JOC The Journal of Organic Chemistry

pubs.acs.org/joc Article

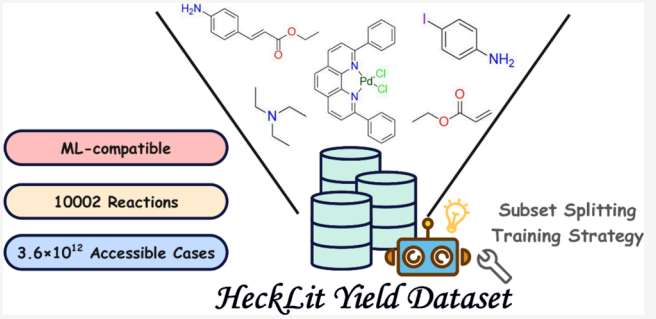
Optimizing Model Learning Performance on a Challenging Heck Reaction Yield Data Set

Shen Wang, Yining Liu, Weiren Zhao, and Yang Li*

Cite This: J. Org. Chem. 2025, 90, 12768−12777 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The development of machine learning (ML) in organic synthesis is limited due to the lack of available data sets. The ML-compatible literature data set for Heck reaction yields, named HeckLit, has been established. With 10,002 cases, the data set spans multiple reaction subclasses and covers a larger chemical space compared to high-throughput experimentation data sets, including nearly 3.6 × 10^{12} accessible cases. HeckLit accelerates the advancement of ML-driven synthesis research; however, it suffers from the same dilemma as other literature-based data sets, sparse distribution and high-yield preference, leading to limited model learning ability on the test set, with R^2 = 0.318. Thus, feature distribution smoothing (FDS) and subset splitting training strategy (SSTS) are utilized to tackle this issue. Despite no improvement when using FDS, SSTS boosted the R^2 to 0.380. This optimization approach relies on the subset division. Therefore, we suggest a criterion for splitting. The SSTS opens up a new avenue for tackling the challenge of learning from large-scale data sets.



1. INTRODUCTION

A specialized domain within computational chemistry, referred to as computer-assisted synthesis prediction (CASP), employs machine learning (ML) methodologies to predict the feasibility, yield, and optimal reaction conditions of chemical processes.[1−9] Owing to its significant benefits, including reduced labor costs and enhanced experimental efficiency and precision, this approach has garnered substantial interest among the chemistry community. Crucially, the efficacy of ML applications in CASP is dependent on the authenticity and integrity of the reaction data utilized.[10]

The majority of existing ML studies in CASP rely on high-throughput experimentation (HTE) data sets (with test set R^2 > 0.7), which are well designed, usually with about a few hundred cases of data, and result in remarkably accurate prediction outcomes.[11−14] However, this type of data set can only be applied to a subset of cases for a given reaction, resulting in very poor generalizability.

Large-scale data sets, covering a chemical space consisting of millions of reactions and typically containing around 10,000 cases of data, are widely favored in pretraining and transfer learning strategies due to the diversity of the reactions they contain.[15−18] Reaction databases such as Reaxys[19] and Open Reaction Database[20] provide reliable literature data; however, these databases suffer from a lack of standardization of reaction conditions and yield data, which makes it difficult to be recognized by ML models, greatly hampering the advancement of CASP. In addition, model learning is challenging in large-scale literature data sets (test set R^2 around 0.2) due to the wide chemical space with sparse and imbalanced data distribution.[21−26] Multiple methods have been proposed to address this challenge. After removing reactivity cliffs and uncertain reactions, Liu et al. improved R^2 by 0.06 on the challenging amide coupling reaction data set. However, the removed data still needs to be predicted.[27] Ma et al. stated that the prediction challenge of literature data sets is from the human biases in experiments and reporting of high-yield results. They utilized cost-sensitive reweighting methods and label distribution smoothing methods to handle this issue, though they succeeded in enhancing over 7.3% in few-shot data prediction, the model performance dropped for all test data.[28] Therefore, developing a reasonable method to improve the model's learning ability on challenging large-scale literature data sets is urgently needed.

Among various homogeneous catalytic reactions, the Heck coupling reaction stands out as a mainstream method for molecular skeleton construction. Since its independent discovery by Heck and Morizoki in the late 1960s, continuous refinement of catalytic systems and reaction parameters has expanded its applicability across diverse chemical transformations, establishing it as a cornerstone methodology for

Received: July 3, 2025
Revised: August 7, 2025
Accepted: August 27, 2025
Published: September 3, 2025

https://doi.org/10.1021/acs.joc.5c01641
J. Org. Chem. 2025, 90, 12768−12777

C−C bond formation.[29] This historical evolution has generated a substantial volume of documented reactions in scientific literature, making it particularly suitable for literature-derived data set development. Notably, multiple Heck reaction data sets have been reported thus far. For instance, Wang et al. constructed a data set for product prediction, comprising 9599 reaction entries.[30] Also, Ree et al. constructed a data set with 14,240 cases for regioselectivity calculation.[31] However, their data sets lack critical information such as detailed reagent specifications, catalyst information, and yield data − essential elements for systematic exploration of the Heck reaction space − thereby limiting its utility in predictive modeling studies (see Section 1.1 in the SI for detailed differences).

As illustrated in Figure 1, to accelerate the development of ML techniques in the field of organic synthesis, we construct

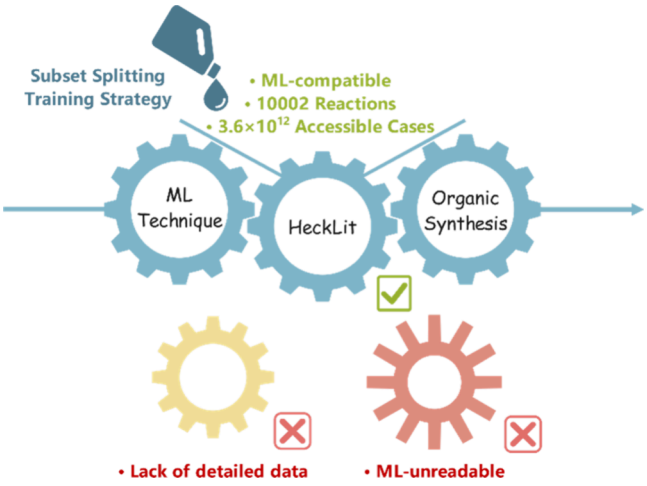


Figure 1. Schematic illustration of this work.

HeckLit, an ML-readable, literature-based, large-scale Heck reaction yield data set, by collecting and carefully processing data from Reaxys. It contains 10,002 Pd-catalyzed Heck reactions with detailed catalytic conditions, covering a large chemical space estimated at 3.6×10^{12} accessible reactions. Multiple statistical methods are used to analyze the data set and compare HeckLit with HTE data sets to study its data distribution and chemical space. Eleven ML methods are used to set a benchmark for HeckLit. To enhance the learning of the models on HeckLit, we start by looking at the data distribution to cope with the poor model learning phenomenon, by using feature distribution smoothing (FDS) and subset splitting training strategy (SSTS). HeckLit is expected to alleviate the shortage of organic data sets available for ML techniques. And the study of 2 optimization methods provides insight into the challenging large-scale data set learning.

2. METHODS

2.1. Data Processing. Figure 2a shows the process of data collecting and cleaning for HeckLit as follows:

1. Data collection: Using the keyword "Heck", we searched on the Reaxys,[19] and as of Nov. 2024, 60043 relevant reactions were available. We collected 52,415 Heck reactions from this publicly accessible database.
2. Reactant and product processing: (i) Screening to obtain reactions with the number of reactants less than or equal to 2 and a unique product; (ii) Normalization of the reactants and products in simplified molecular input line entry system (SMILES) format,[32] the specific workflow of SMILES normalization is displayed in Figure 2b; (iii) Heck reaction template matching based on the RDKit Python package[33] was utilized to check the reaction, as shown in Figure 2c.
3. Reaction yield handling: As illustrated in Figure 2d, reactions without yield data, or the output representing enantioselectivity value or mass, were filtered.
4. Reagent (catalyst, additive, solvent) processing: (i) Convert the reagents from IUPAC name into SMILES format using PubChem[34] and Cactus[35] (specific process is demonstrated in Figure 2e); (ii) Normalization of the reagents' SMILES; (iii) Filtering of reactions without Pd catalyst.
5. Reactions with the same substrates, products, and reagents are considered as identical cases. We averaged the temperature, time, and yield data for these reactions.

In this study, we also used 2 HTE data sets, Buchwald-Hartwig (B−H) data set,[36] and Suzuki−Miyaura (S−M) data set[37] for comparative studies with HeckLit. To prove the applicability of the ML approach to the Heck reaction, Das et al.'s data set was utilized.[38] According to the literature description, we converted all chemicals to SMILES for the study. The data structure of these relevant data sets is listed in Section 1.2 of the SI.

2.2. Machine Learning Techniques. The unsupervised learning methods reaction fingerprint (RXNFP)[39] generated by bidirectional encoder representations from a BERT model, as well as differential reaction fingerprint (DRFP)[40] based on the hash processing of molecular substructures, were used to featurize the reactions, respectively. Four traditional ML models were tested: random forest (RF), XGBoost, support vector machine (SVM), and *k*-nearest neighbor (KNN). Three deep learning models, artificial neural networks (ANN), graph attention network (GAT), and multimodal homogeneous reaction predictor (MMHRP-GCL)[41] were utilized for data set learning and yield predicting. The fact that HeckLit contains a fluctuating number of reactants makes most advanced graph neural network models, such as CGR-GCNN,[42] unusable for this data set. The detailed model structure and parameter settings are listed in Section 2 in the SI.

Based on the diversity and sparsity of the data set, which makes model training challenging, hence, we allocated more data to the train set, with a train/test set of 80%/20%. The model learned data properties on the train set, and the model's performance was evaluated on the test set. For each method, we performed 5 times random train/test set splits. The coefficient of determination (R^2), root-mean-square error (RMSE), and mean absolute error (MAE) were utilized to assess the performance of the model in learning or predicting. The closer the R^2 value is to 1 and the lower the RMSE and MAE values, the better the model is. The equations for the evaluation metrics R^2, RMSE and MAE are displayed in eq 1−3:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y}_i)} \tag{1}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{2}$$

$$\mathrm{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{3}$$

where y_i denotes the ground truth, $\hat{y}_i$ is the predicted value of the model and $\overline{y}_i$ denotes the average of the true values of the whole data set.

2.3. Feature Distribution Smoothing. Feature distribution smoothing (FDS), proposed by Yang et al., is an algorithm for deep learning models.[43] The method transfers the feature statistics between nearby target bins and thus enhances the prediction ability of the model on imbalanced data. The concrete algorithm is displayed in Section 3 in the SI.

2.4. Subset Splitting Training Strategy. The subset splitting training strategy (SSTS) is designed based on the idea of divide-and-

https://doi.org/10.1021/acs.joc.5c01641
J. Org. Chem. 2025, 90, 12768−12777

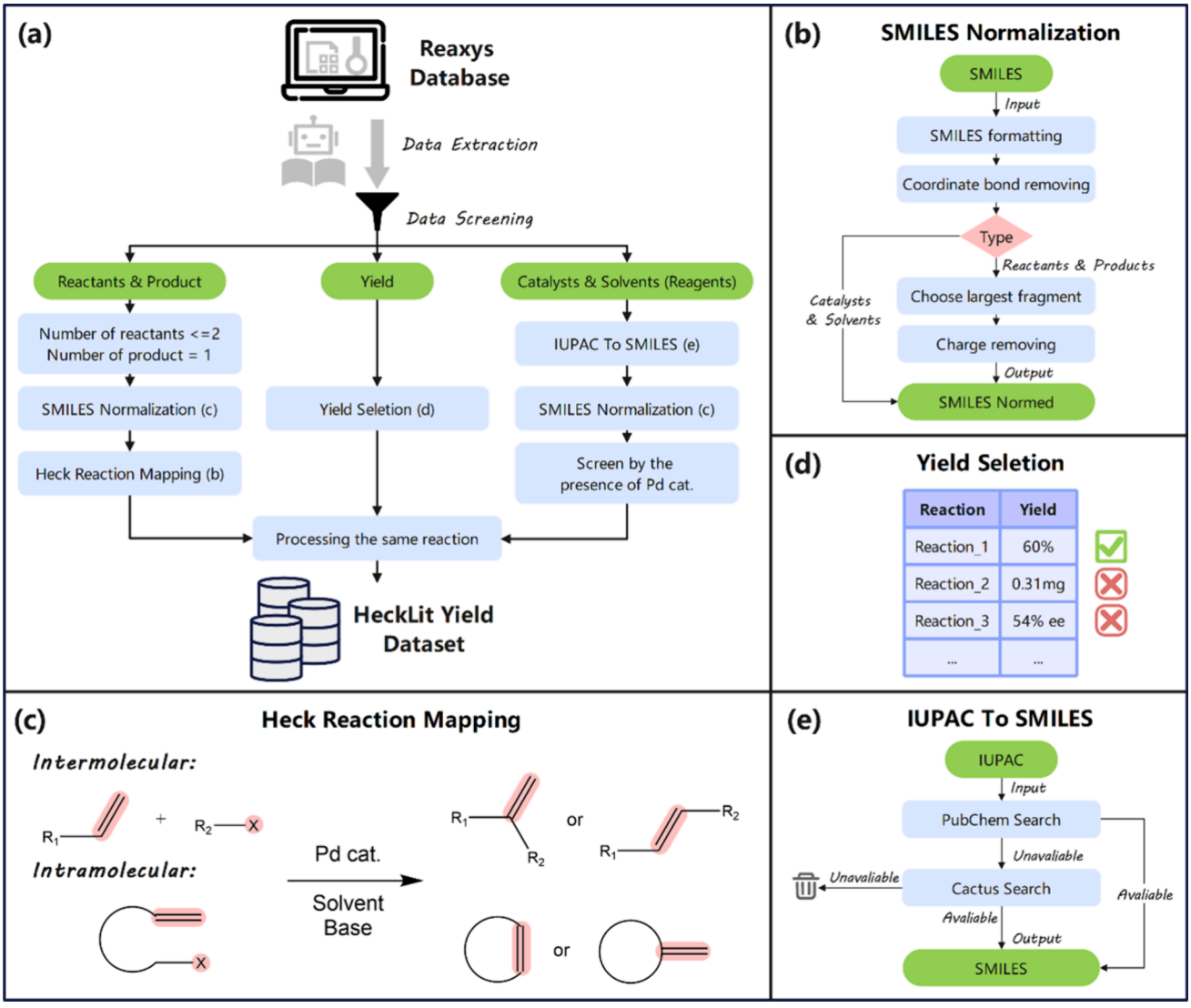


Figure 2. Data cleaning workflow for HeckLit. (a) Overview of cleaning procedure. (b) Reaction mapping. (c) SMILES normalization workflow. (d) Yield selection rule. (e) IUPAC to SMILES conversion procedure.

conquer calculations. The method first splits the data set into multiple subgroups and then uses ML models to learn the properties of each subset. It is worth noting that when testing this method, to ensure the consistency of the ratio between the train/test set, we will first conduct the subset splitting and then divide the train/test set. The detailed algorithm is shown below.

Divide the data set A into several subgroups according to the properties of the data, where $A = \{a_1, \cdots, a_i, \cdots, a_n\}$ and each $a_i = (x_{ai}, y_{ai})$. In the statistical division, we first mapped the data set to *2-dimensions* using principal component analysis (PCA) dimensionality reduction technique and later defined the subsets manually using *K*-means. In the chemical division, the subsets are determined based on the reaction characteristics. For each subgroup a_i in the train set, the model$_i$ is trained for it:

$$\text{criterion}(y_{ai}, \text{model}_i(x_{ai})) = \text{Loss} \rightarrow \text{model}_i \tag{4}$$

where criterion$(\cdot)$ denotes the loss function. Loss is the value calculated by the loss function. Right arrow represents training of the model by loss values.

When making predictions on the external data set B, it is also divided into several subsets based on data characteristics: $B = \{b_1, \cdots, b_i, \cdots, b_n\}$, where $b_i = (x_{bi})$. The label prediction is performed using models corresponding to each subgroup:

$$y_{\text{predict}} = \text{model}_i(x_{bi}) \tag{5}$$

where y_{predict} denotes the predicted value of the model.

3. RESULTS AND DISCUSSION

3.1. Statistical Study of HeckLit. After data processing, as listed in Figure 3a, HeckLit contains 10,002 reactions of reaction yield data (8630 of intermolecular reactions and 1372 of intramolecular reactions), involving 6718 reactants, 8891 products, 338 catalysts (45 Pd catalysts), and 60 solvents. Preliminary estimating, the data set covers the chemical space with almost 3.6×10^{12} accessible reactions (see Section 4.1 in the SI for calculations).[44] The data structure of HeckLit is displayed in Figure 3b, including (i) Reactive substances: including reactants, products, catalysts, and solvents information, which are stored in SMILES. (ii) Reactant conditions: reaction temperature (°C) and time (h). (iii) Yield: yield of the chemical reaction (%). (iv) Search information: Reaction ID in Reaxys and reference or patent information.

In organic chemical synthesis, Heck reactions are systematically classified into intermolecular and intramolecular subsets. As shown in Figure 4, the reactions can be further classified into

https://doi.org/10.1021/acs.joc.5c01641
J. Org. Chem. 2025, 90, 12768−12777

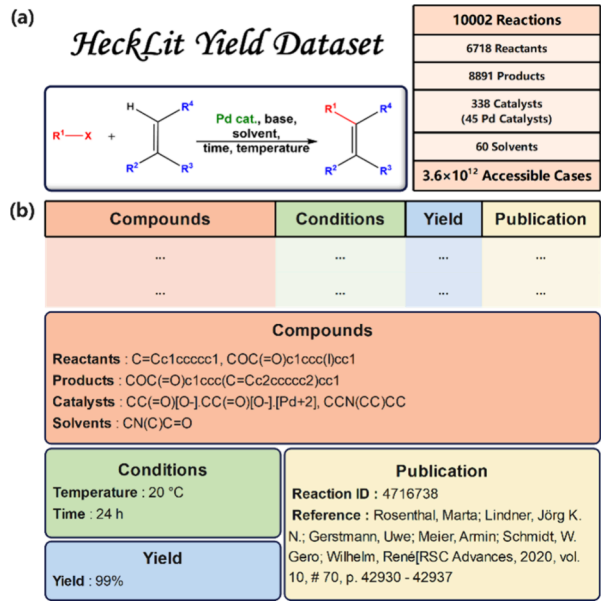


Figure 3. Introduction to HeckLit yield data set. (a) Data set composition. (b) Data structure in HeckLit.

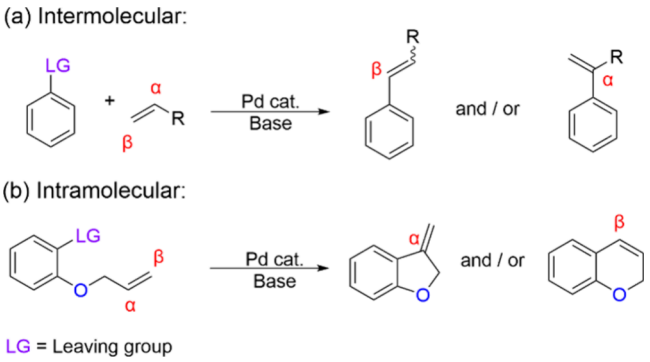


Figure 4. Illustration of Heck reaction types. (a) Intermolecular reaction. (b) Intramolecular reaction.

subsubsets according to the chemical characteristics: the regioselectivity of C═C bond (α/β-position insertion), leaving group (LG) type, and the size of the ring formed by intramolecular reaction, which are represented in red, green, and blue, respectively, in Figure 5a.

The proportion of each subsubset is listed in Figure 5a. The height of the red bars for the two subsets indicates that Heck intramolecular reactions favor α-position insertion, while intermolecular reactions favor β-position insertion. This can be attributed to the fact that the α-position insertion can form 5- or 6-membered stable rings in intramolecular reactions, while there is less spatial resistance in intermolecular reactions for β-position insertion. The height of the green bars suggests the substitution of leaving group tendency of the Heck reaction pair, which prefers the insertion of C−Br and C−I bonds because of lower bond energies, leading to a lower energy barrier in the oxidative addition step. The height of the blue bars indicates that Heck intramolecular reactions are common in constructing 5- and 6-membered rings since they are thermodynamically more stable than the other rings. The wide range of reaction subsubsets demonstrates that the application of HeckLit is anticipated to accelerate the research and development for Heck reactions.

Figure 5b shows the distributions of reaction time, temperature, and yield for two subsets. Statistical analysis shows that the intramolecular Heck reaction requires shorter reaction times than the intermolecular reaction, which corresponds to the greater rate of intramolecular reaction. Moreover, the yield distributions of intramolecular and intermolecular reactions are similar, but the more sluggish reactions give a lower yield for intermolecular reactions. This is due to the high-yield tendency of reported literature data. This preference may trigger high model-predicted values for low-yield reaction yields and is one of the challenges for literature-based data sets.[45]

To investigate the distribution of reagent information in the data set, we counted the number of occurrences of each reagent and sorted them in descending order. Figure 5c shows the distribution of the diversity of each reagent in HeckLit and its two subsets. The x-axis indicates the top N reagents of the occurrence count, and the y-axis represents the proportion of occurrences of these reagents to all reagents. The closer the slope of the plotted curve is to a constant, the more uniformly the reagents are distributed.[46] The results show that the slope of the curve decreases with increasing top N reagent type, suggesting an uneven distribution of reagents in HeckLit, which is the reason that learning large-scale data sets is difficult.

Additionally, a statistical analysis of solvent effects and catalysts in HeckLit is conducted, as shown in Section 4.2 in the SI.[47−49] The study demonstrates the potential of HeckLit in the Heck reaction study.

3.2. Chemical Information: HTE vs HeckLit. Learning the inherent differences in chemical diversity and yield distributions between the HeckLit and HTE data sets helps us gain a deeper insight into large-scale data. To explore the difference in chemical space, we projected reactions into the same space using the multidimensional scaling (MDS) method, as shown in Figure 6a. Compared to B−H and S−M HTE data sets,[36,37] HeckLit covers a larger chemical space. In Figure 6b, the density of the chemical space is compared by calculating the cosine similarity in pairs of cases. The results show that most of the reactions in HTE data sets are in the high similarity, ranging from 0.4 to 0.8, while that in HeckLit are in the low similarity region, by 0.0−0.4, suggesting that the data in HeckLit are sparsely distributed in the chemical space and it is challenging for model learning. HeckLit, a large-scale data set, is generalization-oriented, leveraging its diverse chemical space to enhance model generalizability; however, the sparsity may lead to poor predictive accuracy, making it particularly suited for massive searches, such as reaction condition exploration or catalyst screening scenarios. In contrast, HTE data sets are accuracy-centric, where their confined chemical space enables high model precision, rendering them more effective for well-defined catalysis investigations. Figure 6c shows the yield distribution of the 3 data sets. It can be seen that HTE data sets possess more low-yield reactions, while HeckLit includes a higher proportion of high-yield cases. Exploring data sets with different yield distributions raises the prospect that these data will be useful for ML models.

3.3. Model Benchmarks. Table 1 lists the learning performance of models on Das et al.'s data set[38] and HeckLit. To demonstrate the applicability of the ML techniques for the Heck reaction yield prediction task, we first evaluated the performance of 11 methods on Das et al.'s data set. ANN + DRFP method stands out among all methods, with R^2 by 0.707, RMSE by 5.68% and MAE by 3.95%, even better than the original work using experimental and quantum chemical

https://doi.org/10.1021/acs.joc.5c01641
J. Org. Chem. 2025, 90, 12768−12777

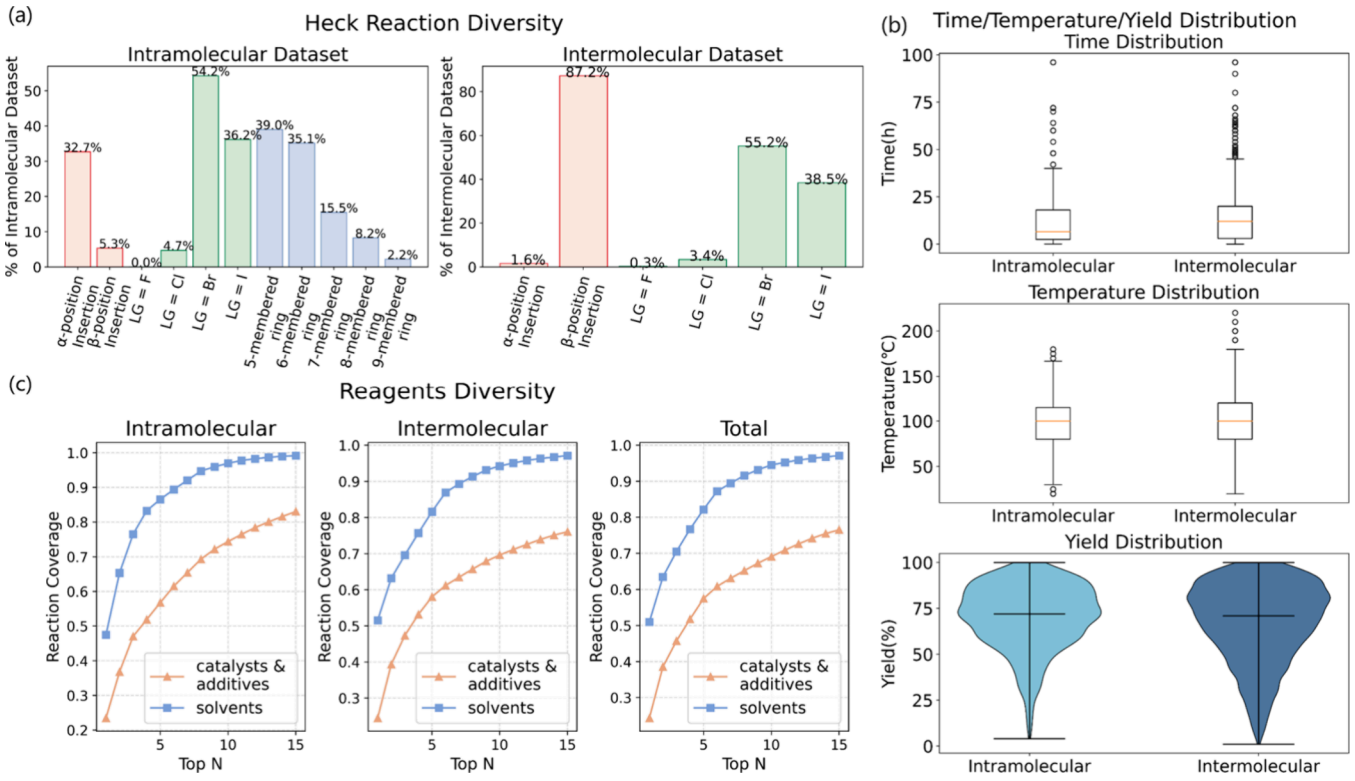


Figure 5. Data distribution analysis of HeckLit. (a) Reaction diversity analysis. (b) Reaction time, temperature, and yield distribution plots. (c) Reagent diversity analysis.

calculation parameters, indicating the usability of DRFP in Heck reaction feature learning. Still, ANN + DRFP outperforms other methods in HeckLit learning, with R^2 by 0.318, RMSE by 17.60%, and MAE by 13.06%, which is slightly better than that of RF + DRFP. Thus, in all other experiments, we have taken one of these 2 methods. In Section 4.3 in the SI, we plot the cosine similarity distribution for Das et al.'s data set and HeckLit, showing that the sparser chemical space serves as the primary reason for the model's inferior learning performance on HeckLit compared to Das et al.'s data set. It is also worth noting that GAT and MMHRP-GCL trail behind in performance on challenging data set learning, with R^2 ranging from 0.553−0.590 and 0.207−0.239 on the 2 data sets, respectively. We interpret this result as the advanced graph-based model or multimodal model's tendency to fall into part of the data, making it poorly predictive of the entire data set, or the atomic features used that do not adequately capture the reaction property.[50]

To prove that the reagent data that we carefully cleaned for HeckLit plays a pivotal role in yield prediction, we conducted an ablation study by using DRFP generated without reagent information. As displayed in Table 1, ANN + DRFP without reagent information exhibits worse learning ability than using DRFP, with R^2 lower than 0.106, suggesting that the reagent information can enhance model learning performance in reaction prediction. Further, we conducted the ablation experiments on catalysts and additives, and solvents. The model performance of DRFP w/o catalyst and additives is worse than that of DRFP w/o solvent, with an R^2 lower than 0.054, demonstrating the higher importance of catalyst and additive information.

We also set a benchmark using RF + DRFP for each subsubset, which is displayed in Section 4.4 in the SI. The model learning performance shows that for small-sized subsubsets, the learning ability of the ML model is not stable, with R^2 generally fluctuating widely on the test set, suggesting that the subsubsets are sparse. However, the RMSE of each subset is within 18.57%, indicating that although the model is difficult to accurately predict the reaction yields in the HeckLit subsubsets, it is still able to make effective judgements on the high and low reaction yields.

3.4. Optimization for Model Learning. A multitude of studies stated that the yield distribution might lead to poor model performance.[28,45,51] To demonstrate that this problem also exists in HeckLit, we divided the data set into 10 yield-based subsets and categorized each as few-shot, medium-shot, or many-shot according to its sample size (see Section 4.5 in the SI for details). Figure 7 illustrates the yield distribution and error distribution of the model trained on HeckLit, which shows that the subset with fewer reactions has greater prediction errors.

To handle this issue, we utilized the FDS method, which distributes the feature statistics across neighboring label bins, hence performing distribution smoothing on the feature space. The experimental details are demonstrated in Section 4.6 in the SI. Table 2 lists the performance of ANN + DRFP for the whole test set and each shot as a baseline, compared with the performance using FDS. Among them, FDS improves performance on few-shot and medium-shot with RMSE being reduced by 0.29% and 0.15%, respectively. However, this method drops the model performance on the entire data set, with RMSE being increased by 0.08%. We also split the data set into 100 yield-based subsets and tested with FDS, yielding consistent conclusions (See Section 4.6 in the SI).

In the previous analysis, HeckLit exhibits sparse feature distributions and complex composition, which may contribute to the poor model performance. Hinton et al. suggested that it makes sense to train many specialized models, each of which is

https://doi.org/10.1021/acs.joc.5c01641
J. Org. Chem. 2025, 90, 12768−12777

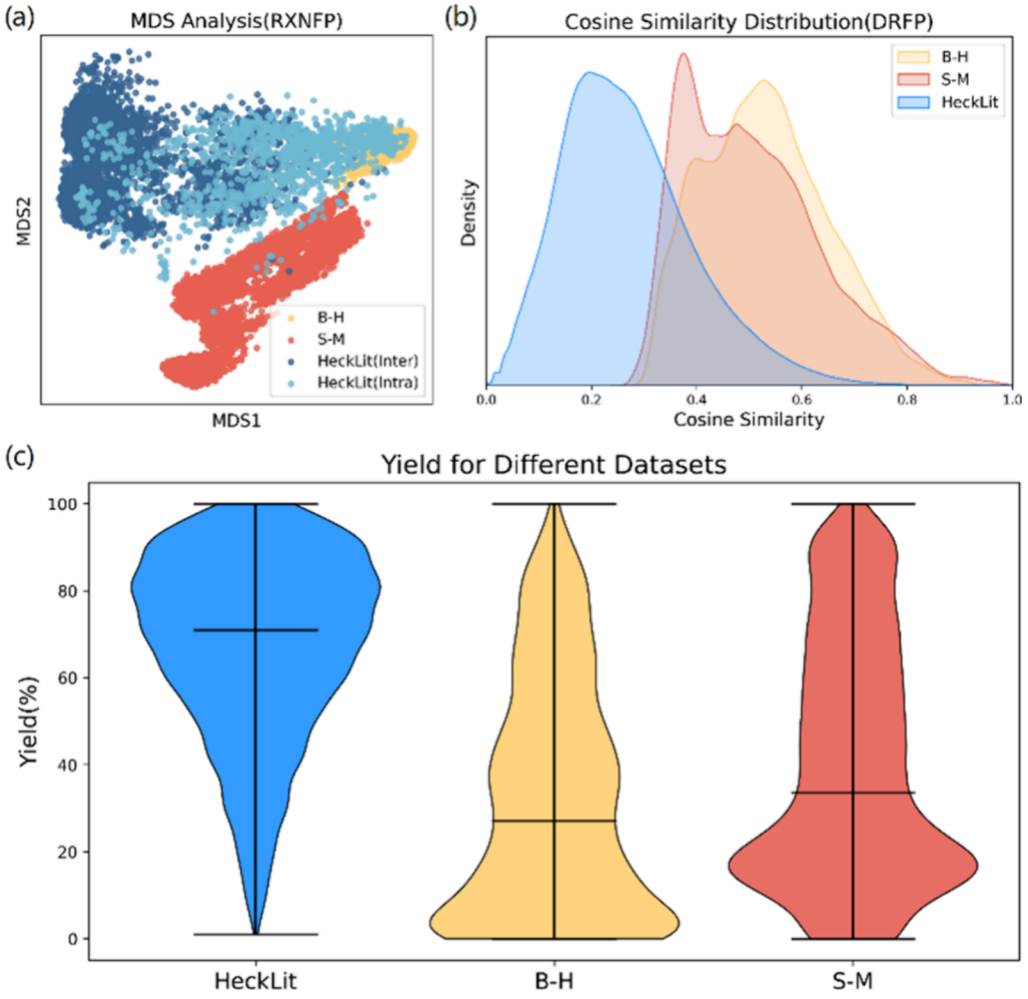


Figure 6. Chemical space and yield distribution study of HeckLit compared with B−H & S−M data sets. (a) MDS analysis. (b) Cosine similarity distribution between reactions. (c) Yield distribution.

Table 1. Model Performances on Das et al.'s Data Set and HeckLit[a]

data set		Das et al.'s data set			HeckLit		
model	descriptor	R^2	RMSE	MAE	R^2	RMSE	MAE
RF	137 features	0.522 ± 0.139	7.37 ± 3.05	4.56 ± 1.07			
RF	RXNFP	0.300 ± 0.227	8.49 ± 1.80	5.50 ± 0.65	0.238 ± 0.010	18.60 ± 0.19	14.66 ± 0.13
RF	DRFP	0.601 ± 0.044	6.59 ± 1.84	4.25 ± 0.53	0.318 ± 0.011	17.60 ± 0.28	13.08 ± 0.14
XGBoost	RXNFP	0.295 ± 0.373	8.35 ± 1.92	5.47 ± 0.89	0.174 ± 0.022	19.36 ± 0.21	14.87 ± 0.12
XGBoost	DRFP	0.391 ± 0.233	7.72 ± 0.95	4.87 ± 0.36	0.271 ± 0.016	18.19 ± 0.35	13.68 ± 0.15
SVM	RXNFP	0.235 ± 0.059	9.26 ± 3.24	6.57 ± 0.99	0.169 ± 0.018	19.42 ± 0.25	15.17 ± 0.15
SVM	DRFP	0.289 ± 0.161	8.64 ± 1.85	6.63 ± 0.87	0.309 ± 0.007	17.71 ± 0.21	13.77 ± 0.08
KNN	RXNFP	0.183 ± 0.094	9.49 ± 2.98	6.41 ± 0.27	0.115 ± 0.014	20.04 ± 0.19	16.13 ± 0.15
KNN	DRFP	0.311 ± 0.046	8.83 ± 3.19	5.48 ± 0.62	0.124 ± 0.012	19.93 ± 0.13	15.93 ± 0.12
ANN	DRFP	**0.707 ± 0.020**	**5.68 ± 1.66**	**3.95 ± 0.58**	**0.318 ± 0.010**	**17.60 ± 0.24**	**13.06 ± 0.15**
ANN	DRFP (w/o reagents)				0.212 ± 0.019	18.90 ± 0.21	14.20 ± 0.23
ANN	DRFP (w/o catalysts and additives)				0.248 ± 0.016	18.48 ± 0.18	13.80 ± 0.21
ANN	DRFP (w/o solvents)				0.302 ± 0.004	17.79 ± 0.15	13.23 ± 0.05
GAT	Graph	0.590 ± 0.119	6.71 ± 2.40	4.65 ± 0.69	0.239 ± 0.006	18.58 ± 0.17	14.30 ± 0.20
MMHRP-GCL	Graph+Text	0.553 ± 0.112	7.22 ± 2.79	4.82 ± 1.08	0.207 ± 0.005	18.97 ± 0.20	14.59 ± 0.37

[a]Note: "w/o" denotes "without". **Bold**-labeled numbers indicate the best performance.

trained on data that is highly enriched in examples.[52] Taking this as a breakthrough, we utilized SSTS methods by using multiple models responsible for learning a portion of the features of the data set. A variety of methods are used for subset division, including PCA + *K*-means to set subsets based on statistical technique, inter/intramolecular subsets based on reaction type, α/β-position insertion, and no regioselectivity subsets based on regioselectivity, and subsets that take into account both reaction type and regioselectivity (see Section 4.7 in the SI).

https://doi.org/10.1021/acs.joc.5c01641
J. Org. Chem. 2025, 90, 12768−12777

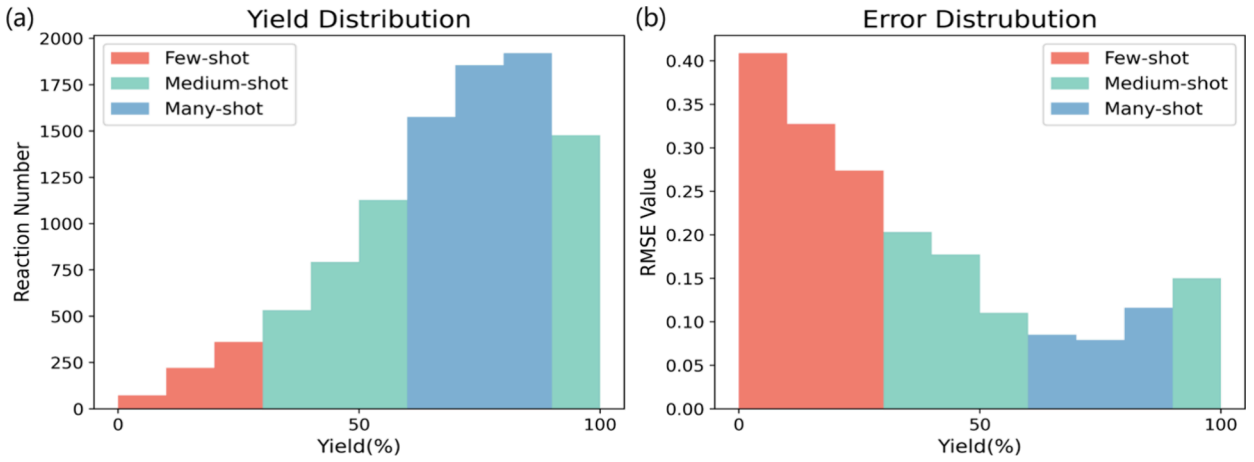


Figure 7. Comparison between (a) yield distributions and (b) test error distributions when model learning on HeckLit.

Table 2. FDS Method for Model Learning Performance Improvement

evaluation metrics	RMSE				MAE			
shot	all	few	medium	many	all	few	medium	many
baseline	17.60	35.15	19.22	12.46	13.06	29.89	15.38	9.38
FDS	17.68	34.86	19.07	12.62	13.12	29.50	15.23	9.44
vs	**+0.08**	*-0.29*	*-0.15*	**+0.16**	**+0.06**	*-0.39*	*-0.15*	+0.06

Note: **Bold**-labeled numbers indicate performance decreases; *Italic*-labeled numbers represent improvement.

Table 3. SSTS Method for Model Learning Performance Improvement

		evaluation metrics		
subset splitting method	subset number	R^2	RMSE	MAE
baseline	1	0.318 ± 0.010	17.60 ± 0.24	13.06 ± 0.15
PCA + *K*-means	3	0.305 ± 0.009	17.75 ± 0.21	13.21 ± 0.08
PCA + *K*-means	4	0.307 ± 0.008	17.75 ± 0.18	13.18 ± 0.06
PCA + *K*-means	5	0.307 ± 0.006	17.73 ± 0.13	13.19 ± 0.06
inter/Intra molecular	2	**0.380 ± 0.009**	**16.62 ± 0.42**	**12.25 ± 0.35**
α/β-position Insertion	3	0.315 ± 0.008	17.64 ± 0.22	13.11 ± 0.14
intra/intermolecular & α/β-position insertion	6	0.326 ± 0.009	17.49 ± 0.24	12.96 ± 0.11

Note: **Bold**-labeled numbers indicate the best performance.

Table 3 demonstrates the model performance of ANN + DRFP using SSTS in different subset splitting methods compared to the baseline. By splitting the data set into inter/intra molecular, the model boosts the R^2 value from 0.318 to 0.380, and the statistical significance testing between the baseline and the optimal performance shows p values of the evaluation metrics are ranging from 0.0004 to 0.0031 (<0.05), indicating that SSTS is a practical method to effectively optimize model learning on a large-scale data set, and the improvement is not derived from randomness. It is pertinent to point out that the PCA + *K*-means method to set subsets fails to improve the model performance, and the remaining 2 chemical knowledge-based subset segmentation methods do not effectively improve the model learning capability. This involves a game between a general model and multiple specialized models. Although diverse expert models can address distinct aspects of complex chemical issues and may lead to better outcomes than a single model could achieve, they fail to effectively integrate knowledge across multiple subdomains and sometimes exhibit inferior performance compared to the general model (i.e., the baseline).[53,54] Hence, the method for subset splitting is the decisive factor in determining whether the SSTS approach can enhance data set learning. We propose a subset splitting criterion for SSTS: (1) The subset should be large enough since small-scale data will increase the difficulty for deep learning model training. It is recommended that the subset contains at least 500 cases and involves at least 100 compounds.[55,56] (2) A significant difference in chemical aspects among subsets is necessary, for subtle differences between subsets may lead to weak cross-subdomains capabilities of the model.

4. CONCLUSIONS

Aiming to alleviate the scarcity of organic reaction data sets, in this work, we built an ML-compatible literature-based large-scale yield data set, HeckLit, by collecting and cleaning the data from Reaxys. It contained 10,002 reactions, covering a large chemical space with more than 3.6×10^{12} accessible reactions. Statistical analysis of HeckLit showed the data set was chemically diverse, but also suffered from a high-yield preference. Comparison between HeckLit and HTE data sets found that HeckLit represented a larger chemical space but had a sparse and imbalanced data distribution, implying that model learning on this large-scale literature data set was challenging.

https://doi.org/10.1021/acs.joc.5c01641
J. Org. Chem. 2025, 90, 12768−12777

Subsequently, 11 models were first evaluated to be suitable for Heck reaction prediction and then utilized to set an ML benchmark. ANN + DRFP, with R^2 of 0.318, RMSE of 17.60%, and MAE of 13.06%, outperformed other methods. Also, the ablation study on reagents' information was conducted, suggesting that detailed chemical information in HeckLit was one of the strengths. However, the low R^2 value indicated that current ML models still struggle in learning large-scale data sets, due to the uneven yield distribution and sparse chemical space coverage. Thus, we applied FDS and SSTS to handle this issue. Although the improvement of FDS was limited, with RMSE values decreasing on few-shot and medium-shot data by 0.29% and 0.15%, respectively, increasing by 0.08% on the whole test set, SSTS boosted R^2 from 0.318 to 0.380. Owing to the cross-subset knowledge integration barriers, SSTS relies on the subset splitting method; therefore, we point out that subsets need to satisfy the criterion of (1) not being too small (at least 500 cases and 100 compounds) (2) having significant differences in chemical properties.

Despite the incremental progress we have made in large-scale reaction data sets, there are still many problems and challenges to be solved. Our analysis reveals yield-reporting bias and data sparsity, impeding large-scale reaction data set learning. We recommend documenting negative results and resolving data scarcity through automated laboratories.[57−59] Additionally, considering the variability in experimental yield data, taking the yield prediction task into multiclass classification (high, medium, low, etc.) is possibly more suitable for large-scale data sets. Moreover, the prediction performance and generalization ability of current models need to be improved. While large language models such as ChatGPT or DeepSeek, with large-scale parameters and reasoning ability, may be better in large-scale data learning.[60] To address current limitations in reaction data accessibility and model generalizability, interdisciplinary collaborations must prioritize the creation of ML-readable organic reaction repositories and adaptive learning frameworks, which are critical for unlocking AI-driven innovations in synthetic chemistry.

■ ASSOCIATED CONTENT

Data Availability Statement

The data underlying this study are openly available in GitHub at https://github.com/AIChem-ShenWang/HeckLit-Code.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.joc.5c01641.

Illustration of Wang et al.'s, Ree et al.'s, B−H, S−M, and Das et al.'s data sets, and HeckLit; framework and parameter settings of machine learning models, artificial neural network, and graph attention network; performance of models on Das et al.'s data set and HeckLit; statistical analysis of solvation effects and Pd catalysts; cosine similarity distribution of Daset al.'s data set and HeckLit; test performances on each sub-subset; parameter settings of the FDS method; model performance using FDS; model performance using FDS with 100 subsets; chemical space of subsets in different splitting methods; and statistical significance testing for SSTS with the inter/intra molecular subset splitting method (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Yang Li − *State Key Laboratory of Fine Chemicals, Dalian University of Technology, Dalian 116024, China; School of Chemical Engineering, Ocean and Life Sciences, Dalian University of Technology, Panjin 124221, China;* orcid.org/0000-0002-5719-9044; Email: chyangli@dlut.edu.cn

Authors

Shen Wang − *State Key Laboratory of Fine Chemicals, Dalian University of Technology, Dalian 116024, China; Leicester International Institute, Dalian University of Technology, Panjin 124221, China;* orcid.org/0009-0008-4174-4301

Yining Liu − *State Key Laboratory of Fine Chemicals, Dalian University of Technology, Dalian 116024, China; School of Chemical Engineering, Ocean and Life Sciences, Dalian University of Technology, Panjin 124221, China*

Weiren Zhao − *State Key Laboratory of Fine Chemicals, Dalian University of Technology, Dalian 116024, China; School of Chemical Engineering, Ocean and Life Sciences, Dalian University of Technology, Panjin 124221, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.joc.5c01641

Author Contributions

S.W.: Conceptualization, methodology, investigation, data curation, and writing—original draft, review, and editing. Y.L.: Validation and writing—review and editing. W.Z.: Writing—review and editing. Y.L.: Project administration.

Funding

Y.L. is supported by the National Science Foundation of China (21903010). S.W. is supported by the Fundamental Research Funds for the Central Universities (DUT24BK047).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors would like to thank the funding sources mentioned above.

■ REFERENCES

(1) Yang, L.-C.; Li, X.; Zhang, S.-Q.; Hong, X. Machine learning prediction of hydrogen atom transfer reactivity in photoredox-mediated C-H functionalization. *Org. Chem. Front.* **2021**, *8*, 6187−6195.

(2) Nakajima, H.; Murata, C.; Noto, N.; Saito, S. Database Construction for the Virtual Screening of the Ruthenium-Catalyzed Hydrogenation of Ketones. *J. Org. Chem.* **2025**, *90*, 1054−1060.

(3) Zhang, S.-Q.; Xu, L.-C.; Li, S.-W.; Oliveira, J. C. A.; Li, X.; Ackermann, L.; Hong, X. Bridging Chemical Knowledge and Machine Learning for Performance Prediction of Organic Synthesis. *Chem.−Eur. J.* **2023**, *29*, No. e202202834.

(4) Felten, S.; He, C. Q.; Emmert, M. H. C−H Aminoalkylation of 5-Membered Heterocycles: Influence of Descriptors, Data Set Size, and Data Quality on the Predictiveness of Machine Learning Models and Expansion of the Substrate Space Beyond 1,3-Azoles. *J. Org. Chem.* **2025**, *90*, 2613−2625.

(5) Li, S.-W.; Xu, L.-C.; Zhang, C.; Zhang, S.-Q.; Hong, X. Reaction performance prediction with an extrapolative and interpretable graph model based on chemical knowledge. *Nat. Commun.* **2023**, *14*, 3569.

(6) Andronov, M.; Voinarovska, V.; Andronova, N.; Wand, M.; Clevert, D.-A.; Schmidhuber, J. Reagent prediction with a molecular transformer improves reaction data quality. *Chem. Sci.* **2023**, *14*, 3235−3246.

(7) 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.
(8) Cong, S.; Zhang, M.; Song, Y.; Chang, S.; Tian, J.; Zeng, H.; Ji, H. Graph-sequence enhanced transformer for template-free prediction of natural product biosynthesis. *Patterns* **2025**, *6*, No. 101259.
(9) Liu, Y.; Yang, Q.; Li, Y.; Zhang, L.; Luo, S. Application of Machine Learning in Organic Chemistry. *Chin. J. Org. Chem.* **2020**, *40*, 3812−3827.
(10) Voinarovska, V.; Kabeshov, M.; Dudenko, D.; Genheden, S.; Tetko, I. V. When Yield Prediction Does Not Yield Prediction. *J. Chem. Inf. Model.* **2024**, *64*, 42−56.
(11) Mahjour, B.; Shen, Y.; Cernak, T. Ultrahigh-Throughput Experimentation for Information-Rich Chemical Synthesis. *Acc. Chem. Res.* **2021**, *54*, 2337−2346.
(12) Benavides-Hernández, J.; Dumeignil, F. From Characterization to Discovery: Artificial Intelligence, Machine Learning and High-Throughput Experiments for Heterogeneous Catalyst Design. *ACS Catal.* **2024**, *14*, 11749−11779.
(13) Li, B.; Su, S.; Zhu, C.; Lin, J.; Hu, X.; Su, L.; Yu, Z.; Liao, K.; Chen, H. A deep learning framework for accurate reaction prediction and its application on high-throughput experimentation data. *J. Cheminf.* **2023**, *15*, 72.
(14) Xu, Y.; Gao, Y.; Su, L.; Wu, H.; Tian, H.; Zeng, M.; Xu, C.; Zhu, X.; Liao, K. High-Throughput Experimentation and Machine Learning-Assisted Optimization of Iridium-Catalyzed Cross-Dimerization of Sulfoxonium Ylides. *Angew. Chem., Int. Ed.* **2023**, *62*, No. e202313638.
(15) Tu, Z.; Stuyver, T.; Coley, C. W. Predictive chemistry: machine learning for reaction deployment, reaction development, and reaction discovery. *Chem. Sci.* **2023**, *14*, 226−244.
(16) Zhang, C.; Zhai, Y.; Gong, Z.; Duan, H.; She, Y.-B.; Yang, Y.-F.; Su, A. Transfer learning across different chemical domains: virtual screening of organic materials with deep learning models pretrained on small molecule and chemical reaction data. *J. Cheminf.* **2024**, *16*, 89.
(17) Shi, R.; Yu, G.; Huo, X.; Yang, Y. Prediction of chemical reaction yields with large-scale multi-view pre-training. *J. Cheminf.* **2024**, *16*, 22.
(18) Shi, R.; Yu, G.; Chen, L.; Yang, Y. YieldFCP: Enhancing Reaction Yield Prediction via Fine-grained Cross-modal Pre-training. *Artificial Intelligence Chemistry* **2025**, *3*, No. 100085.
(19) Reaxys. https://www.reaxys.com (accessed Mar 30, 2015).
(20) 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.
(21) Lowe, D. M. *Extraction of Chemical Structures and Reactions from the Literature*; University of Cambridge, 2012.
(22) NextMove. https://nextmovesoftware.com (accessed Mar 7, 2022).
(23) Lowe, D. *Chemical reactions from US patents (1976−2016)*; **2017**.
(24) 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.
(25) Baraka, S.; Kerdawy, A. M. E. Multimodal Transformer-based Model for Buchwald-Hartwig and Suzuki-Miyaura Reaction Yield Prediction. *arXiv* **2022**, 14062. http://arxiv.org/abs/2204.14062
(26) Jiang, S.; Zhang, Z.; Zhao, H.; Li, J.; Yang, Y.; Lu, B.-L. When SMILES Smiles, Practicality Judgment and Yield Prediction of Chemical Reaction via Deep Chemical Language Processing. *IEEE Access* **2021**, *9*, 85071−85083.
(27) Liu, Z.; Morozbcd, Y. S.; Isayev, O. The challenge of balancing model sensitivity and robustness in predicting yields: a benchmarking study of amide coupling reactions. *Chem. Sci.* **2023**, *14*, 10835−10846.
(28) Ma, Y.; Huang, X.; Nan, B.; Moniz, N.; Zhang, X.; Wiest, O.; Chawla, N. V. Are we Making Much Progress? Revisiting Chemical Reaction Yield Prediction from an Imbalanced Regression Perspective. In *WWW' 24 Companion*; **2024**, pp 790 - 793.
(29) Xu, B.; Wang, Q.; Fang, C.; Zhang, Z.-M.; Zhang, J. Recent advances in Pd-catalyzed asymmetric cyclization reactions. *Chem. Soc. Rev.* **2024**, *53*, 883−971.
(30) Wang, L.; Zhang, C.; Bai, R.; Li, J.; Duan, H. Heck reaction prediction using a transformer model based on a transfer learning strategy. *Chem. Commun.* **2020**, *56*, 9368−9371.
(31) Ree, N.; Göller, A. H.; Jensen, J. H. What the Heck?—Automated Regioselectivity Calculations of Palladium-Catalyzed Heck Reactions Using Quantum Chemistry. *ACS Omega* **2022**, *7*, 45617−45623.
(32) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(33) *RDKit: Open-source chemoinformatics and machine learning*. http://www.rdkit.org (accessed Mar 30, 2025).
(34) PubChem. https://pubchem.ncbi.nlm.nih.gov (accessed Mar 30, 2025).
(35) *CADD Group Chemoinformatics Tools and User Services*. http://cactus.nci.nih.gov (accessed Mar 30, 2025).
(36) 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.
(37) 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**, *359*, 429−434.
(38) Das, M.; Sharma, P.; Sunoj, R. B. Machine learning studies on asymmetric relay Heck reaction—Potential avenues for reaction development. *J. Chem. Phys.* **2022**, *156*, 114303.
(39) Schwaller, P.; Probst, D.; Vaucher, A. C.; Nair, V. H.; Kreutter, D.; Laino, T.; Reymond, J.-L. Mapping the space of chemical reactions using attention-based neural networks. *Nat. Mach. Intell.* **2021**, *3*, 144−152.
(40) Probst, D.; Schwaller, P.; Reymond, J.-L. Reaction classification and yield prediction using the differential reaction fingerprint DRFP. *Digital Discovery* **2022**, *1*, 91−97.
(41) Wang, S.; Zhao, W.; Liu, Y.; Li, Y. Multi-modal Homogeneous Chemical Reaction Performance Prediction with Graph and Chemical Language Information. *Chin. J. Chem.* **2025**, *43*, 1230−1238.
(42) Heid, E.; Green, W. H. Machine Learning of Reaction Properties via Learned Representations of the Condensed Graph of Reaction. *J. Chem. Inf. Model.* **2022**, *62*, 2101−2110.
(43) Yang, Y.; Zha, K.; Chen, Y.-C.; Wang, H.; Katabi, D. Delving into Deep Imbalanced Regression. *arXiv* **2021**, No. 09554. http://arxiv.org/abs/2102.09554
(44) Schleinitz, J.; Langevin, M.; Smail, Y.; Wehnert, B.; Grimaud, L.; Vuilleumier, R. Machine Learning Yield Prediction from NiCOlit, a Small-Size Literature Data Set of Nickel Catalyzed C-O Couplings. *J. Am. Chem. Soc.* **2022**, *144*, 14722−14730.
(45) Maloney, M. P.; Coley, C. W.; Genheden, S.; Carson, N.; Helquist, P.; Norrby, P.-O.; Wiest, O. Negative Data in Data Sets for Machine Learning Training. *Org. Lett.* **2023**, *25*, 2945−2947.
(46) Fitzner, M.; Wuitschik, G.; Koller, R.; Adam, J.-M.; Schindler, T. Machine Learning C-N Couplings: Obstacles for a General-Purpose Reaction Yield Prediction. *ACS Omega* **2023**, *8*, 3017−3025.
(47) Amatore, C.; Jutand, A. Role of dba in the Reactivity of Palladium(0) Complexes Generated in Situ from Mixtures of $Pd(dba)_2$ and Phosphines. *Coord. Chem. Rev.* **1998**, *178*, 511.
(48) Bruno, N. C.; Tudge, M. T.; Buchwald, S. L. Design and Preparation of New Palladium Precatalysts for C-C and C−N Cross-Coupling Reactions. *Chem. Sci.* **2013**, *4*, 916−920.
(49) Cong, M.; Fan, Y.; Raimundo, J.-M.; Tang, J.; Peng, L. Pd(dba)2 vs Pd2(dba)3: An in-Depth Comparison of Catalytic Reactivity and Mechanism via Mixed-Ligand Promoted C-N and C−S Coupling Reactions. *Org. Lett.* **2014**, *16*, 4074−4077.
(50) Saebi, M.; Nan, B.; Herr, J. E.; Wahlers, J.; Guo, Z.; Zuranski, A. M.; Kogej, T.; Norrby, P.-O.; Doyle, A. G.; Chawla, N. V.; Wiest, O. On the use of real-world datasets for reaction yield prediction. *Chem. Sci.* **2023**, *14*, 4997−5005.
(51) Toniato, A.; Vaucher, A. C.; Laino, T.; Graziani, M. Negative chemical data boosts language models in reaction outcome prediction. *Sci. Adv.* **2025**, *11*, No. eadt5578.

https://doi.org/10.1021/acs.joc.5c01641
J. Org. Chem. 2025, 90, 12768−12777

(52) Hinton, G.; Vinyals, O.; Dean, J. Distilling the Knowledge in a Neural Network. *arXiv* **2015**, No. 02531.
(53) Nunn, R. Many-models Medicine: Diversity as the Best Medicine. *J. Eval. Clin. Pract.* **2012**, *18*, 974−978.
(54) Ibarz, B.; Kurin, V.; Papamakarios, G.; Nikiforou, K.; Bennani, M.; Csordás, R.; Dudzik, A.; Bošnjak, M.; Vitvitskyi, A.; Rubanova, Y.; Deac, A.; Bevilacqua, B.; Ganin, Y.; Blundell, C.; Veličković, P. A Generalist Neural Algorithmic Learner. *arXiv* **2022**, 11142.
(55) Zantvoort, K.; Nacke, B.; Görlich, D.; Hornstein, S.; Jacobi, C.; Funk, B. Estimation of minimal data sets sizes for machine learning predictions in digital mental health interventions. *npj Digit. Med.* **2024**, *7*, 361.
(56) Siemers, F. M.; Feldmann, C.; Bajorath, J. Minimal data requirements for accurate compound activity prediction using machine learning methods of different complexity. *Cell Rep. Phys. Sci.* **2022**, *3*, No. 101113.
(57) Raghavan, P.; Haas, B. C.; Ruos, M. E.; Schleinitz, J.; Doyle, A. G.; Reisman, S. E.; Sigman, M. S.; Coley, C. W. Dataset Design for Building Models of Chemical Reactivity. *ACS Cent. Sci.* **2023**, *9*, 2196−2204.
(58) 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. A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **2019**, *365*, No. eaax1566.
(59) Song, T.; Luo, M.; Zhang, X.; Chen, L.; Huang, Y.; Cao, J.; Zhu, Q.; Liu, D.; Zhang, B.; Zou, G.; Zhang, G.; Zhang, F.; Shang, W.; Fu, Y.; Jiang, J.; Luo, Y. A Multiagent-Driven Robotic AI Chemist Enabling Autonomous Chemical Research On Demand. *J. Am. Chem. Soc.* **2025**, *147*, 12534−12545.
(60) Das, M.; Ghosh, A.; Sunoj, R. B. Advances in machine learning with chemical language models in molecular property and reaction outcome predictions. *J. Comput. Chem.* **2024**, *45*, 1160−1176.

https://doi.org/10.1021/acs.joc.5c01641
J. Org. Chem. 2025, 90, 12768−12777