimic the industrial setting and with complete disregard to the database composition, we conclude that the reagent diversity of the databases is too biased to properly cover such a use-case. We propose this is a major obstacle for creating a general-purpose yield-prediction model, even though the amount of available reactions in databases seems large.

The mismatch is also clearly observed in the yield, which was the target variable in our case (see Figure 4b). In agreement with everyday experience in the lab, the plates mostly result in low-yielding reactions. On the other hand, the database is skewed to high-yielding reactions. We also note here that we tried oversampling techniques[42] to counter both the yield mismatch and the reagent diversity, but none of these improved the model performance (see the Supporting Information). We can also see that selecting only reactions performed on a high scale (second column in Figure 4) effectively has an even worse reagent diversity and yield skew. This is likely due to the lack of incentive to publish or even perform a low-yielding reaction on a high scale. While this, of course, is reasonable from an experimental viewpoint, it is a further obstacle to obtaining suitable reaction data for ML. In a similar recent study, Strieth-Kalthoff et al.[43] also found the lack of negative data (which are, in our case, low yielding reactions) to be an obstacle for predictive modeling.

Further evidence for the mismatch between the database and the plate reactions can be obtained via dimensionality reduction. In Figure 5 we visualize all chemical reactions from these sets in a two-dimensional space via t-distributed stochastic neighbor embedding (t-SNE).[44] The three panels are obtained by considering either the computed features of (i) all involved compounds, (ii) only the reactants, or (iii) only the reagents. While in the first panel, it seems that the plate reactions are rather intermixed with the database reactions, this is the opposite if only considering reactants or reagents. In particular, the reagent visualization (third panel) reveals that there are a few big clusters in the databases (which correspond to the few most commonly used conditions), and the plate conditions seem rather isolated from them. Recently, Beker et al.[45] found similarly that an ML model trained to predict the best reaction conditions for Suzuki−Miyaura couplings cannot outperform the baseline spanned by simply picking the most popular choices, which suggests the model could only learn the pre-induced reagent popularity of the training data. Together with the yield bias, we mainly attribute our models' generalization failure to this reagent bias.

There is also an aspect that was not investigated further by us: the databases contain reactions from thousands of different labs and experimenters. While the measurement of a yield is already a procedure that is rarely accurate to the third decimal place, this will introduce further noise on the yields. Take, for instance, the type of yield that was measured (isolated or from liquid chromatography−mass spectrometry), which was not distinguishable in the data fields (perhaps this is possible by analyzing reaction procedure texts with a lot of additional effort). We have worked with the assumption that these effects will likely induce an upper bound to the model performance. However, these "hidden" experimental parameters could have a much stronger impact than anticipated. Further work to quantify and remedy such error sources is needed.

4. CONCLUSIONS

In this work, we assessed whether the combined entries from three of the largest reaction data providers can be used to train a generalizable yield prediction model for C−N cross-couplings. With three disparate modeling approaches, we assessed the performance on different data splits, including by database, by patent-type, or by publication time. Considering the difficulty of the overall task of yield prediction, a promising performance was achieved (best ranking correlations around 0.75 and best R^2 around 0.65), which suggests creating general-purpose models for yield prediction is within reach.

However, putting the resulting models to the test on a dedicated experimental set of plate reactions created in our lab, the performance could not be transferred. We attribute this result to the lack of reagent and yield diversity as uncovered by our investigation. Improving the data situation for the approach chosen in this work is likely difficult. The data we used resulted from several decades of lab work. While we know that the amount of reactions added to the databases grows exponentially,[30] it is unlikely that any improvement to the reaction diversity is made quickly. The incentive to publish low-yielding or even failed reactions is generally low, and we hope the result of our study contributes to improve this. It will be interesting to see what advances promising initiatives such as the open reaction database[46] can bring in this regard.

Another important lesson from this work is that judging reaction yield prediction models can be extremely deceiving when only literature data is involved. Dedicated experimental test sets are rare, and subsequently, researchers in the field apply certain data splits to mimic realistic test sets. After the results displayed earlier, we believe this is an impossible task and the resulting test sets will always retain hidden biases from the literature data that are not present in typical lab experiments. It is very likely that this finding also applies to other fields where data comes from experiments with numerous lab-related parameters. We hope the plate data set provided with this work can be utilized by other researchers as a benchmark set to train or test their own approaches.

Overall, it seems that alternative data-driven approaches for the optimization of reaction conditions are favorable. While works which focused on smaller chemical spaces reported more favorable outcomes, general-purpose yield predictions from literature data seem unrealistic at the moment, even if these data are larger by orders of magnitude. Furthermore, HTE studies avoid the noise coming from experiments being performed in different environments by default. Therefore, it seems that labs that aim to integrate ML models need to gather their own data sets that are created with the required chemical diversity in mind. Recent reports by Xu et al.[47] and Rinehart et al.[48] also reinforce the viability of this route. The chemical and pharmaceutical industries would greatly benefit from pre-competitive collaboration where high quality and high-diversity reaction data are shared in an intellectual-property-preserving manner, such as federated learning.[49,50]

DATA AVAILABILITY

The reaction data used as the experimental test set in this work is provided in the Supporting Information. Our custom code for generating RDKit-inspired reaction-site features is available online at https://github.com/Scienfitz/RDKit-reactive-site-features. The code release associated with this paper has been uploaded to Zenodo https://doi.org/10.5281/zenodo.7296590. The reaction extension for attentivefp can be found at https://github.com/Scienfitz/attentivefp_reac with a Zenodo snapshot at https://doi.org/10.5281/zenodo.7296594.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.2c05546.

Reaction data for the experimental test set (XLSX)

Visualization of the literature data, additional metrics for the literature model, oversampling for yield prediction, comparison of custom featurization performances, and an active learning approach to combine plate and literature data (PDF)

AUTHOR INFORMATION

Corresponding Author

Georg Wuitschik − *Roche Pharma Research and Early Development, pCMC Process Research, Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd, CH-4070 Basel, Switzerland;* orcid.org/0000-0001-6682-1213; Email: georg.wuitschik@roche.com

Authors

Martin Fitzner − *Roche Pharma Research and Early Development, pRED Informatics, Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd, CH-4070 Basel, Switzerland; Merck KGaA, Science and Technology Office, Digital Chemistry, 64293 Darmstadt, Germany;* orcid.org/0000-0001-6790-4301

Raffael Koller − *Roche Pharma Research and Early Development, pCMC Process Research, Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd, CH-4070 Basel, Switzerland*

Jean-Michel Adam − *Roche Pharma Research and Early Development, pCMC Process Research, Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd, CH-4070 Basel, Switzerland*

Torsten Schindler − *Roche Pharma Research and Early Development, pRED Informatics, Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd, CH-4070 Basel, Switzerland*

Complete contact information is available at:
https://pubs.acs.org/10.1021/acsomega.2c05546

Author Contributions

J.-M.A., T.S., and M.F. conceived the research project. M.F. performed the data preparation, ML modeling, figure creation, and writing of the draft. G.W. performed the experiments. All authors contributed to the interpretation of the data.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank Jean-Louis Reymond for comments and suggestions on the manuscript and Elsevier and CAS for the help and data provided by them. In particular, the authors thank CAS for supplying the reaction data for this project. M.F. was funded by the Roche pRED Operations Advanced

Analytics Postdoctoral Fellowship Program, which is aligned with the pRED Postdoctoral Fellowship Program.

REFERENCES

(1) Williams, W. L.; Zeng, L.; Gensch, T.; Sigman, M. S.; Doyle, A. G.; Anslyn, E. V. The Evolution of Data-Driven Modeling in Organic Chemistry. *ACS Cent. Sci.* **2021**, *7*, 1622−1637.

(2) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547−555.

(3) Mater, A. C.; Coote, M. L. Deep learning in chemistry. *J. Chem. Inf. Model.* **2019**, *59*, 2545−2559.

(4) Goh, G. B.; Hodas, N. O.; Vishnu, A. Deep learning for computational chemistry. *J. Comput. Chem.* **2017**, *38*, 1291−1307.

(5) Segler, M. H.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555*, 604−610.

(6) Strieth-Kalthoff, F.; Sandfort, F.; Segler, M. H.; Glorius, F. Machine learning the ropes: principles, applications and directions in synthetic chemistry. *Chem. Soc. Rev.* **2020**, *49*, 6154−6168.

(7) 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.

(8) Segler, M. H.; Waller, M. P. Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chem.—Eur. J.* **2017**, *23*, 5966−5971.

(9) Ruiz-Castillo, P.; Buchwald, S. L. Applications of palladium-catalyzed C−N cross-coupling reactions. *Chem. Rev.* **2016**, *116*, 12564−12649.

(10) Gao, H.; Struble, T. J.; Coley, C. W.; Wang, Y.; Green, W. H.; Jensen, K. F. Using machine learning to predict suitable conditions for organic reactions. *ACS Cent. Sci.* **2018**, *4*, 1465−1476.

(11) Genheden, S.; Mårdh, A.; Lahti, G.; Engkvist, O.; Olsson, S.; Kogej, T. Prediction of the chemical context for Buchwald-Hartwig coupling reactions. *Mol. Inf.* **2021**, *41*, 2100294.

(12) Li, J.; Eastgate, M. D. Making better decisions during synthetic route design: leveraging prediction to achieve greenness-by-design. *React. Chem. Eng.* **2019**, *4*, 1595−1607.

(13) 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.

(14) Häse, F.; Roch, L. M.; Kreisbeck, C.; Aspuru-Guzik, A. Phoenics: a Bayesian optimizer for chemistry. *ACS Cent. Sci.* **2018**, *4*, 1134−1145.

(15) Häse, F.; Aldeghi, M.; Hickman, R. J.; Roch, L. M.; Aspuru-Guzik, A. Gryffin: An algorithm for Bayesian optimization of categorical variables informed by expert knowledge. *Appl. Phys. Rev.* **2021**, *8*, 031406.

(16) Dong, J.; Peng, L.; Yang, X.; Zhang, Z.; Zhang, P. XGBoost-based intelligence yield prediction and reaction factors analysis of amination reaction. *J. Comput. Chem.* **2022**, *43*, 289−302.

(17) Probst, D.; Schwaller, P.; Reymond, J.-L. Reaction Classification and Yield Prediction using the Differential Reaction Fingerprint DRFP. *Digital Discovery* **2021**, *1*, 91.

(18) Nielsen, M. K.; Ahneman, D. T.; Riera, O.; Doyle, A. G. Deoxyfluorination with sulfonyl fluorides: navigating reaction space with machine learning. *J. Am. Chem. Soc.* **2018**, *140*, 5004−5008.

(19) 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*, No. eaau5631.

(20) Sandfort, F.; Strieth-Kalthoff, F.; Kühnemund, M.; Beecks, C.; Glorius, F. A structure-based platform for predicting chemical reactivity. *Chem* **2020**, *6*, 1379−1390.

(21) Skoraczyński, G.; Dittwald, P.; Miasojedow, B.; Szymkuć, S.; Gajewska, E.; Grzybowski, B. A.; Gambin, A. Predicting the outcomes of organic reactions via machine learning: are current descriptors sufficient? *Sci. Rep.* **2017**, *7*, 3582.

(22) Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J.-L. Prediction of chemical reaction yields using deep learning. *Mach. Learn.: Sci. Technol.* **2021**, *2*, 015016.

(23) Schwaller, P.; Gaudin, T.; Lányi, D.; Bekas, C.; Laino, T. "Found in Translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chem. Sci.* **2018**, *9*, 6091−6098.

(24) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS Cent. Sci.* **2019**, *5*, 1572−1583.

(25) Kosugi, M.; Kameyama, M.; Migita, T. Palladium-catalyzed aromatic amination of aryl bromides with N, N-di-ethylamino-tributyltin. *Chem. Lett.* **1983**, *12*, 927−928.

(26) Guram, A. S.; Buchwald, S. L. Palladium-catalyzed aromatic aminations with in situ generated aminostannanes. *J. Am. Chem. Soc.* **1994**, *116*, 7901−7902.

(27) Paul, F.; Patt, J.; Hartwig, J. F. Palladium-catalyzed formation of carbon-nitrogen bonds. Reaction intermediates and catalyst improvements in the hetero cross-coupling of aryl halides and tin amides. *J. Am. Chem. Soc.* **1994**, *116*, 5969−5970.

(28) Guram, A. S.; Rennels, R. A.; Buchwald, S. L. A simple catalytic method for the conversion of aryl bromides to arylamines. *Angew. Chem., Int. Ed. Engl.* **1995**, *34*, 1348−1350.

(29) Louie, J.; Hartwig, J. F. Palladium-catalyzed synthesis of arylamines from aryl halides. Mechanistic studies lead to coupling in the absence of tin reagents. *Tetrahedron Lett.* **1995**, *36*, 3609−3612.

(30) 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*, 13085−13093.

(31) Reaxys. https://www.reaxys.com (accessed Jan 31, 2020).

(32) Copyright 2020 Elsevier Limited except certain content provided by third parties. Reaxys is a trademark of Elsevier Limited.

(33) Chemical Abstract Services. https://www.cas.org (accessed Jan 31, 2020).

(34) United States Patent and Trademark Office. https://www.uspto.gov (accessed Jan 31, 2020).

(35) Lowe, D. *Chemical reactions from US patents (1976-Sep2016)*. 2017.

(36) Xiong, Z.; Wang, D.; Liu, X.; Zhong, F.; Wan, X.; Li, X.; Li, Z.; Luo, X.; Chen, K.; Jiang, H.; et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *J. Med. Chem.* **2019**, *63*, 8749−8760.

(37) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: New York, NY, USA, 2016; pp 785−794.

(38) Landrum, G. *RDKit: A Software Suite for Cheminformatics, Computational Chemistry, and Predictive Modeling*. 2013.

(39) Diorazio, L. J.; Hose, D. R.; Adlington, N. K. Toward a more holistic framework for solvent selection. *Org. Process Res. Dev.* **2016**, *20*, 760−773.

(40) Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A. V.; Gulin, A. CatBoost: unbiased boosting with categorical features. **2017**, arXiv preprint arXiv:1706.09516.

(41) Dorogush, A. V.; Ershov, V.; Gulin, A. CatBoost: gradient boosting with categorical features support. **2018**, arXiv preprint arXiv:1810.11363.

(42) Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321−357.

(43) 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.* **2022**, *61*, No. e202204647.

(44) van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579−2605.

(45) 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.

(46) 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.

(47) Xu, J.; Kalyani, D.; Struble, T.; Dreher, S.; Krska, S.; Buchwald, S. L.; Jensen, K. F. Roadmap to Pharmaceutically Relevant Reactivity Models Leveraging High-Throughput Experimentation. **2022**, ChemRxiv:10.26434/chemrxiv-2022-x694w.

(48) Rinehart, N. I.; Saunthwal, R. K.; Wellauer, J.; Zahrt, A. F.; Schlemper, L.; Shved, A. S.; Bigler, R.; Fantasia, S.; Denmark, S. E. Development and Validation of a Chemoinformatic Workflow for Predicting Reaction Yield for Pd-Catalyzed C-N Couplings with Substrate Generalizability. **2022**, ChemRxiv:10.26434/chemrxiv-2022-hspwv.

(49) Li, T.; Sahu, A. K.; Talwalkar, A.; Smith, V. Federated learning: Challenges, methods, and future directions. *IEEE Signal Process. Mag.* **2020**, *37*, 50−60.

(50) Yang, Q.; Liu, Y.; Cheng, Y.; Kang, Y.; Chen, T.; Yu, H. Federated learning. *Synth. Lect. Artif. Intell. Mach. Learn.* **2019**, *13*, 1−207.