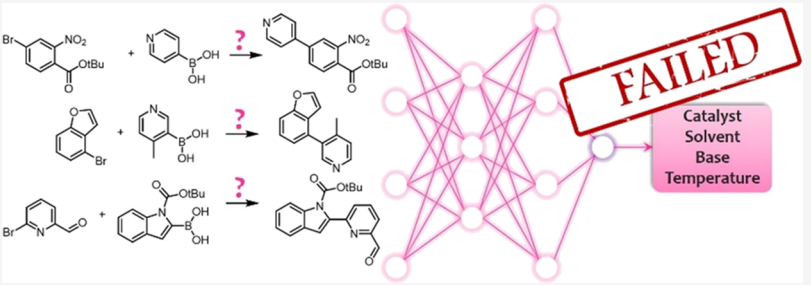
JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
pubs.acs.org/JACS
Article

Machine Learning May Sometimes Simply Capture Literature Popularity Trends: A Case Study of Heterocyclic Suzuki–Miyaura Coupling

Wiktor Beker, Rafał Roszak, Agnieszka Wołos, Nicholas H. Angello, Vandana Rathore, Martin D. Burke,* and Bartosz A. Grzybowski*

Cite This: J. Am. Chem. Soc. 2022, 144, 4819–4827
Read Online
ACCESS | Metrics & More | Article Recommendations | Supporting Information

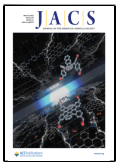


ABSTRACT: Applications of machine learning (ML) to synthetic chemistry rely on the assumption that large numbers of literature-reported examples should enable construction of accurate and predictive models of chemical reactivity. This paper demonstrates that abundance of carefully curated literature data may be insufficient for this purpose. Using an example of Suzuki–Miyaura coupling with heterocyclic building blocks—and a carefully selected database of >10,000 literature examples—we show that ML models cannot offer any meaningful predictions of optimum reaction conditions, even if the search space is restricted to only solvents and bases. This result holds irrespective of the ML model applied (from simple feed-forward to state-of-the-art graph-convolution neural networks) or the representation to describe the reaction partners (various fingerprints, chemical descriptors, latent representations, etc.). In all cases, the ML methods fail to perform significantly better than naive assignments based on the sheer frequency of certain reaction conditions reported in the literature. These unsatisfactory results likely reflect subjective preferences of various chemists to use certain protocols, other biasing factors as mundane as availability of certain solvents/reagents, and/or a lack of negative data. These findings highlight the likely importance of systematically generating reliable and standardized data sets for algorithm training.

INTRODUCTION

Machine learning (ML) is making an impact on many fields of research with remarkable successes in areas in which learning is based on well-defined rules (e.g., game theory[1,2]) or large and high-quality data sets (e.g., protein folding[3]). In contrast, when the data are of lesser quality and involve features not properly captured by ML models, the predictions can be less impactful.[4] This is also the case in chemistry where the limitations of data-driven AI are now being recognized.[5,6] On the one hand, when reaction data sets include sufficiently large numbers of mechanistically well-defined reactions, ML models have been able to predict reactivity patterns more accurately than either heuristic or even QM methods and, with physically meaningful descriptors, can extrapolate to compound classes outside of the training sets. For instance, we have demonstrated such ability in predicting regio-, site-, and diastereoselectivity patterns of Diels–Alder cycloadditions,[7] Hong and co-workers showed high fidelity of ML models in assessing radical C-H functionalizations of heterocycles,[8] whereas Seeberger and Gilmore[9] and separately Reymond[10] demonstrated highly accurate models of glycosylation stereoselectivity. On the other hand, when sometimes idiosyncratic human choices or hard-to-control variables come into play, ML methods fare significantly worse. One example is synthesis planning where ML methods have been limited to simple targets,[11,12] often suggest chemically implausible transformations,[13] and cannot emulate more far-sighted thinking of human experts over multiple steps (of note, such multistep "strategizing" has been successfully implemented in "hybrid" systems[14,15] combining knowledge-base and ML approaches, as recently demonstrated by

Received: November 14, 2021
Published: March 8, 2022



https://doi.org/10.1021/jacs.1c12005
J. Am. Chem. Soc. 2022, 144, 4819–4827

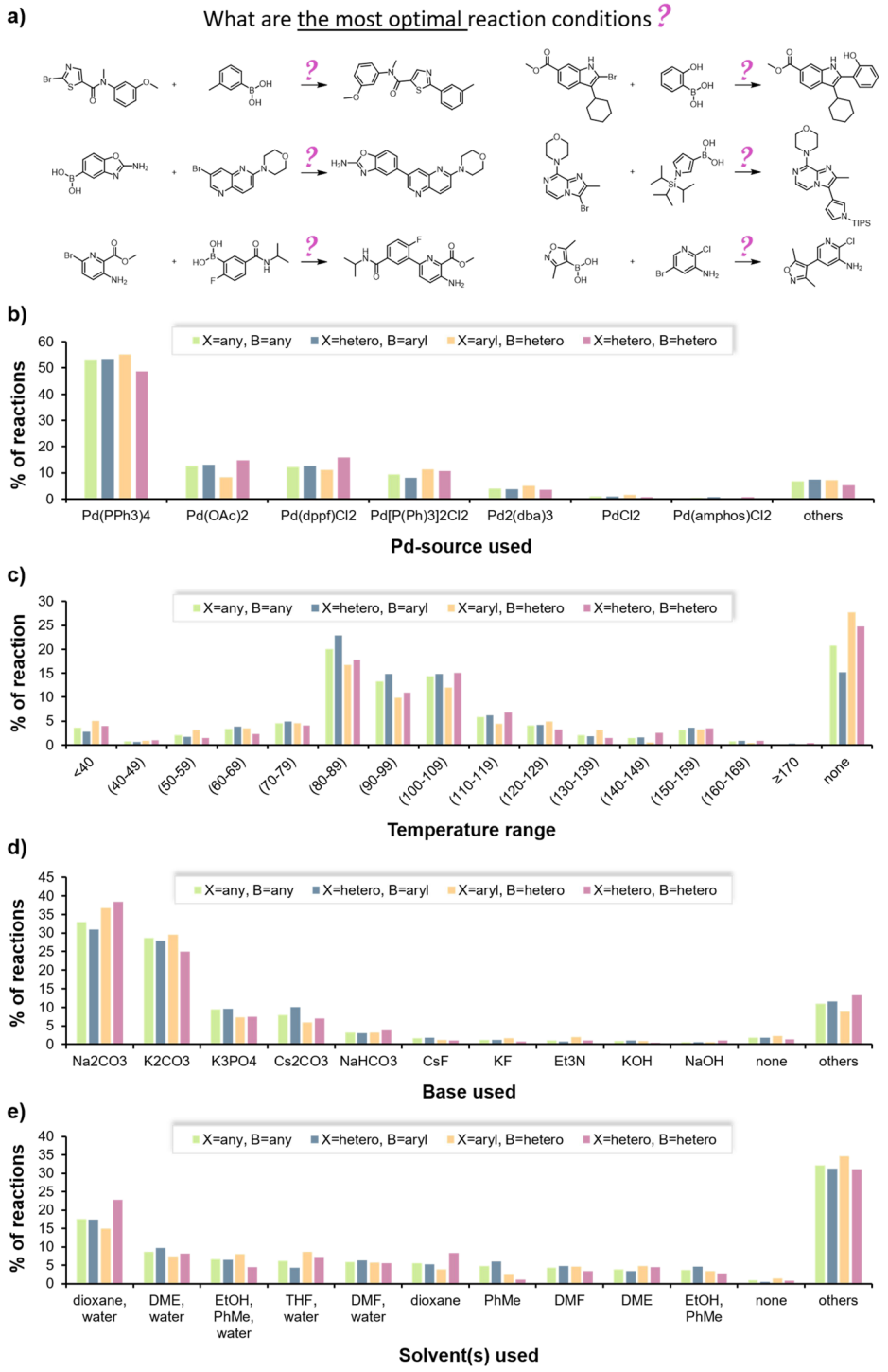


Figure 1. Formulation of the prediction problem and literature-based statistics of reaction conditions. (a) Data set of literature-reported reactions we consider comprises heteroaryl-heteroaryl and aryl-heteroaryl Suzuki couplings (additionally restricted to only bromides and boronic acids). The objective is to use AI models to predict "optimal" reaction conditions for a given pair of substrates. Literature-based statistics of (b) most common Pd sources used in heteroaromatic Suzuki couplings [>50% of all published reactions used $Pd(PPh_3)_4$ as a catalyst]; (c) reaction temperatures (almost 50% reactions were performed between 80 and 109 °C; ~20% of the records do not report temperature); (d) bases (five most common bases cover >80% of reaction space; additionally, carbonate bases were used in almost 70% of reactions); (e) solvents and solvent mixtures (five most common solvent mixtures cover only 45% of reaction space). Legends color-code the specific types of substrates used: "X = any"—any type of halide; "X = hetero"—heteroaromatic halide; "X = aryl"—aryl halide; "B = any"—any type of boronic acid; "B = hetero"—heteroaromatic boronic acid; "B = aryl"—aryl boronic acid.

machine-designed syntheses of drugs and complex natural products[16−18]). Another example is prediction of reaction yields for which data-driven methods perform poorly,[19] especially on diverse data sets; in this case, the limited predictability likely reflects the fact that yields can vary perceptibly depending on human or environmental factors, for example, chemist's skill, minute difference in manual procedures, or even time of the year (for yield variability in

https://doi.org/10.1021/jacs.1c12005
J. Am. Chem. Soc. 2022, 144, 4819−4827

reactions performed over the years for the same compounds by the same group of chemists, see the Supporting Information of ref 20).

Another important problem, tackled herein, deals with the prediction of optimal conditions for a particular reaction in which there are generally multiple viable choices of solvents or reagents. Several works[21−24] have attempted to use ML for the prediction of reaction conditions, and the overall message they seem to convey is that ML can, in fact, offer accurate predictions provided adequate numbers of literature examples on which to build the models (but see also critical ref 6). However, here, we demonstrate with a case study that this may have been an overoptimistic interpretation, and that even with large quantities of carefully curated literature data, ML approaches may not perform considerably better than estimates based on the popularity of reaction conditions reported in the literature. In other words, these ML models do not provide significantly more insights than just suggesting the most popular conditions which could be obtained by simple statistics over literature examples[25,26] and no "machine intelligence."

As a case in point, we consider the problem of predicting reaction conditions most suitable for a given pair of substrates engaging in heteroaryl-heteroaryl or aryl-heteroaryl Suzuki coupling. With >10,000 reaction examples with full condition information, this reaction seemed to provide reaction statistics that would be sufficient for successful ML. After categorizing the solvents, bases, temperatures, and sources of palladium, we apply various neural network (NN) approaches (feed-forward and graph convolution) as well as word-embedding and positive-unlabeled (PU) learning techniques to develop predictive models. Alas, all of these models offer only low accuracy of prediction, not significantly exceeding naïve baseline in which reaction conditions are assigned as those most popular in the literature. Moreover, the same and largely negative outcomes are observed when models described by others to predict reaction conditions are applied to this same data set—in all cases, they do not perform much better than the literature popularity measures.

Overall, the fact that numerous state-of-the-art ML approaches fail to identify a predictive link between the structures of substrates and most suitable reaction conditions suggests that such a link may be inaccessible based on published data alone.

The result reminds us that in synthetic chemistry, data are heavily influenced by nonscientific factors such as chemist's subjective preference for certain protocols or even current availability of chemicals in one's laboratory—there are no "descriptors" to capture these factors within ML models. We advocate that the path forward for chemical ML is to use robotized protocols[27,28] to generate standardized data sets and, in particular, multiple repeats of reactions carried out under different conditions, such that objective comparisons and learning of good vs bad conditions become possible.

RESULTS AND DISCUSSION

Reaction Data Set and Classes of Reaction Conditions. We considered Suzuki coupling[28−30] between heteroaryl-heteroaryl and heteroaryl-aryl partners (Figure 1a). These reactions were retrieved from Reaxys repository.[31] We excluded reactions not reporting yields, those in which no source of palladium was provided, and those coming from patents (which are not peer-reviewed). We have focused on the Reaxys data set because it has higher quality than the machine-extracted reaction set from patents (though we also provide analyses for patent reactions). The details of data curation are included in the Supplementary Information, Section S1. These procedures left a set of 16,748 reactions for which catalyst, base, and solvent were reported and 13,337 for which temperature was also given. A total of 1037 reactions had the same substrates and products but differed in reaction conditions used (after categorization into the classes detailed below, there were 511 such examples). The Reaxys reaction IDs for entire data set are provided at http://doi.org/10.5281/zenodo.4652819. Because these reactions use a variety of solvents and reagents, we first performed statistical analyses to categorize them into broader classes. Figure 1b shows that 92% of reactions use five sources of Pd, predominantly $Pd(PPh_3)_4$. In terms of reaction temperatures, the most popular ones are between 80 and 109 °C (Figure 1c). Regarding the bases, the five most popular ones cover 82% of cases, with carbonates being most widely used (Figure 1d). The least consensus seems to be in the use of solvents and solvent mixtures for which five most popular types account for only 45% of all reported reactions (Figure 1e). Based on these trends and additional analyses given in the Supporting Information (e.g., that counterions present in the bases have no systematic effect on reaction yields; see Figure S8), and in the effort to limit the space of parameters to predict, we focused on the prediction of solvents and bases.

Reflecting the statistics, the bases were categorized, according to popularity, as carbonates, phosphates, fluorides, hydroxides, amines, acetates, and other/miscellaneous. For solvents, we tested two types of categorizations. The more detailed one comprised 13 classes, ranked in the order of decreasing popularity as water/ethers, ethers, water/alcohols/aromatics, water/amides, alcohols/aromatics, aromatics, amides, water/aromatics, low boiling polar aprotic solvents/water, water/alcohols, water, alcohols, and other. The more "coarse-grained" classification distinguished six solvent types: {alcohols, water/polar solvents, water/alcohols, water/amides, water, amides}, {water/aromatics, alcohols/aromatics, water/alcohols/aromatics}, {aromatics}, {ethers}, {water/ethers}, {other}. In this way, we defined either $7 \times 13 = 91$ or $7 \times 6 = 42$ classes of reaction conditions, the latter less accurate but in principle easier to predict, should the finer classification prove challenging.

Models Based on Standard NNs. With these preliminaries, our main task was to develop ML models to predict which of the base/solvent class should be used for a given pair of substrates engaging in a Suzuki reaction. To make such predictions, we first used a standard feed-forward architecture with two hidden layers (130 and 15 neurons) with exponential linear unit (ELU) activation functions and softmax for the last layer. The NN had two outputs—one for the predicted base and another for solvent class. Each output gave a ranked list of, respectively, bases and solvents. Inputs were pairs of substrates for which we tested four representations:

(1) Morgan fingerprints with 512-bit length and radius 3;
(2) Chemical descriptors from the RDKit library[32] (200 descriptors for each substrate);
(3) Vectors combining the said Morgan fingerprints and RDKit descriptors;
(4) 20 dimensional latent/compressed representation obtained from an autoencoder (AE) comprising three

https://doi.org/10.1021/jacs.1c12005
J. Am. Chem. Soc. 2022, 144, 4819−4827

hidden layers (30, 20, and 30 neurons with rectified linear units (ReLU) or exponential activation functions) and using the Morgan fingerprint as input representation. The output of the second hidden layer was used as latent representation of substrates. Although this is not a representation *per se*, the pretraining of AE can help in removing unimportant (redundant) variables and regularizing the model, as well as providing a denser representation to the following classification layers.[33] The last feature is of particular importance for the fingerprint input, which is very sparse by its construction.

Each of these models was evaluated by fivefold cross-validation repeated five times (each time with a random 80:20 test/train split).

The results are summarized in Table 1a,b and give top-n accuracies for all models (i.e., probabilities that the base and solvent used in a particular literature-reported reaction would also be among the n-top predictions of the NN). One conclusion to make is that these accuracies do not vary perceptibly with the representation used. In addition, the accuracies are satisfactory for base prediction (which is heavily dominated by carbonates; see Figure 1e) but significantly less so for solvents, for which top-1 predictions are only ∼42−51% correct for simplified six-solvent categorization and ∼36−43% for 13 solvent classes. In fact, the accuracies are often on par with a very naïve "model" in which reactions are simply assigned the *n*-most popular bases or solvents (e.g., for top-1, each reaction is assigned carbonate as base and mixture of water and ether as solvent). This means that our NN models are not performing significantly better than a simple condition "popularity" baseline. As a side note, we observe that the AE model, while not providing better accuracy, provides additional regularization, as indicated by learning curves (Figures S15 and S16).

Table 1. Summary of Accuracies Obtained by Standard Feed-Forward Networks[a]

(a)

input	prediction accuracy of base (7 classes)			prediction accuracy of solvent (6 classes)		
	top-1	top-2	top-3	top-1	top-2	top-3
"popularity" baseline	76.8	89.6	93.8	29.8	57.4	75.5
Morgan fingerprint	80.6 (3.1)	91.0 (2.7)	94.4 (1.9)	51.7 (7.8)	69.4 (5.0)	81.2 (2.8)
RDKit descriptors	74.8 (2.2)	88.6 (1.9)	92.8 (1.6)	42.6 (5.4)	62.9 (4.4)	76.9 (4.3)
Morgan + descriptors	76.9 (3.3)	89.1 (2.1)	93.0 (1.9)	45.2 (7.3)	64.4 (6.0)	78.1 (4.4)
autoencoder	77.7 (2.7)	90.2 (1.6)	93.5 (1.3)	42.2 (5.5)	62.3 (3.7)	77.2 (2.3)

(b)

input	prediction accuracy of base (7 classes)			prediction accuracy of solvent (13 classes)		
	top-1	top-2	top-3	top-1	top-2	top-3
"popularity" baseline	76.8	89.6	93.8	29.7	41.4	52.6
Morgan fingerprint	79.8 (3.4)	90.1 (2.5)	94.1 (1.1)	43.3 (8.6)	57.4 (7.2)	67.0 (6.4)
RDKit descriptors	77.1 (3.8)	88.7 (2.1)	92.9 (1.7)	36.7 (7.0)	51.6 (6.1)	62.2 (5.3)
Morgan + descriptors	78.4 (3.3)	88.5 (2.6)	92.4 (2.4)	39.6 (7.7)	54.0 (7.3)	63.6 (6.1)
autoencoder	77.2 (3.1)	89.7 (1.5)	93.8 (1.5)	36.0 (4.5)	50.6 (4.5)	60.8 (4.1)

[a]Top-k accuracy metric is the probability (in %) of finding the actual class within top-k classes ordered according to model's predictions (values in parentheses are standard deviations from fivefold cross-validation). Part (a) is for the model taking into account six solvent classes. Part (b) is for 13 solvent classes.

Advanced NN Models. The failure of the abovementioned models could be reasonably ascribed to a simplistic NN architecture. Accordingly, we examined the performance of state-of-the-art graph convolutional neural networks (GCNNs)[34] and statistical correction proposed by Elkan and co-workers[35,36] (we apply this correction to the NN classifier, denoting it PU-NN model). GCNNs process learn directly from molecular graphs rather than from a predefined set of substructures or descriptors and have been successfully applied to predict, for instance, pKa values of C−H acids[37] and other molecular properties.[38,39] Statistical correction present in PU-NN, on the other hand, aims to solve the so-called PU problem. In our case, PU means that for particular substrates, the fact that certain reaction conditions were not reported does not mean that they were unsuitable for the reaction ("negative") but only that they were untested. In other words, the reaction might still be feasible under unreported conditions and, at best, it can be assumed that the literature-reported conditions are *close* to optimal. In technical terms, this means that we now face a *multilabel* rather than a *multiclass* binary classification problem. This means that for each pair of substrates, all possible solvents/bases have their own 0/1 labels assigned independently, and more than one solvent/base can be deemed suitable for the reaction (in contrast, in a multiclass classification analyzed previously, each substrate pair could be assigned *only to one* out of many solvent/base classes). The prediction of "the best" solvent/base for a given pair of substrates is then performed by choosing the class with the highest prediction probability (a probability that the class "matches").

For testing of these two architectures, we focused on the problem of solvent selection (for six coarse-grained solvent classes) for which literature-based distribution is less dominated by a single class than in the case of bases (see Figure 1d,e), and which has proven more problematic for feed-forward NNs (see Table 1). The relevant entries in Table 2 indicate that the top-1 accuracies are, again, below 50% and the top-3 ones are not much better than the naïve, popularity-based baseline.

For the completeness of comparisons, we also tested a feed-forward architecture with substrate fingerprints (as before) but with multilabel instead of multiclass classification. Furthermore, we explored two modifications to this model's input: (i) addition of the base class (to verify if solvent is in any way correlated to the base); and (ii) Mol2Vec representation of fingerprints. The Mol2Vec technique is inspired by language processing and casts the fingerprints into a 300-dimensional space, whereby the mutual proximity of points is expected to reflect the "chemical" similarity between compounds (the construction of such a space itself is based on statistical properties inferred from a large "corpus" such as the ZINC data set[40]). Unfortunately, the "feed-forward" entries in Table 2 evidence that none of these models improved the accuracy of prediction perceptibly.

https://doi.org/10.1021/jacs.1c12005
J. Am. Chem. Soc. 2022, 144, 4819−4827

Table 2. Coarse-Grained Solvent Classification by Advanced NN Models[a]

model architecture	input	top-1	top-2	top-3
"popularity"-based baseline		29.2	53.8	73.1
GCNN	molecular graph of substrates	40.6 (6.3)	61.0 (5.2)	74.7 (3.4)
PU-NN	ECFP6 of substrates	42.1 (6.1)	60.9 (4.6)	74.0 (2.5)
feed-forward	ECFP6 of substrates	45.8 (6.5)	63.5 (5.5)	75.9 (4.1)
feed-forward	ECPF6 of substrates + base class	46.4 (5.6)	64.2 (5.1)	76.6 (5.3)
feed-forward	Mol2Vec[42] embedding of substrates	34.9 (3.9)	54.9 (3.1)	70.1 (2.7)

[a]GCNN: Graph convolutional neural network.[34] PU-NN: NN classifier with PU correction.[35,36] ECFP6: Extended connectivity fingerprints with diameter 6.[41] Top-*k* accuracy metric is the probability (in %) of finding the actual class within top-*k* classes ordered according to model's predictions (values in parentheses are standard deviations from fivefold cross-validation). The baseline values refer to ordering produced by the corresponding frequency in the literature. Note that to mitigate class imbalance, all models used sample weights inversely proportional to class frequency (e.g., if a given solvent class was rarely used in the literature, the error of corresponding "matching" examples was multiplied according to the class size. This adjustment is meant to consider large and small classes on equal footing, without size-induced bias).

Augmenting Models with the Information about Yields. In a further effort to improve prediction accuracies, we decided to take advantage of the yield information, which, in principle, should help the models to distinguish between good and bad reaction conditions with greater precision. The logic here is to first teach the AI models to predict reaction yields for all possible reaction conditions and then, for a given pair of substrates, select as optimal those conditions that correspond to maximal yield. We began by training a regressor having a general feed-forward architecture. The inputs were vectors concatenating 512-bit Morgan fingerprints of the substrates, temperature (°C), as well as several vectorized forms of reaction conditions, either (i) one-hot encoded classes of solvents and bases or (ii) the so-called "learnable embedding"—a technique from natural-language processing—of conditions, in which ligands, solvents/solvent mixtures, and bases were first tokenized and then transformed into multidimensional vectors (for details, see caption to Table 3). In all cases, the NN had two hidden layers (40 and 10 neurons), and activation functions for the layers were ELU, linear, and ReLU.

Results summarized in Table 3 demonstrate that irrespective of the vectorization scheme used, the mean absolute errors (MAEs) of yield prediction were similar, around 16%. Also similar were the predicted spreads of the yields of reactions performed under different conditions; however, the "best" and "worst" conditions were predicted to vary by 5−10%, which is much lower than ∼20−30% observed in experiments. This finding means that our regressors are largely insensitive to reaction conditions. In this light, it is not surprising that the top-*k* values, that is, conditions' assignments based on the prediction of the highest-yielding, second-highest-yielding, and so forth reactions, are very poor. Significantly, these predictions are again worse than a frequency-based baseline (even if the model is additionally penalized for incorrect predictions of the same substrates in different conditions; see Section S3).

Table 3. Accuracy of Yield Prediction Using Feed-Forward Neural Networks with Different Input Representations[a]

input data	loss	MAE	top-1	top-2	top-3	Mdiff
popularity-based baseline		16.3	25.1	44.7	59.4	
fine classes	MSD	16.2 (2.3)	0.8 (0.4)	0.9 (0.4)	1.1 (0.6)	9.4
fine classes (with ligand)	MSD	16.0 (1.9)	1.5 (0.7)	1.8 (0.9)	2.5 (1.7)	6.1
"coarse-grained" classes	MSD	16.3 (2.2)	0.6 (0.7)	0.8 (0.7)	1.1 (0.8)	6.1
"coarse-grained" (with ligand)	MSD	15.6 (2.0)	1.0 (0.8)	1.8 (1.6)	3.1 (2.8)	4.6
embedded conditions	MSD	16.3 (2.7)				
embedded coarse-grained classes	MSD	16.6 (2.4)	7.6 (11.7)	12.9 (12.4)	14.7 (11.3)	5.4
classifier			37.0	48.8	56.9	

[a]MAE = mean absolute error; top-*k* values as in Tables 1 and 2 in %; values in parentheses are standard deviations from fivefold cross-validation; Mdiff—mean difference between conditions predicted to be the best and the worst for particular coupling partners. Popularity baseline is defined according to most popular literature-reported conditions (though, unlike in Table 2, here both base and solvent are considered). The last entry labeled as "classifier" refers to the combined predictions of separate base and solvent classifiers based on fingerprint representation. "Learnable embedding" was performed separately for each of three components (ligand, solvent, and base). Tokenization took place before NN training and involved selection of top-X (54 solvents, 72 bases, and 81 ligands) most frequent entries in the literature data, and they were assigned a number (index in the model's "dictionary")—usually one of those numbers covered all less significant, null, or unknown entries. Bases, ligands, and solvents were each assigned single tokens, whereas solvent mixtures, up to four components, were represented by tuples of four tokens representing pure solvents (and ordered according to predominance in mixture and with null/zero tokens used to denote "missing" solvents in binary and tertiary mixtures). The embedding layer in the NN kept a "dictionary" translating each token into a D-dimensional vector, whose components were optimized during training. Here, each token was assigned a 3D vector, resulting in a 24D representation of reaction conditions (a concatenation of two 3D vectors for ligand and base, as well as four 3D vectors for solvent components).

In order to compare those results with the aforementioned classification approach, we trained separate fingerprint-based models for base and solvent classification (in multilabel formulation) and used them to predict the best conditions. The condition class probabilities (required to sort the predictions from best to worst) were taken as a product of corresponding base/solvent class probabilities. This model, as can be seen in the last row of Table 3, exceeded the naïve baseline in top-1 and top-2 accuracies but was still largely unsatisfactory (e.g., top-1 < 40%).

Last but not least, we consider a model extreme in its naiveté—namely, assigning average yield (77%) independently on the input substrates and conditions. Such baseline has an MAE of 16.3%—comparable with even the best regression models (15.6%), especially when standard deviation from cross-validation (typically ∼2%) is taken into account. Yet again, this means that the AI models do not offer any major advantages over simplistic measures based on literature statistics.

https://doi.org/10.1021/jacs.1c12005
J. Am. Chem. Soc. 2022, 144, 4819−4827

Table 4. Accuracy of Condition Prediction Using Previously Reported Models[a]

task type	data source	Reaxys				USPTO			
	input data metric	top-1	top-2	top-3	MAE	top-1	top-2	top-3	MAE
	popularity-based baseline	25.1	44.7	59.4	16.3	29.8	51.8	62.7	21.1
classification	reaction conditions recommender[22]	38.7	46.1	50.7		26.4	31.0	34.0	
classification	Rel-GAT[42]	39.6	53.6	62.6		46.3	60.9	70.6	
regression	yield-BERT[43]	13.3	14.1	14.7	14.1	5.6	8.0	10.9	19.2

[a]Top-k values as in Tables 1−3 in %. Popularity baseline is defined according to most popular literature-reported conditions (though, unlike in Table 2, here, both base and solvent are considered).

Predictions of Previously Described Models. As mentioned in the Introduction section, several prior works reported ML models as relatively accurate in predicting reaction conditions. We tested performance of three such state-of-the-art approaches applied to the Suzuki coupling problem: Reaction Conditions Recommender (RCR) developed by Gao et al.,[22] Yield-BERT predicting reaction yields based on SMILES-represented reaction and associated reaction conditions,[43] and Rel-GAT (Relational Graph Attention Neural Network) previously evaluated on Suzuki and several other coupling reactions.[24]

The RCR was used with the NN parameters provided by the authors as trained on the entire Reaxys data set (i.e., encompassing our own data set). For each reaction, we collected top-10 recommendations and translated them into our coarse-grained solvent and base classes. We note that palladium catalyst—usually $Pd(PPh_3)_4$—was present in 82.3% of the top-1 recommendations (and 94.1% of all top-10 proposals), indicating that the model correctly recognized Suzuki coupling reaction. On the other hand, for the solvent and base prediction problem, RCR did comparably to our own classifiers and the popularity baseline (RCR's top-1, 2, and 3 scores were 38.7, 46.1, and 50.7%, respectively).

Regarding the Yield-BERT[43] model, we re-trained and tested (5 × CV) it on our data set using the same hyperparameters as the authors (attempts to optimize those hyperparameters did not improve the model; see Section S2.4). The model was originally trained on a significantly smaller and less diverse data set of Suzuki couplings (5760 reactions from ref 44 differing in halide and boronate substrates as well as reaction conditions but all yielding the same product) and achieved MAE of 8.1%. On our larger and more diverse set, the MAE was 14%, that is, only slightly better than our simple regressor. Importantly, when the model was used to score different reaction conditions, the top-1 accuracy was 13.3% which is again better than our regressor (7.6%) but well below the literature popularity baseline (25.1%, see Table 4).

On the other extreme, the Rel-GAT[24] was originally evaluated on a significantly broader data set, that is, various types of couplings and, within the Suzuki coupling, on all such examples (i.e., not only the more synthetically challenging[45,46] aryl-heteroaryl couplings but also aryl-aryl). Here, by introducing chemically relevant classes of solvents and bases (instead of explicit classification used in ref 43), we create a more difficult classification problem (consider, for instance, that out of four most popular bases, three are carbonates), especially when class imbalance is taken into account (see Figure S15 in Section S5.5). In Rel-GAT, this imbalance problem was not addressed, whereas we applied sample weights to address this issue; we note that in a recent study on toxicity prediction, this technique turned out to outperform other approaches to balance the data set.[47] Still, even with these precautions, the model performed similarly to our GCNN discussed earlier and achieved the top-1, 2, and 3 accuracies of condition prediction of, respectively, 39.6, 53.6, and 62.6%, (with standard deviations of the mean ∼4%; see Table 4 and further details in Table S13).

Next, we investigated whether our results were in any way peculiar to the Reaxys data set. To this end, the experiments described above were repeated using 5434 reactions from the USPTO[48] collection and deposited at https://github.com/rmrmg/SuzukiConditions/blob/master/uspto/dataset/suzuki_USPTO_with_hetearomatic.txt (for details of the extraction procedure, see Section S1.2). The results summarized in the right portion of Table 4 evidence that in terms of top-1,2,3 metrics, all tested models offer accuracies comparable to the Reaxys data set. The only model that outperforms the popularity baseline is the rel-GAT classifier. However, it should be noted that the popularity baseline is higher for USPTO than for Reaxys—this effect can be explained by the lower diversity of condition classes in USPTO (see Section S5.5 and Figure S18) with regard vs Reaxys (Figure S15). Furthermore, it has to be stressed that automatically curated reactions in USPTO are generally of lower quality (see Section S1.3). Indeed, careful evaluation of randomly sampled 50 entries from each database revealed that as much as half of the USPTO entries may be compromised, whereas in the case of Reaxys, this estimate is at the level of ca. 10% (this is also in line with our recent estimates, spanning all reaction types, of erroneous entries in these repositories[49]). The errors in USPTO records are particularly evident in the case of solvent entries (see Section S1.3.1 and Table S2), plausibly because this collection is dominated by the "other" solvent class (Figure S18), causing corrupted solvent entries to fall out of our classification criteria. With this evidence, the real-world performance of any model trained on such low-quality data, even with the highest possible values of performance metrics, is at least questionable. Further analyses of the NN models trained on USPTO are provided in Sections S8 and S9.

Finally, to better understand why even the state-of-the-art ML models offer such limited accuracies, we compared their performance on the pairs of reactions involving the same substrates under different reaction conditions. To this end, we selected all 316 pairs of reactions from Reaxys that satisfy the following criteria: (a) they use the same pair of substrates; (b) they were performed under different conditions (according to our classification of solvents and bases); and (c) the difference of their yields is greater than 10%. If more than two reactions met these conditions, we chose a pair that maximizes the yield difference. Our expectation here was that on such pairs, a good model should correctly recognize which conditions are "better" (i.e., provide higher yield) for a given reaction. The simplistic popularity metric orders these pairs correctly in 43.7% cases.

https://doi.org/10.1021/jacs.1c12005
J. Am. Chem. Soc. 2022, 144, 4819−4827

This is on par with RCR (35.4%), Yield-BERT (51.3%), and Rel-GAT (47.0%). This result suggests that both of these ML models are relatively insensitive to reaction conditions and capture only some crude correlations between the structure of the reactants and the preferred reaction conditions. Some additional analyses are provided in Sections S5.4 and S5.5.

CONCLUSIONS

In summary, we applied a range of ML techniques, from simple to state-of-the-art, to answer a seemingly simple question—that is, which reaction conditions should be chosen for substrates engaging in a reaction of a particular type. Even though we used a large, diverse, and carefully curated data set of Suzuki–Miyaura couplings, all of these models gave largely unsatisfactory prediction accuracies (especially for the solvent prediction subproblem), not significantly higher than the popularity baseline. At first sight, this may be surprising given that ML models are expected to offer accurate predictions if trained on large enough, high-quality reaction data sets. This might be true when the descriptors used to construct the model capture the chemical essence of a problem in question, for example, when trying to predict reaction outcomes given its substrates, the structural, steric, and electronic descriptors are generally sufficient.[7−10,50] The condition prediction problem, however, is markedly different because in addition to the structural features of reactants, products, and reagents, it entails several "human" factors: Conditions are often chosen based on the query of relevant literature, ultimately selecting those most frequently reported (this may explain why popularity-based metrics worked nearly as well as ML). In addition, mundane factors of instantaneous availability of specific reagents/solvents in one's laboratory or even "historical" preference for certain choices (i.e., conditions commonly used in one's laboratory) might come into play. In other words, chemistry often propagates its own practices/routines, and these factors are hardly quantifiable as "descriptors" of sorts. A way around this problem is to begin to augment the available literature data by systematic and standardized experiments in which reactions are repeated under multiple conditions such that meaningful conclusions about better vs worse ones can be learned. Several years ago, such augmentation of thousands upon thousands of reactions would hardly be possible as human chemists lack incentives to repeat—just for the sake of generating more data—a successful reaction under multiple other and likely worse yielding conditions. The recent progress in synthesis automation, however, should make such an effort feasible, at least for some more popular classes of reactions. Until such multiple-condition data become available, we advocate that ML models are always accompanied by and compared against popularity-based baselines which are known, by themselves, to capture certain reactivity trends.[25,26]

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacs.1c12005.

Experimental details, additional details of neural networks, hyperparameters selection, custom loss function, and further statistical analyses (PDF)

AUTHOR INFORMATION

Corresponding Authors

Martin D. Burke − *Department of Chemistry and Department of Biochemistry, Institute for Genomic Biology, Carle Illinois College of Medicine, and Beckman Institute, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, United States;* orcid.org/0000-0001-7963-7140; Email: mdburke@illinois.edu

Bartosz A. Grzybowski − *Allchemy, Inc., Highland, Indiana 46322, United States; Institute of Organic Chemistry, Polish Academy of Sciences, Warsaw 01-224, Poland; Center for Soft and Living Matter, Institute for Basic Science (IBS), Ulsan 44919, Republic of Korea; Department of Chemistry, Ulsan Institute of Science and Technology (UNIST), Ulsan 44919, Republic of Korea;* orcid.org/0000-0001-6613-4261; Email: nanogrzybowski@gmail.com

Authors

Wiktor Beker − *Allchemy, Inc., Highland, Indiana 46322, United States; Institute of Organic Chemistry, Polish Academy of Sciences, Warsaw 01-224, Poland*

Rafał Roszak − *Allchemy, Inc., Highland, Indiana 46322, United States; Institute of Organic Chemistry, Polish Academy of Sciences, Warsaw 01-224, Poland*

Agnieszka Wołos − *Allchemy, Inc., Highland, Indiana 46322, United States; Institute of Organic Chemistry, Polish Academy of Sciences, Warsaw 01-224, Poland*

Nicholas H. Angello − *Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, United States;* orcid.org/0000-0001-6436-3669

Vandana Rathore − *Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/jacs.1c12005

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the Defense Advanced Research Projects Agency under the Accelerated Molecular Discovery Program (Cooperative Agreement No. HR00111920027 dated August 1, 2019). The content of the information presented in this work does not necessarily reflect the position or the policy of the Government.

REFERENCES

(1) Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; Chen, Y. T.; Lillicrap, T.; Hui, F.; Sifre, L.; van den Driessche, G.; Graepel, T.; Hassabis, D. Mastering the Game of Go without Human Knowledge. *Nature* **2017**, *550*, 354−359.

(2) Silver, D.; Hubert, T.; Schrittwieser, J.; Antonoglou, I.; Lai, M.; Guez, A.; Lanctot, M.; Sifre, L.; Kumaran, D.; Graepel, T.; Lillicrap, T.; Simonyan, K.; Hassabis, D. A General Reinforcement Learning Algorithm that Masters Chess, Shogi, and Go Through Self-Play. *Science* **2018**, *362*, 1140−1144.

(3) Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C. L.; Zidek, A.; Nelson, A. W. R.; Bridgland, A.; Penedones, H.; Petersen, S.; Simonyan, K.; Crossan, S.; Kohli, P.; Jones, D. T.; Silver, D.; Kavukcuoglu, K.; Hassabis, D. Improved Protein Structure Prediction Using Potentials from Deep Learning. *Nature* **2020**, *577*, 706−710.

(4) Roberts, M.; Driggs, D.; Thorpe, M.; Gilbey, J.; Yeung, M.; Ursprung, S.; Aviles-Rivero, A. I.; Etmann, C.; McCague, C.; Beer, L.; Weir-McCall, J. R.; Teng, Z.; Gkrania-Klotsas, E. A.-C.; Rudd, J. H. F.; Sala, E.; Schönlieb, C. B. Common Pitfalls and Recommendations for Using Machine Learning to Detect and Prognosticate for COVID-19 Using Chest Radiographs and CT Scans. *Nat. Mach. Intell.* **2021**, *3*, 199−217.
(5) Artrith, N.; Butler, K. T.; Coudert, F. X.; Han, S.; Isayev, O.; Jain, A.; Walsh, A. Best Practices in Machine Learning for Chemistry. *Nat. Chem.* **2021**, *13*, 505−508.
(6) Chuang, K. V.; Keiser, M. J. Comment on "Predicting Reaction Performance in C−N Cross-Coupling Using Machine Learning". *Science* **2018**, *362*, No. eaat8603.
(7) Beker, W.; Gajewska, E. P.; Badowski, T.; Grzybowski, B. A. Prediction of Major Regio-, Site-, and Diastereoisomers in Diels-Alder Reactions by Using Machine-Learning: The Importance of Physically Meaningful Descriptors. *Angew. Chem., Int. Ed.* **2019**, *58*, 4515−4519.
(8) Li, X.; Zhang, S.; Xu, L.; Hong, X. Predicting Regioselectivity in Radical C−H Functionalization of Heterocycles through Machine Learning. *Angew. Chem., Int. Ed.* **2020**, *59*, 13253−13259.
(9) Moon, S.; Chatterjee, S.; Seeberger, P. H.; Gilmore, K. Predicting glycosylation stereoselectivity using machine learning. *Chem. Sci.* **2021**, *12*, 2931−2939.
(10) Pesciullesi, G.; Schwaller, P.; Laino, T.; Reymond, J.-L. Transfer learning enables the molecular transformer to predict regio- and stereoselective reactions on carbohydrates. *Nat. Commun.* **2020**, *11*, 3878.
(11) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555*, 604−610.
(12) Coley, C. W.; Thomas, D. A.; Lummiss, J. A. M.; Jaworski, J. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H.; Hicklin, R. W.; Plehiers, P. P.; Byington, J.; Piotti, J. S.; Green, W. H.; Hart, A. J.; Jamison, T. F.; Jensen, K. F. A Robotic Platform for Flow Synthesis of Organic Compounds Informed by AI Planning. *Science* **2019**, *365*, No. eaax1566.
(13) Borrelli, W.; Schrier, J. Evaluating the Performance of a Transformer-based Organic Reaction Prediction Model. *ChemRxiv.* **2021**, 3nqv9.
(14) Badowski, T.; Gajewska, E. P.; Molga, K.; Grzybowski, B. A. Synergy Between Expert and Machine-Learning Approaches Allows for Improved Retrosynthetic Planning. *Angew. Chem., Int. Ed.* **2020**, *59*, 725−730.
(15) Molga, K.; Szymkuć, S.; Grzybowski, B. A. Chemist ex Machina: Advanced Synthesis Planning by Computers. *Acc. Chem. Res.* **2021**, *54*, 1094−1106.
(16) Szymkuć, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A. Computer-Assisted Synthetic Planning: The End of the Beginning. *Angew. Chem., Int. Ed.* **2016**, *55*, 5904−5937.
(17) Klucznik, T.; Mikulak-Klucznik, B.; McCormack, M. P.; Lima, H.; Szymkuć, S.; Bhowmick, M.; Molga, K.; Zhou, Y.; Rickershauser, L.; Gajewska, E. P.; Toutchkine, A.; Dittwald, P.; Startek, M. P.; Kirkovits, G. J.; Roszak, R.; Adamski, A.; Sieredzińska, B.; Mrksich, M.; Trice, S. L. J.; Grzybowski, B. A. Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory. *Chem* **2018**, *4*, 522−532.
(18) Mikulak-Klucznik, B.; Gołębiowska, P.; Bayly, A. A.; Popik, O.; Klucznik, T.; Szymkuć, S.; Gajewska, E. P.; Dittwald, P.; Staszewska-Krajewska, O.; Beker, W.; Badowski, T.; Scheidt, K. A.; Molga, K.; Młynarski, J.; Mrksich, M.; Grzybowski, B. A. Computational Planning of the Synthesis of Complex Natural Products. *Nature* **2020**, *588*, 83−88.
(19) Skoraczyński, G.; Dittwald, P.; Miasojedow, B.; Szymkuć, S.; Gajewska, E. P.; Grzybowski, B. A.; Gambin, A. Predicting the Outcomes of Organic Reactions via Machine Learning: Are Current Descriptors Sufficient? *Sci. Rep.* **2017**, *7*, 3582.
(20) Emami, F. S.; Vahid, A.; Wylie, E. K.; Szymkuć, S.; Dittwald, P.; Molga, K.; Grzybowski, B. A. A Priori Estimation of Organic Reaction Yields. *Angew. Chem., Int. Ed.* **2015**, *54*, 10797−10801.
(21) 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.
(22) 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.
(23) Zuranski, A. M.; Alvarado, J. I. M.; Sjields, B. J.; Doyle, A. G. Predicting Reaction Yields via Supervised Learning. *Acc. Chem. Res.* **2021**, *54*, 1856−1865.
(24) Maser, M. R.; Cui, A. Y.; Ryou, S.; DeLano, T. J.; Yue, Y.; Reisman, S. E. Multilabel Classification Models for the Prediction of Cross-Coupling Reaction Conditions. *J. Chem. Inf. Model.* **2021**, *61*, 156−166.
(25) Kowalczyk, B.; Bishop, K. J. M.; Smoukov, S. K.; Grzybowski, B. A. Synthetic Popularity Reflects Chemical Reactivity. *J. Phys. Org. Chem.* **2009**, *22*, 897−902.
(26) Soh, S.; Wei, Y.; Kowalczyk, B.; Gothard, C. M.; Baytekin, B.; Gothard, N.; Grzybowski, B. A. Estimating Chemical Reactivity and Cross-Influence from Collective Chemical Knowledge. *Chem. Sci.* **2012**, *3*, 1497−1502.
(27) Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Martinez Alvaro, J. I.; Janey, J. M.; Adams, R. P.; Doyle, A. G. Bayesian Reaction Optimization as a tool for chemical synthesis. *Nature* **2021**, *590*, 89−96.
(28) Li, J.; Ballmer, S. G.; Gillis, E. P.; Fuji, S.; Schmidt, M. J.; Palazzolo, A. M. E.; Lehmann, J. W.; Morehouse, G. F.; Burke, M. D. Synthesis of many different types of organic small molecules using one automated process. *Science* **2015**, *347*, 1221−1226.
(29) Kinzel, T.; Zhang, Y.; Buchwald, S. L. A new Palladium precatalyst allows for the fast Suzuki−Miyaura coupling reactions of unstable polyfluorophenyl and 2-heteroaryl boronic acids. *J. Am. Chem. Soc.* **2010**, *132*, 14073−14075.
(30) Lennox, A. J. J.; Lloyd-Jones, G. C. Selection of boron reagents for Suzuki−Miyaura coupling. *Chem. Soc. Rev.* **2014**, *43*, 412−443.
(31) Reaxys.com. 2021. Reaxys. [online] Available at: <https://www.reaxys.com/> (Accessed March 2, 2021).
(32) RDKit: Open-source cheminformatics Available at: <http://www.rdkit.org> (Accessed March 2, 2021).
(33) Erhan, D.; Bengio, Y.; Courville, A.; Manzagol, P.-A.; Vincent, P.; Bengio, S. J. Why Does Unsupervised Pre-training Help Deep Learning? *Mach. Learn. Res.* **2010**, *11*, 625−660.
(34) Defferrard, M.; Bresson, X.; Vandergheynst, P. *Advances in Neural Information Processing Systems*; 30th Annual Conference on Neural Information Processing Systems; The MIT Press, 2016
(35) Li, W.; Guo, Q.; Elkan, C. A Positive and Unlabeled Learning Algorithm for One-Class Classification of Remote-Sensing Data. *IEEE Trans. Geosci. Remote Sens.* **2011**, *49*, 717−725.
(36) PuLearn: Positive-unlabeled learning with Python. Available at: <https://pulearn.github.io/pulearn> (Accessed July 8, 2021).
(37) Roszak, R.; Beker, W.; Molga, K.; Grzybowski, B. A. Rapid and Accurate Prediction of pKa Values of C−H Acids Using Graph Convolutional Neural Networks. *J. Am. Chem. Soc.* **2019**, *141*, 17142−17149.
(38) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370−3388.
(39) Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.* **2018**, *10*, 370−377.
(40) Sterling, T.; Irwin, J. J. ZINC 15—Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324−2337.
(41) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

https://doi.org/10.1021/jacs.1c12005
J. Am. Chem. Soc. 2022, 144, 4819−4827

(42) Jaeger, S.; Fulle, S.; Turk, S. Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. *J. Chem. Inf. Model.* **2018**, *58*, 27−35.
(43) 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.
(44) Perera, D.; Tucker, J. W.; Brahmbhatt, S.; Helal, C. J.; Chong, A.; Farrell, W.; Richardson, P.; Sach, N. W. A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. *Science* **2018**, *26*, 429−434.
(45) Kudo, N.; Perseghini, M.; Fu, G. C. A Versatile Method for Suzuki Cross-Coupling Reactions of Nitrogen Heterocycles. *Angew. Chem., Int. Ed.* **2006**, *45*, 1282−1284.
(46) Bhaskaran, S.; Padusha, M. S. A.; Sajith, A. M. Application of Palladium Based Precatalytic Systems in the Suzuki-Miyaura Cross-Coupling Reactions of Chloro-Heterocycles. *ChemistrySelect* **2020**, *5*, 9005.
(47) Bae, S. Y.; Lee, J.; Jeong, J.; Lim, C.; Choi, J. Effective data-balancing methods for class-imbalanced genotoxicity datasets using machine learning algorithms and molecular fingerprints. *Comput. Toxicol.* **2021**, *20*, No. 100178.
(48) Lowe, D. *Chemical Reactions from US Patents (1976-Sep 2016)*. 2017 https://figshare.com/articles/dataset/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873 (accessed January 2, 2022).
(49) Szymkuć, S.; Badowski, T.; Grzybowski, B. A. Is Organic Chemistry Really Growing Exponentially? *Angew. Chem., Int. Ed.* **2021**, *60*, 26226−26232.
(50) Moskal, M.; Beker, W.; Szymkuć, S.; Grzybowski, B. A. Scaffold-Directed Face Selectivity Machine-Learned from Vectors of Non-covalent Interactions. *Angew. Chem., Int. Ed.* **2021**, *60*, 15230.

https://doi.org/10.1021/jacs.1c12005
J. Am. Chem. Soc. 2022, 144, 4819−4827